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, 10.07.2024

An Overview of 3D Object Reconstruction Diffusion Models

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

ABSTRACT

This paper examines the latest advancements in 3D object reconstruction diffusion models, focusing on methodologies like convolutional neural networks, generative adversarial networks, and neural implicit functions. Key models are evaluated for their performance and accuracy in generating 3D meshes from single images as well as their limitations. Through a comparative analysis, the paper highlights the performance, visual quality, and potential applications of these models in fields such as entertainment and medical imaging.

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. Another exemplary field is real estate development, where architects and engineers utilize Building Information Modeling (BIM). BIM improves collaboration and efficiency in construction projects by integrating various aspects of a building's lifecycle, from initial design through maintenance and deconstruction, into 3D models (Azhar, Khalfan, and Maqsood 2020).

With modern interactive media, the demand for high-quality and numerous 3D models has significantly increased. Triple-A games and blockbuster movies heavily rely on detailed 3D models to create true-to-life experiences.

However, creating these models is a time-consuming and costly process that requires skilled artists and designers, especially when compared to capturing images or videos with cameras or smartphones. Recent advancements in Generative 3D AI have made it possible to generate 3D models from a single image, a process known as 3D object or mesh reconstruction.

While the generated models are currently not yet at the level required for Triple-A games or the high standards of the entertainment industry, it holds significant potential for transforming and partially automating the creation and use of 3D models. At present, these technologies are more suitable for smaller prototyping projects or indie productions.

Some examples of 3D mesh reconstruction include the newly released TripoSR (Tochilkin et al. (2024)) as well as established models like Zero-1-to-3 (R. Liu et al. (2023)) and One-2-3-45 (M. Liu 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, visual results and training based on 3D-model datasets such as Objaverse-XL (Deitke et al. 2023). 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.

2 CONCEPTS AND FUNCTIONALITY

To understand the process of 3D mesh reconstruction, it is essential to explore the various techniques employed to generate 3D models from 2D images.

3D mesh reconstruction involves several methodologies, each contributing uniquely to the overall objective. Early techniques like Shape from Shading (Horn and Brooks 1989) utilized shading information to infer the 3D structure of objects. Multi-view Stereo (MVS) employed multiple images from different viewpoints to estimate 3D structure through triangulation methods.

The introduction of Convolutional Neural Networks (CNNs) significantly improved the accuracy of 3D reconstructions by providing robust tools for image analysis and feature extraction. Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) further enhanced these capabilities by offering a superior training method. Combining CNNs and GANs has led to advanced models such as Pixel2Mesh++ (Wen et al. 2019), enabling the accurate estimation of complete three-dimensional models from single images.

Lastly, Neural Implicit Functions, such as NeRF (Mildenhall et al. 2021), have further improved 3D reconstruction by using neural networks to model continuous volumetric functions, generating high-quality 3D reconstructions with smooth surfaces.

The following sections examine each technique in detail, exploring their specific mechanisms, innovations, and contributions to the field.

2.1 Shape from Shading

Shape from Shading (SFS) is an early technique used to estimate the shape of an object from a single image, dating back to the late 1980s Horn and Brooks (1989). This method aims to reconstruct the 3D shape of an object by analyzing shading information in a 2D image.

The principle of SFS is to use variations in shading to infer the 3D geometry of an object's surface. This method operates under several key assumptions:

1. Lambertian Reflectance: The surface of the object reflects light uniformly in all directions, meaning the intensity of reflected light depends only on the angle between the light source and the surface normal.

2. Single Light Source: The object is illuminated by a single distant light source.

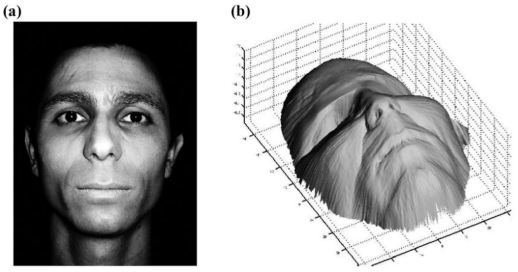

Figure 1: (a) A real face image. (b) Surface recovered from (a) by the generic SFS algorithm with the perspective model, with the light source located at the optical center (taken from He and Chen (2018)).

