Unordered Navigation to Multiple Semantic Targets in Novel Environments

Bernadette Bucher[1]* Katrina Ashton[1]* Bo Wu[1] Karl Schmeckpeper[1]
Siddharth Goel[2] Nikolai Matni[1] Georgios Georgakis[3] Kostas Daniilidis[1]

1. Introduction

Consider the problem of finding a set of objects in a novel environment. In many practical cases, the order in which the objects are found does not matter. However, the order does matter from the perspective of efficiency. If we anticipate finding certain objects near each other based on known semantic relationships, we should plan to search for them sequentially. For example, if two objects will probably be found in a kitchen and a third will not, then we should look for both of the objects in the kitchen before moving to find another object in a different room. Furthermore, if the environment is novel, we need to balance exploiting these contextual semantic priors with exploring the specific unknown layout of the environment. In this work, we propose a method for reasoning over these type of semantic relationships in order to efficiently find a set of unordered objects.

The challenge of this problem is two-fold. First, from the perspective of developing long horizon sequential plans, the agent needs to reason over an optimal ordering of targets while taking into account uncertainty in their positions. Second, to navigate efficiently, the agent needs to leverage contextual semantic priors in order to develop target-driven navigation plans without prior access to a map. Prior work in multi-object navigation considers ordered [2, 10–12, 16] and unordered [14] cases in which objects are spawned randomly in their environment, removing the challenge of leveraging contextual semantic priors. Semantically meaningful target locations were introduced recently in a task for ordered multi-object navigation [18]. We consider an unordered navigation scenario in which objects are typically found in semantically meaningful fixed locations.

2. Approach

To solve our proposed multi-object navigation task, we need to trade off the exploitation of prior contextual semantic understanding with exploration of the novel environment. To enable this planning approach, we explicitly predict unseen semantic maps and estimate uncertainty over those predictions with an approach leveraged in prior novel environment navigation tasks [3–5]. This prediction model f is defined as an ensemble of hierarchical segmentation models to predict a semantic map of the unseen environment. The input is an incomplete occupancy region $o'_t \in \mathbb{R}^{|C^o| \times h \times w}$ and a ground-projected semantic segmentation $\hat{\chi}_t \in \mathbb{R}^{|C^\chi| \times h \times w}$ at time-step t. The output is a top-down semantic local region $\hat{m}_t \in \mathbb{R}^{|C^\chi| \times h \times w}$, where C^o is the set of occupancy classes containing *unknown*, *occupied*, and *free*, C^χ is the set of semantic classes, and h, w are the dimensions of the local crop. Following prior work [4], we construct f as an ensemble of models for which the the mean estimates $P(m_t|o'_t, \hat{\chi}_t)$ and the variance approximates the uncertainty of model predictions.

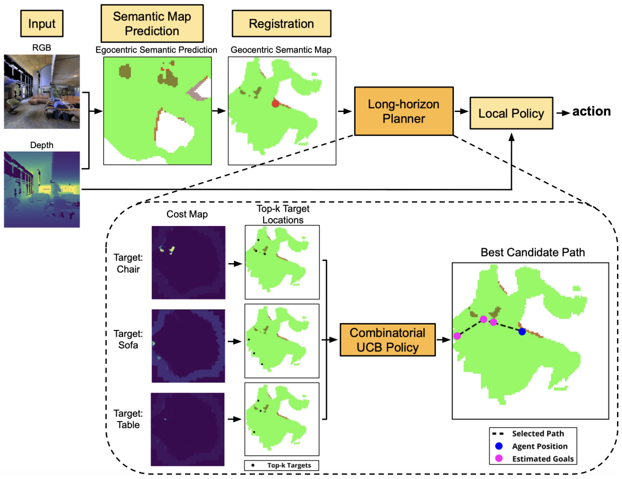

Figure 1. Our novel framework for navigating to an unordered set of target objects. We explicitly predict unobserved semantic maps over which we estimate uncertainty and register those predictions in a global semantic map. Then, our long-horizon planner creates a cost-map for each target object based on their predicted locations and associated uncertainty from which the top-k locations are selected for each target. Finally, it combinatorially constructs paths from these candidate locations and selects the best path with our objective in Equation 1.

We use this semantic map predictor to estimate an optimal ordered path to find a given set of unordered target objects via an upper confidence bound (UCB) policy.

