Abstract—We aim for mobile robots to function in a variety of common human environments. Such robots need to be able to reason about the locations of previously unseen target objects. Landmark objects can help this reasoning by narrowing down the search space significantly. More specifically, we can exploit background knowledge about common spatial relations between landmark and target objects. For example, seeing a table and knowing that cups can often be found on tables aids the discovery of a cup. Such correlations can be expressed as distributions over possible pairing relationships of objects. In this paper, we propose an active visual object search strategy based on our introduction of the Semantic Linking Maps (SLiM) model. SLiM simultaneously maintains the belief over a target object’s location as well as landmark objects’ locations, while accounting for probabilistic inter-object spatial relations. Based on SLiM, we describe a hybrid search strategy that selects the next best view pose for searching for the target object based on the maintained belief. We demonstrate the efficiency of our SLiM-based search strategy through comparative experiments in simulated environments. We further demonstrate the real-world applicability of SLiM-based search in scenarios with a Fetch mobile manipulation robot.

I. INTRODUCTION

Being able to efficiently search for objects in an environment is crucial for service robots to autonomously perform tasks [9], [27], [7]. When asked where a target object can be found, humans are able to give hypothetical locations expressed by spatial relations with respect to other objects. For example, a cup can be found “on a table” or “near a sink”. Table and sink are considered landmark objects that are informative for searching for the target object cup. Robots should be able to reason similarly about objects locations, as shown in Figure 1.

Previous works [10], [13], [26] assume landmark objects are static, in that they mostly remain where they were last observed. This assumption can be invalid for dynamic landmark objects that change their location over time, such as chairs, food carts and toolboxes. Temporal assumptions can mislead the search process if the prior on the landmarks’ locations is too strong. Further, there also exists uncertainty in the spatial relations between landmark objects and the target object, and between landmark objects themselves. For example, a cup can be “in” or “next to” a sink.

Considering the problem of dynamic landmarks, we propose the Semantic Linking Maps (SLiM) model to account for uncertainty in the locations of landmark objects during object search. Building on Lorbach et al. [18], we model inter-object spatial relations probabilistically via a factor graph. The marginal belief on inter-object spatial relations inferred from the factor graph is used in SLiM to account for probabilistic spatial relations between objects.

Using the maintained belief over target and landmark objects’ locations from SLiM, we propose a hybrid search strategy for active object search. We select the next best view pose, which guides the robot to explore promising regions that may contain the target and/or landmark objects. Previous works [30], [6], [25], [2] have shown the benefit of purposefully looking for landmark objects (Indirect Search) before directly looking for the target object (Direct Search). The proposed hybrid search strategy draws insights from both indirect and direct search. We demonstrate the efficiency of the proposed hybrid search strategy in our experiments.

In this paper, we describe the Semantic Linking Maps model as a Conditional Random Field (CRF). Our description of SLiM as a CRF allows us to simultaneously maintain the belief over target and landmark object locations with probabilistic modeling over inter-object spatial relations. We also describe a hybrid search strategy based on SLiM that draws upon ideas from both indirect and direct search representations. This SLiM-based search makes use of the maintained belief over objects’ locations by selecting the next best view pose based on the current belief. In our experiments, we show that the proposed object search approach is more robust to noisy priors on landmark locations by simultaneously maintaining belief over the locations of target and landmark objects.
II. RELATED WORK

Existing works have studied object search with different assumptions on prior knowledge of the environment. Some assume priors on landmark objects’ locations in the environment, and utilize the spatial relations between the target object and landmark objects to prioritize regions to search. Kollar et al. [10] utilize object-object co-occurrences extracted from image tags on Flickr.com to infer target object locations. Kunze et al. [13] expanded the generic notion of co-occurrences to more restrictive spatial relations (e.g. “in front of”, “left of”), which provide more confined regions to search, thus improving the search efficiency. Toris et al. [26] proposed to learn a temporal model on inter-object spatial relations to facilitate search. These methods assume the landmark objects to be static, however, we believe accounting for the uncertainty in landmark objects’ locations is important for object search.

