Towards Self-Supervised Semantic Representation with a Viewpoint-Dependent Observation Model

Yuri Feldman and Vadim Indelman

***Abstract*—In the context of semantic SLAM, we propose to represent the semantic information attached to objects (or generally, scenes) as continuous vectors in a latent space induced by a learned predictive observation model. We propose two observation models relating spatial changes in semantic measurements of an object to the latent object representation, and show how they can be used for joint inference of geometry and semantics free of discrete variables by maintaining a posterior over robot trajectory, geometric object and environment properties, and the learned object latent semantic representation. Both models relax assumptions on ground truth required to learn them w.r.t. existing analogues. In particular, one of the models is a residual-style formulation which can be learned in a weakly-supervised manner, under relatively mild assumptions of locally correct odometry and data association between some object detections in consecutive frames, not requiring prior knowledge of candidate object categories, or an object coordinate system to be defined.**

I. Introduction

Semantic SLAM methods aim to construct a representation of the robot environment that captures information beyond geometric, and localize the robot within it. A common approach, known as "Object-Level SLAM" [9], is to represent objects detected in the robot environment as semantic landmarks and estimate their poses and discrete category labels jointly with robot trajectory. Inference of object 6dof poses is enabled by offline fitting of a viewpoint-dependent measurement model for each object category, requiring representative measurements and ground truth poses of camera relative to the objects be known at training time, which limits the number of object categories that can be thus represented. Further, category labels carry limited and specialized semantic information, as generally the definition of categories is not unique. Adding additional categorical information to object state leads to a combinatorial increase in inference complexity, and requires hand-tailored observation models.

II. Notations and Problem Formulation

Consider a robot traversing an unknown environment, taking observations of different scenes. Robot motion between times t_k and t_{k+1} is initiated by a control input $\mathcal{U}_k$, that may originate from a human user, or be determined by a motion planning algorithm. We denote the robot pose at time instant k by x_k, and by $\mathcal{X}_{0:k} = \{\mathcal{X}_0, \dots, \mathcal{X}_k\}$ the sequence of poses up to that time. We denote by $\mathcal{Z}_k = \{z_{k,(i)}\}$ all observations obtained at time k. We denote with $\mathcal{Z}_k^g \subseteq \mathcal{Z}_k$ (correspondingly, $\mathcal{Z}_k^g = \{z_{k,(i)}^g\}$) geometric detections, including detections of landmarks, object bounding boxes (coordinates) and centroids. We denote with $\mathcal{Z}_k^s \subseteq \mathcal{Z}_k$ (correspondingly, $\mathcal{Z}_k^s = \{z_{k,(i)}^s\}$) the semantic measurements, i.e. bounding box RGB images.

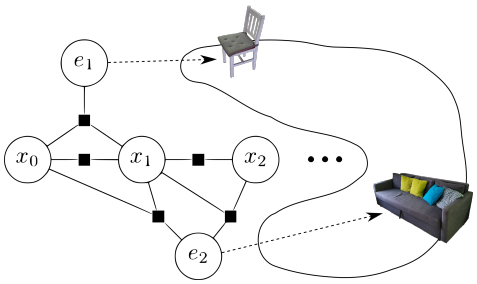


Fig. 1: An example factor graph with two objects described by latent representations e_1 and e_2 observed from consecutive poses $x_{0:2}$. Ternary factors correspond to the learned semantic observation model Eq. (15)

We further denote the observation and user control history up until time k as $\mathcal{H}_k = \{\mathcal{U}_{0:k-1}, \mathcal{Z}_{0:k}\}$.

