Reasoning and Planning about Unobserved Objects with Memory Models

Yixuan Huang

1 Motivation

For robots to assist humans in their daily lives in roles such as home assistants and caregivers for elders, they must be able to reason about objects not observed in their current perceptual data. For example, if asked to retrieve an apple stored inside a cabinet, the robot should remember where it was placed and that it must first open the cabinet to grasp the object. Further if the robot picks up a box containing objects it previously placed inside the container, it should know that these objects will move with the box, while those sitting near the box will not. We desire for robots to maintain these models over many separate task assignments from their operators in order to enable long-term autonomy.

2 Objectives

We advocate for using an unsupervised video object segmentation (UVOS)[1] algorithm to explicitly manage our object-oriented memory. Specifically, we examine incorporating an explicit UVOS-based memory model into the framework from Huang et al. [2,3], which was an effective framework to learn relational dynamics across varying object and environments. Key to its success is the ability to encode a variable number of objects for a given observation using a graph neural net [2] or transformer-based encoder [3]. Prediction of relations enables the model's use in logic-based task planning [4], where relations have proved an effective means of communication between robots and humans [2,5–9]. However, the existing framework assumes all relevant objects to be observable. We propose two ways to integrate predictions from the video tracker of [10] with the relational dynamics prediction model of [3]. Both approaches augment the current state estimate with information from currently unobserved objects for use in predicting inter-object relations and action effects. One approach directly augments the latent space of the dynamics model by concatenating the previously predicted latent state tokens for unobserved objects with those currently observed. We term this *Latent Occluded Object Memory* (LOOM). The second method, termed *Direct Occluded Object Memory* (DOOM), directly augments the input point cloud with the previously observed object point cloud transformed based on its previously predicted pose estimate.

3 Related work

Reasoning about object permanence is an important capability for robot manipulation [11–14]. Xu et al. [11] and Ebert et al. [12] are the first ones to propose deep learning models to reason about occluded objects and do downstream planning with the learned dynamics models. However, both of these approaches assume goal images for planning which may not always be available from human operators. They additionally examine only planar pushing tasks with only a single moving object at each time. Our work examines a much more diverse set of tasks and skills, while also requiring only logic-based goal representations. Curtis et al. [14] propose a system with modules that can estimate affordances and properties to perform multi-step manipulation tasks with unknown and occluded objects. However, they assume complete object shape and have many engineered modules including the affordance module.

Table 1: Real World Planning Success.

Objects	4	5	6	7	all	Distractors	1	2	3	all
DOOM	5/5	5/5	5/5	4/5	**19/20**	DOOM	5/5	4/5	5/5	**14/15**
LOOM	5/5	4/5	5/5	5/5	**19/20**	LOOM	4/5	5/5	4/5	**13/15**
[2,3]	-	-	-	-	0/20	[2,3]	-	-	-	0/15

4 Approach

We assume the robot perceives the world as a point cloud, Z_t, at each timestep t. The robot then takes action A_t and receives subsequent observation Z_{t+1}. At new observation Z_{t+1}, some objects may have become occluded and other, new objects may appear. Based on the history of observations and actions $Z_{0:t}$, $A_{0:t}$ we would like the robot to plan to achieve a goal, potentially involving previously observed, but currently occluded objects. We define the goal as a logical conjunction of M desired object and environment relations, $\mathbf{g} = r_1 \wedge r_2 \wedge ... \wedge r_M, r_j \in \mathcal{R}$, where $\mathbf{g}$ denotes the goal conjunction, r_j represents each goal relation, and $\mathcal{R}$ denotes all possible relations. Our robot is given a set of L parametric action primitives $\mathcal{A} = \{A_1, \ldots, A_L\}$ where A_l defines a skill, which has associated continuous skill parameters θ_l. For example, a push skill is defined with parameters encoding the push direction and length, or a pick-and-dump skill is defined with parameters encoding the grasp pose and dump pose. The robot's planning task is defined as finding skills and skill parameters $\tau = ((A_0, \theta_0), \ldots, (A_{H-1}, \theta_{H-1}))$ that, when sequentially executed, transform the objects so they satisfy all desired object and environment relations in the goal $\mathbf{g}$. To solve this problem we propose a novel memory-based neural network framework. Instead of taking the entire history of observations as input, the model takes the current observation, Z_t, current action (A_t, θ_t) and a compressed memory of the previous observations $Z_{0:t-1}$ and actions $A_{0:t-1}$, and predicts the resulting relations r'_{t+1} and object poses p'_{t+1}. This enables our framework to remember the pose of disappeared objects after several actions. By chaining together predictions, we can effectively perform multi-step planning. We propose two different implementations of this framework called DOOM and LOOM, which respectively use a point cloud-based encoding and latent space encoding to represent the memory, Q. For the details of our approaches, please refer to [15].

