Semantic Mapping with Simultaneous Object Detection and Localization

Zhen Zeng Yunwen Zhou Odest Chadwicke Jenkins Karthik Desingh

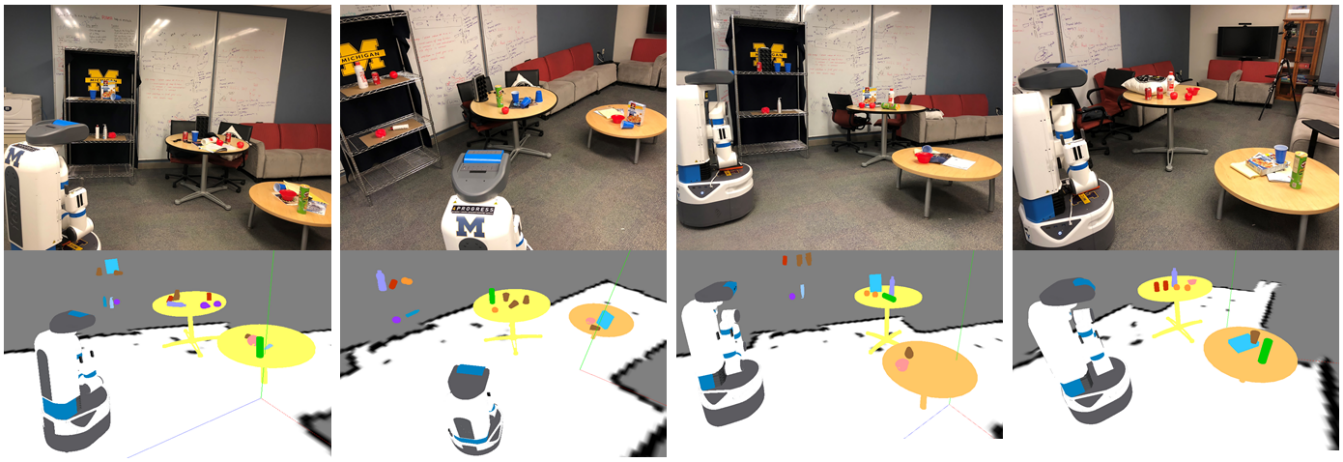

Fig. 1: Robot semantically maps a student lounge in four different visits. Each column shows an RGB snapshot of the environment, together with the corresponding semantic map composed by the detected and localized objects. We propose Contextual Temporal Mapping (*CT-Map*) method to simultaneously detect objects and localize their 6 DOF pose given streaming RGB-D observations. To achieve this, we probabilistically formulate semantic mapping problem as a problem of belief estimation over object classes and poses. We use Conditional Random Field (CRF) to model contextual relations between objects and temporal consistency of object poses. (Best viewed in color)

***Abstract*— We present a filtering-based method for semantic mapping to simultaneously detect objects and localize their 6 degree-of-freedom pose. For our method, called Contextual Temporal Mapping (or *CT-Map*), we represent the semantic map as a belief over object classes and poses across an observed scene. Inference for the semantic mapping problem is then modeled in the form of a Conditional Random Field (CRF). *CT-Map* is a CRF that considers two forms of relationship potentials to account for contextual relations between objects and temporal consistency of object poses, as well as a measurement potential on observations. A particle filtering algorithm is then proposed to perform inference in the *CT-Map* model. We demonstrate the efficacy of the *CT-Map* method with a Michigan Progress Fetch robot equipped with a RGB-D sensor. Our results demonstrate that the particle filtering based inference of *CT-Map* provides improved object detection and pose estimation with respect to baseline methods that treat observations as independent samples of a scene.**

I. Introduction

For robots to effectively operate and interact with objects, they need to understand not only the metric geometry of their surroundings but also its semantic aspects. When requested to organize a room or search for an object, robots must be able to reason about object locations and plan goal-directed mobile manipulation accordingly. We aim to enable robots to semantically map the world at the object level, where the representation of the world is a belief over object classes and poses. With the recent advances in object detection via neural networks, we have stronger building blocks for semantic mapping. Yet, such object detections are often times noisy in the wild, due to biases and insufficient diversity in training dataset. In our work, we aim to be robust to false detections from such networks. We model the object class as part of our hidden state for generative inference, rather than making hard decisions on class labels as given by the detector.

Given streaming RGB-D observations, our goal is to infer object classes and poses that explain observations, while accounting for **contextual** relations between objects and **temporal** consistency of object poses. Instead of assuming that every object is independent in the environment, we aim to explicitly model the *object-object* contextual relations during semantic mapping. More specifically, objects from the same category (e.g., food category) are expected to co-occur more often than objects that belong to different categories. Additionally, physical plausibility should be enforced to prevent objects from intersecting with each other, as well as floating in the air.

Temporal consistency of object poses also plays an important role in semantic mapping. Objects could stay where they were observed in the past, or gradually change their semantic locations over time. Under cases of occlusion, modeling temporal consistency can potentially help the localization of partially observed objects. Through temporal consistency modeling, the robot could gain a notion of object permanence, i.e., believing that objects continue to exist even when

Z. Zeng, Y. Zhou, O.C. Jenkins, K. Desingh are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA, 48109-2121. [zengzhen|ywchow|ocj|kdesingh]@umich.edu

they are not being directly observed.