We are interested in maintaining a posterior

$$\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{L}, \mathcal{O}, \mathcal{E} \mid \mathcal{H}_k), \tag{1}$$

over robot track, geometric landmarks $\mathcal{L}$, object geometric information $\mathcal{O} = \{o_i\}$ (centroids in the simplest case, other representation possibilities exist including ellipsoids [7] or cubes [15]), and per-object continuous variables $\mathcal{E} = \{e_i\}$ capturing object semantic information. In the following, we

III. Approach

The update equation for the target posterior at time k can be generally written by applying Bayes rule as

$$\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{L}, \mathcal{O}, \mathcal{E} \mid \mathcal{H}_k) = \tag{2}$$

$$\eta \cdot \mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_{0:k}, \mathcal{L}, \mathcal{O}, \mathcal{E}, \mathcal{H}_k^-) \cdot \mathbb{P}(\mathcal{X}_k \mid \mathcal{X}_{k-1}, \mathcal{U}_{k-1}) \tag{3}$$

$$\cdot \mathbb{P}(\mathcal{X}_{0:k-1}, \mathcal{L}, \mathcal{O}, \mathcal{E} \mid \mathcal{H}_k), \tag{4}$$

where $\mathcal{H}_k^- \doteq \mathcal{H}_k \setminus \{\mathcal{Z}_k\}$, and η is a constant normalization factor. In the above, the motion model $\mathbb{P}(\mathcal{X}_k \mid \mathcal{X}_{k-1}, \mathcal{U}_{k-1})$ is assumed known, and priors for the map variables $\mathcal{L}, \mathcal{O}, \mathcal{E}$ may be present or assumed uninformative. We further split the measurement update term into a "geometric" and a "semantic" part

$$\mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_{0:k}, \mathcal{L}, \mathcal{O}, \mathcal{E}, \mathcal{H}_k^-) = \tag{5}$$

$$\mathbb{P}(\mathcal{Z}_k^g \mid \mathcal{X}_k, \mathcal{L}, \mathcal{O}) \cdot \mathbb{P}(\mathcal{Z}_k^s \mid \mathcal{X}_{0:k}, \mathcal{O}, \mathcal{E}, \mathcal{H}_k^-). \tag{6}$$

The former (geometric) term is comprised of geometric models e.g. projection factors (and thus given variables does not depend on measurement history). The latter term is the semantic observation model. Neglecting interactions among object

detections (e.g. occlusions) and assuming that data association is known we can split the semantic term into per-detection components

$$\mathbb{P}(\mathcal{Z}_k^s \mid \mathcal{X}_{0:k}, \mathcal{O}, \mathcal{E}, \mathcal{H}_k^-) = \prod_i \mathbb{P}(z^s_{k,(o_i)} \mid \mathcal{X}_{0:k}, o_i, e_i, \mathcal{H}_k^-), \tag{7}$$

where by a slight abuse of notation $z^s_{k,(o_i)}$ is an observation corresponding to object o_i.

In a factor graph representation, the above components correspond to semantic factors, involving a semantic measurement, variables describing the measured object geometry (o_i) and semantics (e_i) and robot poses when measuring the object.

In the below, we consider two forms for the semantic measurement model Eq. (7) and show how they can be learned under relatively mild requirements on ground-truth data.

A. Viewpoint-Dependent Model

Viewpoint-dependent semantic observation models allow joint inference of semantics and geometry (robot track, object localization). Commonly ([14, 2, 13, 4, 12]), such models capture the distribution of the output of a detector or a classifier conditioned on the pose relative to the object, and thus are of the form

$$\mathbb{P}(\cdot \mid c, x^{(rel)}), \tag{8}$$

where c is the object class (discrete variable) and $x^{(rel)}$ is the pose relative to the object from which the measurement was taken. In practice, a separate model of this form needs to be fit for each class of interest, in particular requiring ground truth classification for the training set.

We propose to consider a semantic model similar to Eq. (8), replacing the discrete class variable with a per-object semantic representation vector ε_i, i.e. - explicitly writing the term from Eq. (7)

$$\mathbb{P}(z^s_{k,(o_i)} \mid e_i, \mathcal{X}_{0:k}, o_i, \mathcal{H}_k^-) = \mathbb{P}(z^s_{k,(o_i)} \mid x_k, o_i, e_i) \tag{9}$$

$$= \mathbb{P}(z^s_{k,(o_i)} \mid e_i, \mathcal{X}_k^{(rel)}), \tag{10}$$

with $\mathcal{X}_k^{(rel)}$ - the robot pose relative to the object at time k. Similarly to Eq. (8) under this model the relative pose is required to be known at training time, but contrary to it, the requirement for ground truth classification can be relaxed to a requirement of data association among measurements of the same instance, as we show below.

