Extended Abstract Track

Learning Effective NeRFs and SDFs Representations with 3D GANs for Object Generation

Abstract

We present a solution for 3D object generation of ICCV 2023 OmniObject3D Challenge. In recent years, 3D object generation has made great process and achieved promising results, but it remains a challenging task due to the difficulty of generating complex, textured and high-fidelity results. To resolve this problem, we study learning effective NeRFs and SDFs representations with 3D Generative Adversarial Networks (GANs) for 3D object generation. Specifically, inspired by recent works, we use the efficient geometry-aware 3D GANs as the backbone incorporating with label embedding and color mapping, which enables to train the model on different taxonomies simultaneously. Then, through a decoder, we aggregate the resulting features to generate Neural Radiance Fields (NeRFs) based representations for rendering high-fidelity synthetic images. Meanwhile, we optimize Signed Distance Functions (SDFs) to effectively represent objects with 3D meshes. Besides, we observe that this model can be effectively trained with only a few images of each object from a variety of classes, instead of using a great number of images per object or training one model per class. With this pipeline, we can optimize an effective model for 3D object generation. This solution is among the top 3 in the ICCV 2023 OmniObject3D Challenge.

Keywords: 3D Object Generation, Generative Adversarial Networks, NeRFs, SDFs

1. Introduction

3D object generation aims at generating meaningful 3D surfaces and synthesis images of 3D objects given random inputs (Wu et al., 2023; Jiang et al., 2023). Inspired by the great success of 2D object generation (Karras et al., 2019; Esser et al., 2021), there have been many studies (Lunz et al., 2020; Gadelha et al., 2017) extending 2D methods to 3D generation. Recent 3D generation approaches focus more on effective 3D representation learning (Pavllo et al., 2023; Mo et al., 2019) and advanced generation strategy (Karras et al., 2020b; Chan et al., 2022; Wang et al., 2023). Despite the great progress and promising results in recent years, 3D object generation is still a challenging task due to the difficulty of generating complex, textured and high-fidelity results.

To devise a solution for 3D object generation in the ICCV 2023 OmniObject3D Challenge, we simultaneously consider effective 3D representation learning and advanced generation strategy. In terms of 3D representation learning, textured mesh (Choi et al., 2023) and Neural Radiance Fields (NeRFs) (Mildenhall et al., 2021) embrace the great potential to effectively encode 3D object features. In particular, textured mesh representations can support individual texture map generation, which allows us to easily replace surface textures, while NeRFs-based representations excel at rendering high-fidelity images. Furthermore, when considering effective shape representation, there are a variety of options, ranging from explicit representations like point-based (Qi et al., 2017), mesh-based (Ben-Hamu et al., 2018) and voxel-based (Wang et al., 2021) approaches to implicit ones such as the Signed Distance Functions (SDFs) (Park et al., 2019; Liu et al., 2023) and occupancy

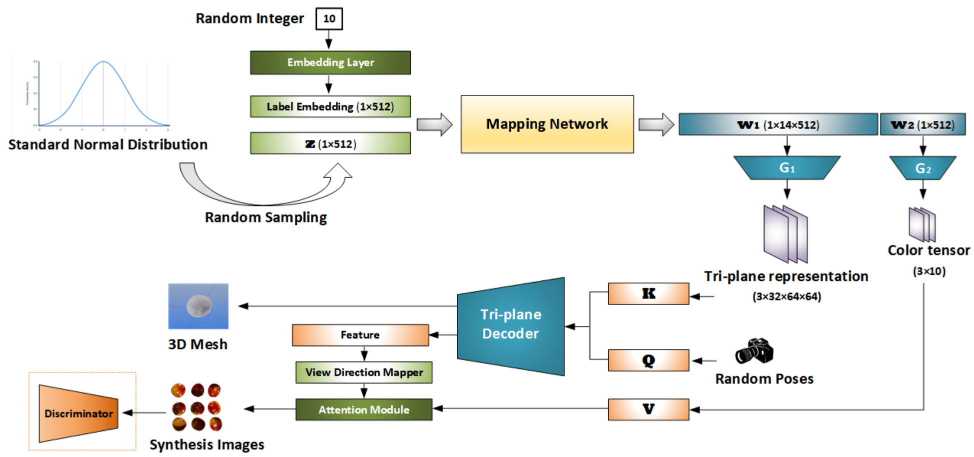

Figure 1: An overview of the framework of our solution for 3D object generation.