3. Known Light Source Direction: The direction of the light source relative to the camera is known.

When all assumptions are met, the intensity of the reflected light in relation to the incoming light direction can be used to estimate the surface normals at each point of the object. By integrating these normals, the depth of the object can be estimated, providing a 3D representation.

While Shape from Shading is a robust technique, it has limitations. It requires accurate knowledge of the light source direction and assumes Lambertian reflectance, which does not hold for many real-world objects, resulting in significant deviations from real light behavior (Wolff 1996). This limitation is not unique to SFS but is present in many computer graphics and computer vision applications, as a compromise for computational efficiency and simplicity. Additionally, SFS is sensitive to noise and may produce inaccurate results in complex scenes.

2.2 Multi-view Stereo

Multi-view Stereo (MVS) uses multiple images from different angles to estimate an object's 3D structure (see Figure 2). This concept, known as Structure from Motion (SfM), is derived from various works rather than a single publication. Ullman and Brenner (1997) was among the first to describe the process, noting that the structure of four non-coplanar points can be recovered from three orthographic projections.

SfM involves:

1. Camera Calibration: Estimating intrinsic and extrinsic camera parameters, including focal length, principal point, and distortion.

2. Feature Detection and Matching: Using techniques like Difference of Gaussians (DoG) and Harris corner detection. Furukawa and Ponce (2010), for example, proposed a robust algorithm based on these techniques.

3. Pose Estimation: Determining the relative positions and orientations of cameras based on features.

4. Triangulation: Using poses and features to estimate 3D points.

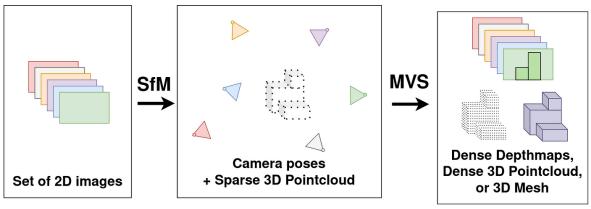

Figure 2: Multi-view Stereo Pipeline (taken from Rouch (2023))

5. Dense Reconstruction: Interpolating sparse 3D points to create a solid model.

MVS is also significant in augmented and virtual reality for real-time environment reconstruction and mapping.

Despite its effectiveness, MVS has limitations, such as requiring multiple images from different viewpoints, which may not always be available. It can also struggle with textureless surfaces and repetitive patterns, making feature matching challenging. These limitations have driven the development of deep learning methods, which can overcome some challenges by learning from large datasets of images and 3D models.

2.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are deep learning models especially useful for image data. They have significantly impacted computer vision and found applications in 3D mesh reconstruction.

CNNs process data with convolutional layers that capture spatial hierarchies of features through a series of filters in a grid-like structure. This method is effective for image data, where the spatial arrangement of pixels is crucial for understanding content.

Compared to traditional methods like MVS, CNNs can learn complex features from images and automatically extract relevant information for 3D reconstruction. They are also more robust to noise and variations in lighting conditions.

AlexNet by Krizhevsky, Sutskever, and Hinton (2012) was one of the first successful applications of CNNs to image classification, achieving substantial performance improvements over previous methods. AlexNet reduced top-1 error rates to 37.5% and top-5 error rates to 17.0%, significantly better than the previous state-of-the-art.

Top-1 and top-5 error rates are common metrics for evaluating image classification models. The top-1 error rate is the percentage of images for which the correct label is not the top prediction, while the top-5 error rate is the percentage of images for which the correct label is not among the top five predictions.

2.4 Generative Adversarial Networks

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) are a type of deep learning model that consists of two neural networks: a generator and a discriminator. The generator creates new data samples, while the discriminator distinguishes between real and generated samples. The two

networks are trained together in competition, where the generator tries to produce realistic data to fool the discriminator, and the discriminator tries to identify real versus generated data.

