Incorporating Foundation Model Priors in Modeling Novel Objects for Robot Instruction Following in Unstructured Environments

Moksh Malhotra, Aman Tambi*, Sandeep S. Zachariah*, P. V. M. Rao and Rohan Paul
Indian Institute of Technology Delhi, India

I. Introduction

This paper addresses the problem of acquiring rich object models to support high-level task execution in unstructured *a-priori* unknown environments. Consider tasking a robot to explore an *a-priori* unknown area as follows: *"Robot, explore the path in front, which objects are there on the ground?", "Commander, I see a truss, a pipe, a crate and a tree branch on the ground", "Robot, clear away the pipe and the tree branch to the sides"*. A core ingredient of such an interaction is a metric-semantic representation of the environment that includes knowledge of which objects are present and maintain their geometric extent and position; enabling future interactions such as moving a specific object of interest. Following an instruction such as "clearing away a tree branch" involves grounding which object, determining how to grasp as well as determining how to transport it (without collisions with other objects) and safely place them in free regions and correcting for any pose errors while placing. As opposed to small objects that fit within the form closure of the gripper, interactions with large objects with *a-priori* unknown geometries (e.g., a truss, a tree branch or a pipe) requires significant reasoning in each phase of grasping, transport and placement. Hence, for a generalist manipulation agent capable of performing a range of sequential interactions with novel objects requires a mechanism to rapidly form a metric and semantic representation of the environment online in a zero-shot manner without explicit prior training.

In this work, we propose an approach that sequentially fuses prior knowledge from pre-trained foundation models with raw point clouds acquired online to arrive at rich object models to facilitate sequential manipulation tasks. Our approach is structured as follows. First, we leverage pre-trained VLMs with coarse-to-fine prompts to arrive at object masks which can be fused from multiple views to extract object geometry. Next, we observe that directly projecting the object masks on the 3D point cloud results in irregular/noisy object geometries which leads to poor grasp proposals. Hence, we incorporate depth priors from foundation models trained for monocular depth estimation via guided filtering over the 3D point cloud. We observe that the resulting models possess higher geometric accuracies and lead to more reliable grasps as well as robust collision-free transport (avoiding other objects) in the scene. Finally, in order to support sequential interactions, we incorporate a rapid method to locally build the map region where the object was extracted from (often revealing other objects underneath) as well as accurately model the region where the object was placed (also correcting for pose due to errors during transport and placing). Experimental evaluation demonstrates high-quality acquisition of 3D object models in relation to alternate approaches. Project webpage: https://reail-iitdelhi.github.io/3DObjectModels.github.io/

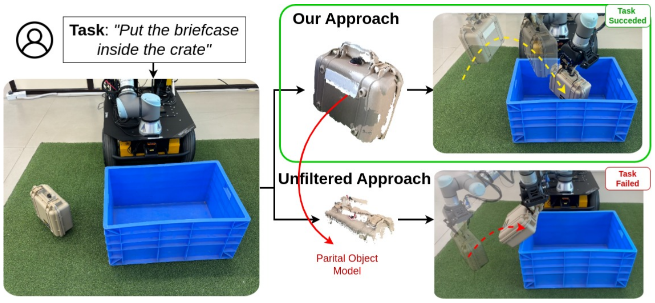

Fig. 1. Highlighting the importance of extracting object models for instruction following. Directly extracting object models from unfiltered point cloud results in collisions. Our approach utilizes depth priors from foundational models in combination with multi-view reconstruction resulting in a higher goal-reaching rate.

II. Related Work

Traditional efforts for object modeling for outdoor manipulation rely on the availability of a prior geometric model which is aligned with the 3D point cloud acquired yielding the object pose and a segmented out object [1], [4]. Others attempt to infer objects directly from a SLAM process by modeling associations between features typically found on common objects [11], [10]. Their success is primarily in settings where objects are well separated and application to setting where objects can be contained within or supported by other objects are more challenging to deal with. Finally, other efforts attempt to identify objects by directly segmenting or explicitly recognizing certain classes within the 3D point cloud based on geometric feature similarly [6]. The key limitation of the former approach is the absence of language alignment (preventing future referring interactions) and the former suffers from generalization beyond the class distribution the model is trained on.

Recent emergence of large visual and language models led to a number of efforts exploring the zero-shot capability of such models. A popular approach is to leverage vision-language models to form object masks in the 2D visual space and then project the masks into 3D forming an object models.