function (Mescheder et al., 2019). In light of these cutting-edge techniques, we mainly focus on exploring effective NeRFs representations for 3D object generation while still optimizing SDFs representations to effectively represent objects with 3D meshes. Moving on to advanced generation strategy, there are usually different options, such as latent generation (Wang et al., 2023) and direct generation (Müller et al., 2023). In particular, latent generation methods aim to generate the representation of the desired output in the latent space and decode the latent code back to the 3D space, while direct generation focus on directly generating the desired 3D representation without utilizing the latent space. Besides, from the perspective of generation process, generation strategies can be categorized into Generative Adversarial Networks (GANs) (Karras et al., 2020b; Chan et al., 2022), Variational AutoEncoder (VAE) (Petrovich et al., 2021), flow-based (Han et al., 2019), Diffusion Models (DMs) (Wang et al., 2023), *etc.*. In our solution, we employ the efficient geometry-aware 3D GANs (Chan et al., 2022) as the backbone for 3D object generation. The overall framework of our solution is depicted in Fig. 1. Specifically, we aim to learn effective NeRFs and SDFs representations with 3D Generative Adversarial Networks (GANs) for 3D object generation. Inspired by the recent success of efficient geometry-aware 3D GANs, *i.e.*, EG3D (Chan et al., 2022; Pavllo et al., 2023), we employ it as the backbone and incorporate label embedding and color mapping (Pavllo et al., 2023) to enable model training on different taxonomies simultaneously. Next, we generate NeRFs-based representations for rendering high-fidelity synthetic images and SDFs to effectively represent objects with 3D meshes through a decoder. In addition, we empirically find that we can effectively train this model with only a few images per object from various classes, rather than training with a great number of images per object or optimizing one model per class. With this pipeline, we can optimize an effective model for 3D object generation. This solution is one of the final top-3-place solutions in the ICCV 2023 OmniObject3D Challenge.

2. Technical Solution

Motivation of Using GANs instead of DMs. Although diffusion models have recently gained an increasing popularity in 3D object generation, most existing works (Müller et al., 2023; Wang et al., 2023) have only been verified to work under specific category conditions (e.g., chairs (Müller et al., 2023) or avatars (Wang et al., 2023)). Besides, since the OmniObject3D dataset (Wu et al., 2023) is a collection of objects from hundreds of taxonomies, following the pipeline of existing DMs would necessitate the use of hundreds of pre-trained models, which is impractical. Hence, we employ a method that enables simultaneous training on all the taxonomies in the OmniObject3D dataset. Moreover, given the success of StyleGAN2 (Karras et al., 2020b) in representing NeRFs in latent space, we have chosen latent generation for our pipeline.

Solution Pipeline. The framework of our solution is depicted in Fig. 1. Overall, our solution adopts a GAN-based generation process. Specifically, with EG3D (Chan et al., 2022) as the backbone, firstly, we generate a one-dimensional latent code, $\mathbf{z}$ (1×512), which is obtained by randomly sampling 512 values from a standard normal distribution. Meanwhile, we randomly generate an integer, indicating the object label, within the range of the total category number of OmniObject3D. The integer is then transformed into a label embedding by an embedding layer. This enables the model to be trained on different taxonomies simultaneously. Next, we forward the label embedding and latent code $\mathbf{z}$ into a mapping network and generate an intermediate latent code $\mathbf{w}$ $(1 \times 15 \times 512)$ with promoted dimensions. Here, the first 14 dimensions of $\mathbf{w}$ ($\mathbf{w}_1$ in the figure) are used as the input to the object generator ($\mathbf{G}_1$ in the figure) while the last dimension of $\mathbf{w}$ ($\mathbf{w}_2$ in the figure) are employed as the input to the color generator ($\mathbf{G}_2$ in the figure). $\mathbf{G}_1$ and $\mathbf{G}_2$ generate the tri-plane representation $(3 \times 32 \times 64 \times 64)$ and the color tensor (3×10), respectively. Following (Pavllo et al., 2023), we use the color generator which allows convenient texture conversion. After that, we determine queries, $\mathbf{Q}$, based on the randomly generated camera pose. The tri-plane representation is then taken as keys, $\mathbf{K}$, and transmitted into the tri-plane decoder together with $\mathbf{Q}$. Then, we use a decoder to generate the SDF of the object, which can be sampled to create a 3D mesh. This decoder also produces a feature as the input to a view direction mapper, from which the output is multiplied by the color tensor, $\mathbf{V}$ (the values), in the attention module. Finally, the attention module produces the 2D synthesis image and a discriminator (Chan et al., 2022) which only functions during training is used for adversarial training.