* Denotes equal contribution.
1 GRASP Laboratory, University of Pennsylvania, 2 Amazon
3 NASA Jet Propulsion Laboratory, California Institute of Technology
Corresponding authors: kashton@seas.upenn.edu, bbucher@theaiinstitute.com

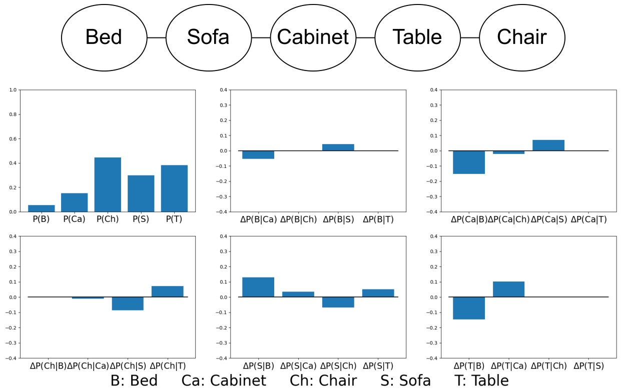

Figure 2. Pair-wise conditional probability of navigation success over 3-object navigation. P(A) in this figure is the probability of finding object A given that at least one other object is found. $\Delta P(A|B) := P(A|B) - P(A)$, where $P(A|B)$ is the probability of finding A given that B is found. We abbreviate the object classes Bed, Cabinet, Chair, Sofa, and Table as B, Ca, Ch, S, and T respectively. The graphical model shows experimentally observed semantic relationships between different object classes.

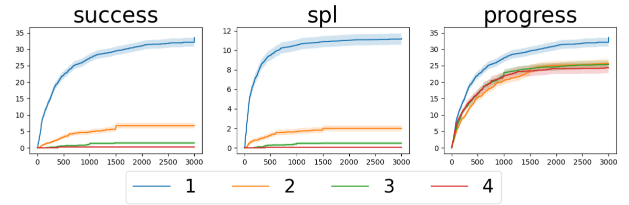

Figure 3. Success, SPL, and Progress over 3000 time steps for 1-, 2-, 3-, and 4-object navigation methods. An episode is considered successful if the agent is able to navigate to all the target object categories within the maximum number of steps before ending the episode. Progress is the percent of objects the agent successfully finds within the episode time limit. Success weighted by path length (SPL) downweights the success rate by the ratio of the shortest possible path from the agent to all the target objects to the path taken by the agent.

We denote f_c as the ensemble-based map prediction function f which only returns the values for a given target class c. We denote $\sigma_c(o'_t, \hat{\chi}_t) = \sqrt{\text{Var}\, f_c(o'_t, \hat{\chi}_t; \theta)}$ as the standard deviation of the target class probability computed over models in the ensemble, and we denote the mean as $\mu_c(o'_t, \hat{\chi}_t) = \frac{1}{N}\sum_{i=1}^{N} f_c(o'_t, \hat{\chi}_t; \theta_i)$. Let ζ be the collection of n distinct semantic targets for which classes are not repeated. For each class $c \in \zeta$, we choose the top k values of $\mu_c(o'_t, \hat{\chi}_t) + \alpha_1 \sigma_c(o'_t, \hat{\chi}_t)$, the upper bound of $P_c(m_t|o'_t, \hat{\chi}_t)$, the conditional probability class c is at location i. We use these upper bounds and associated locations to construct and score a set of candidate paths γ. Candidate paths are defined by the n estimated target locations assigned the ordering resulting in the shortest Euclidean distance starting from the current agent position to each estimated target location. We select the candidate path with the highest score and begin traversing it. On a fixed interval set by a hyperparameter, we replan, selecting the path again by using the set of semantic targets which have not yet been successfully reached.

We now explain how we construct and score paths from the top k candidate locations for each target class. First, paths are constructed by choosing every combination of locations where one candidate for each class is selected. We select the ordering of each set of object locations that has the minimal Euclidean distance, generating kn path candidates. We denote j as the node index in the path γ_i. Since each planned path γ_i is now ordered, we can downweight the contribution of later nodes γ_i, following the temporal logic planning literature [6–9] to form the objective

$$\arg\max_{\gamma_i \in \gamma} \sum_{j=0}^{n-1} \alpha_0^{-j} \left(\mu_j(o'_t, \hat{\chi}_t) + \alpha_1 \sigma_j(o'_t, \hat{\chi}_t) - \alpha_2 d_{j,j+1} \right) \tag{1}$$

where $d_{j,j+1}$ is the Euclidean distance between nodes j and $j+1$ and α_0, α_1 and α_2 are hyperparameters. To perform target-driven navigation, local predicted semantic regions are registered to a global map which is used during planning. We use DD-PPO [17] to navigate to the first map location in the path selected by our UCB policy.

