Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/112/2024.17919

A Review of the State of the Art 3D Generative Models and Their Applications

Zimeng Qiu[1,a,*]

[1]*Department of Computing Science, University of Alberta, Edmonton, Canada*
a. qzimeng@ualberta.ca
**corresponding author*

Abstract: Ever since 2022, there has been a large number of 3D generative models that have been devised and published, such as AvatarGen, CityDreamer, and HOLOFUSION. Generally speaking, these models can perform tasks such as generating a 3D human model, creating an unbounded city scene, and constructing a 3D object. And it is not a surprise that 3D generative models are very popular these years because there has been a witness of huge need for 3D models in the global market and the models themselves also serve as both convenient and productive tools for the relevant industries. For instance, 3D generative models can utilize a combination of Generative Adversarial Network (GAN) and Multi-Layer Perceptron (MLP) or Neural Radiance Field (NeRF) or Diffusion Model to produce 3D human model; Autoregressive Model or Feature Extraction + Volume Rendering to generate 3D scenes; Diffusion Model or GAN + MLP to produce 3D objects. This paper tries to present a taxonomy of the main 3D generative models from the angle of the kinds of outputs and strategies employed by different models.

Keywords: 3D Generative Model, 3D Human Model, 3D Scene, 3D Object.

1. Introduction

Before referring to the definition of "3D Generative Model", there should be a crystal-clear understanding of the word "Generative Model". Examples like ChatGPT can receive data format such as text input to generate the corresponding text output as the answer; Pix2Pix can generate a colored picture given a black and white picture. From the two examples above, researchers know that the generation model can perform tasks such that to create a certain form of data out of the data that is given by the user/trainer. To explain further, a generative model has the functionality to dig out the pattern hidden inside the input data and generate data formats like text, pictures, 3D models, etc. based on the discovered pattern. And "3D Generative Model" is exactly one of the most popular kinds among all the generative models as there is an exceeding need for assistance by the 3D generative model. According to the survey carried out by GRAND VIEW SEARCH, there is expected to be an increment of 14.7% from 2023 to 2030 in the North American 3D Mapping & 3D Modelling Market and there is already a global market size of $660.3M in 2022. In addition, countries like China not only possess a huge 3D model market with a size of $193M in 2023 but also witnessed a rapid increment in the number of patent applications regarding 3D models jumping from less than 300 in 2014 to over 1100 in 2021. Thus, it will only become a trend that the 3D generative models will be adopted by more fields and sectors. Consequently, it is necessary to have a state-of-the-art review

DOI: 10.54254/2755-2721/112/2024.17919

regarding the 3D generative models as they are changing the traditional working flow of industries such as games and AR/VR and since 3D generative models can now generate 3D human models or 3D scenes, people working in these two industries, or any industry related with 3D modeling will have to readapt themselves to the prompted new technologies to enhance their productivity. Therefore, this paper aims to provide useful information for the professionals who work in the relevant industries and expect to get assistance from the 3D generative models or researchers who want to have an overall knowledge about the current state-of-the-art 3D generative models. So far in the 3D generative model area, there have been several representative works, such as the Get3D model invented by Toronto AI Lab which can extract 3D surface mesh and colors to construct a 3D model; the EG3D model carried out by Stanford University and Nvidia that produce not only high-res multi-view 2D pictures, but high-quality 3D geometry of the model as well; Get3DHuman model invented by Tencent can generate 3D model by dividing the normal workflow of generating 3D model into producing geometry and texture separately and combine them at the end of the workflow; AG3D model that generates 3D human model with a canonical space so that it can make free deformations to the human model.

This paper will be organized as follows: First, it will introduce what are the basic strategies employed by the 3D generative models. Then, the specific models will be analyzed based on the taxonomy of their outputs, which will be regarding 3D human generation, 3D scene generation, and 3D object generation, and the strategies they employed. Next, the paper will argue the performance of each model by listing out the specific quantitative evaluation of each model and making proper comparisons in between. Then, the paper will discuss the future expectations and limits of the 3D generative model. Finally, the paper will finish with briefly concluding the content covered.

