Semantic Linking Maps for Active Visual Object Search (Extended Abstract)

Zhen Zeng[1] , **Adrian Röfer**[2] , **Odest Chadwicke Jenkins**[1]
[1]Department of Electrical Engineering and Computer Science, Robotics Institute, University of Michigan, USA
[2]Department of Computer Science, University of Bremen, Germany
{zengzhen, ocj}@umich.edu, aroefer@uni-bremen.de

Abstract

We aim for mobile robots to function in a variety of common human environments, which requires them to efficiently search previously unseen target objects. We can exploit background knowledge about common spatial relations between landmark objects and target objects to narrow down search space. In this paper, we propose an active visual object search strategy method through 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.

1 Introduction

Being able to efficiently search for objects in an environment is crucial for service robots to autonomously perform tasks [Khandelwal *et al.*, 2017; Veloso *et al.*, 2015; Hawes *et al.*, 2017]. 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 [Kollar and Roy, 2009; Kunze *et al.*, 2014; Toris and Chernova, 2017] assume landmark objects are static. This assumption is invalid for dynamic landmark objects such as chairs. 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*.

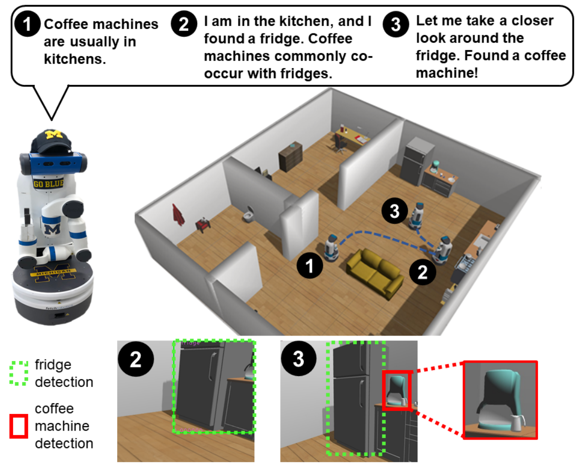

Figure 1: Robot tasked to find a coffee machine.

We propose the Semantic Linking Maps (*SLiM*) model to account for uncertainty of objects locations as well as inter-object spatial relations during object search. Building on Lorbach et al. [Lorbach *et al.*, 2014], we model inter-object spatial relations probabilistically via a factor graph. Inferred marginal belief from the factor graph is used in *SLiM* to account for probabilistic spatial relations between objects. We describe *SLiM*as a Conditional Random Field (CRF) model to simultaneously maintain the belief over target and landmark object locations with probabilistic modeling over inter-object spatial relations.

Using the maintained belief over target and landmark objects' locations from *SLiM*, we propose a hybrid 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 [Wixson and Ballard, 1994; Garvey, 1976; Sjöö *et al.*, 2012; Aydemir *et al.*, 2011] 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. In our experiments, We demonstrate the robustness of *SLiM*to uncertainty of object locations and inter-object spatial relations, as well as the efficiency of the proposed hybrid search strategy.

2 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 $R_{ij} \in$ {*In, On, Contain, Support, Proximity, Disjoint*}. For example, the relation $R_{ij} = In$ 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 $\mathcal{B}(R_{ij})$.

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 $U : \tau \mapsto \mathbb{R}$. τ specifies the 6 DOF of camera view pose. During object search, 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.

3 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 [Zeng *et al.*, 2018], 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

$$p(O_{0:T} | x_{0:T}, z_{0:T}) = \frac{1}{Z} \prod_{t=0}^{T} \prod_{i=1}^{N} \phi_p(o_t^i, o_{t-1}^i) \phi_m(o_t^i, x_t, z_t) \prod_{i,j} \phi_{c,\mathcal{B}(R_{ij})}(o_t^i, o_t^j) \quad (1)$$

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.