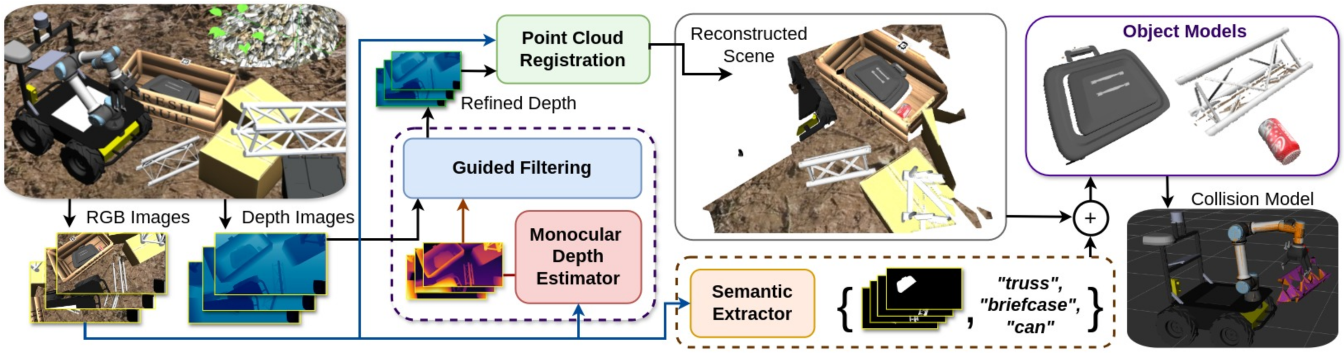

Fig. 2. The proposed method sequentially builds the global point cloud using a sequence of posed RGBD images. A guided filter employing depth priors from foundation models is used to refine the noisy depth images. The semantic extractor detects all objects and generates the segmentation mask for each object, which is fused with the global point cloud to extract the 3D model of the objects.

Wherein, leading efforts in this category [3] show zero-shot grounding of instructions to the 3D point cloud for object fetching tasks the resulting object models are limited to small regular shaped objects and do not further consider object transfer or updating the map when the object moves. Other efforts focus on learning priors to facilitate object interaction such as grasping [2], placing [7].

Our approach employs VLMs as the front end to provide object labels and masks and fuses depth priors for zero-shot acquisition of object models. Classical methods such as ICP [12] are used for multi-view scene reconstruction and local scene repair during sequential task execution. Subsequent tasks such as grasping, motion planning and placement use the proposed model.

III. Technical Approach

Consider the scenario where the robot workspace is populated by an *a-priori* set of objects $o \in \mathcal{O}$. The robot is instructed by a human to perform a task as given by the natural language instruction $\lambda \in \Lambda$. The robot is equipped with an eye-in-hand camera capable of capturing RGB images $\mathcal{I}_{o:t}$ and depth images $\mathcal{D}_{0:t}$. The robot's overall task is to synthesize the sequence of semantic actions $\pi = [a_0, a_1, ..., a_n]$ such as picking, transporting and placing of objects for a given instruction.

Objects can have a large geometric extent and possess complex intra-object relationships. E.g., an object can be positioned inside, supported by or lean over an another object. For the safe transport (without colliding with other objects) of these objects in the workspace, an accurate 3D model of the objects is required. However, directly using the point clouds does not suffice for an accurate 3D model due to the limited visibility of the workspace. Hence, we propose a pipeline to estimate the model of the objects $\mathcal{M} = \{m_j\}_{j=1...N}$ directly from the point cloud of the scene $\mathcal{Z}_t$ in a zero-shot manner. Each object in the scene consists of a 3D model in point cloud space (z_j), $SE(3)$ pose in the global frame (θ_j), a semantic label (l_j), and grasp-poses (ϕ_j) respectively as given by $m_j = \langle z_j, \theta_j, l_j, \phi_j \rangle$.

Semantic Mask and Label Extraction: The first stage in the pipeline is to extract the semantics of the object $\mathcal{S} = (s_j, l_j)$, where s_j is the segmentation mask. For this, a stage-wise method is employed, where the objects are first detected followed by the extraction of segmentation masks. Object detection is accomplished through a combination of a Visual Question Answering model (GPT-4V) and a phrase grounding model [9]. A class-agnostic foundational model-based segmentor [8] is used to extract the masks.

$$\mathrm{SemanticExtraction}(\mathcal{I}_{\mathrm{0:t}}) \rightarrow \mathcal{S}_t \qquad (1)$$

Model Refinement using Depth Priors: To remove noise from raw depth images induced due to reflections, textureless surfaces, we employ a custom guided filtering approach $\mathrm{GuidedFilter}(\mathcal{G}_i, \mathcal{D}_i)$ built on [5] which takes as input a guidance image $\mathcal{G}_i$ and the noisy depth data $\mathcal{D}_i$ to produce $\mathcal{D}_i^{'}$. $\mathcal{D}_i^{'}$ is a depth image that exhibits less noise compared to $\mathcal{D}_i$ while also inheriting structural information from $\mathcal{G}_i$. $\mathcal{G}_i$ is a high-fidelity relative depth map that is estimated from monocular depth estimation foundation models [13].