Considering both contextual and temporal factors in semantic mapping, we propose the **Contextual Temporal Mapping** (***CT-MAP***) method to simultaneously infer object classes and poses. Examples of semantic maps generated by *CT-Map* are shown in Figure 1. To avoid deterministically representing the world as a collection of recognized objects with poses, we maintain a belief over the object classes and poses across observations.

For generative inference, *CT-MAP* probabilistically formalizes the semantic mapping problem in the form of a Conditional Random Field (CRF). Dependencies in the CRF model capture the following aspects: 1) compatibility between the latent semantic mapping variables and observations, 2) contextual relations between objects, and 3) temporal consistency of object poses. We propose a particle filtering based algorithm to perform generative inference in *CT-MAP*, inspired by Limketkai et al [18].

We evaluate the proposed semantic mapping method *CT-MAP* with the Michigan Progress Fetch robot. The performance of *CT-MAP* is quantitatively evaluated in terms of object detection and pose estimation accuracy. We show that *CT-MAP* is effective in simultaneously detecting and localizing objects in cluttered scenes. We demonstrate object detection performance superior to Faster R-CNN [27], and accurate 6 DOF object pose estimation compared to 3D registration methods such as ICP, and FPFH [28]. We also highlight examples in which our method benefits from modeling temporal consistency of object poses and object contextual relations.

II. Related Work

Our work semantically maps the world through simultaneous object detection and 6 DOF object pose estimation. Contextual relations between objects and temporal consistency of object poses are being modeled for better scene understanding. Here we discuss the related works in a) semantic mapping, b) object detection and pose estimation, c) object contextual relations, and d) object temporal dynamics modeling.

a) Semantic Mapping: Considering the plethora of work [17] in the field of semantic mapping which vary in semantic representations, we limit our focus to the works that provide object-level semantics. Works in semantic SLAM [3], [30], [5] demonstrated SLAM at the object level. Similarly, we aim at providing a semantic map of the world at the object level, and we focus on mapping while making use of existing metric slam method (e.g., ORB-SLAM [23]) to stay localized.

A widely used approach for semantic mapping is to augment 3D reconstructed map with objects. Civera et al. [6] ran an object detection thread parallelly with a monocular SLAM thread. They registered objects to the map by aligning the object faces relying on the SURF features. Ekvall et al. [7] actively recognized objects based on SIFT features, and integrated object recognition with SLAM for triangulation of object locations. But Civera et al. and Ekvall et al. did not deal with false detections, and their experiments were carried out in environments with no clutter.

To be robust to false detections, Pillai et al. [25] proposed aggregating object evidence over multiple frames to get better detection, compared to single frame object detection. But their method relied on 3D geometric segmentation that singulates objects from the background, which is vulnerable when dealing with clutter. Sünderhauf et al. [36] combined object detection over multiple frames and 3D geometric segmentation to get reasonable object boundaries. They produced 3D reconstructed map with object instance segments as central semantic entities. But their method did not provide object pose information, which is critical for robotic manipulation tasks.

Other works have focused on scene labeling of 3D map as a parallel SLAM thread is running in the background. People have proposed different methods for single frame scene labeling [44], [33], [21], [41], and fused labels across multiple frames to generate a dense 3D semantic map. Our work focus on detecting and localizing object entities in the environment, instead of dense labeling of every surfel or voxel in the reconstructed 3D map.

b) Object Detection and Pose Estimation: Deep neural network based object detectors [26], [20], [27] are nowadays widely adopted for focusing attention in region of interest given an image. Works in object pose estimation adopt these object detectors to get prior on object locations. Zhen et al. [43] generated scene hypotheses based on object detections returned by R-CNN [10], and they used Bayesian based bootstrap filter to estimate object poses. Similarly, Sui et al. [35] and Narayanan et al. [24] proposed generative approach for object pose estimation given RGB-D observation. Discriminative object pose estimation methods use local [14], [28] or global [29], [1] descriptors to estimate object poses via feature matching. However, feature-based methods are sensitive to the clutterness in the environment. Our work takes the generative approach and builds on Zhen et al. [43] for object pose estimation through Bayesian filtering, while [43] modeled objects independently and took single image at input, we model the contextual dependencies between objects and temporal consistency of each object instance given streaming data.

Works that simultaneously detect and localize objects are highly related to our work. Xiang et al. [42] proposed PoseCNN as a novel network for object detection and 6 DOF object pose estimation given a RGB image. Tremblay et al. [40] and Tekin et al. [38] converted the problem of simultaneous object detection and pose estimation into a problem of detecting the vertices of object bounding cuboid. Unlike these works that take single image as input and outputs deterministic estimate of object poses, our work maintains a belief over object classes and poses across observations.