$\phi_p(o_t^i, o_{t-1}^i)$ is the *prediction potential* that models objects to remain static or move with temporal coherence (varies across object classes) during the search. $\phi_m(o_t^i, x_t, z_t)$ is the *measurement potential* that accounts for the observation model, and $z_t = \{z_t^i | i = 1, \cdots, N\}$ are (potentially noisy) detections for each object o^i at time t. We model the spatial relations between objects with *context potential* $\phi_{c,\mathcal{B}(R_{ij})}$. Here, we extend ϕ_c from our previous work by parameterizing it with the belief $\mathcal{B}(R_{ij})$ over the inter-object spatial relation between o^i, o^j,

$$\phi_{c,\mathcal{B}(R_{ij})} = \sum_r \mathcal{B}(R_{ij} = r) \phi_{c,r}(o_t^i, o_t^j, R_{ij} = r) \quad (2)$$

where r can take any value in the set of possible relations {*In, On, Contain, Support, Proximity, Disjoint*}. Please refer to the full paper [Zeng *et al.*, 2020] for more details on each potential function.

3.1 Inference

We propose a particle filtering inference method for maintaining the belief over object locations. Instead of estimating the posterior of the complete history of object locations

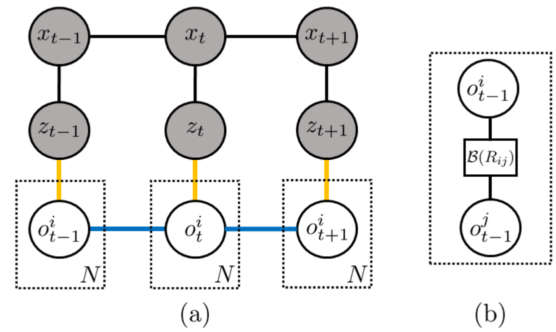

Figure 2: CRF-based *SLiM* model: (a) Known: $\{x^t\}$ robot poses, $\{z^t\}$ sensor observations; Unknown: $O_t = \{o_t^1, \cdots, o_t^N\}$. (b) Plate notation: at time t, the spatial relations between objects o^i, o^j is parameterized by the belief over their spatial relations $\mathcal{B}(R_{ij})$.

$p(O_{0:T} | x_{0:T}, z_{0:T})$, we recursively estimate the posterior probability of each object $o_t^i \in O_t$, similarly to [Zeng *et al.*, 2018; Limketkai *et al.*, 2007]. Please refer to the full paper [Zeng *et al.*, 2020] for more details on the particle filtering algorithm.

3.2 Probabilistic Inter-Object Spatial Relations

To get the belief over inter-object spatial relations $\mathcal{B}(R_{ij})$ for each object pair $o^i, o^j \in O$, we use a factor graph by building on preceding work by Lorbach et al [Lorbach *et al.*, 2014]. We generalize [Lorbach *et al.*, 2014] by relaxing the assumption on known spatial relations between landmark objects.

The factor graph $G : \{\mathbb{V}, \mathbb{F}, \mathbb{E}\}$ consists of variable vertices $\mathbb{V} = \{R_{ij} | \forall_{i \neq j}\, o^i, o^j \in O\}$, factor vertices $\mathbb{F} = \{F_{CS}, F_{LC}\}$ and edges $\mathbb{E}$ which connect factor vertices with variable vertices. Specifically, $F_{CS} : R_{ij} \mapsto \mathbb{R}$ is a unary factor that considers *commonsense knowledge* on spatial relation between objects, $F_{CS}(R_{ij}) = \text{Frequency}(R_{ij})$, extracted from online image search engine (e.g. Flickr) by counting the frequency of Rij, similarly to [Lorbach *et al.*, 2014]. F_{LC} : $(R_{ij}, R_{ik}, R_{jk}) \mapsto \{0, 1\}$ is a triplet factor that considers *logical consistency* between a triplet of objects o^i, o^j, o^k, with 1 if consistent and 0 otherwise. 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. We infer the marginal belief over inter-object relations $\mathcal{B}(R_{ij})$ through Belief Propagation [Kschischang *et al.*, 2001].

4 Search Strategy

Based on the belief over 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 maintained particles of the target object o in section 3, we fit Gaussian Mixture Models (GMMs) with auto selecting the number of clusters [Figueiredo and Jain, 2002],

$$\langle o_t^{(k)}, \alpha_t^{(k)} \rangle \sim \langle \mathcal{N}(x_n, \Sigma_n), \omega_n \rangle \quad (3)$$

