Manipulation-Enhanced Spatial Mapping via Belief Prediction Models

Nils Dengler[1]∗ Joao Marcos Correia Marques[2]∗ Jesper Mücke[1] Tobias Zaenker[1]
Shenlong Wang[2] Kris Hauser[2] Maren Bennewitz[1]

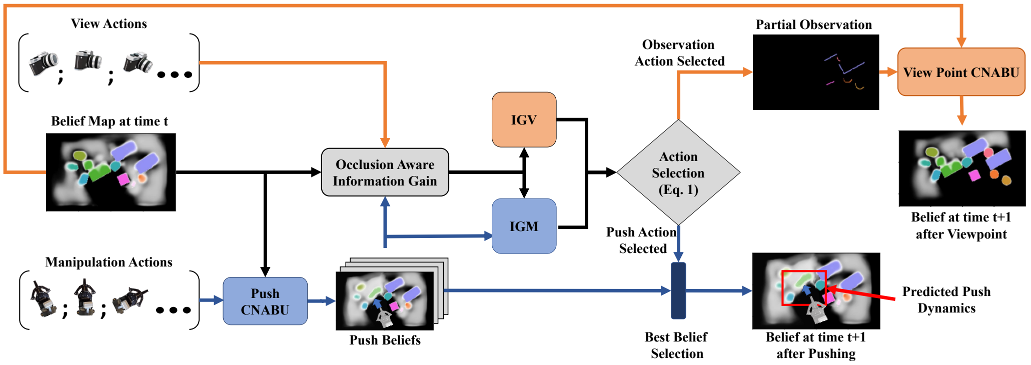


Fig. 1: From a prior map belief, our pipeline predicts the potential map belief from pushes, evaluates information gain from multiple views, and selects the action with the highest gain.

I. Introduction

Searching for objects in cluttered environments requires selecting efficient viewpoints and manipulation actions to remove occlusions and reduce uncertainty in object locations, shapes, and categories. In this work, we address the problem of manipulation-enhanced semantic mapping, where a robot has to efficiently reason about objects in a cluttered shelf. Although Partially Observable Markov Decision Processes (POMDPs) are standard for decision-making under uncertainty, representing unstructured interactive worlds remains challenging in this formalism, due to its state space complexity. To tackle this difficulty, we define a POMDP whose belief is summarized by a metric-semantic grid map and propose a novel framework that uses neural networks to perform map-space belief updates to reason efficiently and simultaneously about object geometries, locations, categories, occlusions, and manipulation physics. We call these networks Calibrated Neural Accelerated Belief Update (CNABU) networks and show that they can generalize to novel scenarios and transfer well sim-to-real in zero-shot fashion.

II. Methods

A. Overview

In this work, we consider a confined environment with movable objects of varying sizes and orientations, where some objects may be unobservable from any viewpoint due to occlusions. We aim to determine the most informative sequence of actions for a robot, within a given action budget, that minimizes the difference between the robot's internal map belief and the true environment configuration using a similarity metric, such as IoU. A robotic arm, equipped with a wrist-mounted RGB-D camera and a gripper, aims to build an accurate map of the current workspace configuration C_W [1] after a sequence of actions, which can be either taking an RGB-D image or performing a manipulation (i.e., a push) to move objects and reveal occluded areas.

Let Φ^t represent the robot's internal environment map at time t. When manipulating the environment, it causes a transition on the workspace configuration space from $c^t \mapsto c^{t+1} \in C_W$ according to the environment's dynamics. Further, whenever the robot chooses to take another RGB-D observation, it updates its internal environment representation according to its belief update, $\Phi^t \rightarrow \Phi^{t+1}$. Over time, the agent must update its belief after each individual action. However, traditional POMDP updates are impractical due to the high dimensionality of the belief space [2]. We propose using uncertainty-aware evidential deep learning [3] to predict a factorized belief distribution that aligns with plausible configurations while maintaining compactness.

B. Solving the POMDP

We solve the POMDP using a k-step receding horizon greedy planner (Fig. 1) and approximating the reward function with Volumetric Information Gain (VIG) [4]. For computational efficiency, we use a horizon of $k = 2$.