Given streaming data, Salas-Moreno et al. [30] assumed repeated object instances in the environment to effectively recognize and localize objects, but their model lacks inter-object dependences. Tateno et al. [37] incrementally seg-

mented 3D surface reconstructed by an underlying SLAM thread, then 3D segments were recognized as objects and object poses were estimated via 3D descriptor matching. Their work is similar to our work in terms of the output, but they depend on 3D geometric segmentation which is not guaranteed to segment objects out in clutter. In addition, they require dense SLAM with small voxel size which is hard to scale.

c) Object Contextual Relations: Contextual relations play a key role in modeling spatial relations between objects for scene understanding. Koppula et al. [16] showed semantic labeling on point clouds using co-occurrence and geometric relations between objects. Jiang et al. [13] explored indirectly modeling object contextual relations by hallucinating human interactions with the environment. Similarly, [9], [11], [15], [8], [2] have proven modeling *object-object* and *object-place* contextual relations to be useful in place recognition, object detection and object search tasks. In our work, we mainly utilize *object-object* contextual relations in terms of co-occurrence and geometric relations.

d) Object Temporal Dynamics Modeling: We need to maintain the belief over object poses even when objects are not being observed. Different types of the objects share different characteristics of dynamics. For example, structural objects such as furnitures tend to stay approximately at the same location, while small objects such as food items can often be moved from one place to another. Bore et al. [4] proposed to learn long-term object dynamics over multiple visits of the same environment. Russel et. al. [39] proposed a temporal persistence model to predict the probability of an object staying at the location where it is last observed after certain time period. We are inspired by the temporal persistence model proposed in [39], and we reason about the possible locations of an object observed in the past based on the contextual relations between objects.

III. Problem Formulation

We focus on semantic mapping at the object level. Our proposed *CT-Map* method maintains a belief over object classes and poses across an observed scene. We assume that the robot stays localized in the environment through an external localization routine (e.g., ORB-SLAM [23]). The semantic map is composed by a set of N objects $O = \{o^1, o^2, \cdots, o^N\}$. Each object $o^i = \{o^c, o^g, o^\Psi\}$ contains the object class $o^c \in \mathcal{C}$, object geometry o^g, and object pose o^Ψ, where $\mathcal{C}$ is the set of object classes $\mathcal{C} = \{c_1, c_2, \cdots, c_n\}$.

At time t, the robot is localized at x_t. The robot observes $z_t = \{I_t, S_t\}$, where I_t is the observed RGB-D image, and S_t are semantic measurements. The semantic measurements $s_k = \{s_k^s, s_k^b\} \in S_t$ are returned by an object detector (as explained in section V-A), which contains: 1) a object detection score vector s_k^s, with each element in s_k^s denoting the detection confidence of each object class, and 2) a 2D bounding box s_k^b.

We probabilistically formalize the semantic mapping problem in the form of a CRF, as shown in Figure 2. Robot pose x_t and observation z_t are known. The set of objects O are

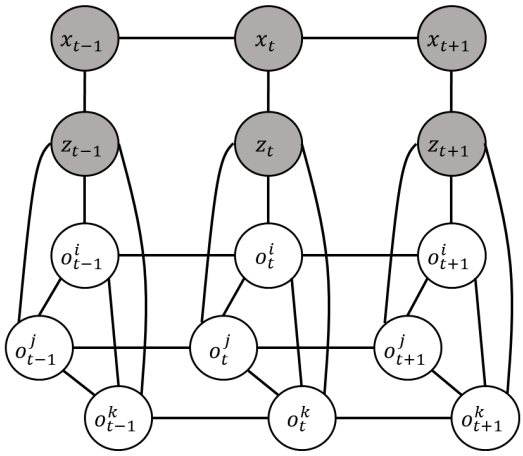

Fig. 2: Graphical model of the semantic mapping problem. Observed variables are robot poses x_t and observations z_t. Unknown variables are objects $\{o^1, o^2, \cdots, o^N\}$. We compute the posterior over objects while modeling contexual relations between all pairs of objects at each time point, and temporal consistency of each object across consecutive time points.

unknown variables. We model the contextual dependencies between objects and the temporal consistency of each individual object over time. The posterior probability of the semantic map is expressed as:

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

where Z is a normalization constant, and action applied to object o^i at time t is denoted by u_t^i. ϕ_p is the *prediction potential* that models the temporal consistency of the object poses. ϕ_m is the *measurement potential* that accounts for the observation model given 3D mesh of objects. ϕ_c is the *context potential* that captures the contextual relations between objects.

A. Prediction Potential

We use two different prediction models for predicting object pose, depending on whether the object is in the field of view or not. If the object is being observed, we model the action u as a continuous random variable that follows a Gaussian distribution with zero mean and small variance Σ. This assumption leads to prediction of small object movements in 3D to be modeled as:

$$o_t^\Psi \sim \mathcal{N}(o_{t-1}^\Psi, \Sigma)$$

which allows us to express the prediction potential as:

$$\phi_p(o_t^i, o_{t-1}^i, u_{t-1}^i) = \exp(-(o_t^\Psi - o_{t-1}^\Psi)^T\Sigma^{-1}(o_t^\Psi - o_{t-1}^\Psi)) \quad (2)$$

