

Object and Contact Point Tracking in Demonstrations Using 3D Gaussian Splatting

Michael Büttner
Bielefeld University, Germany
mbuettner@techfak.uni-bielefeld.de

Jonathan Francis
Bosch Center for AI, USA

Helge Rhodin
Bielefeld University, Germany

Andrew Melnik
Bremen University, Germany

Abstract: This paper introduces a method to enhance Interactive Imitation Learning (IIL) by extracting touch interaction points and tracking object movement from video demonstrations. The approach extends current IIL systems by providing robots with detailed knowledge of both where and how to interact with objects, particularly complex articulated ones like doors and drawers. By leveraging cutting-edge techniques such as 3D Gaussian Splatting and FoundationPose for tracking, this method allows robots to better understand and manipulate objects in dynamic environments. The research lays the foundation for more effective task learning and execution in autonomous robotic systems.

Keywords: 3D Gaussian Splatting, Contact Points, Tracking

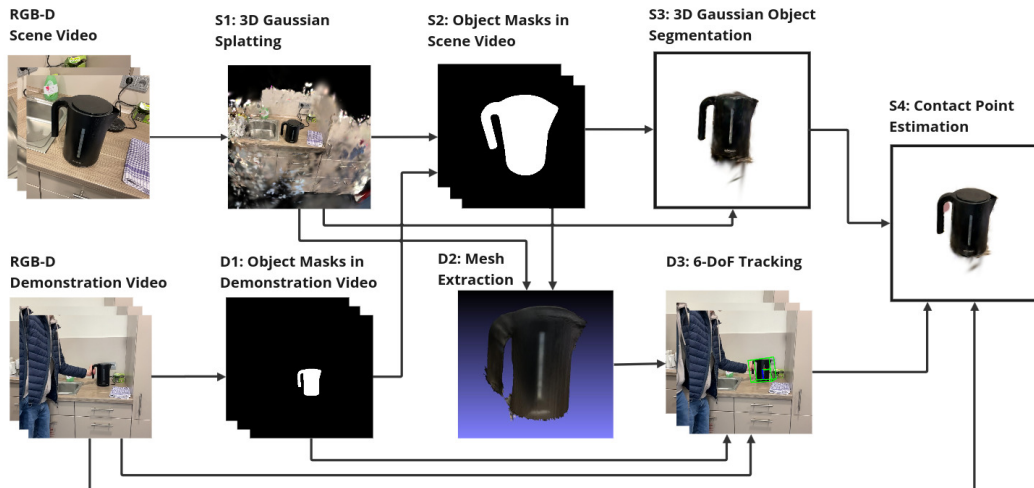


Figure 1: Overview of the pipeline. We start with RGB-D recordings of the scene and the demonstration. We train a 3D Gaussian Splatting [1] Scene on the scene video, and do object masking on the demonstration video using RAFT [2] and SAM 2 [3]. These masks are used to create object masks of the scene video, which in turn are used to create a mesh using GS2Mesh [4] and a Gaussian object segmentation using SAGS [5]. The mesh is used to do 6-DoF tracking with FoundationPose [6], which in turn is used to estimate contact points. Here, the mesh is visualized by MeshLab [7].

1 Introduction

Autonomous robotic systems have advanced significantly [8, 9], but challenges remain in manipulating novel objects and articulated structures like doors or cabinets. Robots must accurately identify,

grasp, and move objects within dynamic environments. Reinforcement Learning (RL) [10] and Imitation Learning (IL) [11] are widely used approaches, with RL often requiring extensive training and careful reward design, while IL benefits from expert demonstrations but struggles with generalization to unseen scenarios [12, 13]. Interactive Imitation Learning (IIL) improves IL by incorporating real-time human feedback [14, 15], allowing robots to adjust during execution, but still faces difficulties in tasks requiring nuanced object interaction [16].

To address this, we propose a method for extracting touch interaction points, or contact points, and tracking object movement from video demonstrations, providing robots with detailed information for manipulating articulated objects. By combining video-based learning with 3D Gaussian Splatting [1] for 3D scene reconstruction, we create a task-relevant representation of the environment, enabling improved robot performance in complex tasks. This method enhances the IIL framework by offering precise interaction data, laying the groundwork for more autonomous robotic manipulation.

2 Related Works

Recent advancements in robotic grasping leverage deep learning models to enable robots to interact robustly with objects in real-world settings. Among foundational approaches, Dex-Net 2.0 introduced a large-scale synthetic grasping dataset to train convolutional neural networks for grasp quality prediction, setting a baseline for efficient grasp point selection through simulated grasping attempts [17]. Building on these concepts, Contact-GraspNet expanded 6-DoF grasp prediction to real-world scenarios using RGB-D data, anchoring 3D points in the point cloud as potential contact points and simplifying the grasp representation into a 4-DoF framework for streamlined inference [18]. Our work builds on these methods by integrating multi-stage estimation for 6-DoF pose tracking and contact prediction, designed to support seamless interaction within a unified, sequential pipeline.

