Multi-Skill Manipulation-Enhanced Mapping with Evidential Learning in Confined Environments

Yitian Shi, Nils Dengler, Jesper Mücke, Sicong Pan, Rania Rayyes, Maren Bennewitz

I. Introduction

We propose Multi-Skill Manipulation-Enhanced Mapping (MS-MEM), a hierarchical evidential framework for uncertainty-aware occlusion mapping that jointly reasons over active viewpoint selection, non-prehensile pushing, and prehensile grasping. MS-MEM combines a scene-level metric-semantic belief map with a local grasp representation based on a full-evidential extension of vMF-Contact (FE-vMF), which models both grasp affordance and directional uncertainty. Within a POMDP formulation, the framework predicts the outcomes of candidate actions and evaluates them using Disturbance- and Occlusion-aware Information Gain (DOIG), a unified objective that balances expected visibility improvement against scene disturbance across heterogeneous action skills. In this way, MS-MEM enables localized and controllable occluder removal, improving visibility while minimizing unnecessary global disturbance to the scene. The demonstration video is at: https://www.youtube.com/watch?v=c32Qt-9ZIf44

II. Methodology

A. System Overview

As illustrated in Fig. 2, MS-MEM is organized as a closed-loop hierarchical framework, in which all subsequent modules operate on a shared evidential belief Φ_t of the scene.

At each time step, the framework considers three action branches conditioned on the current evidential scene belief Φ_t: Viewpoint selection with Disturbance- and Occlusion-aware Information Gain (DOIG), Uncertainty-informed Push Selection (**UPS**), and Uncertainty-informed Grasp Selection (**UGS**).

Initially, all action candidates are compared within their individual modules by evaluating their post-action beliefs and the DOIG objective to ensure fair comparisons across different action skills. Within the **UGS** module (illustrated in Fig. 3), the belief Φ_t is used to infer uncertainty-aware grasp hypotheses with the proposed Full-Evidential vMF-Contact (*FE-vMF*). These hypotheses are further accumulated over time by the novel Full Evidential Uncertainty-guided Multi-view Grasp Fusion (*FE-UMGF*). The **UGS** module evaluates these uncertainty-aware grasp candidates by their respective DOIGs,

This work is supported by the German Federal Ministry of Research, Technology, and Space (BMFTR) under the Robotics Institute Germany (RIG), the DFG SFB-1574-471687386 project, and the Ministry of Science, Research and Arts of the Federal State of Baden-Württemberg within the InnovationCampus Future Mobility.

M. Bennewitz, S. Pan, and N. Dengler are additionally with the Lamarr Institute for Machine Learning and Artificial Intelligence and the Center for Robotics, Bonn, Germany.

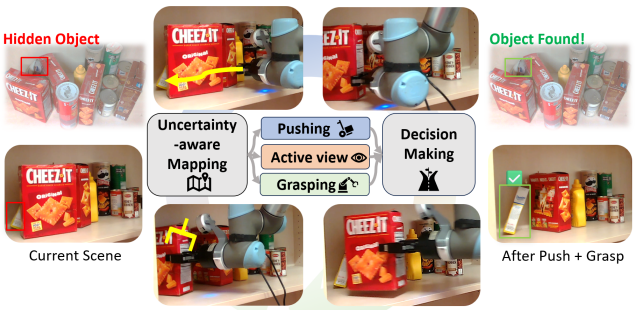

Fig. 1. Overview of MS-MEM.

