FH Salzburg
MultiMediaTechnology

An Overview of 3D Object Reconstruction Diffusion Models

Seminar Paper

Author: Felix Beer
Advisor: DI Gerlinde Emsenhuber

Repository: https://github.com/felixbeer/3d-diffusion-models-paper

Salzburg, Austria, dd.mm.yyyy

An Overview of 3D Object Reconstruction Diffusion Models

Felix Beer
fbeer.mmt-b2022@fh-salzburg.ac.at
Salzburg University of Applied Sciences

Abstract

1 Introduction

In today's world, nearly every industry uses 3D models to visually represent objects or environments. Whether for entertainment, development, or research, 3D models are essential tools for understanding complex concepts and ideas.

Creating these models is a time-consuming and costly process that requires skilled artists and designers, especially when compared to taking pictures or recording videos with cameras or smartphones. However, recent advancements in Generative 3D AI have made it possible to generate 3D models from a single image. This process, known as 3D object- or mesh-reconstruction, has the potential to transform and partially automate the creation and use of 3D models.

Some examples include the newly released TripoSR (Tochilkin et al. (2024)) as well as established models like Zero-1-to-3 (Liu, Wu, et al. (2023)), One-2-3-45 (Liu, Xu, et al. (2023)), and One-2-3-45++ (Liu, Shi, et al. (2023)).

This paper provides an overview of the various approaches and models used for 3D mesh reconstruction, including a comparison of their performance and visual results. Furthermore, it discusses the different strategies like voxel-based (Zhirong Wu et al. 2015), point cloud-based (Charles et al. 2017), and mesh-based (Wang et al. 2018) methods. It also explores the underlying methods and concepts like convolutional neural networks and neural radiance fields (Mildenhall et al. 2021) used in modern neural network-based models. Finally, it offers a brief look at use cases and applications that benefit the most as well as 3D-model datasets such as Objaverse-XL (Deitke et al. 2023) and their impact on model training.

2 Concepts and Functionality

To understand the process of 3D mesh reconstruction, it is essential to understand the underlying concepts and techniques used in the field. This section provides a quick overview of the different methods and approaches used to generate 3D models from 2D images.

2.1 Shape from Shading

Starting with shape from shading, this is the most basic technique used to estimate the shape of an object from a single image, which dates back to the late 80s (Horn and Brooks (1989)). The basic idea is to use the shading information in the image to infer the 3D structure of the object. This is done by assuming that the object gets illuminated by a single light source and that the surface of the object is Lambertian, meaning that the surface reflects light uniformly in all directions. This by itself is a strong assumption, as most real-world objects are not perfectly Lambertian. Estimating surface normals at each point on the object involves using the intensity of the reflected light and then integrating the normals to estimate the object's depth. This technique was one of the first to show that basic 3D shape information could be recovered from a single image and has since inspired many other techniques for 3D reconstruction.

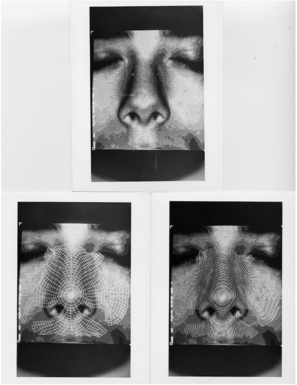

Figure 1: Shape from shading (Horn and Brooks 1989)

2.2 Multi-view Stereo

Before the appearance of deep learning, multi-view stereo was the most common technique used for 3D reconstruction. The basic idea is to use multiple images of an object taken from different angles to estimate the 3D structure of the object. This is done by first estimating the camera parameters for each image and then using these parameters to triangulate the 3D points in the scene. Furukawa and Ponce (2010)

2.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have empowered the field of computer vision and have found applications in many areas, including 3D mesh reconstruction. CNNs are a type of deep learning model that are especially good at working with image data. They are designed to automatically and adaptively learn spatial hierarchies of features from the data. A very popular CNN architecture is the AlexNet (Krizhevsky, Sutskever, and Hinton 2012). Krizhevsky, Sutskever, and Hinton (2012)

2.4 Voxel-based Approaches

Zhirong Wu et al. (2015)

2.5 Point Cloud-based Approaches

Charles et al. (2017)

2.6 Mesh-based Approaches

Wang et al. (2018)

2.7 Neural Implicit Functions

Park et al. (2019) Mildenhall et al. (2021)

3 Models

3.1 Models

In the recent years several models have been developed to generate 3D models from 2D images. Some of the most prominent models include:

3.1.1 One-2-3-45

This model was developed by Liu, Shi, et al. (2023)...

3.1.2 Zero-1-to-3

Zero-1-to-3 is a model developed by Liu, Wu, et al. (2023)...

3.1.3 TripoSR

TripoSR is a model developed by Tochilkin et al. (2024)...

3.2 Comparison

Result comparison between models both visually and in terms of performance.

4 Applications of 3D Mesh Reconstruction

As the field is still relatively new, no mainstream applications have been established yet. However, the potential is great and some possible applications have already been identified.