Fitting the Viewpoint-Dependent Model

A model of the form Eq. (10) can be learned from data as the decoder in a Conditional-VAE [11] framework. We assume a training dataset comprised of measurements (RGB bounding boxes) $\mathcal{Z}^s_{0:k}$. To simplify notations, in this subsection we drop the superscript and assume a single measurement per time index, i.e. $\mathcal{Z}_k^s \equiv \mathcal{Z}_k = \{z_k\}$ (and thus, use $\mathcal{Z}_k$ and z_k interchangeably). We assume that we are given the corresponding poses relative to the object instance $\mathcal{X}^{(rel)}_{0:k}$ viewed at each time step (might be a different one at each time index k), and the observed object instance identifiers in discrete variables $\beta_{0:k}$.

Formally, we wish to maximize the joint posterior over object measurements and instance identifiers given relative viewpoints (we neglect correlations among time steps)

$$\log \mathbb{P}(\mathcal{Z}_{0:k}, \beta_{0:k} \mid \mathcal{X}^{(rel)}_{0:k}) = \sum_k \log \mathbb{P}(z_k, \beta_k \mid \mathcal{X}_k^{(rel)}). \tag{11}$$

We choose the standard Gaussian prior for the latent semantic representation $\mathbb{P}(e)$ and write the evidence lower bound as[1]

$$\log \mathbb{P}(z_k, \beta_k \mid \mathcal{X}_k^{(rel)}) \geq \tag{12}$$
$$-KL(q_\phi(e \mid z_k, \mathcal{X}_k^{(rel)}) \parallel \mathbb{P}(e)) +$$
$$\mathop{\mathbb{E}}_{e \sim q_\phi(\cdot \mid z_k, \mathcal{X}_k^{(rel)})} \{\log \mathbb{P}_\theta(z_k, \beta_k \mid e, \mathcal{X}_k^{(rel)})\},$$

with standard notations of encoder q_ϕ and decoder p_θ and

$$\mathbb{P}_\theta(z_k, \beta_k \mid e, \mathcal{X}_k^{(rel)}) = p_\theta(z_k \mid e, \mathcal{X}_k^{(rel)}) \cdot \mathbb{P}(\beta_k \mid e) \tag{13}$$

$$\propto p_\theta(z_k \mid e, \mathcal{X}_k^{(rel)}) \cdot \mathbb{P}(e \mid \beta_k), \tag{14}$$

the latter proportionality true by assuming uninformative priors on e and β_k. The term $\mathbb{P}(e \mid \beta_k)$ determines how likely a given representation is for a given instance. In practice, this term pulls together the latent vectors obtained for various viewpoints of the same instance. We model it with a Gaussian (in latent space) with a set covariance (a parameter), fitting a per-instance mean vector. This term can be used at test time to determine the instance most likely to be the one having produced a measurement. This usage however is only valid for instances seen at training time, and is different from the more general task of classification (or "concept grounding"), which we address in Sec. IV-A.

B. Viewpoint-Predictive Model

Fitting a model of the form Eq. (10) requires training-time knowledge of the pose relative to the object $\mathcal{X}^{(rel)}$. Implicitly, it also requires a coordinate system for the object to be defined. Such an object coordinate system is in general not unique and application - specific. Worse, while it makes sense to define a coordinate system for individual object instances, it can be less straightforward to extend across different instances of the same class (as implied by the observation model), accommodating for intra-class variations in appearance and possibly functionality. Finally, object coordinate system needs to be (essentially, manually) defined across class instances for each class of interest, limiting the general applicability of the approach.