2. Basic Strategies Employed by the 3D Generative Models

Before looking into the details of each 3D generative model that will be introduced later, some basic strategies should be learned first as the foundational and essential part of understanding how a 3D Generative Model works behind the scenes. Thus, this section will divide “3D Generative Model” into two keywords “3D” and “Generative” respectively to discuss what are the common strategies utilized in these two aspects. First, the following methods will often be used in expressing or observing 3D models:

- NeRF: can receive a collection of multi-view 2D images of an object as input and use a radiance field to represent features (color and volume density) for each point in the space, then it will reconstruct the complete 3D scene via volume rendering [1].
- Signed Distance Function (SDF): is a function that represents the distance from any point to the closest surface of the 3D object in a space, with a sign indicating if the point is inside or outside of the object. By SDF, models can handle more precise and detailed operations on 3D models [2].
- Skinned Multi-Person Linear model (SMPL): is a model that can parametrize a 3D human model so that its shapes and poses can be adjusted accordingly [3].
- Super Resolution Model (SR): is a model that runs a reconstruction process to convert low-resolution images into high-resolution images [4]. The model is often used to raise the quality of textures on the generated 3D human model.
- Bird’s Eye View (BEV): is often used in 3D scene representation for its efficient and expressive performance. Additionally, BEV includes a height map and a semantic map that can be encoded into scene features for volume rendering [5].
- The strategies above are mainly adopted to express or observe 3D models, though some of them like NeRF also have the responsibility to generate a 3D model and BEV can provide features for volume rendering, they are placed inside the keyword “3D” for the reason that strategies in the

DOI: 10.54254/2755-2721/112/2024.17919

keyword “Generative” cannot express 3D models as the methods above can. Next, the following methods are used to generate 3D models:

- GAN: is often combined with MLP to generate 3D models, the role GAN plays in model generation is producing implicit features for 3D models, and in 3D human model generation, it also creates a tri-plane space.
- MLP: is as described above, MLP will map the features generated by GAN which are two-dimensional data into three-dimensional space so that after volume rendering, the 3D model can be produced as desired.
- Diffusion Model: is the process of adding noises to and subtracting noises from the 3D featured latent code to produce 3D models. It is often used when the input is text and the output is a 3D model, where the Diffusion Model helps to convert from text to images, then from images to desired 3D model.
- Autoregressive Model: produces 3D models step by step based on the original pattern or virtual token prediction.

The strategies above usually do not generate 3D models on their own, instead, they will often combine other different methods numbered from 1 to 2 and since different models have their branched strategies, hence this section only picks the ones that are most commonly adopted in every model.

Now that the strategies for both “3D” and “Generative” have been introduced, the next section will cover the representative state-of-the-art 3D generative models that are grouped by their forms of outputs and strategies they utilize during the generation.

3. 3D Generative Models categories and introduction

3.1. 3D Human Model Generation

3.1.1. Combine GAN and MLP to generate the 3D human model

AvatarGen is a model that receives latent code and camera pose as a pair of inputs, which will go through MLP to capture the features within so that the encoder (based on GAN) can generate the captured features in a tri-plane space with a standard human model [6]. At the same time, the human pose and the human shape will be put into a pair of data that will be mapped to the canonical space to deform the human model with the guidance of sample points from the observation space and SMPL model. During the deformation process, the human model will be transformed from its original pose and position in the observation space to a standardized version in the canonical space. The deformed human model will then be rendered into a low-resolution featured image via a volume renderer based on SDF. Lastly, the low-resolution image will be transformed into a high-resolution image via the SR model. GAN is also adopted in this model to improve the performance of the generator, camera discriminator, and model’s geometry condition.