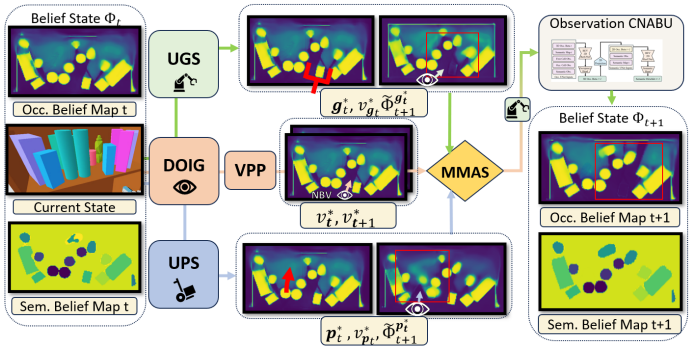

Fig. 2. Overall pipeline of MS-MEM.

while **UPS** evaluates push candidates with the same regime as [1]. Moreover, viewpoint actions are assessed via View Point Planning (**VPP** [2]), allowing their direct comparison to manipulation actions by Multi-Manipulation Action Selection (**MMAS**). Finally, after the execution of the chosen actions, the observation CNABU σ_o updates the evidential map with the newly acquired observation o_{t+1}.

B. Disturbance- and Occlusion-aware Information Gain (DOIG)

As an extension of volumetric information gain (IG) [3], DOIG serves as the common utility for all three action branches. Before its calculation, for manipulation actions $a_t \in \{\boldsymbol{p}_t, \boldsymbol{g}_t\}$, the system first predicts the post-action belief $\tilde{\Phi}^{a_t}_{t+1}$, and evaluates the volumetric IG of a subsequent NBV: $v_{t+1} \in \mathcal{V}$; For a pure view-change action v_t, the utility is evaluated as the sum of the immediate IG under the current belief Φ_t and the best subsequent IG after incorporating that observation o_{t+1} (referred to as View Point Planning (VPP) [2]). Formally, the OIG [2] is defined as

$$\mathrm{OIG}(a_t) := \begin{cases} \zeta_o \Delta \mathrm{H}(\Phi_t, \tilde{\Phi}^{a_t}_{t+1}) + \max\limits_{v_{t+1}\in\mathcal{V}} \mathrm{IG}(v_{t+1}|\tilde{\Phi}^{a_t}_{t+1}), & a_t \notin \mathcal{V}, \\ \mathrm{IG}(a_t|\Phi_t) + \max\limits_{v_{t+1}\in\mathcal{V}} \mathrm{IG}(v_{t+1}|\tilde{\Phi}^{a_t}_{t+1}), & a_t \in \mathcal{V}. \end{cases}$$

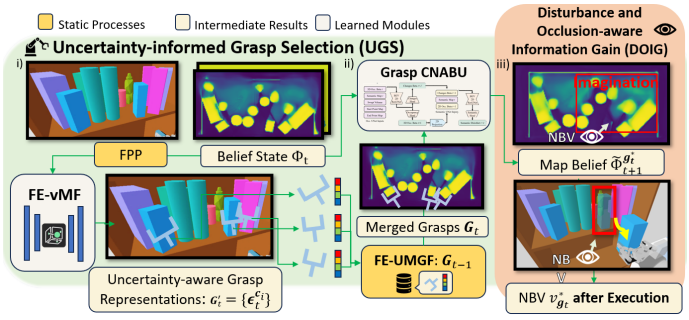

Fig. 3. Uncertainty-informed grasp selection (**UGS**) pipeline

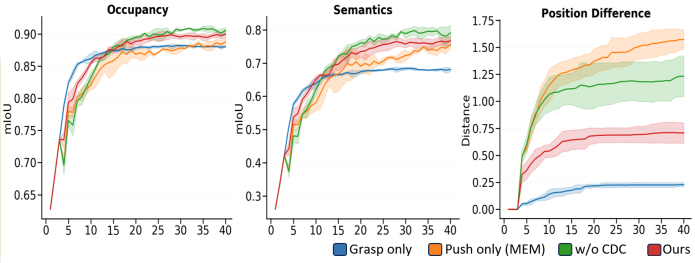

Fig. 4. Simulation experiments.