Affordance-based modeling has also provided valuable insights into robotic contact and hand pose estimation. For instance, HRP (Human Affordances for Robotic Pre-Training) annotates "affordance labels" (e.g., 2D contact points, hand poses, and active objects) by analyzing human-object interactions, which enables predictive models for robotic behavior cloning [19]. Leveraging models like the 100DOH hand-object detector [20] and FrankMocap [21], HRP improves hand-object contact point labeling for robotics, showing the potential of human-inspired affordance prediction for grasp and contact point accuracy. While HRP primarily addresses video-based affordance prediction in 2D images, our approach extends this by integrating contact detection into a 3D reconstruction of the object, enhancing robustness to occlusions and object rotations.

Additionally, Gaussian Splatting has gained traction in robotic applications by enhancing photorealism and creating interactive 3D models in real-time environments. While SplatSim [22] uses Gaussian Splats in simulators to achieve realistic visual transfer for Sim2Real training, Physically Embodied Gaussian Splatting integrates Gaussian splats with physical priors, dynamically reconstructing 3D environments to enable robotic adaptation to scene changes [23]. In our pipeline, Gaussian Splatting supports updates to 3D object positions, assisting with pose estimation and maintaining model integrity.

3 Methods

The basic input consists of two RGB-D videos shot using the Spectacular Rec app [24]. The first video is dynamic, capturing the scene from multiple viewpoints with a focus on the object to be manipulated. The second video, called the demonstration video, is a static shot of a human manipulating the object from a fixed camera position. These videos allow us to perform 3D Gaussian Splatting [1] to reconstruct the scene and track the object's 6-DoF pose using FoundationPose [6]. Contact points are identified through the depth images and object poses. An simple overview can be

found in Figure 1 and a detailed overview can be found in Figure A.1 in the Appendix. Specifically, the algorithm proceeds as follows:

1. In the first step, we use Spectacular AI client to obtain the camera poses and a sparse point cloud from the dynamic scene video. These estimations are then combined with the original 3D Gaussian Splatting training to reconstruct a Gaussian Splatting model of the environment, including the object of interest (Figure 1 S1).
2. In the second step, we export the demonstration video into a format that can be used by FoundationPose. As this tracking method requires object masks for accurate tracking, we first find a bounding box of the manipulated object using RAFT optical flow [2], as shown in Figure A.2 in the Appendix. We accumulate potential bounding boxes over the video and use a clustering mechanism to find the most probable bounding box. This is used as an input for SAM 2 [3] to generate these masks (Figure 1 D1).
3. Next, we create masks of the object from the perspective of the scene camera poses (Figure 1 S2), which can be obtained using the mask from the demonstration video perspective, as well as its camera pose in the scene, which can be obtained using COLMAP [25, 26]. The process can be seen in Figure A.3 in the Appendix.
4. With these masks, we create a mesh of the object using GS2Mesh [4] (Figure 1 D2) and a segmentation of the Gaussians using SAGS [5] (Figure 1 S3).
5. After obtaining the mesh, FoundationPose is used to track the object’s 6DoF pose in camera space (Figure 1 D3). These poses are translated into the scene space for further use. Using the object’s poses and the segmentation obtained by SAGS, the object’s trajectory can be visualized in the GS scene. Inaccurate demonstration camera pose estimation can be corrected by shifting the object’s estimated starting position to its position in the 3DGS scene.
6. By utilizing depth images and hand mask data from the demonstration video, along with simple distance thresholding, contact points on the object can be identified, as seen in Figure 2. We accumulate these points for up to 10 frames of contact and select the most occurring points as the final contact points (Figure 1 S4).

The processing was carried out on an NVIDIA RTX 3060 Ti GPU of 8GB VRAM. Due to the GPU’s computational constraints, the chosen models balance accuracy and memory efficiency.

4 Experimental Results

The video material used in this experiment was recorded utilizing an iPad Pro (11”, 3rd Generation) to capture human-object interaction sequences. The primary focus was to track human touch points on a stationary object during manipulation. The videos were recorded in a controlled environment with consistent lighting, and the camera was positioned at a fixed angle. The video resolution was initially set at 1920x1440 pixels and later reduced to 960x720 pixels to speed up processing without significant loss of detail. Manipulations of 12 different objects were recorded, including articulated objects and portable objects. These videos were recorded to test out the successes and edge cases of this approach. The recordings introduced various challenges to assess edge cases:

- The jug and mug, which have metallic surfaces, present difficulties due to reflections.
- The jug, mug and the pincher, also have smooth, textureless surfaces, which adds complexity.
- The detergent bottle includes a transparent section, making its demonstration video more challenging, especially with significant hand occlusion.
- The detergent bottle demonstration also includes a large occlusion, as the demonstrator’s hand covers a large section of the bottle.
- The shoe and the brush demonstrations include heavy rotations.