3.1.2. Utilize the NeRF model to produce the 3D human model

In this section, the paper argues EVA3D as the representative of using the NeRF model to produce the human model. Firstly, EVA3D uses a collection of 2D images as input to the NeRF model to get the implicit 3D representation of the human model, which helps to do the volume rendering and uses SDF to represent the implicit geometry of the model [7]. Then EVA3D uses SMPL to specify the human shape and pose so that it can use inverse “Linear Blend Skinning (LBS)”, which is an algorithm embedded in SMPL, to transform the human model from its original position and pose in the observation space to a standardized position and pose in the canonical space. Also, SMPL divides the human model into 16 local bounding boxes, where each box is assigned a subnetwork (developed

DOI: 10.54254/2755-2721/112/2024.17919

based on GAN) and each subnetwork will be queried to generate the final human model by integrating the results of the queries.

3.1.3. Utilize the Diffusion model to produce the 3D human model

Using the diffusion model to generate a human model, HumanNorm stands out to be one of the state-of-the-art models. Similar to Get3DHuman the paper mentioned in the introduction, HumanNorm also generates geometry and texture separately. During the generation, HumanNorm employs a normal-adapted diffusion model, a depth-adapted diffusion model, and a normal-aligned diffusion model based on the guidance of the text input [8]. In geometry generation, the normal-adapted diffusion model can guide the rendered normal maps produced by the DMTet model, and the depth-adapted diffusion model can guide the rendered depth maps produced by the DMTet model. DMTet is a 3D generative model that creates 3D shapes conditioned on the user input [9]. Under the guidance of the normal-adapted and depth-adapted diffusion models, the rendered human model will gradually approach the high-fidelity result via SDS loss. Next, the normal-aligned diffusion model in the texture generation stage will leverage the normal maps as guidance to make sure that the generated texture aligns with the geometry and eventually produces the desired human model after several loops of adjustments.

3.2. 3D Scene Generation

3.2.1. Utilize the Autoregressive model to produce the 3D scene

SGAM is a model that can generate an unbounded 3D world derived from a single snapshot of a scene [10]. To explain how it works: firstly, the model receives a single representation of the scene along with a new query viewpoint and generates a realistic RGB-D image of the 3D scene. Then the mapping module will update the scene representation based on the RGB-D information of that newly generated image from a specific camera position. By repeating this process iteratively, the module can construct the 3D scene continuously and consistently.

3.2.2. Combine Feature Extraction + Volume Rendering to produce the 3D scene

SceneDreamer is a model that takes in some noise and a style code (originated from a collection of 2D images) and eventually generates an unbounded 3D scene where the model can render 2D images as output [5]. To explain its workflow: SceneDreamer firstly exploits the height map and semantic map from BEV scene representation (mapped from the received noise) to query the scene features so that the generative hash grid can be adapted to compute and organize the space-varied and scene-varied features from the scene features. Then a volume render will combine the style code with the generated features to render 2D images as output, which will go through a discriminator for Real/Fake judgment to improve the future performance of the model.

3.3. 3D Object Generation

3.3.1. Utilize the Diffusion model to produce the 3D object

HoloDiffusion is a unique model that can receive continuous frames from a specific collection of videos containing an object as input and eventually generate the same object with a consistent view but in a brand-new camera location that never appears in the input video frames [11]. Firstly, the model takes in a series of videos where each video provides an RGB image and the corresponding camera pose so that the model can observe the object in the scene from different angles. Next, a 3D

DOI: 10.54254/2755-2721/112/2024.17919

voxel grid will be exploited to represent the latent features of the object's shape and appearance so that the requested angle of the object can be rendered by taking the 3D grid and projecting it to a 2D image based on the required camera pose. Finally, the model makes the refinements to the object's 3D representation by comparing features in the generated grid with those in each frame of the video to gradually adjust the representation to a more accurate representation.

