6-DoF Tracking and 3D Reconstruction using Neural Radiance Fields

Saahil Athrij
University of Washington
Seattle, WA 98195
sathrij@uw.edu

Aiswarya Janardhanan
University of Washington
Seattle, WA 98195
ajanard5@uw.edu

Riya Jain
University of Washington
Seattle, WA 98195
riyaj26@uw.edu

Abstract

This research project focuses on the domain of 6-DoF (six degrees of freedom) tracking and 3D reconstruction of unknown objects using Neural Radiance Fields (NeRFs). The primary objective is to overcome the challenge of accurately determining the position and orientation of arbitrary rigid objects without relying on pre-existing 3D models or prior knowledge of the object's visual characteristics. Deep learning techniques are employed to leverage their capabilities in segmentation and robust feature extraction, enabling the handling of objects with diverse visual textures and complex environmental conditions. A fundamental aspect of our methodology involves the integration of memory-augmented pose graph optimization, which ensures spatiotemporal consistency during the tracking process. This optimization technique synergistically combines the advantages of memory-based approaches with pose graph optimization, resulting in enhanced accuracy and reliability in estimating and tracking the pose of the object. To assess the performance and efficacy of our approach, we utilize two widely recognized benchmark datasets: YCBInEOAT and NOCS. These datasets encompass a wide array of scenes and objects, thereby facilitating a comprehensive evaluation of the system's capabilities. The implementation for the method can be found here: https://github.com/sahil-athrij/BundleNeRF

1 Introduction

In recent years, significant advancements have been made in computer vision, particularly in the fields of object tracking and 3D reconstruction. The accurate determination of position and orientation in 3D space is crucial for applications like augmented reality, robotics, and autonomous systems [1, 20]. However, traditional approaches heavily depend on pre-existing 3D models or prior knowledge of an object's visual characteristics, which restricts their usefulness to known objects and controlled environments.

In the context of robot manipulation, precise pose estimation plays a vital role in successful manipulation tasks. While forward kinematics (FK) can offer pose estimates in certain cases, its reliability decreases when dealing with factors such as slippage or compliance of the manipulator's end-effector [21, 9, 3]. To overcome this challenge, visually-based single-image 6D pose estimation and tracking methods have been developed. However, these methods usually assume access to a specific object's 3D model, limiting their ability to handle novel and unseen instances [15, 16]. Recent research efforts have aimed to address this limitation by focusing on category-level 3D models or even relaxing the dependence on 3D models altogether. These advancements have paved the way for more generalized and robust pose estimation and tracking techniques [15, 16].

In this research project, we focus on addressing this challenge by leveraging the power of Neural Radiance Fields (NeRFs) and deep learning techniques. Our primary objective is to develop a robust

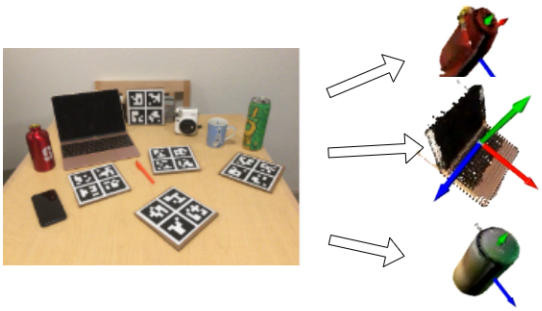

Figure 1: The project focuses on causal 6-DoF object tracking and 3D reconstruction using monocular RGB-D sequences. Notably, the method excels in generalization without prior knowledge of the object or interaction agent. It can handle challenging conditions such as flat and untextured surfaces, specular highlights, thin structures, severe occlusion, and various interaction agents like human hands, bodies, or robotic arms. The output gives accurate pose tracking of the objects and reconstruction using NeRFs.

framework for 6-DoF (six degrees of freedom) tracking and 3D reconstruction of unknown objects, without relying on pre-existing 3D models or prior knowledge of their visual properties. By harnessing the capabilities of deep learning, we aim to handle objects with diverse visual textures and complex environmental conditions.