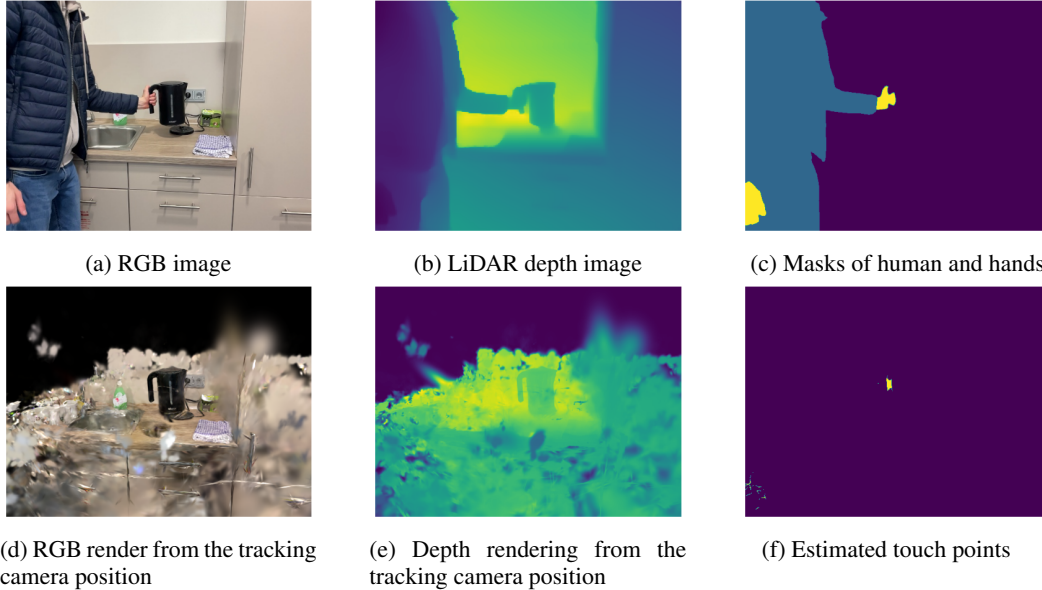


Figure 2: In order to estimate touch points, we calculate the absolute difference between the depth image of the LiDAR camera and the rendered depth image, threshold it, and apply the hand mask that was generated using Grounding DINO [27] + SAM 2 [3]. This is done for the first 10 frames of contact and then accumulated to find the most probable points.

- The shoe demonstration also shows the shoe’s underside, which is not seen in the scene.
- The cookies box presented a scenario with large displacement, as it was moved far away from the camera during the demonstration.
- The fridge demonstration video includes a person walking in front of the fridge to test partial occlusion.
- The fridge is not fully visible, as the top end of the fridge is outside the camera’s view.
- The cabinet demonstration video has the objects on the counter rearranged compared to the scene video.

Each component of the pipeline was evaluated separately to assess individual performance and isolate failure points. Since models operate sequentially, inaccuracies in earlier stages could potentially propagate through the pipeline and affect subsequent outcomes. To mitigate this, we performed manual interventions when possible, enabling isolated evaluation of each step. For 6-DoF pose estimation using FoundationPose, we assessed two approaches: (1) standard pose tracking, which relies on the previous frame’s estimated pose for current pose predictions, and (2) per-frame pose estimation, which independently estimates poses in each frame using object masks. Although standard tracking achieves near-real-time performance, per-frame estimation incurs an 11x slowdown, trading off speed for accuracy. Given the lack of ground-truth object poses, contact point estimations were derived from estimated poses. We calculated the cumulative success by combining the steps from the previous approaches.

The method produced the most reliable results when applied to large objects with visually distinct features on all sides, such as kettles and cappuccino canisters, as this aids in more consistent pose estimation across frames. Similarly, articulated objects with large flat surfaces, such as fridges or dishwashers, allowed for smoother tracking. These characteristics helped minimize ambiguity in object recognition, leading to more accurate results. Contact point estimation yielded equivalent results on both pose tracking and per-frame pose estimation. Even among episodes where tracking was lost shortly after the manipulation starts, correct contact points were estimated. Visual examples can be found in Figure 3.