We propose to sidestep the definition of an object coordinate frame by using a "viewpoint-predictive" model of the form

$$\mathbb{P}(z^s_{k,(o_i)} \mid \mathcal{X}_{0:k}, o_i, e_i, \mathcal{H}_k^-) = \mathbb{P}(z^s_{k+1} \mid z^s_k, e_i, \Delta\mathcal{X}_k), \tag{15}$$

relating object measurement from the current time $z^s_k \in \mathcal{H}_k^-$ to the next measurement z^s_{k+1} of an object described by semantic representation vector e_i, with $\Delta\mathcal{X}_k = \mathcal{X}_{k+1} \ominus \mathcal{X}_k$ the camera

[1]The full derivation of the bound is provided in [5] Sec. A.

motion between the time steps, and the underlying assumption of known association between z^s_k and z^s_{k+1}. Conditioning of the model on camera motion both obviates the need to define an object coordinate frame, and may allow the model to be learned in a weakly-supervised manner, only requiring locally correct odometry, as we will show next.

Fitting the Viewpoint-Predictive Model

Adopting once more the notations from Sec. III-A, our goal is to maximize the joint posterior over measurements

$$\mathbb{P}(\mathcal{Z}_{0:k} \mid \mathcal{U}_{0:k-1}) = \tag{16}$$

$$\int \mathbb{P}(\mathcal{Z}_{0:k} \mid \Delta\mathcal{X}_{0:k}) \cdot \mathbb{P}(\Delta\mathcal{X}_{0:k} \mid \mathcal{U}_{0:k-1}) \; d\{\Delta\mathcal{X}_{0:k}\} \tag{17}$$

$$\overset{(1)}{\approx} \mathbb{P}(\mathcal{Z}_{0:k} \mid \Delta\widehat{\mathcal{X}}_{0:k}) \overset{(2)}{\propto} \prod_k \mathbb{P}(z_{k+1} \mid z_k, \Delta\widehat{\mathcal{X}}_k), \tag{18}$$

where $\Delta\mathcal{X}_k \doteq \mathcal{X}_{k+1} \ominus \mathcal{X}_k$ is the robot motion between time indexes k and $k+1$ and

$$\Delta\widehat{\mathcal{X}}_{0:k} \doteq \arg\max_{\Delta\mathcal{X}_{0:k}} \mathbb{P}(\Delta\mathcal{X}_{0:k} \mid \mathcal{U}_{0:k-1}), \tag{19}$$

i.e. a maximum likelihood single-sample approximation (1) in Eq. (18) and a uniform prior assumption on $\mathbb{P}(z_0)$ in proportionality (2) (in the same equation). The target function thus obtained only requires knowledge of (local) camera motion and implicitly - of data association, as measurements z_k, z_{k+1} are assumed to originate in the same object. In particular, the product (1) in Eq. (18) in practice splits into independent per-object products (omitted here for simplicity of presentation).

Finally, note that although the assumption of known robot motion appears formally equivalent to the assumption of known localization w.r.t. the initial pose, in practice it is much weaker, as the estimate Eq. (19) is only required to be locally correct, and there is no accumulated drift.

We can now write down the lower bound for optimizing the objective Eq. (18). As before, we develop for a single term - the summand, in the log posterior expression, or a factor in the product as it appears in Eq. (18).

$$\mathbb{P}(\mathcal{Z}_{k+1} \mid \mathcal{Z}_k, \Delta\mathcal{X}_k) \geq \tag{20}$$
$$-KL(q_\phi(e \mid \mathcal{Z}_k) \parallel \mathbb{P}(e)) +$$
$$\mathop{\mathbb{E}}_{e\sim q_\phi(e|\mathcal{Z}_k)} \{\log p_\theta(\mathcal{Z}_{k+1} \mid e, \mathcal{Z}_k, \Delta\mathcal{X}_k) + \log p_\theta(\mathcal{Z}_k \mid e)\} + K,$$

where K is (a negative) constant w.r.t. the optimization. The full derivation is provided in [5], Sec. B. The obtained expression does not depend on ground truth beyond data association and robot motion between consecutive time indexes and so can be fit in a weakly-supervised manner.