4.1 View Pose Generation

For each Gaussian component $\mathcal{N}(\mathbf{x}_n, \Sigma_n)$, we generate a set of camera view pose candidates $\{\tau_n^i = (\mathbf{c}_n^i, \psi_n^i)\}$, where $\mathbf{c}_n$ and ψ_n denote the translation and the rotation of the camera respectively. Initially, we sample $\mathbf{c}_n$ evenly from a circle with a fixed radius around the center $\mathbf{x}_n$ of the Gaussian component, and assign a default value to ψ_n. Then, we optimize each sampled view pose via

$$\underset{\tau_n=\tau_n}{\operatorname{argmin}}\ 1-\mathbf{v}_n\cdot\frac{\mathbf{x}_n-\mathbf{c}_n}{\|\mathbf{x}_n-\mathbf{c}_n\|}\quad \text{s.t}\ \mathbf{x}_n\in E_{\tau_n},\quad c(\tau_n)>0 \tag{4}$$

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

4.2 View Pose Selection

We propose two different utility functions that trade off between navigation cost and the probability of search success.

Direct Search Utility

$\mathbf{U}_{\mathrm{DS}}$ enables **direct** object search, by encouraging the robot to explore promising areas that could contain the target object while accounting for navigation cost,

$$\mathbf{U}_{\mathrm{DS}}(\tau_k)=\omega_n+\alpha\frac{1}{\arctan(\sigma d_{nav})} \tag{5}$$

where ω_n is the weight of the Gaussian component (as in (3)) that τ_k is generated from, and d_{nav} is the navigation distance from the current robot location to view pose τ_k. Parameter α trades off between the probability of finding the target object and the navigation cost. Parameter σ determines how quickly the $\arctan(\sigma d_{nav})$ plateaus. 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 $3m$.

Hybrid Search Utility

$\mathbf{U}_{\mathrm{HS}}$ enables **hybrid** object search, by encouraging the robot to explore promising areas that could contain the target object and/or any landmark object, while accounting for navigation cost

$$\begin{aligned}\mathbf{U}_{\mathrm{HS}}(\tau_k)=\omega_n+\alpha\frac{1}{\arctan(\sigma d_{nav})}\\ +\beta\max_{j,n}\ \mathrm{CoOccur}(o,o^j)\omega_n^j\mathbf{I}_n^j\end{aligned} \tag{6}$$

where the additional term compared to $\mathbf{U}_{\mathrm{DS}}$ 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 $\mathrm{CoOccur}(o,o^j)$, inspired by *indirect object search* strategy [Garvey, 1976; Wixson and Ballard, 1994]. Specifically, $\mathrm{CoOccur}(o,o^j)=(1-\mathcal{B}(R_{target,j}=Disjoint))$, and ω_n^j is the weight of the n-th Gaussian component of GMMs fitted to landmark object o^j particles. And $\mathbf{I}_n^j$ is 1 if the n-th Gaussian of object o^j is within the effective observation region at camera pose τ_k, otherwise 0.

5 Experiments

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

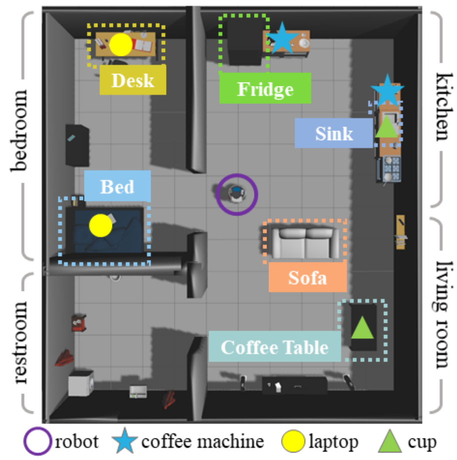

Figure 3: An simulated apartment with four rooms. There are 6 landmark objects and 3 target objects: *coffee machine, laptop, cup*. Each target object has two equally possible locations.

Simulation Experiments

The simulation experiments are performed in an apartment-like environment (10mx11m) in the Gazebo, as shown in Figure 3. Object detector returns an object detection, if the object is in view, not fully occluded, and within the effective observation range. For large objects (e.g. sofa, fridge), mid-sized objects (e.g. table, sink), and small objects (e.g. cup, laptop, coffee machine), the effective observation range is 5m, 4m, 2.5m respectively. We benchmark following methods.

UDS. Uninformed direct search (Eq.5). The robot does not account for the spatial relations between the target and landmark objects (omitting Eq. 2 in *SLiM*). This baseline represents a naive approach for object search.