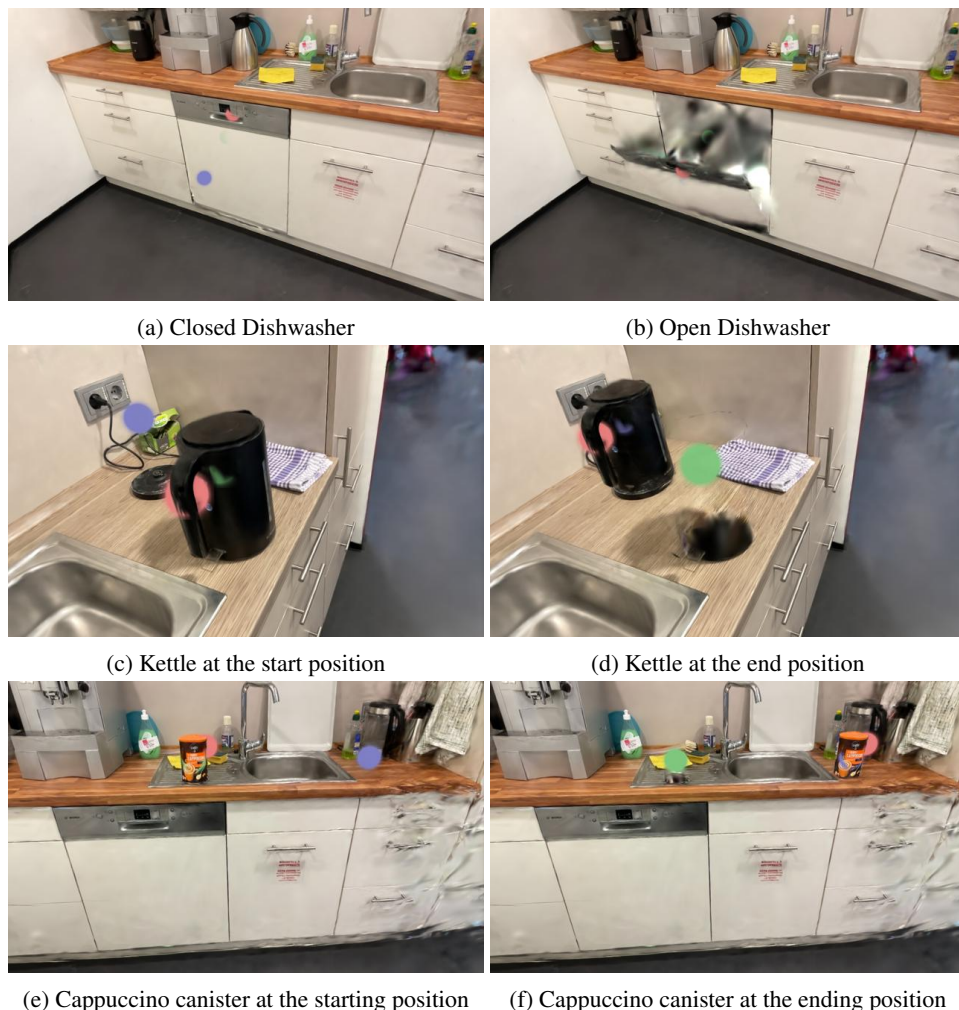


Figure 3: Visuals for successful tracking and contact point estimation episodes. The red spheres stand for the identified contact points, the green spheres for the starting positions, and the blue spheres stand for the end position.

However, since multiple models are applied sequentially, errors or imprecisions in one stage can propagate through the pipeline and negatively impact subsequent steps. While 3D Gaussian Splatting combined with Spectacular AI’s pose estimation is generally robust, light reflections on surfaces, such as on the jug and the mug, can introduce artifacts in the texture of the mesh, which in turn affect the accuracy of the 6DoF pose estimation. When applying RAFT to the demonstration video, the model had difficulty detecting motion on plain surfaces and frequently produced false positives when shadows were present, although outliers in single frames could be filtered out by the bounding box clustering approach. This approaches struggled with doors such as the dishwasher door, as there is less movement on the section near the hinges, leading to smaller bounding boxes. Even when bounding boxes were correctly defined, SAM 2 occasionally failed to consistently generate object masks, with the segmentation breaking down at certain time steps. In the fridge scene, SAM 2 would segment the fridge door’s surface while leaving out the handle, which is a critical omission later in the pipeline when contact points are estimated. Even manual input annotation would not lead to successful cases, where both the door and the handle are segmented. GS2Mesh, which uses stereo vision for depth prediction, sometimes produced faulty depth maps on textureless or reflective surfaces, leading to unnatural deformations or dents in the mesh that could disrupt the 6DoF tracking.

Furthermore, SAGS’s voting system for determining which Gaussians belong to the object requires the entire object to be within the camera’s field of view at each frame; otherwise, parts of the object may be excluded, which happened with the upper section of the fridge door. Conversely, relaxing this restriction to give Gaussians outside the camera’s frustum neutral votes results in the inclusion of irrelevant Gaussians far from the object.