Existing works have also explored known priors on spatial relations between landmark and target objects. Given exact spatial relations between landmark and target objects, Sjöö et al. [25] used an indirect object search strategy [30], [6], where the robot first searches for landmarks objects, and then searches for a target object in regions satisfying given spatial relations. On the other hand, given a probabilistic distribution over the spatial relations between objects, Aydemir et al. [2] formulate the object search problem as a Markov Decision Process. In our work, we learn the probabilistic inter-object spatial relations by building on ideas of Lorbach et al. [18], where inter-object relations are being probabilistically modeled via a factor graph.

There are also works that do not assume prior knowledge of the environment. Researchers have explored object search with visual attention mechanisms [23], [24], [19], such as saliency detection. Similar to [10], [13], other research [17], [4], [8] utilizes object-object co-occurrences to guide the search for a target object. Positive and negative detections of landmark objects will result in an updated belief over the target object. We expand object-object co-occurrences to finer-grained spatial relations between objects, i.e., “in”, “on”, “proximity”, “disjoint”, which specify more confined regions for object search.

Other literature [29], [12], [28] has also explored object-place relations to facilitate object search. Wang et al. [29] build a belief road map based on object-place co-occurrences for efficient path planning during object search. Kunze et al. [12] bootstraps commonsense knowledge on object-place co-occurrences from the Open Mind Indoor Common Sense (OMICS) dataset. Samadi learned similar knowledge by actively querying the World Wide Web (WWW). Our work also takes object-place co-occurrences into account. Aydemir et al. [1] made use of place-place co-occurrences to infer the type of the room next door, as the robot explores an environment during search. Manipulation-based object search, as in [32], [31], [14], is not within the scope of this paper.

III. PROBLEM STATEMENT

Let \( O = \{o^i | i = 1, \cdots, N\} \) be the set of objects of interest, including landmark objects and the target object for search. Given observations \( z_{0:T} \) and robot poses \( x_{0:T} \), we aim to maintain the belief over object locations \( P(O_T | x_{0:T}, z_{0:T}) \), while accounting for the probabilistic spatial relations \( R_{ij} \) between objects \( o^i, o^j \in O \). For this work, we consider the set of spatial relations to be \( R_{ij} \subset \{\text{In}, \text{On}, \text{Contain}, \text{Support}, \text{Proximity}, \text{Disjoint}\} \). For example, the relation \( R_{ij} = \text{In} \) indicates that object \( o^i \) is inside object \( o^j \). The probabilistic spatial relations between object \( o^i, o^j \) is represented by the belief over \( R_{ij} \), denoted as \( \text{B}(R_{ij}) \).

Based on the maintained belief \( P(O_T | x_{0:T}, z_{0:T}) \), the robot searches for the target object by selecting the next best view pose ranked by an utility function \( U : \tau \mapsto \mathbb{R} \). \( \tau \) specifies the 6 DOF of camera view pose. The utility function \( U \) trades off between navigation cost and the probability of search success. Upon a user request to find a target object, the robot iterates between the belief update of objects’ locations and view pose selection, until the target object is found or the maximum search time is reached.

IV. SEMANTIC LINKING MAPS

For Semantic Linking Maps (SLiM), we consider inter-object spatial relations, while maintaining the belief over target and landmark objects’ locations. Building on our previous work [33], we probabilistically formalize the object location estimation problem via a Conditional Random Field (CRF). The model is now extended to account for probabilistic inter-object spatial relations, as shown in Figure 2.

The posterior probability of object locations \( O \) history is

\[
p(O_{0:T} | x_{0:T}, z_{0:T}) = \frac{1}{Z} \prod_{t=1}^{T} \prod_{i=1}^{N} \phi_p(o^i_t, o^i_{t-1}) \phi_m(o^i_t, x_t, z_t) \prod_{t,j} \text{B}(R_{ij})(o^i_t, o^j_t)
\]

where \( Z \) is a normalization constant. Robot pose \( x_t \) and observation \( z_t \) are known. We assume that the robot stays localized given a metric map of the environment.