3. Results

We execute multi-object navigation experiments with the Matterport3D dataset (MP3D) [1] in the Habitat simulator [13, 15] to enable interaction with the 3D residential home reconstructions provided by MP3D. Unordered navigation is performed with combinations from the following 5 semantic object categories: *bed, sofa, chair, table, cabinet*. The action space consists of MOVE_FORWARD by 25cm, TURN_LEFT by 10°, TURN_RIGHT by 10°, and STOP. An object is successfully found by the agent if the agent is within 1 meter of the target object. The semantic map prediction model uses the pre-trained weights from prior work [4].For all of our experiments, we use an ensemble size of 4. Figure 3 shows success metrics.

In our task definition, we emphasize the need to have semantically meaningful object locations to match real-world robotic use case scenarios. One reason this task design is important is so that methods can exploit these semantic relationships for increased performance over long-horizon planning. We demonstrate that our method successfully achieves this capability by computing the conditional probability of success over 3-object navigation. In our problem, two objects are semantically associated if they frequently co-occur in our test scenes. Figure 2 demonstrates the semantic relationships visualized in a graphical model which align with our own contextual semantic understanding of home layouts. Our method demonstrates that it exploits these relationships in planning since the directional change in conditional probability of navigation success is consistent for all classes (if $\Delta P(A|B) >= 0$, then $\Delta P(B|A) >= 0$).

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2

[2] Peihao Chen, Dongyu Ji, Kunyang Lin, Weiwen Hu, Wenbing Huang, Thomas H Li, Mingkui Tan, and Chuang Gan. Learning active camera for multi-object navigation. *NeurIPS*, 2022. 1

[3] Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. *International Conference in Robotics and Automation (ICRA)*, 2022. 1

[4] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation, 2022. 1, 2

[5] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15460–15470, 2022. 1

[6] Yiannis Kantaros, Matthew Malencia, Vijay Kumar, and George J Pappas. Reactive temporal logic planning for multiple robots in unknown environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11479–11485. IEEE, 2020. 2

[7] Morteza Lahijanian, Matthew R Maly, Dror Fried, Lydia E Kavraki, Hadas Kress-Gazit, and Moshe Y Vardi. Iterative temporal planning in uncertain environments with partial satisfaction guarantees. *IEEE Transactions on Robotics*, 32(3):583–599, 2016. 2

[8] Scott C Livingston, Richard M Murray, and Joel W Burdick. Backtracking temporal logic synthesis for uncertain environments. In *2012 IEEE International Conference on Robotics and Automation*, pages 5163–5170. IEEE, 2012. 2

[9] Matthew R Maly, Morteza Lahijanian, Lydia E Kavraki, Hadas Kress-Gazit, and Moshe Y Vardi. Iterative temporal motion planning for hybrid systems in partially unknown environments. In *Proceedings of the 16th international conference on Hybrid systems: computation and control*, pages 353–362, 2013. 2

[10] Pierre Marza, Laetitia Matignon, Olivier Simonin, and Christian Wolf. Multi-object navigation with dynamically learned neural implicit representations. *arXiv*, 2022. 1

[11] Pierre Marza, Laetitia Matignon, Olivier Simonin, and Christian Wolf. Teaching agents how to map: Spatial reasoning for multi-object navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1725–1732, 2022. 1

[12] Sriram Narayanan, Dinesh Jayaraman, and Manmohan Chandraker. Long-hot: A modular hierarchical approach for long-horizon object transport. *arXiv preprint arXiv:2210.15908*, 2022. 1

[13] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[14] Fabian Schmalstieg, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Learning long-horizon robot exploration strategies for multi-object search in continuous action spaces. *arXiv preprint arXiv:2205.11384*, 2022. 1

[15] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[16] Saim Wani, Shivansh Patel, Unnat Jain, Angel X Chang, and Manolis Savva. Multion: Benchmarking semantic map memory using multi-object navigation. *NeurIPS*, 2020. 1

[17] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019. 2

[18] Haitao Zeng, Xinhang Song, and Shuqiang Jiang. Multi-object navigation using potential target position policy function. *IEEE Transactions on Image Processing*, pages 1–1, 2023. 1