FoundationPose struggled with objects that lacked distinct features on all sides, such as a jug with a handle. In such cases, it often misestimated the object’s orientation, rotating the handle out of view, which creates problems for touch point prediction. Using standard pose estimation, tracking was lost for smaller objects such as the mug, the pitcher and the brush. While all other objects have accurate estimations before manipulation, the brush’s estimated position deviated immediately from its true position. The resulting estimations appeared to float in midair. Similarly, occlusions was an issue for standard pose tracking, as the fridge’s estimated pose shifted to the right during partial occlusion and the detergent’s pose was lost entirely during manipulation. Using per-frame pose estimation improved the 3D position estimation of these objects, but rotation estimation accuracy declined, as some frames received erratic shifts in their rotation. Large distances to the camera also become a problem in the cookies box scene. While pose estimation is successful for the first half of the video, it declined in the second half as standard pose tracking lost track of the object, while per-frame pose estimation results jumped far into the background. In the shoe scene, tracking was accurate before the underside of the shoe was shown. Afterwards, the 3D position estimation remained close to the object while rotation accuracy varied. For transferring the pose estimation into the Gaussian Splatting scene, COLMAP was used. While not shown in this table, COLMAP failed in environments with insufficient keypoints or when the demonstration camera deviated significantly from the scene camera’s trajectory.

Lastly, while pose estimation was successful for most portable objects, articulated objects received mixed results. While the dishwasher’s contact point was correctly placed on its recessed handle, the other articulated objects’ contact points were placed on the surface near the handle.

When combining these steps, we can evaluate the overall success of the approach. Considering a ”High-Accuracy Success”, where both 3D position and rotation accuracy are highly accurate, two scenes are successful, the kettle scene and the cappuccino scene, resulting in a success rate of 17%. Here, 4 scenes fail as early as the masking stages, 5 scenes at the pose tracking stage and 1 at the contact point estimation stage. When considering a ”Low-Accuracy Success”, where rotation accuracy isn’t considered, the jug, mug and detergent scenes are included, raising the success rate to 42%.

5 Discussion

This method offers an innovative approach for extracting contact points and object trajectories from video demonstrations, enabling robots to manipulate articulated and portable objects like dishwashers and kettles.

However, there are technical and practical limitations to consider when applying this approach. FoundationPose yields state of the art results in 6DoF pose estimation tracking but relies heavily on a high-quality mesh reconstructions, which are often challenging to generate. Factors such as object size, texture, shape, distance from the camera, light reflections, and computational power influence the potential quality of the mesh; all of these factors are prominent in real-world environments. The human user would need to ensure minimal occlusion of the object during demonstration, while ideally moving the object slowly and only showing parts of the object viewable in the scene video. This becomes a problem for articulated objects like doors, that move their outer surface outside the camera’s view while being opened. Given enough computing power, future work could employ BundleSDF [28] to simultaneously track and reconstruct the mesh. The effectiveness of the approach also depends on specific object characteristics, with the approach performing best on rigid, textured objects while presenting challenges for small, textureless or deformable objects.

Table 1: Success for individual components of the pipeline, as well as a combination of all components. "High-Accuracy Success" indicates precise 6DoF pose tracking, while "Low-Accuracy Success" accepts moderate positional alignment without regard for rotational accuracy.

	Fridge	Cabinet	Dish-washer	Kettle	Cappuccino	Pitcher	Cookies Box	Shoe	Jug	Mug	Detergent	Brush	Success Rate
Step-Specific Success													
Dem. Obj. Masking	X*	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	10/12 (83%)
Scene Obj. Masking	(X)	✓	✓	✓	✓	✓	X**	✓	X	✓	✓	✓	9/12 (75%)
Pose Tracking													
Position	✓-	✓	✓	✓	✓	✓	✓X	✓X	✓-	✓	X	X	7/12 (58%)
Rotation	✓	✓	✓	✓	✓	✓	✓X	✓X	✓X	X	X	X	5/12 (42%)
Per-Frame Pose Est.													
Position	✓	✓	✓	✓	✓	✓	✓X	✓X	✓	✓	✓	✓	11/12 (92%)
Rotation	✓X	✓	✓	✓-	✓	✓	✓X	✓X	✓X	X	X	✓X	4/12 (33%)
Contact Point Est.	X	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	9/12 (75%)
Overall Success													
High-Acc. Success w/ Pose Tracking	X	X	X	✓	✓	X	X	X	X	X	X	X	2/12 (17%)
Low-Acc. Success w/ Per-Frame Pose Est.	X	X	X	✓	✓	X	X	X	✓	✓	✓	X	5/12 (42%)

- ✓ success
- ✓- success, but with less accuracy
- ✓X mixed results, with both accurate and highly inaccurate estimations
- X failure
- (X) masks capture the object while leaving out the handle
- X* segmentation mask appears incorrect for the first few frames, but remains correct for the frames after
- X** no masks were generated for earlier frames of the demonstration, but only for later frames