\( \phi_p(o^i_t, o^i_{t-1}) \) is the prediction potential that models the movement of an object over time. We assume objects to remain static or move with temporal coherence (varies across object classes) during the search, i.e.

\[
\phi_p(o^i_t, o^i_{t-1}) = e^{-\|o^i_t - o^i_{t-1}\|^2} S^{-1}(o^i_t - o^i_{t-1})
\]

\( \phi_m(o^i_t, x_t, z_t) \) is the measurement potential that accounts for the observation model, and \( z_t = \{z^i_t | i = 1, \cdots, N\} \) are (potentially noisy) detections for each object \( o^i \) at time \( t \). Because \( z^i_t \) and \( o^i_t \) are independent if \( j \neq i \), we simplify \( \phi_m(o^i_t, x_t, z_t) \) to \( \phi_m(o^i_t, x_t, z^i_t) \) s.t.,

\[
\phi_m(o^i_t, x_t, z^i_t) = \begin{cases} P_{TN}, & \text{if } o^i_t \in E^i_t \text{ but } z^i_t = \emptyset \\ P_{TN}, & \text{if } o^i_t \notin E^i_t \text{ and } z^i_t = \emptyset \\ P_{TP}, & \text{if } \pi(o^i_t) \in z^i_t \\ P_{TP}, & \text{otherwise} \end{cases}
\]

where each \( P \) stands for the probability of false negative, true negative, true positive, and false positive detection. \( E^i_t \) is the
Fig. 2: CRF-based SLiM model: (a) Known: \{x^t\} robot poses, \{z^t\} sensor observations; Unknown: \(O_t = \{o_1^t, o_2^t, \cdots, o_N^t\}\). (b) Plate notation: at time \(t\), the spatial relations between each object pair \(o^i, o^j\) is parameterized by the belief over their spatial relations \(B(R_{ij})\).

effective observation region for \(o^i\) given robot pose at time \(t\). Note, \(E_j^t\) is larger for larger objects, which can be reliably detected from longer distance compared to small objects. \(\pi\) is the camera projection matrix, and \(\pi(o^j) \in z_j^t\) denotes that the projected object lies in the detected bounding box in \(z_j^t\).

We model the spatial relations between objects with context potential \(\phi_{c,B(R_{ij})}\). Here, we extend \(\phi\) from our previous work by parametrizing it with the belief \(B(R_{ij})\) over the inter-object spatial relation between \(o^i, o^j\),

\[
\phi_{c,B(R_{ij})} = \sum_r B(R_{ij} = r) \phi_{c,r}(o^i, o^j, R_{ij} = r)
\]  

Algorithm 1: Inference of objects locations in SLiM.

Input: Observation \(z_t\), Robot pose \(x_t\).

1. Resample \(M\) particles \(o^i_{t-1}\) from \(o^i_{t-1}\) with probability proportional to importance weights \(\alpha_{t-1}^{(k)}\).
2. for \(i = 1, \cdots, n\) do
3. for \(k = 1, \cdots, M\) do
4. Sample \(o^i_{t} \sim \phi_p(o^i_{t-1})\).
5. Assign weight
6. \(\alpha_{t}^{(k)} \propto \phi_m(o^i_{t}, x_t, z_t) \prod_{j \in \Gamma(i)} \phi_{c,B(R_{ij})}(o^i_{t}, o^j_{t-1})\)
7. where \(\phi_{c,B(R_{ij})}(o^i_{t}, o^j_{t-1}) = \sum_r B(R_{ij} = r) \phi_{c,r}(o^i, o^j, R_{ij} = r)\)
8. end
9. end

C representative and divergent particles from the original \(M\) particles (\(C < M\)).

B. Probabilistic Inter-Object Spatial Relations

To get the belief over inter-object spatial relations \(B(R_{ij})\) for each object pair \(o^i, o^j\) \(\in O\), we use a factor graph by building on preceding work by Lorbach et al [18]. We generalize [18] by relaxing the assumption on known spatial relations between landmark objects.