3D Object Model Extraction: Using the global point cloud $\mathcal{Z}_t$, obtained through point cloud registration [12], and the semantics $\mathcal{S}_t$, the object model $\mathcal{M}_t$ is extracted.

$$\mathrm{ObjectModelExtraction}(\mathcal{Z}_{\mathrm{t}}, \mathcal{S}_{\mathrm{t}}) \rightarrow \mathcal{M}_t \qquad (2)$$

For each $m_j \in \mathcal{M}_t$, a coordinate frame is attached at the geometric centroid, and the transformation between this frame and the global frame is represented by $SE(3)$ pose θ_j. Additionally, grasp poses ϕ are computed for each object by a grasp generator $\mathrm{Grasp}(z_j)$, optimizing the pipeline for multiple interactions.

Local Scene Update Post Action Execution: A Large Language Model (LLM)-based planner is utilized to synthesize the plan π for the provided language instruction λ. During the execution of the plan, the object model needs to be updated rapidly. This is accomplished by estimating the object's placement location $\theta_j^{'}$, assuming a rigid transformation between the end-effector and the object once the object is grasped, considering the robot kinematics. The object's 3D model p_j is also updated using the locally updated point cloud $\mathcal{Z}_{t+1}$.

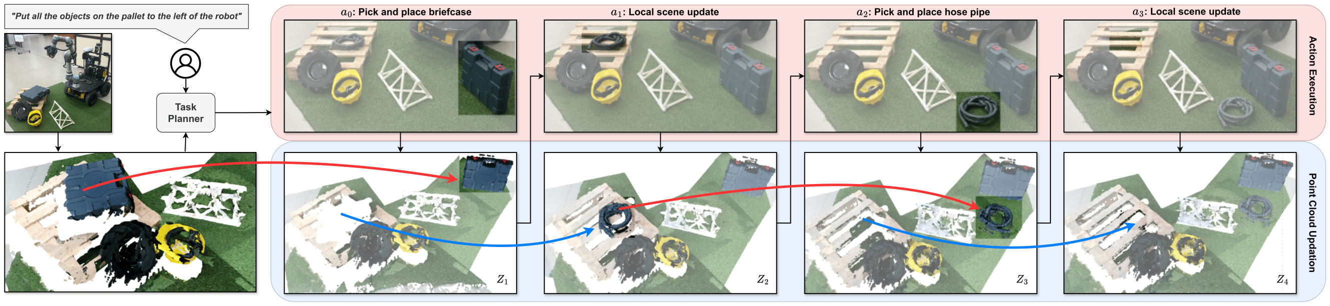
"Put all the objects on the pallet to the left of the robot"
Task Planner
a_0: Pick and place briefcase
a_1: Local scene update
a_2: Pick and place hose pipe
a_3: Local scene update
Action Execution
Point Cloud Updation
Z_1
Z_2
Z_3
Z_4

Fig. 3. Visualization of the plan rollout and scene reconstruction for a scenario involving occlusion. Red arrows indicate the pose update of the object, while blue arrows represent local scene reconstruction. Initially, only the briefcase was visible. Upon removing the briefcase and subsequent local rebuilding, the hose was detected. The figure also demonstrates how the scene is updated when the object is manipulated without requiring global scene reconstruction.

IV. Experiments and Results

The experiments were conducted using a UR5e manipulator mounted on a Husky mobile platform and a Robotiq 3F-gripper end-effector and OAK-D Pro eye-in-hand camera. Evaluation was performed on a dataset containing $15-20$ objects commonly found in outdoor environments, including items such as trusses, construction tools, pipes, barrels, and crates, among others. Ground truth 3D models were acquired using a Revopoint Range 3D Scanner.

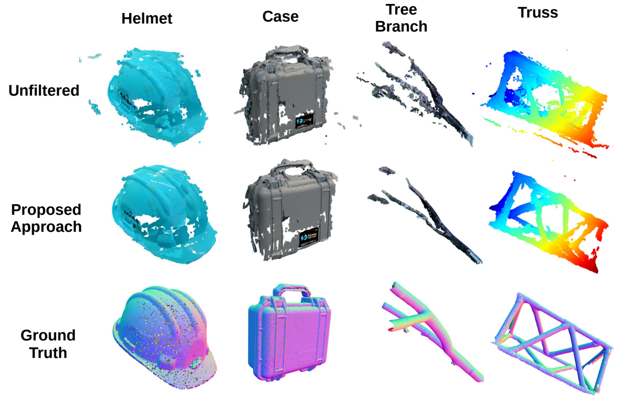

Fig. 4. Qualitative comparison of the 3D model of objects generated using the unfiltered approach (directly masking raw point cloud) and proposed approach with the ground truth model.