When object o^i is not in the field of view for a significant period of time, it can be either located at the same location or moved to a different location due to the actions applied by other agents. As stated by Toris et al. [39], the probability of the object o^i still being at the same location where it was last seen is a function of time. To take into account the fact that

object o^i can be moved to other locations, we model the temporal action u^i with a discrete random variable $\{u_{stay},\ u_{move}\}$. Specifically, u_{stay} denotes no action and the object stays at the same location, and u_{move} denotes a move action is applied and the object is moved to other locations. And these high-level actions follow certain distribution $p(u^i, \Delta t)$,

$$p(u^i = u_{stay}, \Delta t) = r_1 + r_2 \exp(-\frac{\Delta t}{\mu^i}) \tag{3}$$

$$p(u_{stay}, \Delta t) + p(u_{move}, \Delta t) = 1 \tag{4}$$

where r_1, r_2 are constants, and Δt is the time duration that object o^i is not being observed. As Δt increases, the probability of u_{stay} decays, and eventually $p(u_{stay}, \Delta t) = r_1$ as $\Delta t \to \infty$. For different objects o^i, the coefficients μ^i that control the speed of the decay are different. We provide heuristic μ^i for different objects in our experiments, while these coefficients can also be learned as introduced by Toris et al. [39].

B. Measurement Potential

The measurement potential of object o_t^i is expressed as:

$$\phi_m(o_t^i, x_t, z_t) = \begin{cases} \delta, & \text{if } o_t^i \text{ is out of view} \\ g(o_t^i, x_t, z_t), & \text{otherwise} \end{cases}$$

We use non-zero constant δ to account for cases where objects are not in the field of view. $g(o_t^i, x_t, z_t)$ measures the compatibility between the observation z_t and o_t^i, x_t,

$$g(o_t^i, x_t, z_t) = \sum_{s_k \in S_t} h(o_t^i, s_k^s) l(s_k^b, b(o_t^i, x_t)) f(o_t^i, x_t, I_t)$$

where $h(o_t^i, s_k^s)$ is the confidence score of class o_t^c from the detection confidence vector s_k^s. Function l evaluates the intersection over minimum area of two bounding boxes. $b(o_t^i, x_t)$ is the minimum enclosing bounding box of projected o_t^i in image space based on x_t.

We assume known 3D mesh models of objects. Function $f(o_t^i, x_t, I_t)$ computes the similarity between the projected o_t^i and I_t inside bounding box $b(o_t^i, x_t)$, as explained in detail in section V-B. In the case that robot has observed object o^i in the past, and the belief over o^i indicates that it is in the field of current view of the robot. If the robot cannot detect object o^i, then the object could be occluded, in which case we use $g(o_t^i, x_t, z_t) = f(o_t^i, x_t, I_t)$ for the object to be potentially localized.

C. Context Potential

There exist common contextual relations between object categories across all environments. For example, a cup would appear on a table much more often than on the floor, and a mouse would appear besides a keyboard much more often than besides a coffee machine. We refer to these common contextual relations as *category-level* contextual relations. In a specific environment, there exist contextual relations between certain object instances. For example, a TV always stays on a certain table, and a cereal box is usually stored in a particular cabinet. We refer to these contextual relations in a specific environment as *instance-level* contextual relations.

Algorithm 1: Particle filtering in *CT-Map*

Input: Observation z_t, robot pose x_t, particle set for each object $Q_{t-1}^i = \{\langle o_{t-1}^{i(k)}, \alpha_{t-1}^{i(k)} \rangle | k = 1, \cdots, M\}$

Resample M particles $o_{t-1}^{i(k)}$ from Q_{t-1}^i with probability proportional to importance weights $\alpha_{t-1}^{i(k)}$;
for $i = 1, \cdots, N$ **do**
  for $k = 1, \cdots, M$ **do**
    Sample $o_t^{i(k)} \sim \phi_p(o_t^i, o_{t-1}^{i(k)}, u_{t-1})$;
    Assign weight $\alpha_t^{i(k)} \propto \phi_m(o_t^{i(k)}, x_t, z_t) \prod_{j \in \Gamma(i)} \phi_c(o_t^{i(k)}, o_{t-1}^j)$;
  end
end

We manually encode *category-level* contextual relations as prior knowledge to our model, which also can be learned from public scene dataset (e.g., McCormac et al. [22]). Because *instance-level* contextual relations vary across different environments, these relations of a specific environment must be learned over time. The *context potential* is composed by *category-level* potential ϕ_{cat} and *instance-level* potential ϕ_{ins},

$$\phi_c(o_t^i, o_t^j) = w_1 \phi_{cat}(o_t^i, o_t^j) + w_2 \phi_{ins}(o_t^i, o_t^j) \tag{5}$$

We model $\phi_c(o_t^i, o_t^j)$ as mixture of Gaussians, with $\phi_{cat}(o_t^i, o_t^j)$ and $\phi_{ins}(o_t^i, o_t^j)$ each being a Gaussian component.

In our experiments, we manually designed ϕ_{cat} as prior knowledge, and ϕ_{ins} is updated via Bayesian updates. The principle while designing ϕ_{cat} follows two constraints: 1) simple physical constraints such as no object intersection is allowed, and objects should not be floating in the air, and 2) object pairs that belong to the same category co-occur more often than objects from different categories.