The factor graph \(G\) \((\mathcal{V}, \mathcal{F}, \mathcal{E})\) consists of vertices \(\mathcal{V} = \{R_{ij} | i \neq j\ o^i, o^j \in O\}\), factor vertices \(\mathcal{F} = \{F_{CS}, F_{LC}\}\) and edges \(\mathcal{E}\) which connect factor vertices with variable vertices. Specifically, \(F_{CS} : R_{ij} \rightarrow \mathbb{R}\) is a unary factor that considers commonsense knowledge on spatial relation between objects, 

\[
F_{CS}(R_{ij}) = \text{Frequency}(R_{ij})
\]

Similar to [18], we extract commonsense knowledge on \(R_{ij}\) from online image search engine (e.g. Flickr) by counting the frequency of certain spatial relation between objects \(o^i, o^j\). For example, the frequency of \(R_{cup,table} = On\) is computed as the number of search results of a query “cup on the table” divided by the number of search results of a query “on the table”. These extracted frequencies can be noisy. For example, the frequency of “laptop on kitchen” is larger than 0, but it is not a valid expression because it refers to a laptop being on top of the rim geometry of a kitchen. We manually encode the \(F_{CS}(R_{ij})\) for invalid expressions to 0.

\[
F_{LC}(R_{ij}, R_{ik}, R_{jk}) = \begin{cases} 1, & \text{if consistent.} \\ 0, & \text{otherwise.} \end{cases}
\]

For example, if \(o^i\) is in \(o^j\), and \(o^j\) is in \(o^k\), then \(o^i\) should be in \(o^k\) to satisfy logical consistency, i.e., \(F_{LC}(R_{ij}) = In, R_{ik} = In, R_{jk} = In\) = 1. Previous work [18] assumes the spatial relations between landmark objects to be known.
and only relations $R_{\text{target}, j}$ connecting target object $o_{\text{target}}$ and landmark object $o_j$ to be unknown. Their pairwise factor enforcing logical consistency is a binary function $F_{\text{LC}} : (R_{\text{target}, j}, R_{\text{target}, k}) \mapsto \{0, 1\}$. In contrast, our formulation employs a trinary factor $F_{\text{LC}}$ considering all possible combinations of $(R_{ij}, R_{ik}, R_{jk})$ and evaluating their logical consistency.

By applying Belief Propagation [11] on the factor graph formulated as above, we can get the marginal belief over inter-object relations $B(R_{ij})$ between all object pairs. We use the libDAI [20] library for inference. An example of the probabilistic inter-object spatial relations inferred from the factor graph is as shown in Figure 5 and it is used in our experiments.

V. SEARCH STRATEGY

Based on the belief over the object locations, we actively search for the target object, by generating promising view poses and select the best one ranked by a utility function. Given the particle set $\{o_i^{(k)}, \psi_i^{(k)}\}$ of the target object $o$ as being maintained in $[\mathcal{V}]$ we fit Gaussian Mixture Models (GMMs) through Expectation Maximization to the particles by auto selecting the number of clusters [5],

$$\langle o_i^{(k)}, \psi_i^{(k)}\rangle \sim \mathcal{N}(x_n, \Sigma_n), \omega_n$$

(4)

A. View Pose Generation

For each Gaussian component $\mathcal{N}(x_n, \Sigma_n)$, we generate a set of camera view pose candidates $\{\tau_n^i = (c_n^i, \psi_n^i)\}$, where $c_n$ and $\psi_n$ denote the translation and the rotation of the camera respectively.

Initially, we sample the location of the camera $c_n$ evenly from a circle with a fixed radius around the center $x_n$ of the Gaussian component, and assign a default value to rotation $\psi_n$. Note, that these initially sampled view poses can put the robot in collision with the environment, and the camera is not necessarily looking at $x_n$. Thus, we formulate a view pose optimization problem under constraints as below,

$$\arg\min \|x_n - c_n\|_{}$$

(5)

subject to $x_n \in E_{\tau_n}$, $c(\tau_n) > 0$

where $v_n$ is the view direction given $\tau_n$. $E_{\tau_n}$ denotes the effective observation region of the target object at camera pose $\tau_n$, and $c : \tau \mapsto \mathbb{R}$ is a function that computes a signed distance of a configuration $\tau$ to the collision geometry of the environment.