IV. Feasibility of the Approach for Online Semantic SLAM

In the following we address the applicability of our approach in the context of three specialized sub-tasks of online semantic SLAM, namely: inference of semantics (Sec. IV-A), localization and mapping (Sec. IV-B), online operation (Sec. IV-C).

A. "Grounding" of the Semantic Representation

Our formulation thus far made no use of semantic information beyond instance-level data association, yet - the latent representation variables would clearly carry semantic information (as long as instance - level association, or object images can be inferred back from them, as in Eqs. (12, 20, 14). In other words, the described approach could allow to perform semantic mapping without need for ground truth class information, provided that the learned model holds, i.e. that objects encountered at test time are not very different from those seen during training (which in turn requires little ground truth and so could be performed on demand).

Still, for the map to be interpretable by a human operator, given a (task-specific and possibly non-unique) set of candidate classes, a correspondence to the latent space needs to be established. Pirk et al. [8] report the learned latent space (with a different objective function) to reflect semantic structure, i.e. objects that are similar in appearance (and thus, in their case - in function) tend to emergently be close in the learned space (based on appearance only). In such an ideal case, manual tagging of a few example images (which through the encoder induces tagging of a region of the latent space) followed by a nearest neighbor search in the latent space could likely produce satisfactory classification results.

In the general case however, classes of interest could encompass objects with significantly varying appearance which will not be close in the latent space unless explicitly forced. In such a case we can assume a limited amount of classification-tagged instances (a single tagged image implies a tagged instance because of the known data association assumption). For a measurement z_k for which classification is available (denote class c), we can modify the cost function along the lines of Eq. (14) to pull its latent representation (encoder output) closer to that of others of the same class. For a viewpoint-dependent model Sec. III-A the modified term (from Eq. (12)) would be

$$\mathbb{P}(\mathcal{Z}_k, \beta_k, c \mid \mathcal{X}^{(rel)}_k) \geq \tag{21}$$
$$-KL(q_\phi(e \mid \mathcal{Z}_k, \mathcal{X}^{(rel)}_k) \parallel \mathbb{P}(e)) +$$
$$\mathop{\mathbb{E}}_{e\sim q_\phi(\cdot|\mathcal{Z}_k, \mathcal{X}^{(rel)}_k)} \{\log \mathbb{P}_\theta(\mathcal{Z}_k, \beta_k, c \mid e, \mathcal{X}^{(rel)}_k)\},$$

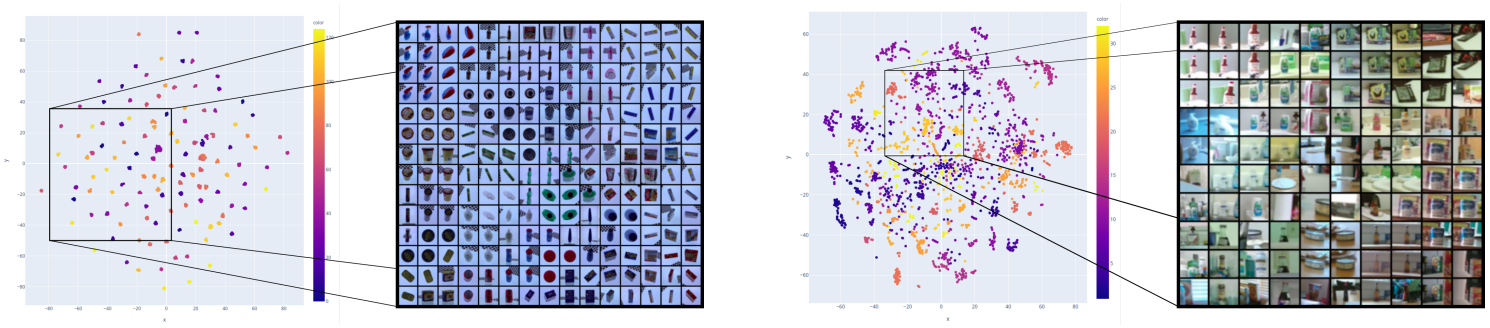
with

$$\mathbb{P}_\theta(\mathcal{Z}_k, \beta_k, c \mid e, \mathcal{X}^{(rel)}_k) \propto \tag{22}$$
$$p_\theta(\mathcal{Z}_k \mid e, \mathcal{X}^{(rel)}_k) \cdot \mathbb{P}(e \mid \beta_k) \cdot \mathbb{P}(e \mid c), \tag{23}$$