4.1 Development and Entertainment

The most prominent application of 3D mesh reconstruction could be in the development and entertainment industry. The ability to generate 3D models from 2D images could revolutionize the asset creation process. This could be especially beneficial for indie developers or small studios that do not have the resources to create high-quality 3D models from scratch. The generated models could be used in video games, movies, animations, and other forms of media. This could significantly reduce the time and cost associated with creating 3D assets, allowing developers and animators to focus on other aspects of their projects. (see Figure 2)

4.2 Medical

4.3 Other Applications

4.3.1 Cultural Heritage

5 Discussion and Future Direction

Discussion of results and their implications. What are the limitations current works? What are the next steps in this research area?

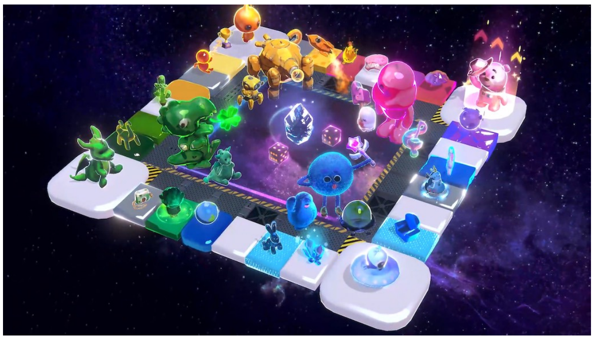

Figure 2: Models generated by One-2-3-45++ (Liu, Shi, et al. 2023)

6 Conclusion

References

Charles, R. Qi, Hao Su, Mo Kaichun, and Leonidas J. Guibas. 2017. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation” [in en]. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 77–85. Honolulu, HI: IEEE, July. ISBN: 978-1-5386-0457-1. https://doi.org/10.1109/CVPR.2017.16.

Deitke, Matt, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, et al. 2023. “Objaverse-XL: A Universe of 10M+ 3D Objects” [in en]. In *Advances in Neural Information Processing Systems 37.*

Furukawa, Yasutaka, and Jean Ponce. 2010. “Accurate, Dense, and Robust Multiview Stereopsis” [in en]. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, no. 8 (August): 1362–1376. ISSN: 0162-8828. https://doi.org/10.1109/TPAMI.2009.161.

Horn, Berthold, and Michael Brooks. 1989. *Shape from Shading.* Vol. 2. MIT Press, January.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” In *Advances in Neural Information Processing Systems,* vol. 25. Curran Associates, Inc.

Liu, Minghua, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. 2023. *One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion.* ArXiv:2311.07885 [cs], November.

Liu, Minghua, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2023. “One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization” [in en]. In *Advances in Neural Information Processing Systems 36.* June.

Liu, Ruoshi, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. "Zero-1-to-3: Zero-shot One Image to 3D Object" [in en], 9298–9309.

Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. "NeRF: representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65, no. 1 (December): 99–106. ISSN: 0001-0782. https://doi.org/10.1145/3503250.

Park, Jeong Joon, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation" [in en]. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 165–174. Long Beach, CA, USA: IEEE, June. ISBN: 978-1-72813-293-8. https://doi.org/10.1109/CVPR.2019.00025.

Tochilkin, Dmitry, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. 2024. *TripoSR: Fast 3D Object Reconstruction from a Single Image.* ArXiv:2403.02151 [cs], March.

Wang, Nanyang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. 2018. "Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images," 52–67.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. "3D ShapeNets: A deep representation for volumetric shapes" [in en]. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 1912–1920. Boston, MA, USA: IEEE, June. ISBN: 978-1-4673-6964-0. https://doi.org/10.1109/CVPR.2015.7298801.

This work has the following word count (counted by texcount):

```
File: body.tex
Encoding: utf8
Sum count: 857
Words in text: 798
Words in headers: 46
Words outside text (captions, etc.): 13
Number of headers: 22
Number of floats/tables/figures: 2
Number of math inlines: 0
Number of math displayed: 0
Subcounts:
  text+headers+captions (#headers/#floats
  237+1+0 (1/0/0/0) Section: Introduction
  42+3+0 (1/0/0/0) Section: Concepts and
  160+3+6 (1/1/0/0) Subsection: Shape fro
  68+2+0 (1/0/0/0) Subsection: Multi-view
  68+3+0 (1/0/0/0) Subsection: Convolutio
  4+2+0 (1/0/0/0) Subsection: Voxel-based
  3+3+0 (1/0/0/0) Subsection: Point Cloud
  3+2+0 (1/0/0/0) Subsection: Mesh-based
  6+3+0 (1/0/0/0) Subsection: Neural Impl
  0+1+0 (1/0/0/0) Section: Models
  50+4+0 (4/0/0/0) Subsection: Models
  11+1+0 (1/0/0/0) Subsection: Comparison
  27+5+0 (1/0/0/0) Section: Applications
  98+3+7 (1/1/0/0) Subsection: Developmen
  0+1+0 (1/0/0/0) Subsection: Medical
  0+4+0 (2/0/0/0) Subsection: Other Appli
  21+4+0 (1/0/0/0) Section: Discussion an
  0+1+0 (1/0/0/0) Section: Conclusion
```