One fundamental aspect of our methodology involves the integration of memory-augmented pose graph optimization. This approach ensures spatiotemporal consistency during the tracking process, enhancing the accuracy and reliability of pose estimation. By combining memory-based approaches with pose graph optimization, we can effectively handle dynamic scenes and accurately track the pose of the object over time.

To evaluate the performance and efficacy of our proposed approach, we utilize two widely recognized benchmark datasets: YCBInEOAT and NOCS. These datasets offer a diverse range of scenes and objects, providing a comprehensive evaluation of our system's capabilities in real-world scenarios. We aim to demonstrate the advancements achieved by our proposed approach.

By overcoming the limitations of traditional methods and leveraging the power of NeRFs and deep learning, our research project aims to contribute to the field of 6-DoF tracking and 3D reconstruction. The ability to accurately determine the position and orientation of arbitrary objects in a 3D space, without relying on pre-existing models or prior knowledge, holds great potential for various applications in computer vision and robotics.

2 Related Work

In this section, we discuss the existing literature related to the problem of estimation and tracking of 6-DoF object poses, as well as the reconstruction of objects.

2.1 6-DoF Object Pose Estimation and Tracking:

State-of-the-art methods in 6-DoF object pose estimation often rely on instance- or category-level object CAD models [17, 2] for offline training or online template matching. However, these approaches are limited to known objects and cannot be applied to novel unknown objects. Some recent works aim to generalize to unseen objects by relaxing the assumption [5] of pre-capturing posed reference views, but they still require pre-training or training on the same objects or object categories [4]. BundleTrack [19] is a notable method that generalizes pose tracking to novel unknown objects, but it lacks the ability to output object shapes. In this project work, we try to leverage this method to create object reconstruction along with pose tracking using Neural Radiance Fields.l

2.2 Simultaneous Localization and Mapping (SLAM):

SLAM addresses a similar problem to our work but focuses on tracking the camera pose in a static environment [12, 14]. Dynamic-SLAM methods track dynamic objects using techniques such as frame-model Iterative Closest Point (ICP) or probabilistic data association [11]. These methods reconstruct models on-the-fly by aggregating observed RGB-D data with newly tracked poses. In contrast, this project tried to leverage a Neural Radiance Field that allows for automatic on-the-fly fusion and dynamically rectifies historically tracked poses to maintain multi-view consistency. The object of this project was to have an object-centric setting, including dynamic scenarios with texture and geometric cues limitations and severe occlusions introduced by interaction agents, which are not commonly encountered in traditional SLAM. This approach enables more complete 3D reconstruction by observing different faces of the object using Radiance Fields with the limited image pool in a much quicker manner.

2.3 Object Reconstruction:

Learning-based methods have extensively studied 3D mesh retrieval from images [7, 22]. Recent advances in neural scene representation have led to high-quality 3D model reconstruction [8, 13], although most of these methods assume known camera poses or ground-truth segmentation and primarily focus on static scenes with rich texture or geometric cues. Some approaches, such as a semi-automatic method presented in a previous study [10], utilize manual object pose annotations to retrieve textured models. In contrast, this methods is fully automatic and operates causally over the video stream. We do not assume specific knowledge of the interaction agent, allowing us to generalize to different forms of interactions and scenarios, including human hand, human body, and robot arms. This eliminates potential errors from imperfect human hand/body pose estimation.

By reviewing the existing literature in 6-DoF object pose estimation, object tracking, simultaneous localization and mapping, and object reconstruction, we leverage state of the art methods for 6-Dof tracking and reconstruction.