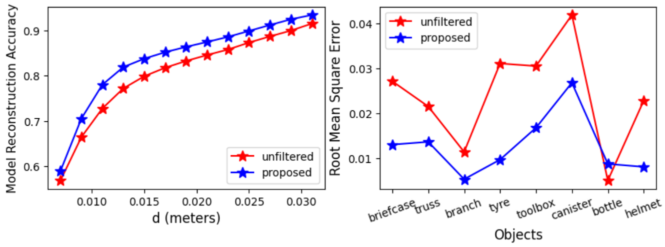

Fig. 5. The proposed method improves over the unfiltered approach in object modelling. (a) Left: Average MRA for a range of allowable error distance d and (b) Right: RMSE across objects dataset.

Accuracy of 3D Object Models: We evaluate both qualitatively and quantitatively the quality of the 3D model extracted for each object. Fig. 4 illustrates that the proposed method, which utilizes depth priors, produces more accurate and complete 3D object models in relation to the unfiltered approach. Specifically, the resulting models show lower noise reduction, smooth surfaces, and capture structural information, which is lacking in the unfiltered method.

We quantitatively evaluate our method using two metrics: (i) Root Mean Squared Error (RMSE): RMSE quantifies the root mean squared distance of each point in the reconstructed model to the nearest point in the ground truth. Results indicate that the proposed approach has a lower error compared to the unfiltered approach with respect to the ground truth. (ii) Model Reconstruction Accuracy (MRA): MRA quantifies the percentage of points in the reconstructed model that are within a distance d to the nearest point in the ground truth. Results indicate that the proposed method has a higher Model Reconstruction Accuracy compared to the unfiltered approach, meaning that a higher percentage of points of the reconstructed model are within tolerance d from the ground truth.

Rapid Scene Reconstruction: Fig. 3 illustrates qualitatively how the scene is rapidly reconstructed during plan rollout without requiring global point cloud registration. Our approach aids in handling cases of occlusion and updating the object model m_j after action execution.

V. Conclusions and Future Scope

In this paper, we introduce a pipeline for the acquisition of 3D models of objects in a zero-shot manner, leveraging priors from foundation models. This enables the robot to robustly execute sequential tasks in scenarios populated by unstructured and *a-priori* objects. Enhancing our method with a more accurate multi-view association and uncertainty-guided exploration will bolster 3D model quality. Furthermore, investigating a depth prior-based hole filling approach holds promise for further refinement. This work relates to the "acquiring 3D geometric information" theme of the ICRA Workshop on 3D Visual Representations for Robot Manipulation and contributes a method for allowing rapid acquisition of object models using foundation model priors for robust instruction following in unstructured environments.

Acknowledgements

We are grateful to Mr. Himanshu Gaurav Singh for assisting in setting up the LLM/LVM infrastructure (while being student at IIT Delhi) and feedback on this work (after graduating). We thank Mr. Mohd. Nadir, and Dr. Piyush Chanana for conceptualizing and realizing the Franka Emika Panda manipulation test bed used for real experiments. We acknowledge Mr. Shailendra Negi and Ms. Sunita Negi for administrative support. We thank Namasivayam K and the IIT Delhi HPC team for assisting with setup and maintenance of high-performance computing infrastructure used for this work. Rohan Paul and P. V. M. Rao acknowledge research funding support from IIT Delhi IRD-Unit, NCAHT-ICMR and DIA-COE.

References

[1] J. Bowkett, S. Karumanchi, and R. Detry. Grasping and transport of unstructured collections of massive objects. *Field Robotics*, 2:385–405, 2022.

[2] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023.

[3] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2023.

[4] J. A. Haustein, K. Hang, J. Stork, and D. Kragic. Object placement planning and optimization for robot manipulators. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7417–7424, 2019.

[5] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2013.

[6] N. Hughes, Y. Chang, and L. Carlone. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. 2022.

[7] D. Kim, N. Oh, D. Hwang, and D. Park. Lingo-space: Language-conditioned incremental grounding for space. *arXiv preprint arXiv:2402.01183*, 2024.

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[9] 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.

[10] L. Nicholson, M. Milford, and N. Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 4(1):1–8, 2018.

[11] K. Ok, K. Liu, and N. Roy. Hierarchical object map estimation for efficient and robust navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1132–1139. IEEE, 2021.

[12] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, pages 145–152, 2001.

[13] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.