Model Training. Since we are simultaneously optimizing NeRFs and SDFs representations, we ensure the generator can initially generate a unit sphere by initializing our model with SDF-pretraining as (Pavllo et al., 2023). Then, the generator is optimized using the Adam optimizer with a learning rate of 0.0025, while the discriminator (Chan et al., 2022) uses a learning rate of 0.002. The batch size is set to 32 and the model is trained for 300,000 iterations. Additionally, we also use adaptive discriminator augmentation (Karras et al., 2020a) to improve model generalization. Overall, the model training objective is:

$$\alpha_0\mathcal{L}_{path} + \alpha_1\mathcal{L}_e + \alpha_2\mathcal{L}_v + \alpha_3\mathcal{L}_{sdf}, \tag{1}$$

where $\mathcal{L}_{path}$ denotes the Path length regulation loss (Karras et al., 2020b), $\mathcal{L}_e$ and $\mathcal{L}_v$ denote the entropy loss and total variation loss (Karras et al., 2020b), respectively, $\mathcal{L}_{sdf}$ represents the SDF eikonal loss (Pavllo et al., 2023), and α_i are weighting parameters. We set α_0=2, α_1=0.05, α_2=0.5 and α_3=0.1.

Figure 2: Images rendered by our solution with label embedding.

Figure 3: Images rendered by our solution without label embedding.

3. Experiments

Dataset. We train our model from scratch for 3D object generation on the training dataset provided by the ICCV 2023 OmniObject3D Challenge. The complete OmniObject3D dataset (Wu et al., 2023) contains 216 classes from a wide range of daily categories and approximately 6k objects in total where each object has 100 images in different views.

Efficient Model Training. Instead of training one model for each class, we train a single model with various types of objects. The model takes each image along with its camera pose, focal length and object type. Besides, we empirically find that we can effectively train this model with only a few images per object from various classes, so we select the first 8 images of each object for model training. This improves the model training efficiency.

Results. Fig. 2 shows some visualization results of synthesis images generated by our solution on OmniObject3D. It can be seen that our approach is capable of generating high-fidelity images from different views for 3D object generation. Besides, Fig. 3 shows results of our solution without the additional label embedding. Comparing Fig. 2 with Fig. 3, we can see that without the additional label embedding, the quality of the synthesis images is significantly worse than the solution with the label embedding. This verifies that the additional label embedding is beneficial for improving model generalization. Furthermore, we compare our method with a state-of-the-art approach proposed in CVPR 2023, NFI (Pavllo et al., 2023), as shown in Table 1, in terms of FID.

Method	FID
NFI (Pavllo et al., 2023)	62.426
Ours	57.624

Table 1: Improvement in terms of FID of our solution after using the additional label embedding.

4. Conclusion

In this paper, we present a solution for 3D object detection of ICCV 2023 OmniObject3D Challenge. The key insight of our solution is to learn effective NeRFs and SDFs Representations with 3D generative adversarial networks for 3D object generation. Overall, this solution is one of the final top-3-place solutions in the ICCV 2023 OmniObject3D Challenge.

References

Heli Ben-Hamu, Haggai Maron, Itay Kezurer, Gal Avineri, and Yaron Lipman. Multi-chart generative surface modeling. *ACM Transactions on Graphics*, 37(6):1–15, 2018.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

Jaehoon Choi, Dongki Jung, Taejae Lee, Sangwook Kim, Youngdong Jung, Dinesh Manocha, and Donghwan Lee. Tmo: Textured mesh acquisition of objects with a mobile device by using differentiable rendering. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16674–16684, 2023.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411. IEEE, 2017.

Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proc. of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019.

Lutao Jiang, Ruyi Ji, and Libo Zhang. Sdf-3dgan: A 3d object generative method based on implicit signed distance function. *arXiv preprint arXiv:2303.06821*, 2023.

Tero Karras, 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*, pages 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020b.

Yibo Liu, Kelly Zhu, Guile Wu, Yuan Ren, Bingbing Liu, Yang Liu, and Shan Jinjun. Mv-deepsdf: Implicit modeling with multi-sweep point clouds for 3d vehicle reconstruction in autonomous driving. In *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2023.

Sebastian Lunz, Yingzhen Li, Andrew Fitzgibbon, and Nate Kushman. Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. *arXiv preprint arXiv:2002.12674*, 2020.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

Ben Mildenhall, 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*, 65(1):99–106, 2021.

Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019.

Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4328–4338, 2023.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. Shape, pose, and appearance from a single image via bootstrapped radiance field inversion. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4391–4401, 2023.

Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proc. of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023.

Xiaogang Wang, Marcelo H Ang, and Gim Hee Lee. Voxel-based network for shape completion by leveraging edge generation. In *Proc. of the IEEE/CVF international conference on computer vision*, pages 13189–13198, 2021.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023.