In 3D mesh reconstruction, CNNs and GANs are often combined to generate 3D models from 2D images. The CNN extracts features from the input image, and the GAN generates the 3D model from these features. An example of this is the Pixel2Mesh++ model by Wen et al. (2019).

2.5 Neural Implicit Functions

Neural Implicit Functions offer an alternative approach to 3D reconstruction by using neural networks to model continuous volumetric functions. Unlike traditional methods that use explicit representations like meshes or voxels, neural implicit functions encode the geometry of 3D objects in a continuous function that can be evaluated at any point in 3D space.

Neural Implicit Functions, such as DeepSDF (Park et al. 2019) and NeRF (Mildenhall et al. 2021), use neural networks to map spatial coordinates to implicit representations of 3D shapes. These models learn a function that outputs a value indicating whether a point lies inside or outside the object (in the case of signed distance functions — DeepSDF) or the density and color at a given spatial location (in the case of radiance fields — NeRF).

Neural Implicit Functions typically work as follows:

1. Data Collection: Collect dataset of 3D shapes and corresponding annotated 2D images. An example is ShapeNet (Chang et al. 2015).

2. Network Training: Train the network to minimize the error between the predicted and actual shape representation using a loss function, similar to the training process of GANs (2.4).

3. Shape Inference: Once trained, the network can infer the 3D shape from new inputs by querying the implicit function at various spatial locations to reconstruct the geometry.

The main advantages of neural implicit functions include their continuous nature, allowing for high-quality reconstructions with smooth surfaces. They also typically require less memory and computation compared to explicit representations like meshes or voxels created by deep convolutional networks (2.3). Neural Radiance Fields (NeRFs) (Mildenhall et al. 2021) have been particularly successful in generating high-quality 3D reconstructions from 2D images (see Figure 3) and are used in many modern neural network-based 3D reconstruction models.

3 Models

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

Figure 3: NeRF - 100 input views to NeRF representation (Mildenhall et al. 2021)

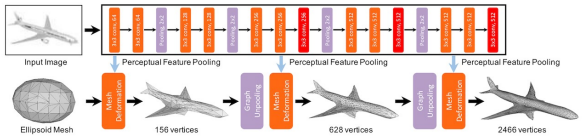

Figure 4: Pixel2Mesh Pipeline (N. Wang et al. 2018)

3.1 Pixel2Mesh

Pixel2Mesh, developed by N. Wang et al. (2018), consists of two main components: an image feature network and a cascaded mesh deformation network. The image feature network, a convolutional neural network (CNN), extracts perceptual features from the input image. The cascaded mesh deformation network is a graph-based convolution network.

A graph-based convolution network operates on a graph rather than a grid, as traditional CNNs do. In Pixel2Mesh, the graph represents the 3D mesh model with vertices and edges.

The Pixel2Mesh model works as follows:

1. The image feature network processes the input image to extract features.

2. The cascaded mesh deformation network initializes with an ellipsoid mesh model.

3. The model uses the extracted features to refine the shape of the mesh.

4. The network refines the mesh iteratively in three blocks, with each iteration refining the shape and increasing the mesh resolution (see Figure 4).

5. The network estimates vertex positions at each step and uses them to look up features from the image feature network for the next iteration.

3.2 Pixel2Mesh++

Pixel2Mesh++ (Wen et al. 2019) adapts the original Pixel2Mesh model to use multiple images, generating a more accurate 3D model. It combines the well-known MVS systems with the Pixel2Mesh model.

Pixel2Mesh++ samples deformations around each vertex of the base mesh and selects the optimal deformation using perceptual features pooled from multiple input images. This multi-view approach enables the model to capture more detailed and accurate geometries, especially in occluded or less visible areas. Single-view methods often fail to capture the

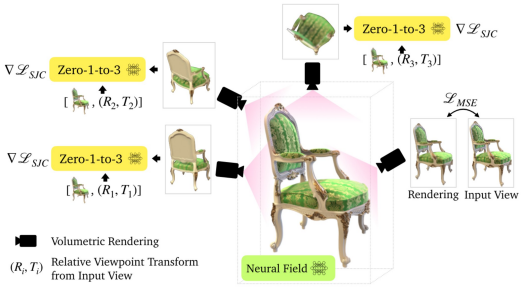

Figure 5: 3D reconstruction with Zero-1-to-3 (R. Liu et al. 2023). The model receives the base image as well as the view parameters R (Rotation) and T (Transform) as input.