IDS-Known-Static. Informed direct search (Eq.5) with a known prior on landmark object locations. The robot assumes that landmark objects are static. This method resembles previous works [Kollar and Roy, 2009; Kunze *et al.*, 2014; Toris and Chernova, 2017].

IDS-Known-Dynamic. Informed direct search (Eq.5) 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 static.

IDS-Unknown. Informed direct search (Eq.5) without prior on landmark object locations. The particles for landmark objects are initialized uniformly across the environment. This method resembles previous works [Loncomilla *et al.*, 2018; Aydemir *et al.*, 2010].

IHS-Unknown. Informed hybrid search (Eq.6) 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. IDS-Known-* methods are provided with a noisy

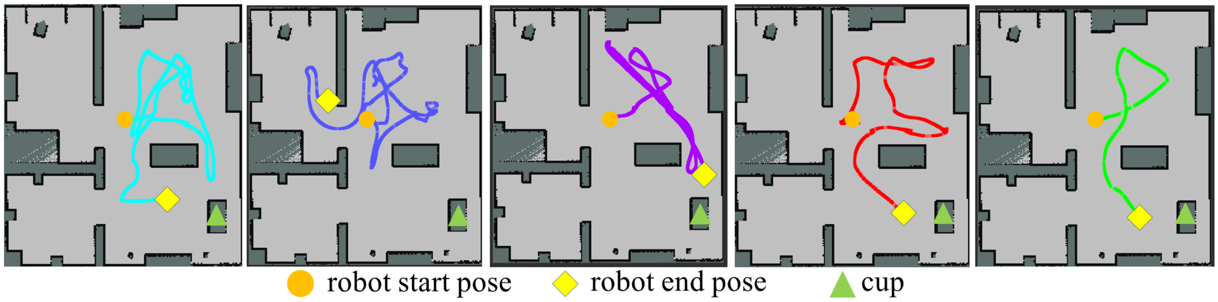

Figure 4: Examples of search paths generated by each method while searching for *cup*. Methods from left to right: UDS, IDS-Known-Static, IDS-Known-Dynamic, IDS-Unknown, IHS-Unknown. (Best viewed in color).

Target Object	Metrics	UDS	IDS known, static	IDS known, dynamic	IDS unknown	IHS unknown
Coffee Machine	Views	7.83	6.17	4.67	6.33	**3.67**
	Search Time (s)	107	76	60	75	**50**
	Search Path (m)	8.68	6.70	5.80	6.74	**4.93**
	Success Rate	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Laptop	Views	11.00	12.50	7.17	5.67	**4.17**
	Search Time (s)	197	222	124	91	**78**
	Search Path (m)	28.27	26.86	13.13	**7.69**	8.40
	Success Rate	0.83	0.50	1.00	1.00	1.00
Cup	Views	13.17	14.50	12.67	11.83	**9.00**
	Search Time (s)	184	229	189	185	**139**
	Search Path (m)	22.64	29.81	23.40	19.68	**13.91**
	Success Rate	0.83	0.33	0.83	0.83	**1.00**

Table 1: Benchmark results for object search in simulation experiments. Among methods that reached 100% success rate, IHS unknown successfully found target objects within the smallest number of views and least search time.

prior on landmark object 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.

For each target object, we run 6 trials per method. In each trial, the robot starts at the same location shown in Figure 3. A trial is successful if the robot finds the target object (i.e. belief converges at the correct location) 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 1. Examples of the resulting search path from each method are depicted in Figure 4. 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.

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 with LIDAR, and navigates with a MPEPC path planner [Park *et al.*, 2012]. 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 repectively. The failure cases were due to false negative detection of the cup due to lighting (we used Faster R-CNN [Ren *et al.*, 2017] trained on COCO dataset [Lin *et al.*, 2014]). Examples of real-world experiments with a Fetch robot is available in online video https://youtu.be/uWWJ5aV6ScE.

6 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.

References

[Aydemir *et al.*, 2010] Alper Aydemir, Kristoffer Sjöö, and Patric Jensfelt. Object search on a mobile robot using relational spatial information. In *Proceedings of International Conference on Intelligent Autonomous Systems*, pages 111–120, 2010.

[Aydemir *et al.*, 2011] Alper Aydemir, Kristoffer Sjöö, John Folkesson, Andrzej Pronobis, and Patric Jensfelt. Search in the real world: Active visual object search based on spatial relations. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 2818–2824. IEEE, 2011.