Notably, MS-MEM introduces the Disturbance and Occlusion-aware Information Gain (DOIG) by applying a Collateral Disturbance Constraint (CDC) to the OIG. This constraint penalizes localized scene changes by identifying confident voxels whose expected semantic classes are altered between Φ_t and $\tilde{\Phi}^{a_t}_{t+1}$: $\mathrm{DOIG}(a_t) := \mathrm{OIG}(a_t) - \zeta_{\mathrm{CDC}}|\mathcal{U}_{\mathrm{diff}}(\Phi_t, \tilde{\Phi}^{a_t}_{t+1})|$.

C. Uncertainty-informed Grasp Selection (*UGS*)

The system overview of **UGS** is depicted in Fig. 3. In general, the **UGS** module first leverages a novel Full-Evidential vMF-Contact *FE-vMF* network which proposes evidential grasp hypotheses (Sec. II-C1). After uncertainty-aware temporal fusion with historical grasps through the *FE-UMGF* buffer (Sec. II-C2), it selects the best executable grasp from the candidates by computing their post-grasp beliefs $\tilde{\Phi}^{g_t}_{t+1}$ and comparing the corresponding DOIGs.

1) Full Evidential Grasp Learning: As the core component of the **UGS** module, to enable uncertainty-aware grasp synthesis and selection, we propose the *Full-Evidential vMF-Contact (FE-vMF)* that quantifies uncertainty of the complete $SE(3)$ grasp configuration.

Unlike prior work [4], we model rotational uncertainty with two decoupled evidential parameterizations for the baseline $\boldsymbol{b}$ and the approach $\hat{\boldsymbol{a}}$ separately: $\hat{\boldsymbol{a}} \sim \mathrm{vMF}(\boldsymbol{\mu}_{\hat{\boldsymbol{a}}}, \kappa_{\hat{\boldsymbol{a}}})$, $\boldsymbol{b} \sim \mathrm{vMF}(\boldsymbol{\mu}_{\boldsymbol{b}}, \kappa_{\boldsymbol{b}})$. Here, $\hat{\boldsymbol{a}}$ is an intermediate estimate used to construct the final approach distribution $\boldsymbol{a}|\boldsymbol{b}, \hat{\boldsymbol{a}} \sim \mathrm{vMF}(\boldsymbol{\mu}_{\boldsymbol{a}}, \kappa_{\boldsymbol{a}})$, while enforcing the orthogonality constraint $\boldsymbol{a} \perp \boldsymbol{b}$ by projecting the distribution of $\hat{\boldsymbol{a}}$ onto the subspace orthogonal to $\boldsymbol{b}$. In addition, we treat the affordance probability q as a random variable following: $q^{\boldsymbol{c}} \sim \mathrm{Beta}(\alpha^{\boldsymbol{c}}, \beta^{\boldsymbol{c}})$. To better capture the full evidential uncertainty, we leverage Point Transformer v3 (PTv3) [5] as the evidential backbone.

*2) Full Evidential Uncertainty-guided Multi-view Grasp Fusion (*FE-UMGF*):* We construct *temporal grasp fusion* by introducing Full Evidential Uncertainty-guided Multi-view Grasp Fusion (*FE-UMGF*) as inspired by [6]. Let $\boldsymbol{G}_{t-1}$ denote the global grasp buffer accumulated up to $t-1$. The current inference results $\boldsymbol{G}'_t = \{\boldsymbol{\epsilon}^{\boldsymbol{c}'_i}\}_{i=1}^{N_{\mathrm{pcd}}}$ represent the newly inferred grasps from the occupancy belief λ^O_t. The overall temporal update is denoted as: $\boldsymbol{G}_t \leftarrow \textit{FE-UMGF}(\boldsymbol{G}_{t-1}, \boldsymbol{G}'_t)$.