5 Significance

We show how well our approaches generalize to a variable number of objects and then we show how the framework's performance in terms of generalization to different number of distractor actions in the real-world evaluation in table. 1. Note that prior works [2,3] would achieve a 0% success rate on this evaluation because prior works cannot plan to achieve goals including occluded objects. We show the diversity of the tasks our approaches can achieve on our website https://sites.google.com/view/rdmemory.

6 Future Work

We identify several areas for future research. One promising idea is to integrate the tracker and planner and train a full system end-to-end. For example, we could leverage the pose prediction from DOOM as a prior for the tracker. We would additionally like to incorporate a mobile base requiring more complicated memory management by coupling with some form of mapping and examine longer-horizon planning tasks.

References

[1] S. Caelles, J. Pont-Tuset, , F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool, "The 2019 davis challenge on vos: Unsupervised multi-object segmentation," *arXiv:1905.00737*, 2019. 1

[2] Y. Huang, A. Conkey, and T. Hermans, "Planning for Multi-Object Manipulation with Graph Neural Network Relational Classifiers," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. [Online]. Available: https://arxiv.org/abs/2209.11943 1, 2

[3] Y. Huang, N. C. Taylor, A. Conkey, W. Liu, and T. Hermans, "Latent space planning for multi-object manipulation with environment-aware relational classifiers," *arXiv preprint arXiv:2305.10857*, 2023. 1, 2

[4] C. R. Garrett, T. Lozano-Pérez, and L. P. Kaelbling, "Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 30, 2020, pp. 440–448. [Online]. Available: https://arxiv.org/abs/1802.08705 1

[5] C. Paxton, C. Xie, T. Hermans, and D. Fox, "Predicting Stable Configurations for Semantic Placement of Novel Objects," in *Conference on Robot Learning (CoRL)*, 11 2021. [Online]. Available: https://arxiv.org/abs/2108.12062 1

[6] W. Liu, C. Paxton, T. Hermans, and D. Fox, "StructFormer: Learning Spatial Structure for Language-Guided Semantic Rearrangement of Novel Objects," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. [Online]. Available: https://sites.google.com/view/structformer 1

[7] Y. Zhu, J. Tremblay, S. Birchfield, and Y. Zhu, "Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021. [Online]. Available: https://arxiv.org/abs/2012.07277 1

[8] R. Li, A. Jabri, T. Darrell, and P. Agrawal, "Towards practical multi-object manipulation using relational reinforcement learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4051–4058. [Online]. Available: https://arxiv.org/abs/1912.11032 1

[9] M. Sharma and O. Kroemer, "Relational learning for skill preconditions," in *Conference on Robot Learning (CoRL)*, 2020. [Online]. Available: https://arxiv.org/abs/2012.01693 1

[10] J. Yuan, J. Patravali, H. Nguyen, C. Kim, and L. Fuxin, "Maximal cliques on multi-frame proposal graph for unsupervised video object segmentation," *arXiv preprint arXiv:2301.12352*, 2023. 1

[11] Z. Xu, Z. He, J. Wu, and S. Song, "Learning 3d dynamic scene representations for robot manipulation," in *Conference on Robot Learning (CoRL)*, 2020. 1

[12] F. Ebert, C. Finn, A. X. Lee, and S. Levine, "Self-supervised visual planning with temporal skip connections." *CoRL*, vol. 12, p. 16, 2017. 1

[13] M. Du, O. Y. Lee, S. Nair, and C. Finn, "Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning," *arXiv preprint arXiv:2205.14850*, 2022. 1

[14] A. Curtis, X. Fang, L. P. Kaelbling, T. Lozano-Pérez, and C. R. Garrett, "Long-horizon manipulation of unknown objects via task and motion planning with estimated affordances," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1940–1946. 1

[15] Y. Huang, J. Yuan, C. Kim, P. Pradhan, B. Chen, L. Fuxin, and T. Hermans, "Out of sight, still in mind: Reasoning and planning about unobserved objects with video tracking enabled memory models," *arXiv preprint arXiv:2309.15278*, 2023. 2