B. View Pose Selection

We propose two different utility functions to rank the view pose candidates:

1) Direct Search utility: $U_{\text{DS}}$ encourages the robot to explore promising areas that could contain the target object while accounting for navigation cost,

$$U_{\text{DS}}(\tau_n) = \omega_n + \alpha \frac{1}{\arctan(\sigma d_{\text{nav}})}$$

(6)

where $\omega_n$ is the weight of the Gaussian component (as in (4)) that $\tau_n$ is generated from, and $d_{\text{nav}}$ is the navigation distance from the current robot location to view pose $\tau_n$. Parameter $\alpha$ trades off between the probability of finding the target object and the navigation cost. Parameter $\sigma$ determines how quickly the $\arctan(\sigma d_{\text{nav}})$ plateaus.

With $U_{\text{DS}}$, the object search is direct because we are directly considering promising areas represented by the GMMs for the target object.

2) Hybrid Search utility: $U_{\text{HS}}$ encourages the robot to explore promising areas that could contain the target object and/or any landmark object, while accounting for navigation cost

$$U_{\text{HS}}(\tau_n) = \omega_n + \alpha \frac{1}{\arctan(\sigma d_{\text{nav}})}$$

(7)

$$+ \beta \max_{j, n} \text{CoOccur}(o, o^j) \omega_n I^j_{\tau_n}$$

where $\beta$ is a small positive constant that encourages the exploration of landmarks.

Fig. 3: Examples of belief updates in SLiM. given observations. Upper: Evolution of particles of fridge, sink, coffee machine over time. Lower: RGB observation (with object detection) over time. (Best viewed in color).
In our experiments, we emulate the common cases where perfect knowledge about object locations which differ from the actual locations, to accurately recognize across the environment. IDS-Known-Unknown represents a naive approach for object search. All methods except for UDS are using the full SLiM model. We assume an effective observation range of 5m, 4m, 2m respectively.

We set up an object detector in simulation that returns a prior on landmark object locations. The particles for the target object are initialized uniformly across the environment is given. We also assume that the room types are as explained in IV-B. CM: coffee machine, CT: coffee table.

IDS-Known-Dynamic: Informed direct search (Eq.6) with a known prior on landmark object locations. This method resembles previous works [10], [13], [26].

IDS-Known-Static: Informed direct search (Eq.6) with a known prior on landmark object locations. The robot assumes that landmark objects are static at the locations provided by the prior. This method resembles previous works [6], [30].

IDS-Known-Direct: Informed direct search (Eq.6) without prior on landmark object locations. The particles for landmark objects are initialized uniformly across the environment. This method resembles previous works [17], [3].

IDS-Known-Hybrid: Informed hybrid search (Eq.7) with a known prior on landmark object locations.

We perform object search tasks in both simulation and real-world environments with a Fetch robot. In the simulation experiments, we quantitatively benchmark various methods, including methods that resemble previous works and our proposed method. In the real-world experiments, we demonstrate qualitatively that the proposed method scales to real-world applications. In both simulation and real-world experiments, the robot accelerates to at most 1m/s and turns at most at 1.7 rad/s.

1) Simulation Experiments: The simulation experiments are performed in an apartment-like environment (10m×11m) setup in the Gazebo simulator, as shown in Figure 4. The landmark objects are initialized uniformly across the environment at camera pose $\tau_o$, otherwise 0.

$U_{HS}$ is inspired by the indirect object search strategy as studied in [6], [30]. Previous studies demonstrated that purposefully looking for an intermediate landmark object helps quickly narrow down the search region for the target object if the landmark object often co-occurs with the target object, thus improving the search efficiency.

With $U_{HS}$, the object search can be considered hybrid because we are considering promising areas represented by GMMs for both the target object (as in direct search) and landmark objects that co-occur with the target object (as in indirect search).

In our experiments, we use a A* based planner to compute $d_{nav}$. We empirically set $\alpha = 0.1$, $\beta = 0.4$, and $\sigma = 0.5$ such that arctan($\sigma d_{nav}$) plateaus as $d_{nav}$ goes beyond 3m.

VI. EXPERIMENTS

Fig. 5: Marginal belief on inter-object spatial relations, as well as object-room relations, inferred from the factor graph as explained in Sec. IV-B.