3.3.2. Combine GAN and MLP to produce the 3D object

Dream3D is an innovative 3D generative model that breaks the 3D model generation into two parts: the model firstly translates text to a 3D model, then it goes through a 3D optimization process to synthesize the generated 3D model with the corresponding text input [12]. To briefly list out the workflow: firstly, the text prompt will be transferred into a rough 2D image representation of the object that appeared in the text prompt via a diffusion model. Then, the latent features within the generated 2D image of the object will be extracted via the StyleGAN model that also combines with SDF to ensure the precise and implicit representation of the 3D object. Next, the extracted features will be fed into an MLP, then the features will be decoded so that the 3D object can be rendered into a high-quality 3D model. Finally, the 3D optimization process will adjust and refine the appearance of the generated 3D model to align with the text description.

4. Experiment data comparisons and analysis

Table 1: Quantitative comparison of the seven 3D generative models that range in descending order

Model	FID↓
HoloDiffusion	94.50
HumanNorm	92.50
Scene Dreamer	76.73
Dream3D	40.83
SGAM	26.60
EVA3D	15.91
AvatarGen	5.25

To gain a more in-depth and precise idea of the 3D generative models the paper has discussed, it is necessary to argue the performance of each model by taking each model under the same evaluation standard. Here, FID (Fréchet Inception Distance) is used as the evaluation metric to assess the quality of the generation results of the seven 3D generative models in Table 1.

It is obvious that though there are three categories of models as the paper discussed: 3D human, 3D scene, and 3D object generative models, their performances vary in a quite drastic way: for the FID values of 3D human generative models AvatarGen, EVA3D, and HumanNorm, they are 5.25, 15.91 and 92.5 respectively that have a maximum difference of more than 87 [6, 7, 8]. The main reason that caused this phenomenon is possibly the strategies they employed to achieve each of their functionality. In terms of the reason why AvatarGen has the highest quality of generation, it may be because AvatarGen generates the 3D human in a continuous workflow compared with HumanNorm which first generates the rough 3D model then covers the model with the generated appearance which breaks the workflow into two parts and that may severely damage the overall quality of the model after combing the geometric model with the appearance [6, 8]. Compared with EVA3D which chooses to divide the human model into several bounding boxes which may affect the continuity of the appearance of the human model at the connection parts, AvatarGen instead generates the 3D

DOI: 10.54254/2755-2721/112/2024.17919

human as a whole which guarantees there will not be glitches that may happen around the connection parts in the EVA3D model [6, 7].

For the performance of 3D scene generative models Scene Dreamer and SGAM, their FID values are 76.73 and 26.60 respectively which indicate there is a huge gap between their performances [5, 9]. The paper argues that it is for a similar reason as discussed before regarding the strategies the two models adopted. As described in the third part of the paper, SGAM uses an autoregressive model to iteratively generate the 3D scene so that it can get a consistent pattern of the desired scene [5]. However, Scene Dreamer exploits the height map and semantic map from BEV to extract the overall features in a bounded BEV scene representation to generate an unbounded scene [9]. Thus, when Scene Dreamer encounters the unbounded area generation, it may lack the ability to generate a good quality 3D scene since it is not pattern-based.

In terms of the reason that caused the difference between the 3D object generative models HoloDiffusion (with FID value of 94.50) and Dream3D (with FID value of 40.83), the paper argues that it may be mainly because of the format of input data that each model receives [10, 11]. What HoloDiffusion has to deal with is a sequence of video frames that contain the object that needs to be viewed from a new angle, while Dream3D takes an input of text that will be transferred into a 2D image, then the image will be used to render the 3D object [10, 11]. The quality of each frame in the video matters to the HoloDiffusion model since it directly derives the 3D object from the video frames, but the quality of each frame in a video cannot be guaranteed as high as a generated 2D image. Therefore, the difference in the format of the input data may well explain why HoloDiffusion has a poorer performance compared with Dream3D.

Overall, the paper observes a huge influence exerted by the different approaches that different models take as well as the format of different input data that the model receives on the performance of the 3D generative models.