IV. Inference

We propose a particle filtering based algorithm to perform inference in *CT-MAP*, as given in Algorithm 1. Nonparametric Belief Propagation [34] [12] is not directly applicable to our problem because we are dealing with high-dimensional data. Sener et al. proposed recursive CRF [31] that deals with discrete hidden state with forward-backward algorithm, while our hidden state is mixed, i.e., object class label in discrete space and object pose in continuous space.

Instead of estimating the posterior of the complete history of objects $O_{1:T}$ as expressed in Equation 1, *CT-Map* can recursively estimate the posterior of each object $o_t^i \in O_t$. This approach to inference is similar to the CRF-filter proposed by Limketkai et al. [18]. We represent the posterior of object o_t^i with a set of M weighted particles, i.e., $Q_t^i = \{\langle o_t^{i(k)}, \alpha_t^{i(k)} \rangle | k = 1, \cdots, M\}$, where $o_t^{i(k)}$ contains object class and pose information as introduced in III-A, and $\alpha_t^{i(k)}$ is the associated weight for the k^{th} particle. In each particle filtering iteration, particles are first resampled based on their associated weights, then propagated forward in time through object temporal consistency, and re-weighted according to the measurement and context potentials.

We associate bounding boxes across consecutive frames based on their overlap. Only if a bounding box has been consistently associated for certain number of frames will we start initiating object class and pose estimation for that bounding box. The initial set of particles given a detected bounding box s_k^b are drawn as following: 1) first we sample the object class o^c based on the corresponding detection confidence score vector s_k^s; 2) then we sample the 6 DOF object pose o^ψ inside s_k^b, by putting the object center around the 3D points at the center region of s_k^b, with orientation uniformly sampled.

To sample the pose of $o_t^{i(k)}$ from $\phi_p(o_t^i, o_{t-1}^{i(k)}, u_{t-1})$ (Step 4 in Algorithm 1), there are two cases as following:

- If $o_{t-1}^{i(k)}$ is within the field of view of the robot, we sample $o_t^{i(k)}$ according to Equation 2.
- If $o_{t-1}^{i(k)}$ is not within the field of view of the robot, we first sample the high-level action $\{u_{stay}, u_{move}\}$ according to Equation 3.
  - If u_{stay} is sampled, then $o_t^{i(k)}$ is sampled based on Equation 2.
  - If u_{move} is sampled, then another object o^j is uniformly sampled from $O \setminus o_i$, which indicates the place that o^i has been moved to. $o_t^{i(k)}$ is then sampled from the region that o^j can physically support.

In step 5 of Algorithm 1, we use $\Gamma(i)$ to denote the indices of objects that are in the neighborhood of object $o_t^{i(k)}$. Because each neighbor object o_{t-1}^j is represented by M particles, it is computationally expensive to evaluate the context potential $\phi_c(o_t^{i(k)}, o_{t-1}^j)$ against each particle of o_{t-1}^j. Thus, we only evaluate the context potential against the most likely particle of o_{t-1}^j.

V. Implementation

A. Faster R-CNN object detector

We deploy Faster R-CNN [27] as our object detector. Given the RGB channel of our RGB-D observation, we apply the object detector and get the bounding boxes from the region proposal network, along with the corresponding class score vector. Then we apply non-maximum suppression to these boxes and merge boxes that have Intersection Over Union (IoU) larger than 0.5. For training, our dataset has 970 groundtruth images for 13 object classes. Each image has around 10 labeled objects. We fine-tuned the object detector based on VGG16 [32] pretrained on COCO [19]. In case of overfitting, we fine-tuned the network for 3000 interations with 0.001 learning rate.

B. Similarity function $f(o_t^i, x_t, I_t)$

We assume as given the 3D mesh model of objects. Thus, we can render the depth image of o_t^i based on its object class o^c and 6 DOF pose o^ψ in the frame of x_t. With rendered depth image $I(o_t^i, x_t)$:

$$f(o_t^i, x_t, I_t) = e^{-\lambda d(I(o_t^i, x_t), I_t)} \tag{6}$$

	Faster R-CNN	*T-Map*	*CT-Map*
mAP	0.607	0.715	0.871

TABLE I: mAP on our scene dataset.

where λ is a constant scaling factor. $d(I(o_t^i, x_t), I_t)$ is the sum of squared differences between the depth values in observed and rendered depth images.

VI. Experiments

We collected our indoor scene dataset with a Michigan Progress Fetch robot for evaluation on our proposed *CT-Map* method. Our indoor scene dataset contains 20 RGB-D sequences of various indoor scenes. We measure the quality of inference for various scenes in terms of 1) object detection and 2) pose estimation. Thus, we follow the mean average precision (mAP) metric and 6 DOF pose estimation accuracy for benchmarking our method. We also show qualitative examples of our semantic maps in Figure 1. More qualitative examples are provided in the video[1].

Across all experiments, we use $w_1 = w_2 = 0.5$ in Equation 5 to treat *category-level* and *instance-level* potentials equally. If an object has not been observed for infinite long period of time, we assume that object has equal probabilities of either staying at the same location or not. Thus, we use $r_1 = r_2 = 0.5$ in Equation 3.