Bounding box estimation presents another challenge. While manual assignment is more effective, automatic detection is the preferred solution. Existing tools like Grounding DINO [27] and the "100 days of hands" hand-object detector [20] struggle with distinguishing between moving and static parts of articulated objects, often encompassing stationary sections in the bounding box. This limitation significantly impacts grasp detection in complex objects like doors or appliances. For that reason, we chose an optical flow approach to determine the bounding box, as it focuses only on the moving parts of objects. However, this approach faces other challenges with articulated objects, as reducing the motion threshold to account for movement near hinges increases its susceptibility to noise. In our tests it failed on the one articulated object that generated correct contact points.

Additionally, the system currently requires several minutes per video for processing steps such as masking, mesh extraction, and pose tracking, especially if per-frame pose estimation is used. This processing time makes real-time application challenging, which may limit its immediate feasibility in scenarios requiring rapid response. However, if meshes are pre-computed and standard pose tracking is employed, near real-time application is possible. Otherwise, improvements in algorithmic efficiency or access to more powerful hardware would be necessary for real-time application on mobile robots.

Environment and lighting conditions also pose limitations, as changes in lighting, reflective surfaces, and cluttered scenes can interfere with depth estimation quality and consequently mesh extraction, impacting downstream processes such as pose tracking and object manipulation. However, as long as an accurate demonstration camera pose is maintained, variations in the placement of background objects between the real-world scene and the 3DGS environment, and potentially different lighting conditions, do not affect tracking performance.

Since multiple models are applied sequentially, the pipeline is subject to error propagation, where inaccuracies in early stages, such as camera pose estimation or object masking, can compound in later stages. These accumulated errors can lead to significant deviations in pose tracking and contact point prediction, limiting the overall reliability of the approach in complex scenarios.

Our system identifies contact points by combining hand masks and depth images. This approach works well with portable objects, as the hand generally avoids occluding other surfaces of the object. However, with larger objects like doors, the hand must cover a substantial portion of the object's

surface to reach handles, leading to occasional misalignment between depth data and masks. This misalignment can create inaccurate contact points, which happened in the cabinet scene. Additionally, our contact point predictions are limited to the visible sections of objects, posing a challenge for robotic applications that require precise, encompassing grasps. For example, when attempting to grasp a cylindrical object like a bottle, predicted contact points may only cover one side, often occluded by the hand, whereas successful grasps require more complete object coverage. Future work could address these challenges by incorporating hand pose estimation to infer occluded areas and establish contact points on parts of the object currently out of view. Additionally, successful application of these contact points assumes precise end-effector calibration on the robot; any misalignment could further reduce grasping accuracy.

6 Applicability

To apply this approach on a mobile robot, a compatible RGB-D camera with IMU recording and a gripper arm would be essential for consistent tracking and object manipulation. The robot could first explore its environment to create a 3D Gaussian Splatting scene, which would serve as a digital twin environment. This mapping would allow the robot to not only locate objects, but also orient itself within the environment. This self-localization could potentially bypass the need for COLMAP for demonstration camera pose estimation, enhancing both efficiency and accuracy when tracking its own movement relative to the objects. Additionally, a photorealistic 3DGS reconstruction of the environment enables the robot to infer computer vision models on parts of the environment outside of its current location.

When tasked with actions like opening a door, the robot could analyze human demonstration data to determine where to grasp and to trace the object’s trajectory for accurate replication. With accurate 6-DoF pose estimation, the robot could also interpret an object’s intended orientation (e.g., ensuring a mug remains upright during transport) and refine its manipulation skills, such as rotating a pitcher to pour liquid into a mug. Moreover, it could track an object’s location across sessions, which would allow it to retrieve items or monitor changes in a dynamically rearranged environment - critical for mobile robotics in dynamic, real-world settings.

7 Conclusion

This paper presents a novel approach utilizing Gaussian Splatting to infer grasp points and object trajectories from video demonstrations, providing a robust framework for robots to interact with both articulated and portable objects in real-world environments. By reconstructing dynamic scenes and identifying critical touchpoints, this method could enable robots to manipulate everyday items, such as doors and appliances, offering valuable applications in home automation and service robotics.

Despite its promise, the method faces challenges, including reliance on high-quality mesh reconstructions, accurate depth estimations, and bounding box inaccuracies. Addressing these limitations - particularly in automatic object segmentation and mesh generation techniques, while keeping a low computational overhead - is crucial for broader practical applications.

Nevertheless, the potential of this approach is clear. As advancements in scene reconstruction, point tracking, and pose estimation continue to develop, this method could significantly improve robotic autonomy and efficiency. Future work should focus on refining these components and exploring new algorithms that minimize computational complexity, allowing for more effective real-time interaction. Ultimately, this approach opens new avenues for research in robotic manipulation and interaction, pushing the boundaries of what is achievable in home robotics today.