5. Future expectations and limitations of 3D generative models

The future of 3D generative models is promising, since it has now achieved the capability to produce highly detailed 3D human models, unbounded 3D scenes, and 3D objects with high fidelity. The recent research area is now aiming to make functionality like photorealism, real-time generation come into reality. Furthermore, there is also a witness of the improvement of quality in texture and geometry alignment and feature details such as human faces and the folding issues of clothes on the generated human bodies. Moreover, combining 3D generative models with the generation of 3D assets in the game and AR/VR industries has become a popular topic in recent years, and it is expected to see deeper and wider cooperation between 3D generative models and relevant industries.

However, there are still some significant problems that remain unsolved. For instance, most of the current 3D generative models can only generate models under a very explicit topic (such as human, scene, or object) without the ability to produce in a more versatile way. In addition, 3D generative models usually require a huge amount of training data as their backbone because they need to study the complex 3D geometries and textures out of the training data. Lastly, current 3D generative models also share the problem of having limitations on the strategies they use. Take GAN that many 3D generative models have been adopted as an example, though GAN can produce outputs in a fast and high-quality fashion, it still lacks enough stability and convergence during the generation process.

6. Conclusion

The paper introduces seven different 3D generative models that are categorized by the type of 3D model they generate and the way they generate the 3D models. Through the descriptions of the 3D generative models, it is easy to see that the recent research on 3D generative models has been in a

DOI: 10.54254/2755-2721/112/2024.17919

very muti-directional and innovative way by taking several different approaches to produce and represent 3D models. In addition, the paper also collects the quantitative data run by each model and makes comparisons among the models in the same category, which turns out that the strategies adopted and input data format received by the models have a huge influence on their performances. Finally, the paper argues that the future of 3D generative models will be promising but some issues still wait to be solved by future researchers.

References

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

[2] *Jones, M. W., Baerentzen, J. A., & Sramek, M. (2006). 3D distance fields: A survey of techniques and applications. IEEE Transactions on visualization and Computer Graphics, 12(4), 581-599.*

[3] *Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2023). SMPL: A skinned multi-person linear model. In Seminal Graphics Papers: Pushing the Boundaries, Volume 2 (pp. 851-866).*

[4] *Farsiu, S., Robinson, M. D., Elad, M., & Milanfar, P. (2004). Fast and robust multiframe super resolution. IEEE transactions on image processing, 13(10), 1327-1344.*

[5] *Chen, Z., Wang, G., & Liu, Z. (2023). Scenedreamer: Unbounded 3d scene generation from 2d image collections. IEEE transactions on pattern analysis and machine intelligence.*

[6] *Zhang, J., Jiang, Z., Yang, D., Xu, H., Shi, Y., Song, G., ... & Feng, J. (2022, October). Avatargen: a 3d generative model for animatable human avatars. In European Conference on Computer Vision (pp. 668-685). Cham: Springer Nature Switzerland.*

[7] *Hong, F., Chen, Z., Lan, Y., Pan, L., & Liu, Z. (2022). Eva3d: Compositional 3d human generation from 2d image collections. arXiv preprint arXiv:2210.04888.*

[8] *Huang, X., Shao, R., Zhang, Q., Zhang, H., Feng, Y., Liu, Y., & Wang, Q. (2024). Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4568-4577).*

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

[10] *Shen, Y., Ma, W. C., & Wang, S. (2022). SGAM: Building a virtual 3d world through simultaneous generation and mapping. Advances in Neural Information Processing Systems, 35, 22090-22102.*

[11] *Karnewar, A., Vedaldi, A., Novotny, D., & Mitra, N. J. (2023). Holodiffusion: Training a 3d diffusion model using 2d images. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 18423-18433).*

[12] *Xu, J., Wang, X., Cheng, W., Cao, Y. P., Shan, Y., Qie, X., & Gao, S. (2023). Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20908-20918).*