Figure 6: Models generated by One-2-3-45++ (M. Liu et al. 2024)

full extent of an object and must estimate the shape by hallucinating the missing parts — hence this model is included in this overview.

3.3 Zero-1-to-3

The Zero-1-to-3 (Zero123) model, developed by R. Liu et al. (2023), generates 3D models from a single image in a zero-shot manner, meaning it can produce accurate 3D reconstructions without extensive per-object training. Unlike Pixel2Mesh, this model creates both the mesh and a colored texture based on the given image, highlighting its robustness and versatility across various types of input images.

The process involves fine-tuning a pre-trained diffusion model, specifically Stable Diffusion, to learn how to control camera extrinsics (the external parameters defining the camera's position and orientation). By using a 3D model dataset, the model learns to manipulate the camera viewpoint, allowing it to create views from different angles — a technique known as novel view synthesis, illustrated in Figure 5.

By merging the input image with the synthesized views, the model effectively learns the 3D structure of the object and generates a high-quality 3D mesh.

The improved version, Zero123-XL, enhances performance by including additional training data from Objaverse-XL (Deitke et al. 2023), a dataset containing over 10 million 3D objects. This larger dataset improves the model's zero-shot capabilities and overall reconstruction accuracy.

3.4 One-2-3-45

One-2-3-45, developed by M. Liu et al. (2023), is a state-of-the-art model designed to generate 3D meshes from single 2D images with exceptional efficiency.

It can reconstruct a full 360-degree 3D textured mesh in just 45 seconds, significantly faster than other models that often require lengthy optimization processes. This rapid processing occurs without the need for per-shape optimization, making it highly suitable for real-time consumer applications. One-2-3-45 uses the Zero123 model (3.3) to generate multi-view images from the input image. These images are compared to the input image for pose estimation, resulting in poses for each of the generated views. The model then feeds these results into a signed distance function-based neural surface reconstruction model (similar to DeepSDF mentioned in 2.5) to generate the finished 3D mesh.

An improved version, One-2-3-45++ (M. Liu et al. 2024), further optimizes the model's performance. The major improvement involves switching from Zero123 to a custom 2D diffusion model. Zero123 predicted each view individually, leading to inconsistencies. The new model uses a fine-tuned 2D diffusion model that attends to all views simultaneously, resulting in more accurate outcomes.

3.5 TripoSR

TripoSR, developed by Tochilkin et al. (2024), is another model for fast 3D object reconstruction from a single image. It was created through a collaboration between Stability AI (the creators of Stable Diffusion) and Tripo AI. Building upon the Large Reconstruction Model (LRM) architecture by Hong et al. (2024), TripoSR uses transformer technology for feed-forward 3D generation, producing high-quality 3D meshes in under 0.5 seconds.

The LRM architecture consists of three main components: an image encoder, a triplane representation, and a NeRF decoder. The image encoder, based on a vision transformer, converts an input image into a set of detailed features called latent vectors. These latent vectors capture important information about the object's shape and appearance in a compressed form.

Next, the triplane representation arranges these latent vectors in a three-dimensional grid, capturing the object's complex shapes and textures. The triplane representation uses layers that help the model focus on different parts of the image and learn their relations.

Finally, the NeRF decoder predicts the color and density of each point in the 3D grid, creating a smooth and detailed 3D model of the object.

TripoSR introduces several key improvements to this architecture. For instance, it uses a hand-selected, higher-quality subset of the Objaverse dataset, which improves generalization and reconstruction quality. It also introduces a mask loss function that penalizes differences between predicted and ground-truth mask images, reducing artifacts and enhancing reconstruction fidelity.