A. Object Detection

We have noisy object detections coming from baseline Faster R-CNN object detector, while *CT-Map* can correct some false detections by modeling the object class as part of our hidden state. To evaluate the object detection performance of *CT-Map*, we take the estimated 6 DOF pose of all objects in the scene at the end of each RGB-D sequence in our dataset, and project them back onto each camera frame in that sequence to generate bounding boxes with class labels. We run two semantic mapping processes by considering different sets of potentials: 1) Temporal Mapping (*T-Map*): we consider prediction potential in the CRF model; 2) Contextual Temporal Mapping (*CT-Map*): we consider both prediction and context potential in the CRF model, which is the proposed method. For both *T-Map* and *CT-Map*, we include the measurement potential on observation.

We use mAP as our object detection metric. As shown in Table I, *T-Map* improves upon the baseline method Faster R-CNN by incorporating prediction and observation potentials, and *CT-Map* improves the performance further by additionally incorporating context potential. Faster R-CNN did not perform quite well on the test scenarios because the training data do not necessarily cover the variances encountered at test time. Though the performance of Faster-RCNN can be further improved by providing more training data, *CT-Map* provides more robust object detection when training remains limited.

In some cases, objects are not being reliably detected by Faster R-CNN due to occlusion. If an object has been

[1] https://youtu.be/W-6ViSlrrZg

observed in the environment in the past, our method makes predictions on locations that objects can go by modeling the temporal consistency of objects. Thus, even if a detection is not fired on the object due to occlusion, our method can still localize the object and claim a detection. However, in cases where an object is severely occluded and the depth observation lacks enough geometric information from the object, our method will not be able to localize the object. Example detection results highlighting the benefits of the proposed method compared to baseline Faster R-CNN are shown in Figure 4.

B. Pose Estimation

For each RGB-D sequence in our dataset, we locate the frames that each object is last seen, and project the depth frame back into 3D point clouds using known camera matrix. We then manually label the ground truth 6 DOF pose of objects. We compare the estimated object poses at the end of each RGB-D sequence against the ground truth.

Pose estimation accuracy is measured as $accuracy = \frac{N_{correct}}{N_{total}}$, where $N_{correct}$ is the number of objects that are considered correctly localized, and N_{total} is the total number of objects that are present in the dataset. If the object pose estimation error falls under certain position error threshold Δt and rotation error threshold $\Delta\theta$, we claim that the object is correctly localized. Δt is the translation error in Euclidean distance, and $\Delta\theta$ is the absolute angle difference in orientation. For symmetrical objects, the rotation error with respect to the symmetric axis is ignored.

We apply the Iterative Closest Point (ICP) and Fast Point Feature Histogram (FPFH) [28] algorithms as our baselines for 6 DOF object pose estimation. For each RGB-D sequence in our dataset, we take the 3D point clouds of the labeled frame, and crop them based on ground truth bounding boxes. These cropped point clouds are given to the baselines as observations, along with object 3D mesh models. ICP and FPFH are applied to register the object model to the cropped observed point cloud. We allow maximum iterations of 50000.

Our proposed method *CT-Map* significantly outperforms ICP and FPFH by a large margin. As our generative inference iteratively samples object pose hypotheses and evaluates them against the observations, *CT-Map* does not suffer from local minima as much as discriminative methods such as ICP and FPFH.

VII. Conclusion

We propose a semantic mapping method *CT-Map* that simultaneously detects objects and localizes their 6 DOF pose given streaming RGB-D observations. *CT-Map* represents the semantic map with a belief over object classes and poses. We probabilistically formalize the semantic mapping problem in the form of a CRF, which accounts for contextual relations between objects and temporal consistency of object poses, as well as measurement potential on observation. We demonstrate that *CT-Map* outperforms Faster R-CNN in object detection and FPFH, ICP in object pose estimation. In the future, we would like to investigate the inference problem of object semantic locations given partial observations of an environment, e.g., inferring a query object to be on a dining table, or in a kitchen cabinet. Ideally, maintaining a belief over object semantic locations can serve as a notion of generalized object permanence, and facilitate object search tasks.

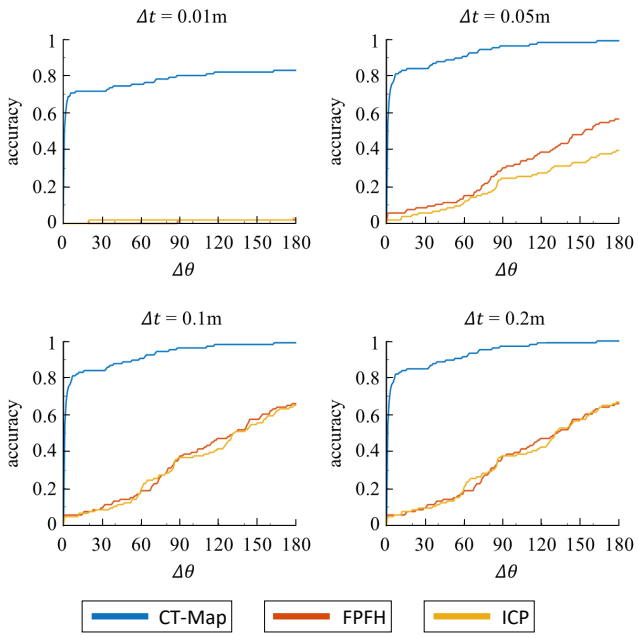

Fig. 3: Object pose estimation of *CT-Map*, compared with FPFH and ICP based baselines. Different plots correspond to different pose estimation correctness criteria defined by position error threshold Δt and rotation error threshold $\Delta\theta$. Our method outperforms FPFH and ICP with a large margin.