To perform an observation action, the robot chooses from $v^i \in \mathbb{V}$ possible views in a fixed array of camera positions $\mathbb{V}$ to which the robot can move. Furthermore, let $\theta_t \in \Theta_t$ be a sampled manipulation action from a set of feasible actions. In our two-step greedy search, we only consider two possible kinds of action sequences: taking two observation actions (v_t, v_{t+1}) or performing a manipulation action followed by an observation (θ_t, v_{t+1}).

∗ These authors contributed equally to this work.;
1. Humanoid Robots Lab, University of Bonn, Germany
2. University of Illinois at Urbana-Champaign, IL, USA.

This work has been supported by the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany.

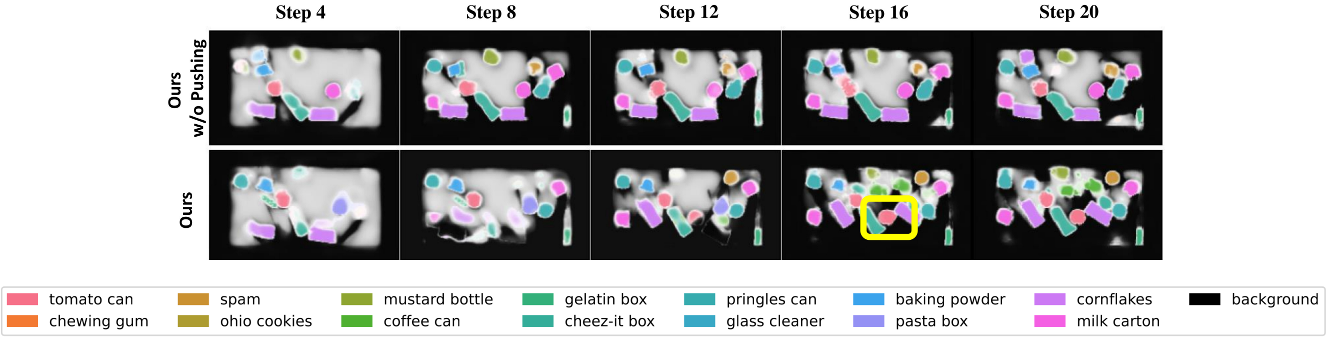
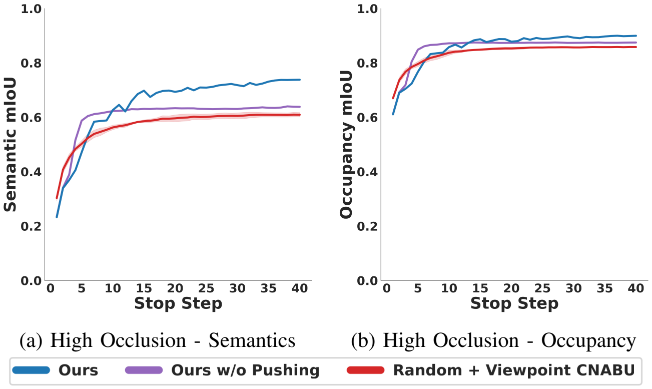


Fig. 3: Real-world experiment results show that VPP and Random baselines struggle with occlusions, while our method explores effectively through manipulations. In Step 16, we highlight a revealed object in yellow.



(a) High Occlusion - Semantics

(b) High Occlusion - Occupancy

Fig. 4: Simulation results in MEM task.

We propose recursively estimating the belief update after an observation action using a deep posterior network, $\sigma_o(\Phi_{t-1}, o_t)$, which we call Calibrated-Neural Accelerated Belief Update (**CNABU**) network to create an implicit Monte-Carlo estimate of the POMDP belief update by $\Phi_t = \sigma_o(\Phi_{t-1}, o_t)$. Similarly, the update after a manipulation action is learned via an action-specific CNABU, $\sigma_m(\Phi_{t-1}, a_t)$.