3.6 Comparison

3.6.1 Model Training

Most models are trained on 3D model datasets. A recent dataset, Objaverse-XL (Deitke et al. 2023), contains over 10 million 3D objects and is used by models like Zero123-XL to improve their zero-shot capabilities. In machine learning, zero-shot refers to a model's ability to handle elements it has never seen before. Similar to training large language models, the 3D models in these datasets are sourced from various locations across the internet, raising ethical concerns about copyright and licensing. In the case of Objaverse-XL, most models are sourced from open-source GitHub repositories and specialized sites like Sketchfab, Thingiverse, and Polycam.

3.6.2 Model Performance

Based on the order of the models listed above, each model's reconstruction accuracy improves compared to the previous one. Thus, TripoSR (3.5), the newest model, is the most accurate and efficient.

Quantitatively comparing models can be done using metrics like Chamfer Distance (CD) and F-score (FS). The Chamfer Distance measures the average distance between points on the generated and ground-truth meshes, while the F-score, the harmonic mean of precision and recall, measures the overlap between the two meshes.

Compared to other models like One-2-3-45 (3.4) and LRM, TripoSR is better in both speed and performance, as indicated by lower Chamfer Distance (CD $\downarrow$) and higher F-score (FS $\uparrow$) metrics on benchmark datasets. TripoSR scores 0.111 for CD and 0.651 for FS, whereas One-2-3-45 scores 0.227 and 0.382, respectively. These values show that TripoSR is about twice as accurate as One-2-3-45.

These metrics can also be visually confirmed by comparing the generated models side by side (see Figure 7).

3.6.3 Model Complexity

Due to the lack of performance testing in this paper, all mentioned results are based on the respective papers.

As mentioned, TripoSR (3.5) is the fastest model, generating high-quality 3D models in under 0.5 seconds. TripoSR is approximately 10 times faster than LRM, the model it is based on, which takes about 5 seconds to generate a 3D object with less accuracy. This is still significantly faster than One-2-3-45++ (3.4), which takes about 60 seconds to generate a 3D object.

The other models compared did not provide exact time measurements.

All these measurements were taken using A100 GPUs, which are state-of-the-art in AI performance with 80 GB of memory.

These graphics cards were also used in the training of these models. LRM was trained on 128 of these GPUs for 3 days, totaling about 9,000 hours at an energy consumption of 250 watts, resulting in 2.3 MWh. While this might sound like a lot, it is relatively low compared to other AI models. For example, the Large Language Model Llama (Touvron et al. 2023) used 2,048 A100 GPUs for 5 months, resulting in an energy usage of 1.8 GWh.

4 Possible Applications

Although the field is still relatively new, the potential applications of 3D mesh reconstruction are broad. Some possible applications have already been identified.

4.1 Development and Entertainment

The most prominent application of 3D mesh reconstruction is in the development and entertainment industry. The ability to generate 3D models from 2D images could revolutionize the asset creation process. This technology is especially beneficial for indie developers or small studios that lack the resources to create high-quality 3D models from scratch. Generated models can be used in video games, movies, animations, and other forms of media, significantly reducing the time and costs associated with creating 3D assets. This allows developers and animators to focus on other aspects of their projects (see Figure 6).

4.2 Medical

With Instantiation-Net, Z.-Y. Wang et al. (2019) demonstrated that 3D mesh reconstruction can be used in the medical field to generate 3D models of organs from medical images. The whole pipeline is shown in Figure 8. These models can be used for diagnosis, treatment planning, and medical education. For example, 3D models of the heart can help plan surgeries and visualize complex anatomical structures.

However, diagnosing based on the 3D model comes with risks, as the model is only an estimation based on training data and not a real representation of the patient's anatomy.

4.3 Other Applications

In the future, 3D mesh reconstruction could also be used in fields like architecture, archaeology, cultural heritage, and many more. Although there have not yet been significant uses of the technology in these fields, it may be necessary to adapt the models to the specific requirements of these disciplines.

5 Conclusion and Future Directions

Of all the models discussed, the most recent TripoSR model stands out in terms of accuracy and efficiency across various scenarios.

While these models generate quite accurate 3D models given the right input images, there is still room for improvement, especially with real-world objects.