[Figueiredo and Jain, 2002] Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):381–396, 2002.

[Garvey, 1976] Thomas D Garvey. Perceptual strategies for purposive vision. Tech. Rep. AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025., 1976.

[Hawes *et al.*, 2017] Nick Hawes, Christopher Burbridge, Ferdian Jovan, Lars Kunze, Bruno Lacerda, Lenka Mudrova, Jay Young, Jeremy Wyatt, Denise Hebesberger, Tobias Kortner, et al. The strands project: Long-term autonomy in everyday environments. *IEEE Robotics & Automation Magazine*, 24(3):146–156, 2017.

[Khandelwal *et al.*, 2017] Piyush Khandelwal, Shiqi Zhang, Jivko Sinapov, Matteo Leonetti, Jesse Thomason, Fangkai Yang, Ilaria Gori, Maxwell Svetlik, Priyanka Khante, Vladimir Lifschitz, et al. Bwibots: A platform for bridging the gap between ai and human–robot interaction research. *The International Journal of Robotics Research*, 36(5-7):635–659, 2017.

[Kollar and Roy, 2009] Thomas Kollar and Nicholas Roy. Utilizing object-object and object-scene context when planning to find things. In *Robotics and Automation (ICRA), 2009 IEEE International Conference on*, pages 2168–2173. IEEE, 2009.

[Kschischang *et al.*, 2001] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, 2001.

[Kunze *et al.*, 2014] Lars Kunze, Keerthi Kumar Doreswamy, and Nick Hawes. Using qualitative spatial relations for indirect object search. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 163–168. IEEE, 2014.

[Limketkai *et al.*, 2007] Benson Limketkai, Dieter Fox, and Lin Liao. Crf-filters: Discriminative particle filters for sequential state estimation. In *Robotics and Automation (ICRA), 2007 IEEE International Conference on*, pages 3142–3147. IEEE, 2007.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[Loncomilla *et al.*, 2018] Patricio Loncomilla, Javier Ruiz-del Solar, and Marcelo Saavedra. A bayesian based methodology for indirect object search. *Journal of Intelligent & Robotic Systems*, 90(1-2):45–63, 2018.

[Lorbach *et al.*, 2014] Malte Lorbach, Sebastian Höfer, and Oliver Brock. Prior-assisted propagation of spatial information for object search. In *Intelligent Robots and Systems (IROS), 2014 IEEE/RSJ International Conference on*, pages 2904–2909. IEEE, 2014.

[Park *et al.*, 2012] Jong Jin Park, Collin Johnson, and Benjamin Kuipers. Robot navigation with model predictive equilibrium point control. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 4945–4952. IEEE, 2012.

[Ren *et al.*, 2017] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2017.

[Sjöö *et al.*, 2012] Kristoffer Sjöö, Alper Aydemir, and Patric Jensfelt. Topological spatial relations for active visual search. *Robotics and Autonomous Systems*, 60(9):1093–1107, 2012.

[Toris and Chernova, 2017] Russell Toris and Sonia Chernova. Temporal persistence modeling for object search. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 3215–3222. IEEE, 2017.

[Veloso *et al.*, 2015] Manuela Veloso, Joydeep Biswas, Brian Coltin, and Stephanie Rosenthal. Cobots: robust symbiotic autonomous mobile service robots. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 4423–4429. AAAI Press, 2015.

[Wixson and Ballard, 1994] Lambert E Wixson and Dana H Ballard. Using intermediate objects to improve the efficiency of visual search. *International Journal of Computer Vision*, 12(2-3):209–230, 1994.

[Zeng *et al.*, 2018] Zhen Zeng, Yunwen Zhou, Odest Chadwicke Jenkins, and Karthik Desingh. Semantic mapping with simultaneous object detection and localization. In *Intelligent Robots and Systems (IROS), 2018 IEEE/RSJ International Conference on*, pages 911–918. IEEE, 2018.

[Zeng *et al.*, 2020] Zhen Zeng, Adrian Röfer, and Odest Chadwicke Jenkins. Semantic linking maps for active visual object search. In *Robotics and Automation (ICRA), 2020 IEEE International Conference on*, pages 1984–1990. IEEE, 2020.