Deviating from [6], *FE-UMGF* aims to generalize the fusion process to account for the full evidential uncertainty. For an old fused grasp $\boldsymbol{\epsilon}^{\boldsymbol{c}}_{t-1} \in \boldsymbol{G}_{t-1}$ located at the cluster center $\boldsymbol{c}$, its accumulated Beta evidence parameters $\alpha^{\boldsymbol{c}}_t$ and $\beta^{\boldsymbol{c}}_t$ are updated by aggregating the evidence from all nearby candidates $\boldsymbol{c}'$ assigned to its neighbouring cluster C_t [1]:

$$\alpha^{\boldsymbol{c}}_t = \gamma\alpha^{\boldsymbol{c}}_{t-1} + \sum_{\boldsymbol{c}' \in C_t} \alpha^{\boldsymbol{c}'}_t, \quad \beta^{\boldsymbol{c}}_t = \gamma\beta^{\boldsymbol{c}}_{t-1} + \sum_{\boldsymbol{c}' \in C_t} \beta^{\boldsymbol{c}'}_t$$

Then, the $\mathbb{E}[q^{\boldsymbol{c}}_t]$ can be calculated following [7]. By decaying historical evidence with decay γ, we implicitly assign higher relative weights to novel belief states Φ_t and retain the grasp outcomes from previous states $\{\Phi_{t-1}, \Phi_{t-2}, ...\}$.

3) Final Grasp Selection: Finally, the optimal grasp $\boldsymbol{g}^*_t$ and associated future viewpoint $v^*_{\boldsymbol{g}_t}$ is selected to maximize the DOIG on the predicted post-grasp belief $\tilde{\Phi}^{\boldsymbol{g}^{\boldsymbol{c}}_t}_{t+1}$.

D. Uncertainty-informed Push Selection (*UPS*)

Similar to **UGS**, the **UPS** module generates targeted non-prehensile pushing actions $\boldsymbol{p}_t$ to uncover occluded regions and their corresponding NBVs $v^*_{\boldsymbol{p}_t}$ using uncertainty-aware visibility corridors [1].

E. Multi-Manipulation Action Selection (*MMAS*)

As the final stage, **MMAS** module serves as the central active decision-making mechanism that selects the optimal action a^*_t by evaluating DOIG (Eq. (II-B)) of all skills.

III. Experiments

A. Simulation Experiments

a) Mapping Performance: Fig. 4 presents the results of our simulation experiments. We compare: (i) *Grasp Only*: pure grasping; (ii) *Push Only*: pure pushing that corresponds to the vanilla MEM [2]; (iii) *w/o CDC*: our method without the CDC penalty, where the standard OIG is used instead of DOIG; and (iv) *Ours*: our full approach. In general, *Ours* achieves high mapping accuracy and low scene disturbance at the same time.

IV. Conclusion

In this paper, we present MS-MEM, an evidential framework for uncertainty-aware mapping via active decision-making under grasp, push and active view selection. By extending MEM with the proposed *FE-vMF* for evidential grasp representation learning, *FE-UMGF* grasp fusion, and the unified DOIG objective, MS-MEM enables direct comparison of heterogeneous action skills under a shared evidential belief representation. Our results show that jointly leveraging pushing and grasping provides clear advantages over single-skill baselines for evidential metric-semantic mapping.

[1] We refer interested readers to [6], for the detail of the grasp fusion.

REFERENCES

[1] N. Dengler *et al.*, "Efficient manipulation-enhanced semantic mapping with uncertainty-informed action selection," in *Humanoids*, 2025.

[2] J. M. C. Marques *et al.*, "Map space belief prediction for manipulation-enhanced mapping," 2025.

[3] J. Delmerico *et al.*, "A comparison of volumetric information gain metrics for active 3d object reconstruction," *Autonomous Robots*, 2018.

[4] Y. Shi *et al.*, "vmf-contact: Uncertainty-aware evidential learning for probabilistic contact-grasp in noisy clutter," in *ICRA*, 2025.

[5] X. Wu *et al.*, "Point transformer v3: Simpler faster stronger," in *CVPR*, 2024.

[6] Y. Shi *et al.*, "Viso-grasp: vision-language informed spatial object-centric 6-dof active view planning and grasping in clutter and invisibility," in *IROS*, 2025.

[7] J. Gao *et al.*, "A comprehensive survey on evidential deep learning and its applications," *TPAMI*, 2025.