One limitation of all the mentioned models is their inability to handle images with backgrounds. Currently, another model or manual intervention is necessary to remove the background from the image.

Additionally, all the models assume Lambertian surfaces for simplicity, resulting in inaccuracies when reconstructing shiny or metallic surfaces.

A significant roadblock in this technology is the inability to see the backside of an object from a single image. Therefore, the estimated areas sometimes appear inaccurate or blurry.

Figure 7: Comparison of TripoSR to other state-of-the-art methods, including One-2-3-45 (Tochilkin et al. 2024)

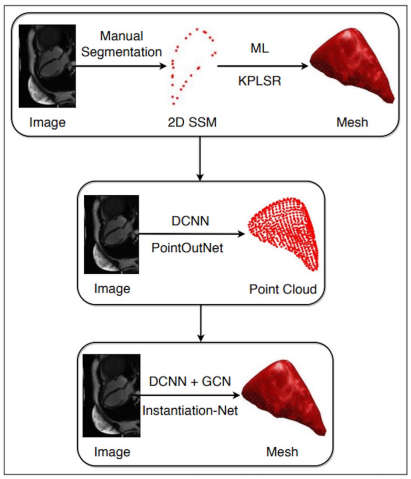

Figure 8: Medical Image Pipeline of Instantiation-Net (Z.-Y. Wang et al. 2019).

REFERENCES

Azhar, Salman, Malik Khalfan, and Tayyab Maqsood. 2020. "Building Information Modeling (BIM): Now and beyond." Publisher: UTS ePress, *The Australasian Journal of Construction Economics and Building* 12, no. 4 (August 20, 2020): 15–28. https://doi.org/10.5130/ajceb.v12i4.3032.

Chang, Angel X., Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, et al. 2015. *ShapeNet: An Information-Rich 3D Model Repository,* December 9, 2015. https://doi.org/10.48550/arXiv.1512.03012.

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 *Advances in Neural Information Processing Systems 37.* https://doi.org/10.48550/arXiv.2307.05663.

Furukawa, Yasutaka, and Jean Ponce. 2010. "Accurate, Dense, and Robust Multiview Stereopsis." *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.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems,* vol. 27. Curran Associates, Inc. https://doi.org/10.48550/arXiv.1406.2661.

He, Yu, and Shengyong Chen. 2018. "Advances in sensing and processing methods for three-dimensional robot vision." *International Journal of Advanced Robotic Systems* 15 (March 20, 2018): 172988141876062. https://doi.org/10.1177/1729881418760623.

Hong, Yicong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2024. *LRM: Large Reconstruction Model for Single Image to 3D*, March 9, 2024. https://doi.org/10.48550/arXiv.2311.04400.

Horn, Berthold, and Michael Brooks. 1989. *Shape from Shading*. Vol. 2. MIT Press, January 1, 1989.

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. https://doi.org/10.1145/3065386.

Liu, Minghua, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. 2024. "One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion," 10072–10083. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.2311.07885.

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 *Advances in Neural Information Processing Systems 36*. June 29, 2023. https://doi.org/10.48550/arXiv.2306.16928.

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," 9298–9309. Proceedings of the IEEE/CVF International Conference on Computer Vision. https://doi.org/10.1109/ICCV51070.2023.00853.

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 17, 2021): 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 165–174. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, June. ISBN: 978-1-72813-293-8. https://doi.org/10.1109/CVPR.2019.00025.

Rouch, Thomas. 2023. "PatchMatch Multi-View Stereo." Medium, April 10, 2023. Accessed June 30, 2024. https://betterprogramming.pub/patchmatch-multi-view-stereo-1-2-fc46e5dfe912.

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*, March 4, 2024. https://doi.org/10.48550/arXiv.2403.02151.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. *LLaMA: Open and Efficient Foundation Language Models*, February 27, 2023. https://doi.org/10.48550/arXiv.2302.13971.

Ullman, S., and Sydney Brenner. 1997. "The interpretation of structure from motion." Publisher: Royal Society, *Proceedings of the Royal Society of London. Series B. Biological Sciences* 203, no. 1153 (January): 405–426. https://doi.org/10.1098/rspb.1979.0006.