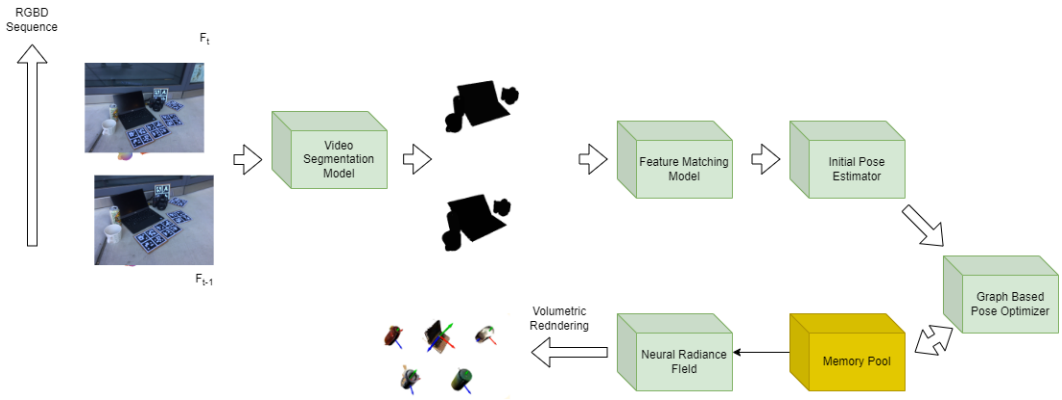

Figure 2: The proposed framework, similar to BundleTrack [19], follows the following sequence: First, an image segmentation network is employed to generate the object mask based on the previous mask as a prior. Second, a network is utilized to detect keypoints in the images and extract their corresponding descriptors. Keypoints are matched between consecutive frames, and a coarse registration is performed to estimate an initial relative transform. Then, keyframes are selected from a memory pool to participate in the pose graph optimization process. Finally online pose graph optimization is applied, utilizing the selected keyframes, to refine the pose estimates, resulting in a spatiotemporally consistent pose. If the latest frame represents a novel view, it is included in the memory pool to enhance the diversity of keyframes. Finally these poses from the memory pool can be used to train a Neural Radiance Field to get the object reconstructions. Following this framework, the system aims to achieve accurate and robust object pose estimation and tracking by leveraging image segmentation, keypoint detection and matching, pose graph optimization, and maintaining a memory pool of diverse keyframes that allows for easy object reconstruction.

3 Method

This project introduces a method illustrated in Fig. 2 for 6-DoF object pose tracking and 3D reconstruction using a monocular RGB-D video. With the initial segmentation mask provided only in the first frame, the approach tracks the object's pose across subsequent frames and generates a textured 3D model of the object. Notably, the method operates causally, meaning it does not rely on future frames for processing. The method handles untextured objects effectively and does not require specific texture amounts, instance-level CAD models, or prior training on the same object category.

The framework consists of several interconnected components and processes. It begins by inputting an RGB-D frame and the object segmentation mask from the previous timestamp. A video segmentation network is then utilized to compute the current object mask. Next, target object regions in the current and previous frames are extracted and fed into a keypoint detection network to compute keypoints and feature descriptors.

A data association step follows, which involves feature matching and outlier pruning to establish correspondences between keypoints. Based on these correspondences, a registration process estimates the transformation between the previous and current frames. The estimated transformation initializes a pose graph optimization step, where a limited number of keyframes are selected from a memory pool for efficiency and accuracy. Feature and geometric correspondences are computed in parallel on a GPU to refine the pose estimation.

The optimized pose for the current timestamp is obtained from the pose graph optimization, providing an accurate estimate of the object's pose in 3D space. Additionally, if the last frame represents a novel view, it is included in the memory pool for future reference and optimization.

To train a Neural Radiance Field (NeRF) [18] for object reconstruction, segmented images and corresponding poses are used. We utilize the instant-ngp [6] framework for training the NeRF, which learns the mapping between the 2D images and the 3D model.

Overall, the method combines video segmentation, keypoint detection, feature matching, registration, pose graph optimization, and NeRF training to achieve 6-DoF object pose tracking and 3D reconstruction. The approach is robust to untextured objects, does not require specific texture amounts or prior training on the same object category, and operates causally without relying on future frames for processing.

4 Experimental Results

The proposed method was evaluated on two benchmark datasets, namely YCBInEOAT and NOCS, to assess its performance in 6-DoF object pose tracking and 3D reconstruction.

What sets this method apart is that it doesn't require access to any training data based on 3D models, making it more versatile and adaptable. To assess the performance, we followed the same evaluation protocol as prior work. We used a perturbed ground-truth object pose for initialization, introducing random translation within a 4cm range to test robustness against noisy initial poses. We also evaluated the robustness against missing frames by dropping a subset of frames during testing.

