Versatile 3D World Reconstruction Using Diffusion Models

Project Background

Three-dimensional (3D) world reconstruction is a long-standing challenge in the field of computer vision and robotics. It is vital to get an accurate and efficient 3D reconstruction method to be practically viable since it is critical for applications that require spatial awareness or interaction with the immediate environment. Autonomous systems, such as self-driving cars, depend on precise 3D maps of their surroundings to avoid obstacles and navigate the roads safely. Furthermore, AR/VR applications require detailed 3D models of the real-world and virtual environments to create an immersive experience for the user.

Most current 3D reconstruction systems rely on well-established techniques like stereo vision, where two images from slightly different viewpoints are used to infer depth, or structure-from-motion (SFM), which reconstructs 3D structure from the motion observed in 2D image sequences. Traditional methods include Simultaneous Localization and Mapping (SLAM), SFM, and stereo vision techniques and have shown promising results in the past when dealing with video or multi-view imagery [1][2]. SLAM allows a camera system to update the map of an unknown environment while keeping track of the camera system inside of the environment. The traditional incremental SFM algorithm is slow, whereas the traditional global SFM algorithm is fast at the cost of accuracy [3]. A hybrid incremental and global SFM model is a great alternative to combine the advantages of both SFM models [3], or SLAM as it can significantly reduce computation time without compromising construction quality [4]. Among the various SLAM approaches, ORB-SLAM3 [5], which built upon ORB-SLAM2, is quite notable.

ORB (Oriented FAST and Rotated BRIEF) combined the FAST key point detector and the BRIEF descriptor to create an efficient and robust feature extraction method [6]. ORB also introduced Rotation-aware BRIEF (rBRIEF) to address the limitations of BRIEF when handling in-plane rotations, thereby improving its overall performance [6]. ORB features are generally favoured for real-time applications because of their efficiency, which is crucial for SLAM systems that are already resource-intensive and often operate with limited computational capabilities.

ORB-SLAM was a real-time, feature-based monocular SLAM system designed for operation in both small and large environments, whether indoors or outdoors. The system demonstrates robustness in the presence of significant motion clutter, supports wide baseline loop closing and relocalization, and includes a fully automatic initialization process [7]. ORB-SLAM2 built on its predecessor by operating with monocular, stereo, and RGB-D cameras, including loop closing, relocalization, and map reuse, and utilizing ORB features for effective tracking and mapping [8]. Most recently, ORB-SLAM3 further enhanced accuracy and robustness through a visual-inertial system based on MAP

estimation and introduces a multi-map system with improved place recognition, allowing it to handle challenging visual conditions effectively [5].

Despite the effectiveness of SLAM, it has limitations such as ORB-SLAM3 failing in low texture environments [5], or the issue of scale ambiguity in Monocular SLAM methods [9]. These challenges highlight the need for advanced methods, such as deep learning models, which offer a novel approach to generating high-quality 3D reconstructions from 2D data, improving efficiency and accessibility. NeuralRecon achieved this by integrating neural networks into the depth estimation process, and utilizing a learning-based TSDF fusion module guided by gated recurrent units learning techniques allowing the system to effectively integrate features from previous video fragments [10].

In recent years, diffusion models have become powerful tools for generating high-quality images, outperforming traditional methods in some areas. Diffusion models have demonstrated exceptional performance in 2D image generation and completion tasks and are now being explored for their potential in 3D modelling. Diffusion models, including frameworks such as DOLCE, offer significant promise for overcoming challenges in image reconstruction from limited viewpoints, showing excellent performance even with minimal data [11]. Similarly, SDFusion, developed for 3D shape reconstruction, leverages multimodal data and an encoder-decoder to learn a diffusion model to enhance shape completion tasks [12]. Diffusion models could be the next step in getting more accurate and robust 3D models of objects and environments.

Project Objective

This project seeks to develop a diffusion-based framework for 3D world reconstruction from both static images and video streams. The primary aim is to provide a flexible and robust method for reconstructing 3D environments that is applicable to various domains, including autonomous navigation, robotics, and AR/VR. There are multiple objectives that the project aims to achieve:

- The system will be able to ensure temporal coherence when processing video streams by maintaining consistency in 3D models across consecutive frames. The diffusion model will be used to refine depth maps and maintain consistency across frames, ensuring that the reconstructed 3D models are both accurate and temporally stable.

- It will enhance accuracy and detail in 3D object reconstruction from single-view images where obtaining multiple views of an object or environment may not be feasible. The system will be able to reconstruct detailed 3D scenes from limited 2D input data, such as single images or video frames by inferring occluded sections using a diffusion model.

- The system will be versatile enough to be used in various applications, including robotics, AR/VR, and environmental monitoring. It could be used for navigation or object manipulation in robotics, creating immersive environments for AR/VR, and generating the environments and landscapes for monitoring using images or video streams.