Acknowledgments

This work was supported by DAAD PPP-USA program under project number 57651674.

References

- [1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL <https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/>.
- [2] Z. Teed and J. Deng. RAFT: recurrent all-pairs field transforms for optical flow. *CoRR*, abs/2003.12039, 2020. URL <https://arxiv.org/abs/2003.12039>.
- [3] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL <https://arxiv.org/abs/2408.00714>.
- [4] Y. Wolf, A. Bracha, and R. Kimmel. Gs2mesh: Surface reconstruction from gaussian splatting via novel stereo views. *arXiv preprint arXiv:2404.01810*, 2024.
- [5] X. Hu, Y. Wang, L. Fan, J. Fan, J. Peng, Z. Lei, Q. Li, and Z. Zhang. Semantic anything in 3d gaussians. *arXiv preprint arXiv:2401.17857*, 2024.
- [6] B. Wen, W. Yang, J. Kautz, and S. Birchfield. FoundationPose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024.
- [7] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. MeshLab: an Open-Source Mesh Processing Tool. In V. Scarano, R. D. Chiara, and U. Erra, editors, *Eurographics Italian Chapter Conference*. The Eurographics Association, 2008. ISBN 978-3-905673-68-5. doi:10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136.
- [8] S. Yenamandra, A. Ramachandran, M. Khanna, K. Yadav, J. Vakil, A. Melnik, M. Büttner, L. Harz, L. Brown, G. C. Nandi, et al. Towards open-world mobile manipulation in homes: Lessons from the neurips 2023 homerobot open vocabulary mobile manipulation challenge. *arXiv preprint arXiv:2407.06939*, 2024.
- [9] A. Melnik, M. Büttner, L. Harz, L. Brown, G. C. Nandi, A. PS, G. K. Yadav, R. Kala, and R. Haschke. Uniteam: Open vocabulary mobile manipulation challenge. *arXiv preprint arXiv:2312.08611*, 2023.
- [10] N. Bach, A. Melnik, M. Schilling, T. Korthals, and H. Ritter. Learn to move through a combination of policy gradient algorithms: Ddpg, d4pg, and td3. In *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part II 6*, pages 631–644. Springer, 2020.
- [11] F. Malato, F. Leopold, A. Melnik, and V. Hautamäki. Zero-shot imitation policy via search in demonstration dataset. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7590–7594. IEEE, 2024.
- [12] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL <http://incompleteideas.net/book/the-book-2nd.html>.
- [13] B. Fang, S. Jia, D. Guo, M. Xu, S. Wen, and F. Sun. Survey of imitation learning for robotic manipulation. *International Journal of Intelligent Robotics and Applications*, 3(4):362–369, Dec 2019. ISSN 2366-598X. doi:10.1007/s41315-019-00103-5. URL <https://doi.org/10.1007/s41315-019-00103-5>.
- [14] S. Milani, A. Kanervisto, K. Ramanauskas, S. Schulhoff, B. Houghton, S. Mohanty, B. Galbraith, K. Chen, Y. Song, T. Zhou, et al. Towards solving fuzzy tasks with human feedback: A retrospective of the miner! basalt 2022 competition. *arXiv preprint arXiv:2303.13512*, 2023.
- [15] Y. Mikami, A. Melnik, J. Miura, and V. Hautamäki. Natural language as polices: Reasoning for coordinate-level embodied control with llms. *arXiv preprint arXiv:2403.13801*, 2024.

- [16] C. Celemin, R. Pérez-Dattari, E. Chisari, G. Franzese, L. de Souza Rosa, R. Prakash, Z. Ajanović, M. Ferraz, A. Valada, and J. Kober. Interactive imitation learning in robotics: A survey, 2022. URL <https://arxiv.org/abs/2211.00600>.
- [17] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. 2017.
- [18] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. 2021.
- [19] M. K. Srirama, S. Dasari, S. Bahl, and A. Gupta. Hrp: Human affordances for robotic pre-training. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.
- [20] D. Shan, J. Geng, M. Shu, and D. Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020.
- [21] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021.
- [22] M. N. Qureshi, S. Garg, F. Yandun, D. Held, G. Kantor, and A. Silwal. Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting, 2024. URL <https://arxiv.org/abs/2409.10161>.
- [23] J. Abou-Chakra, K. Rana, F. Dayoub, and N. Suenderhauf. Physically embodied gaussian splatting: A realtime correctable world model for robotics. In *8th Annual Conference on Robot Learning*, 2024. URL <https://openreview.net/forum?id=AEqOonGrN2>.
- [24] O. Seiskari, P. Rantalankila, J. Ylilammi, V. Kaatrasalo, A. Solin, and J. Kannala. Visual positioning and navigation software for ar & robotics. <https://www.spectacularai.com/>, 2023. Accessed: 2024-10-06.
- [25] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [26] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.
- [27] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [28] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In *CVPR*, 2023.