We benchmark following methods:

- UDS: Uninformed direct search (Eq.6). The robot does not account for the spatial relations between the target and landmark objects (omitting Eq. 3 in SLiM). This baseline represents a naive approach for object search.
- IDS-Known-Static: Informed direct search (Eq.6) with a known prior on landmark object locations. The robot assumes that landmark objects are static at the locations provided by the prior. This method resembles previous works [10], [13], [26].
- IDS-Known-Dynamic: Informed direct search (Eq.6) with a known prior on landmark object locations. This is similar to IDS-Known-Static except that the robot does not assume the landmark objects to remain at the locations expressed in the prior.
- IDS-Known-Direct: Informed direct search (Eq.6) without prior on landmark object locations. The particles for landmark objects are initialized uniformly across the environment. This method resembles previous works [17], [3].
- IHS-Known-Hybrid: Informed hybrid search (Eq.7) without prior on landmark object locations.

All methods except for UDS are using the full SLiM model. We assume that an occupancy-grid map of the environment is given. We also assume that the room types are accurately recognized across the environment. IDS-Known-∗ methods are provided with a noisy prior on landmark object locations which differ from the actual locations, to emulate the common cases where perfect knowledge about landmark locations is not available. For all methods, the particles for the target object are initialized uniformly across the environment at camera pose $\tau_o$, otherwise 0.

$U_{DS}$ acts to encourage the robot to also explore areas that could contain landmark object $o^j$ which co-occurs with the target object $o$ with probability $CoOccur(o, o^j)$. Specifically, $CoOccur(o, o^j) = (1 - B(R_{target, j} = \text{Disjoint}))$, and $\omega^j_n$ is the weight of the n-th Gaussian component of GMMs fitted to the belief over the location of the landmark object $o^j$. And $I^j_n$ is 1 if the n-th Gaussian of object $o^j$ is within the effective observation region at camera pose $\tau_o$, otherwise 0.

The simulation experiments setup in the Gazebo simulator, as shown in Figure 4. The experiments, the robot accelerates to at most 1m/s and turns at most at 1.7 rad/s.
For each target object, we run 6 trials per method. In each trial, the robot starts at the same location, depicted in Figure 4. The object search is terminated if (1) the belief over the target object location has converged, or (2) the maximum search time of 5mins has been exceeded. A trial is successful if the robot finds the target object before timeout. For each target object and each method, we measure the number of view poses, search time, distance travelled by the robot, and search success rate averaged across all trials.

The benchmark result is as shown in Table I. Examples of the resulting search path from each method are depicted in Figure 6. As we can see, UDS is not as efficient because it is not making use of the spatial relations between the target and landmark objects in the environment. Given a noisy prior on landmark object locations, IDS-Known-Dynamic outperforms IDS-Known-Static because it accounts for the uncertainty of the landmark object locations, whereas IDS-Known-Static is misled by the noisy prior.

Given no prior information, IHS-unknown outperforms IDS-unknown because it encourages the robot to explore promising regions that contain the target and/or useful landmark objects, whereas IDS-unknown only considers promising regions that contain the target object. With IHS-unknown, the robot benefits from finding landmark objects which help narrow down the search region for the target object.

2) Real-World Experiments: The real-world experiment is executed in an environment (8mx8m) that consists of a kitchen and a living room. The robot stays localized in the pre-mapped environment based on its LIDAR, and navigates based on a MPEPC based path planner [21]. The target object is a cup, and landmark objects include table, sofa, coffee machine and sink. IHS-Unknown reached average success rate of 0.7 (7 out of 10 trials). The average number of view poses, search time and search path is 4.86, 103s, and 8.32m respectively. The failure cases were due to false negative detection of the cup due to lighting (we used Faster R-CNN [22] trained on COCO dataset [16]). Examples of real-world experiments with a Fetch robot are available in online video [https://youtu.be/uWWJ5aV6ScE].

VII. CONCLUSION

In this paper we present an efficient active visual object search approach through the introduction of the SLiM model. SLiM simultaneously maintains the belief over target and landmark objects locations, while accounting for the probabilistic inter-object spatial relations. Further, we propose a hybrid search strategy that draws insights from both direct and indirect object search. Given noisy or no prior on landmark objects locations, we demonstrate the benefit of modeling landmark objects locations under uncertainty in SLiM, and the hybrid search strategy that encourages the robot to explore promising areas that can contain the target and/or landmark objects in both simulation and real-world experiments.