Acknowledgement

This work was supported in part by NSF award IIS-1638060.

References

[1] A. Aldoma, F. Tombari, R. B. Rusu, and M. Vincze. Our-cvfh–oriented, unique and repeatable clustered viewpoint feature histogram for object recognition and 6dof pose estimation. In *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*, pages 113–122. Springer, 2012.

[2] A. Aydemir, A. Pronobis, M. Göbelbecker, and P. Jensfelt. Active visual object search in unknown environments using uncertain semantics. *IEEE Transactions on Robotics*, 29(4):986–1002, 2013.

[3] S. Y. Bao, M. Bagra, Y.-W. Chao, and S. Savarese. Semantic structure from motion with points, regions, and objects. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2703–2710. IEEE, 2012.

[4] N. Bore, P. Jensfelt, and J. Folkesson. Multiple object detection, tracking and long-term dynamics learning in large 3d maps. *arXiv preprint arXiv:1801.09292*, 2018.

[5] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas. Probabilistic data association for semantic slam. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 1722–1729. IEEE, 2017.

[6] J. Civera, D. Gálvez-López, L. Riazuelo, J. D. Tardós, and J. Montiel. Towards semantic slam using a monocular camera. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 1277–1284. IEEE, 2011.

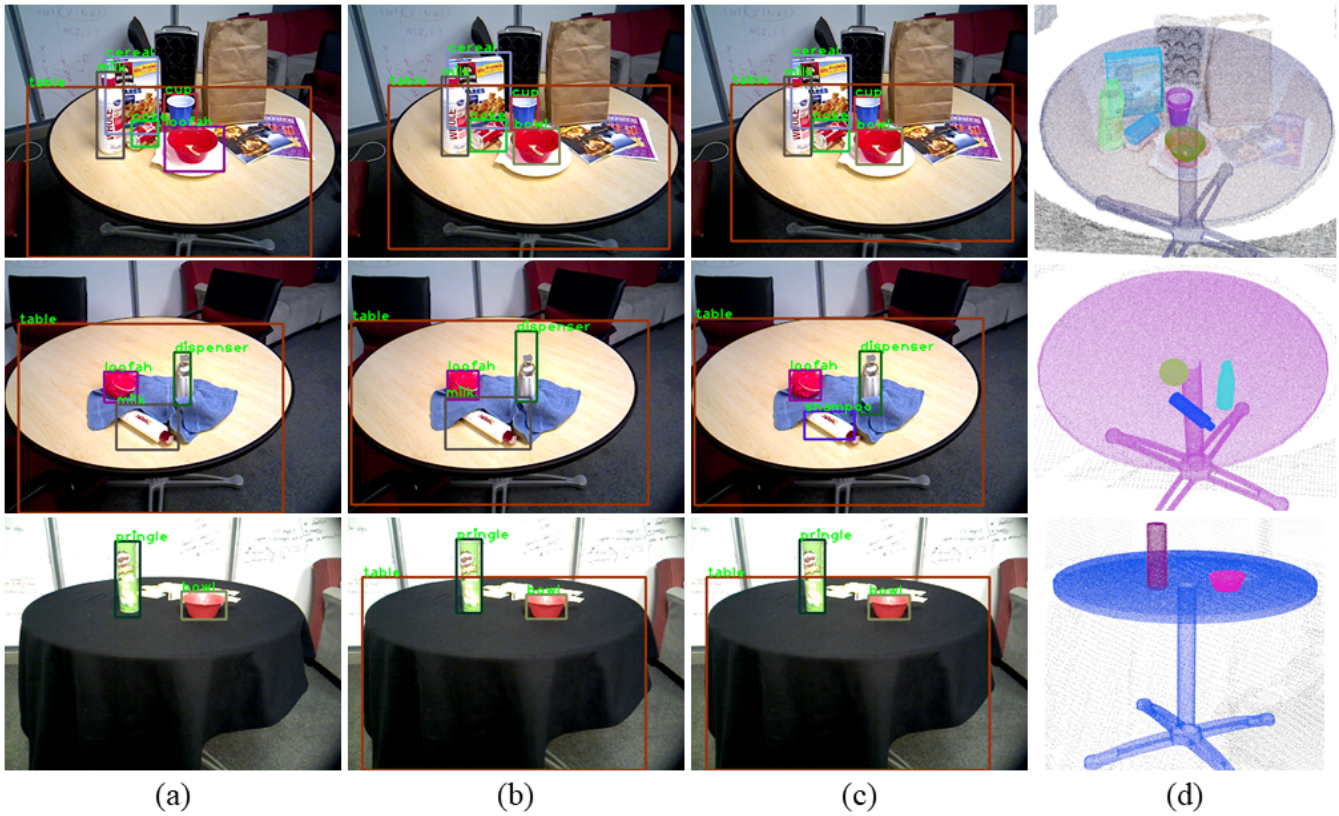

Fig. 4: Mapping examples highlighting detection improvements: (a) raw detection results from baseline Faster R-CNN; (b) detection results from *T-Map* when only considering measurement and prediction potential; (c) detection results *CT-Map* when considering measurement, prediction and context potential; (d) 6 DOF object pose estimates from *CT-Map*. We generate bounding boxes in column (b) and (c) by projecting the localized 3D objects into 2D image space, and finding the minimum enclosing boxes of the projections. The first row shows Faster R-CNN gives false detection on the red bowl as "loofah", while both *T-Map* and *CT-Map* correct the wrong label "loofah" into "bowl". The second row shows Faster R-CNN gives false detection on the shampoo bottle as "milk", and *T-Map* fails to correct the wrong label because the geometry of milk and shampoo is similar, while *CT-Map* successfully corrects the wrong label into "shampoo" based on the context. The third row shows Faster R-CNN does not detect the table due to the appearance change induced by the table cloth, while both *T-Map* and *CT-Map* successfully detect and localize the table. Because the table used to be observed around that location in the past, and our methods benefit from modeling the temporal consistency of object poses. (Best viewed in color)

[7] S. Ekvall, P. Jensfelt, and D. Kragic. Integrating active mobile robot object recognition and slam in natural environments. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 5792–5797. IEEE, 2006.

[8] P. Espinace, T. Kollar, N. Roy, and A. Soto. Indoor scene recognition by a mobile robot through adaptive object detection. *Robotics and Autonomous Systems*, 61(9):932–947, 2013.

[9] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.-A. Fernandez-Madrigal, and J. González. Multi-hierarchical semantic maps for mobile robotics. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 2278–2283. IEEE, 2005.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[11] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *European conference on computer vision*, pages 30–43. Springer, 2008.

[12] M. Isard. Pampas: Real-valued graphical models for computer vision. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003.

[13] Y. Jiang, M. Lim, and A. Saxena. Learning object arrangements in 3d scenes using human context. *arXiv preprint arXiv:1206.6462*, 2012.