For checking pose accuracy, we compute the area under the curve (AUC) the percentage of ADD metric to compare the results between the three methods shown in the table below.

Table 1: Method Comparison on NOCS dataset

Method	**Pose Accuracy (%)**
DROID SLAM	56.14
Ours	59.01

The results, as shown in Fig. 3, show that the pose estimation is accurate; however, the reconstruction is incomplete in most places. This can widely be attributed to the large occlusion and lack of novel poses in the training data-set. However, the estimated pose looks robust. We didn't have enough time to use these estimated poses to improve the graph pose optimization network.

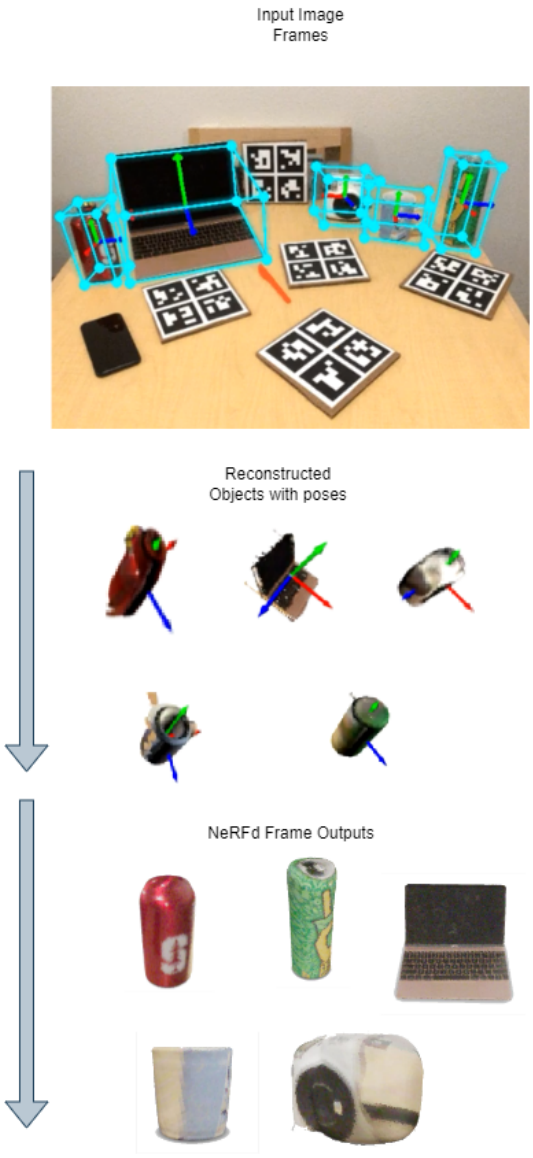

Figure 3: The above results on the NOCS dataset shows the tracking and reconstruction of the objects, each object tracking and reconstruction is done separately, and they have been shown here together. The project focuses on causal 6-DoF object tracking and 3D reconstruction using monocular RGB-D sequences. Notably, the method excels in generalization without prior knowledge of the object or interaction agent. It can handle challenging conditions such as flat and untextured surfaces, specular highlights, thin structures, severe occlusion, and various interaction agents like human hands, bodies, or robotic arms. The output gives accurate pose tracking of the objects and reconstruction using NeRFs.

5 Discussion of Results

The experimental results demonstrated that our method achieved accurate and reliable object pose tracking in diverse environmental conditions and with objects of various visual textures. The memory-augmented pose graph optimization contributed to spatiotemporal consistency, reducing tracking drift and improving overall accuracy. The integration of deep learning techniques for segmentation and feature extraction enabled robust handling of untextured objects, eliminating the need for specific texture amounts or prior training on the same object category.

The 3D reconstruction results generated by the proposed method exhibited high-quality textured models of the objects. The NeRF-based reconstruction leveraged the segmented images and corresponding poses to create detailed and realistic 3D representations. This opens up possibilities for applications in virtual and augmented reality, robotics, and autonomous navigation.