Project Methodology

The proposed methodology will integrate techniques from recent SOTA papers, such as SDFusion and DOLCE [11][12], to achieve the objectives outlined in the section before. The first part will include the data collection and preprocessing, followed by the model architecture and implementation, and finally with testing and validation.

For static image-based reconstruction, we will use publicly available datasets such as ShapeNet, BuildingNet and Pix3D, which provide large collections of 2D images with corresponding 3D models. For video-based reconstruction, datasets like KITTI or ScanNet, which provide video sequences in outdoor and indoor environments. Each video consists of multiple frames, and the dataset contains annotations and reconstructions for these frames, making it suitable for training and evaluating various computer vision models in the context of temporal scene analysis.

For initial depth estimation, traditional methods like SLAM and SFM will be tested to see if they can work in conjunction with the diffusion model for better results by generating a rough depth map from a single-view image or video streams.

The core of the project will be the diffusion model, where variations of diffusion probabilistic models (DPMs) will be implemented to learn and denoise the depth maps. Pre-trained diffusion models, such as those explored in DOLCE [11], will also be looked at to refine these depth maps, improving accuracy and surface detail. The diffusion model will act as a denoising tool, correcting depth artifacts and generating plausible surfaces where data is incomplete or ambiguous. Using 2D images from a single perspective, the diffusion models will attempt to reconstruct complete 3D models of objects, using probabilistic reasoning to infer unseen surfaces.

To ensure temporal consistency, regularization techniques will be looked at and employed to ensure coherent 3D models over time and smooth transitions between frames. This methodology will enable robust 3D modelling from limited data.

Project Schedule and Milestones

Literature Review and Data Collection (1-2 Months ~ October - November):

- Detailed reviews of relevant papers such as DOLCE for CT reconstruction and SDFusion for shape completion.
- Collection of image training datasets for model development.

Traditional Techniques for Depth Map Estimation and Diffusion Model (3-4 Months ~ November – February)

- Look at implementing initial depth estimation techniques to generate coarse 3D maps from 2D data.
- Integrate diffusion models into the 3D reconstruction pipeline.
- Train and fine-tune the diffusion models to improve depth map accuracy.

Final Testing and Application (1-2 Months ~ March – April):

- Apply the developed model to real-world scenarios and test datasets.
- Final testing to ensure a robust and accurate model is created and compared with other techniques

The report will be progressively updated and worked upon alongside the project throughout this tentative timeline.

References

[1] M. Kholil, I. Ismanto, and M. N. Fu'ad, "3D reconstruction using Structure From Motion (SFM) algorithm and Multi View Stereo (MVS) based on computer vision," *IOP Conference Series: Materials Science and Engineering*, vol. 1073, no. 1, p. 012066, Feb. 2021, doi: https://doi.org/10.1088/1757-899x/1073/1/012066.

[2] A. Macario Barros, M. Michel, Y. Moline, G. Corre, and F. Carrel, "A Comprehensive Survey of Visual SLAM Algorithms," *Robotics*, vol. 11, no. 1, p. 24, Feb. 2022, doi: https://doi.org/10.3390/robotics11010024.

[3] L. Gao, Y. Zhao, J. Han, and H. Liu, "Research on Multi-View 3D Reconstruction Technology Based on SFM," *Sensors*, vol. 22, no. 12, pp. 4366–4366, Jun. 2022, doi: https://doi.org/10.3390/s22124366.

[4] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen, "Dense 3D Object Reconstruction from a Single Depth View," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2820–2834, doi: https://doi.org/10.1109/TPAMI.2018.2868195.

[5] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1–17, 2021, doi: https://doi.org/10.1109/tro.2021.3075644.

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, pp. 2564–2571. doi: https://doi.org/10.1109/ICCV.2011.6126544.

[7] R. MurArtal, M. Montiel, and J. D. Tardós, "ORBSLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, doi: https://doi.org/10.1109/TRO.2015.2463671.

[8] R. MurArtal and J. D. Tardós, "ORBSLAM2: An OpenSource SLAM System for Monocular, Stereo, and RGBD Cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, doi: https://doi.org/10.1109/TRO.2017.2705103.

[9] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual SLAM algorithms: a survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 16, 2017, doi: https://doi.org/10.1186/s4107401700272.

[10] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "NeuralRecon: RealTime Coherent 3D Reconstruction from Monocular Video," in *2021 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR), pp. 15593–15602. doi: https://doi.org/10.1109/CVPR46437.2021.01534.

[11] J. Liu *et al.*, “DOLCE: A ModelBased Probabilistic Diffusion Framework for LimitedAngle CT Reconstruction,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10464–10474. doi: https://doi.org/10.1109/ICCV51070.2023.00963.

[12] Y. C. Cheng, H. Y. Lee, S. Tulyakov, A. Schwing, and L. Gui, “SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4456–4465. doi: https://doi.org/10.1109/CVPR52729.2023.00433.