where $\mathbb{P}(e \mid c)$ a Gaussian density with constant covariance and class-specific learnable mean, as in Eq. (14). Similarly for the viewpoint-predictive model Sec. III-B the modified term from Eq. (20) is

$$\mathbb{P}(\mathcal{Z}_{k+1}, c \mid \mathcal{Z}_k, \Delta\mathcal{X}_k) \geq \tag{24}$$
$$-KL(q_\phi(e \mid \mathcal{Z}_k) \parallel \mathbb{P}(e)) +$$
$$\mathop{\mathbb{E}}_{e\sim q_\phi(e|\mathcal{Z}_k)} \{\log p_\theta(\mathcal{Z}_{k+1} \mid e, \mathcal{Z}_k, \Delta\mathcal{X}_k) +$$
$$\log p_\theta(\mathcal{Z}_k \mid e) + \mathbb{P}(e \mid c)\},$$

with $\mathbb{P}(e \mid c)$ defined as before.



(a) Embedding space induced by the Viewpoint-Dependent model fit on the BigBIRD dataset

(b) Embedding space induced by the Viewpoint-Predictive model fit on the Active Vision Dataset

Fig. 2: Learned latent space visualization using tSNE [6]. Left: latent vectors (encoder output) and (roughly) corresponding input images from the BigBIRD dataset [10], while training the Viewpoint-Dependent model from Sec. III-A. Right: the embedding space obtained when fitting the Viewpoint-Predictive model Sec. III-B on pairs of detections from adjacent views of the Active Vision Dataset [1].

B. Improving Localization

A similar principle as was mentioned in the previous clause to adapt the latent space to facilitate classification, could be applied to attempt to facilitate localization. Formally, analogously to the above Sec. IV-A and in the notations of Eq. (1), we could set the optimization objective to be the joint posterior

$$\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{L}, \mathcal{O}, \mathcal{Z}^s_{0:k} \mid \mathcal{Z}^g_{0:k}, \mathcal{U}_{0:k-1}) = \tag{25}$$

$$\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{L}, \mathcal{O}, \mid \mathcal{H}_k) \cdot \mathbb{P}(\mathcal{Z}^s_{0:k} \mid \mathcal{Z}^g_{0:k}, \mathcal{U}_{0:k-1}). \tag{26}$$

Here the second term in Eq. (26) can be developed similarly to e.g. Eq. (18), while the former term is the localization problem from Eq. (1) with a twist: here the model parameters, participating in the semantic factors are variable too, giving a scheme for concurrent learning and inference.

C. Computational Aspects

The results in Sec. V were obtained for RGB detections of size 32x32 and embedding size 128 (a 24x reduction in dimensionality). Since those are detections rather than entire images, a reasonable amount of object appearance detail can in general be captured and consequently modeled. However, directly using the raw detections as measurements would imply a prohibitive factor size of ≈3000 for each semantic observation. One possible approach we currently are considering to tackle the dimensionality problem is using the first few layers of the encoder to reduce the dimensionality of the measurement, then fitting a model to predict the lower-dimensional measurement. Ideally, the dimensionality-reduced measurement would still retain the semantic information relevant to the task, as well as be sensitive to viewpoint changes, to allow for precise localization inference.