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. Proceedings of the European Conference on Computer Vision (ECCV). https://doi.org/10.1007/978-3-030-01252-6_4.

Wang, Zhao-Yang, Xiao-Yun Zhou, Peichao Li, Celia Riga, and Guang-Zhong Yang. 2019. *Instantiation-Net: 3D Mesh Reconstruction from Single 2D Image for Right Ventricle*, September 16, 2019. https://doi.org/10.1007/978-3-030-59719-1_66.

Wen, Chao, Yinda Zhang, Zhuwen Li, and Yanwei Fu. 2019. "Pixel2Mesh++: Multi-View 3D Mesh Generation via Deformation." In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1042–1051. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, October. ISBN: 978-1-72814-803-8. https://doi.org/10.1109/ICCV.2019.00113.

Wolff, Lawrence B. 1996. "Generalizing Lambert's Law for smooth surfaces." In *Computer Vision — ECCV '96*, edited by Bernard Buxton and Roberto Cipolla, 40–53. Berlin, Heidelberg: Springer. ISBN: 978-3-540-49950-3. https://doi.org/10.1007/3-540-61123-1_126.

Appendices

A AI Methodology

AI was mainly used to support grammar and spelling corrections. The tools used were LanguageTool as well as ChatGPT.

A.1 LanguageTool

LanguageTool was used throughout the whole writing process to quickly check for spelling and grammar mistakes. It was especially helpful in the final stages of the paper to ensure that no major mistakes were left. I estimate that about 5% of the paper was corrected by LanguageTool.

A.2 ChatGPT

ChatGPT was used to update some sections of the paper where I did not like the initial wording. Most of these sections were the more technical ones like 2.3 or 2.4. I estimate that about 20% of the paper was improved by ChatGPT. Other than that ChatGPT was used at the beginning to give a brief overview of the topic and to generate a structural idea for this paper.

A.2.1 Prompts

The first 5 prompts were provided from the slides in the supporting course with my own addition of the passive voice correction.

- ”Correct grammar, punctuation, and spelling errors.”
- ”Enhance clarity and coherence by rephrasing unclear or ambiguous sentences.”
- ”Ensure consistency in terminology, tense, and formatting.”
- ”Suggest more precise vocabulary without using overly complex jargon.”
- ”Trim unnecessary words or redundancies to make the text concise.”
- ”Correct any usage of passive voice.”

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

```
File: body.tex
Encoding: utf8
Sum count: 3420
Words in text: 3261
Words in headers: 43
Words outside text (captions, etc.): 108
Number of headers: 25
Number of floats/tables/figures: 8
Number of math inlines: 8
Number of math displayed: 0
Subcounts:
  text+headers+captions (#headers/#floats
  321+1+0 (1/0/0/0) Section: Introduction
  188+3+0 (1/0/0/0) Section: Concepts and
  266+3+33 (1/1/0/0) Subsection: Shape fr
  213+2+8 (1/1/0/0) Subsection: Multi-vie
  195+3+0 (2/0/0/0) Subsection: Convoluti
  117+3+0 (2/0/0/0) Subsection: Generativ
  275+3+10 (2/1/0/0) Subsection: Neural I
  22+1+0 (1/0/0/0) Section: Models
  163+1+5 (1/1/0/0) Subsection: Pixel2Mes
  102+1+0 (1/0/0/0) Subsection: Pixel2Mes
  180+1+25 (1/1/2/0) Subsection: Zero-1-t
  176+1+7 (1/1/0/0) Subsection: One-2-3-4
  231+1+0 (1/0/0/0) Subsection: TripoSR}
  435+7+12 (4/1/6/0) Subsection: Comparis
  23+2+0 (1/0/0/0) Section: Possible Appl
  93+3+0 (1/0/0/0) Subsection: Developmen
  91+1+8 (1/1/0/0) Subsection: Medical
  49+2+0 (1/0/0/0) Subsection: Other Appl
  121+4+0 (1/0/0/0) Section: Conclusion a
```