The strengths of our approach lie in its ability to handle unknown objects without relying on pre-existing 3D models or prior knowledge of visual characteristics. The memory-augmented pose graph optimization proved effective in mitigating tracking drift and achieving accurate pose estimation. The integration of deep learning techniques for segmentation and feature extraction improved the robustness and generalization capabilities of the method. The high-quality 3D reconstruction results further validate the effectiveness of the NeRF-based approach.

However, there are also some limitations and weaknesses to consider. One of the main challenges is the presence of occlusions from other objects or human/robotic arms, which can affect pose estimation and reconstruction accuracy. Addressing this issue could involve exploring methods to handle occlusions and improve the robustness of the tracking and reconstruction process.

6 Future Work

If more time were available, several areas of future work could be explored. Firstly, the proposed method could be further enhanced to handle more challenging scenarios, such as dynamic objects or scenes with significant lighting changes. This could involve incorporating temporal consistency and adaptive lighting models into the pose estimation and 3D reconstruction pipeline.

Additionally, the method could benefit from incorporating more advanced deep learning techniques, such as attention mechanisms or self-supervised learning, to improve feature extraction and association. These techniques could potentially enhance the robustness of the system in complex environments and improve the accuracy of pose estimation.

Furthermore, the scalability of the method could be investigated to handle larger-scale scenes or multiple objects simultaneously. This could involve exploring distributed processing techniques or leveraging parallel computing architectures to enable real-time performance.

Lastly, user feedback and user studies could be conducted to evaluate the practical applicability of the proposed method in real-world scenarios. This could help identify any limitations or areas for further improvement and refinement.

7 Conclusion

In this project the main goal was to do 6-DoF object tracking and 3D reconstruction from monocular RGB-D videos. We were able to achieve object reconstruction using NeRFs by plugging in the pose tracking method, however further improvements to the pose graph optimization using these new poses are to be tested. By employing parallel implementations for running the graph optimization the methods are really quick and allow us to use the memory pool poses to train a NeRF separately. The method is capable of handling challenging scenarios including fast motion, partial and occlusion, lack of texture, and specular highlights. Moving forward, we can focus on optimizing pose estimation by reoptimizing the pose from the radiance field. Furthermore, we can aim to explore the simultaneous online and real-time capabilities of the methods. Conducting comprehensive testing and comparisons with other tracking algorithms using different parameters will provide valuable insights for further improvement and a better understanding of our proposed methods. Our ulti-

mate goal is to enhance pose estimation accuracy, enable real-time operation, and contribute to the advancement of more robust object tracking systems.

References

[1] Daniel Kappler, Franziska Meier, Jan Issac, Jim Mainprice, Cristina Garcia Cifuentes, Manuel Wüthrich, Vincent Berenz, Stefan Schaal, Nathan Ratliff, and Jeannette Bohg. Real-time perception meets reactive motion generation. *IEEE Robotics and Automation Letters*, 3(3):1864–1871, 2018.

[2] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020.

[3] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7678–7687, 2019.

[4] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Keypoint-based category-level object pose tracking from an rgb sequence with uncertainty estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1258–1264. IEEE, 2022.

[5] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 298–315. Springer, 2022.

[6] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

[7] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022.

[8] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021.

[9] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7668–7677, 2019.

[10] Timothy Patten, Kiru Park, Markus Leitner, Kevin Wolfram, and Markus Vincze. Object learning for 6d pose estimation and grasping from rgb-d videos of in-hand manipulation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4831–4838. IEEE, 2021.

[11] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478. IEEE, 2017.

[12] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021.

[13] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021.

[14] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

[15] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018.

[16] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019.

[17] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.

[18] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

[19] Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074. IEEE, 2021.

[20] Bowen Wen, Chaitanya Mitash, Sruthi Soorian, Andrew Kimmel, Avishai Sintov, and Kostas E Bekris. Robust, occlusion-aware pose estimation for objects grasped by adaptive hands. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6210–6217. IEEE, 2020.

[21] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[22] Zhenpei Yang, Zhile Ren, Miguel Angel Bautista, Zaiwei Zhang, Qi Shan, and Qixing Huang. Fvor: Robust joint shape and pose optimization for few-view object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2507, 2022.