V. Experimental Results

We fit a viewpoint-dependent observation model following Sec. III-A to images and ground truth relative poses of objects from the BigBIRD dataset [10]. Each image is cropped using the provided object mask, and resized to 32x32x3. Fig. 2a visualizes the resultant latent space using tSNE [6]. On the left, latent vectors (encoder output) for the various viewpoints for each object show as well-localized clusters of the same color, a single cluster per object - although distinct clusters of very similar colors are present, in practice they correspond to different objects. Right: input images corresponding to the sub-region of the tSNE space highlighted on the right. The images are nearest neighbors to points of an axis-aligned grid in the tSNE space. The same clustered structure of the embedding space shows, with all viewpoints of an object appearing grouped. The observed structure of the embedded space is segregated among clusters corresponding to object instances and is expected to allow the envisioned reasoning using the learned model.

Fig. 2b shows the corresponding result obtained when fitting the viewpoint-predictive model following Sec. III-B with the Active Vision Dataset [1], using pairs of detections of the same instance from adjacent views. Clusters corresponding to the different instances are much less localized, as the data is more challenging, due to occlusions, partial detections, and multiple objects appearing in the same detection.

VI. Related Work

Pirk et al. [8] learn object representations which emergently capture semantics, in an unsupervised manner, for associating detections in consecutive frames, not defining probabilistic models or performing inference. Bloesch et al. [3] perform inference over a latent per-frame residual depth representation. Yu and Lee [16] use a learned model to perform inference over encoded depth measurements (i.e. the learned latent space corresponds to measurements, not to object instances). Both of the latter ([3], [16]) use supervised learning.

References

[1] Phil Ammirato, Patrick Poirson, Eunbyung Park, Jana Kosecka, and Alexander C. Berg. A dataset for developing and benchmarking active vision. In *IEEE*

International Conference on Robotics and Automation (ICRA), 2017.

[2] Israel Becerra, Luis M Valentín-Coronado, Rafael Murrieta-Cid, and Jean-Claude Latombe. Reliable confirmation of an object identity by a mobile robot: A mixed appearance/localization-driven motion approach. *Intl. J. of Robotics Research*, 35(10):1207–1233, 2016.

[3] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslam learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2560–2568, 2018.

[4] Y. Feldman and V. Indelman. Spatially-dependent bayesian semantic perception under model and localization uncertainty. *Autonomous Robots*, 2020.

[5] Y. Feldman and V. Indelman. Towards self-supervised semantic representation with a viewpoint-dependent observation model - supplementary material. Technical report, Technion - Israel Institute of Technology, 2020. URL https://indelman.github.io/ANPL-Website/Publications/Feldman20rss_ws_supp.pdf.

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[7] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters (RA-L)*, 4(1):1–8, 2018.

[8] Sören Pirk, Mohi Khansari, Yunfei Bai, Corey Lynch, and Pierre Sermanet. Online object representations with contrastive learning. *arXiv preprint arXiv:1906.04312*, 2019.

[9] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1352–1359, 2013.

[10] Arjun Singh, James Sha, Karthik S Narayan, Tudor Achim, and Pieter Abbeel. Bigbird: A large-scale 3d database of object instances. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 509–516. IEEE, 2014.

[11] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.

[12] V. Tchuiev, Y. Feldman, and V. Indelman. Data association aware semantic mapping and localization via a viewpoint-dependent classifier model. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019.

[13] WT Teacy, Simon J Julier, Renzo De Nardi, Alex Rogers, and Nicholas R Jennings. Observation modelling for vision-based target search by unmanned aerial vehicles. In *Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1607–1614, 2015.

[14] Javier Velez, Garrett Hemann, Albert S Huang, Ingmar Posner, and Nicholas Roy. Modelling observation correlations for active exploration and robust object detection. *J. of Artificial Intelligence Research*, 2012.

[15] Shichao Yang and Sebastian Scherer. Cubeslam: Monocular 3-d object slam. *IEEE Transactions on Robotics*, 35 (4):925–938, 2019.

[16] HW Yu and Beom Hee Lee. A variational feature encoding method of 3d object for probabilistic semantic slam. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3605–3612. IEEE, 2018.