The term $IGV_t = IG(v_t^*, v_{t+1}^* | \Phi_t^O)$ denotes the highest information gain obtained from two viewpoints, given map representation Φ_t^O, while $IGM_t = IG(v_{\theta_t}^* | \tilde{\Phi}_{t+1}^{\theta_t^*})$ represents the best information gain from a pushing action followed by a viewpoint execution, given the posterior map representation after executing the push action $\tilde{\Phi}_{t+1}^{\theta_t^*}$. Lastly, $Reg_t = \Delta H(\Phi_t, \tilde{\Phi}_{t+1}^{\theta_t^*})$ captures the entropy difference between the current semantic map and the map after the best push action. Our policy decides the action a_t through:

$$a_t = \begin{cases} v_t^* & \text{if } IGV_t > IGM_t + \gamma Reg_t \\ \theta_t^* & \text{otherwise} \end{cases} \tag{1}$$

Here, γ balances VIG and entropy, while ΔH regularizes manipulation to limit unnecessary disturbances. Belief updates depend on the action: $\Phi^{t+1} = \tilde{\Phi}_{t+1}^{\theta_t^*}$ for manipulation or $\Phi^{t+1} = \sigma_o(\Phi^t, o^t)$ for observations.

III. Experimental Results

For experimental evaluation, we set up a shelf scene with a UR5 arm for observation and action execution in PyBullet [5]. The robot is equipped with a Robotiq parallel-jaw gripper and an realsense L515 RGB-D camera for observations. To sample realistic object configurations, a total of 14 different object categories from the YCB dataset are used and sampled in a shelf board of size $(0.8 \times 0.4 \times 0.4)m$.

Quantitative results are shown in Fig. 4. We observe that belief prediction is a powerful approach, leading to excellent scene coverage even without pushing. On highly occluded scenarios, pushing is required to make progress after the visible surfaces are observed. Our method uses pushing to achieve significant higher mIoUs. Note its IoU growth is slower early on, because pushing does not provide information until a viewpoint step is taken. We recreate the simulation scenario in the real world, and compare our approach against two of the strongest baselines in simulation on the MEM task. We evaluate agent performance in 4 categories which score the status of all objects in the scene, depending on whether they were found, found but incorrectly classified, not found or hallucinated. These results are in Tab. I.

In Fig. 3 we show a qualitative result of our agent after efficient viewpoint and push selection. The qualitative results show that with zero-shot transfer from sim-to-real, our proposed method generates a good representation of the scene. Our method (both with pushing and without pushing) is able to identify the majority of objects in the scene, from a total of 16. Further, our complete pipeline reveals several objects that were previously unseen, e.g., a tomato can as highlighted in yellow.

TABLE I: Comparing our method in 10 trials to the strongest baselines in zero-shot transfer to real-world shelves.

Policy	Correctly Found ↑	Missclassified But Found↑	Not Found↓	Hallucinated↓
Random + CNABU	72	47	52	11
Ours w/o Pushing	81	38	52	**6**
Ours	**85**	**52**	**35**	7

IV. Conclusion

In this paper, we presented a POMDP-inspired policy solver, that decides between different action types to generate an uncertainty-aware map-apace dynamics model as belief. Furthermore, our pipeline considers all action types to be equally effective and decides according to the best informative outcome. Our results show the qualitative performance of our system in terms of occupancy and semantics map accuracy and demonstrate that our agent is able to reason about map dynamics and impact of actions to the scene.

REFERENCES

[1] J. Chase Kew, B. Ichter, M. Bandari, T.-W. E. Lee, and A. Faust, "Neural Collision Clearance Estimator for Batched Motion Planning," in *Algorithmic Foundations of Robotics XIV*, S. M. LaValle, M. Lin, T. Ojala, D. Shell, and J. Yu, Eds. Cham: Springer International Publishing, 2021, pp. 73–89.

[2] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1, pp. 99–134, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000437029800023X

[3] D. Ulmer, C. Hardmeier, and J. Frellsen, "Prior and Posterior Networks: A Survey on Evidential Deep Learning Methods For Uncertainty Estimation," *PMLR*, 2023.

[4] J. Delmerico, S. Isler, R. Sabzevari, and D. Scaramuzza, "A comparison of volumetric information gain metrics for active 3D object reconstruction," *Autonomous Robots*, vol. 42, no. 2, pp. 197–208, 2018. [Online]. Available: https://doi.org/10.1007/s10514-017-9634-0

[5] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," http://pybullet.org, 2016–2021.