A Appendix

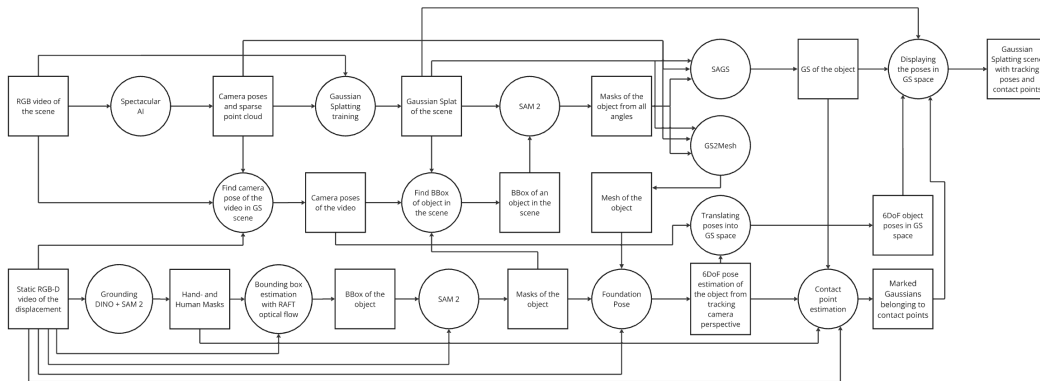


Figure A.1: Detailed overview of the process. Rectangles indicate data while circles indicate processes.

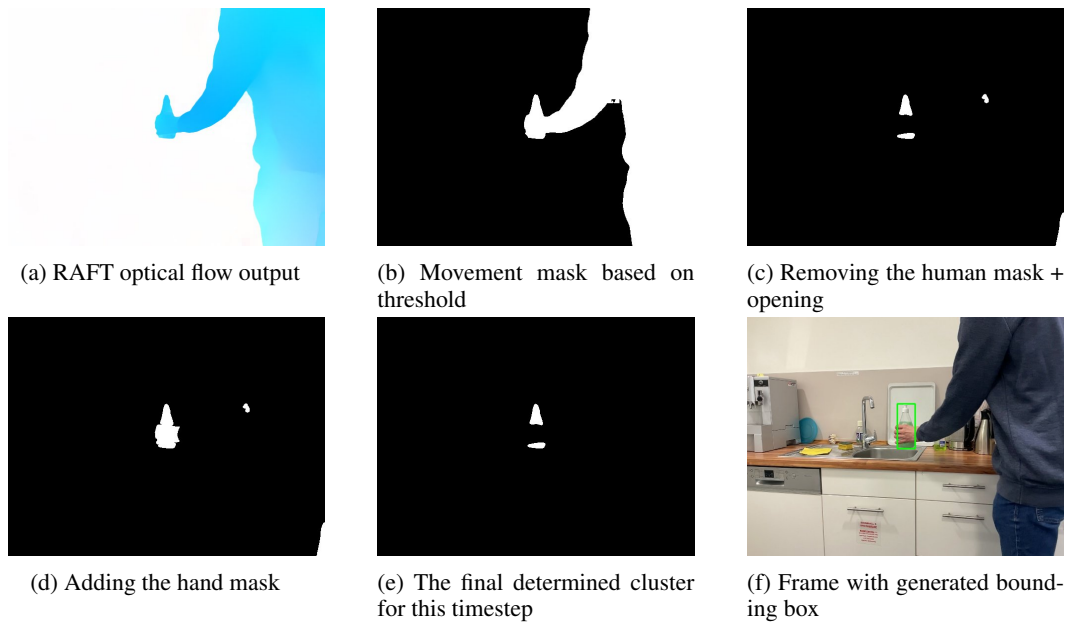


Figure A.2: In order to find the bounding box of the moving object, we use RAFT optical flow on all the frames of the demonstration video with a stride of 6, remove the human from the optical flow mask, cluster the bounding boxes by size and choose the frame whose bounding box is closest to the center of the biggest cluster. The human is filtered out using a mask generated by Grounding DINO [27] and SAM 2 [3].



(a) First frame of the demonstration video, with mask applied to it



(b) Kettle rendered from the closest scene camera view

Figure A.3: In order to find masks of the object in the scene, we take the first frame of the demonstration video and insert it into the scene video trace at the closest camera position. The bounding box from the demonstration video's object mask is taken as an input. To ensure closer alignment to the underlying representation, the scene frames are rendered from the 3DGS scene instead of taken from the video.