[14] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on pattern analysis and machine intelligence*, 21(5):433–449, 1999.

[15] T. Kollar and N. Roy. Utilizing object-object and object-scene context when planning to find things. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 2168–2173. IEEE, 2009.

[16] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *Advances in neural information processing systems*, pages 244–252, 2011.

[17] I. Kostavelis and A. Gasteratos. Semantic mapping for mobile robotics tasks: A survey. *Robotics and Autonomous Systems*, 66:86–103, 2015.

[18] B. Limketkai, D. Fox, and L. Liao. Crf-filters: Discriminative particle filters for sequential state estimation. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 3142–3147. IEEE, 2007.

[19] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[21] J. McCormac, A. Handa, A. Davison, and S. Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 4628–4635. IEEE, 2017.

[22] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. Scenenet

rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. *arXiv preprint arXiv:1612.05079*, 2016.

[23] R. Mur-Artal and J. D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[24] V. Narayanan and M. Likhachev. Discriminatively-guided deliberative perception for pose estimation of multiple 3d object instances. In *Robotics: Science and Systems*, June 2016.

[25] S. Pillai and J. Leonard. Monocular slam supported object recognition. *arXiv preprint arXiv:1506.01732*, 2015.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[27] S. Ren, K. He, R. Girshick, and J. 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.

[28] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. IEEE, 2009.

[29] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 2155–2162. IEEE, 2010.

[30] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1352–1359. IEEE, 2013.

[31] O. Sener and A. Saxena. rcrf: Recursive belief estimation over crfs in rgb-d activity videos. In *Proceedings of Robotics: Science and Systems*, July 2015.

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[33] J. Stückler, B. Waldvogel, H. Schulz, and S. Behnke. Dense real-time mapping of object-class semantics from rgb-d video. *Journal of Real-Time Image Processing*, 10(4):599–609, 2015.

[34] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2003.

[35] Z. Sui, L. Xiang, O. C. Jenkins, and K. Desingh. Goal-directed robot manipulation through axiomatic scene estimation. *The International Journal of Robotics Research*, 36(1):86–104, 2017.

[36] N. Sünderhauf, T. T. Pham, Y. Latif, M. Milford, and I. Reid. Meaningful maps with object-oriented semantic mapping. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 5079–5085. IEEE, 2017.

[37] K. Tateno, F. Tombari, and N. Navab. When 2.5 d is not enough: Simultaneous reconstruction, segmentation and recognition on dense slam. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 2295–2302. IEEE, 2016.

[38] B. Tekin, S. N. Sinha, and P. Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 292–301, 2018.

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

[40] J. Tremblay, T. To, A. Molchanov, S. Tyree, J. Kautz, and S. Birchfield. Synthetically trained neural networks for learning human-readable plans from real-world demonstrations. *arXiv preprint arXiv:1805.07054*, 2018.

[41] Y. Xiang and D. Fox. Da-rnn: Semantic mapping with data associated recurrent neural networks. *arXiv preprint arXiv:1703.03098*, 2017.

[42] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[43] Z. Zeng, Z. Zhou, Z. Sui, and C. J. Odest. Semantic robot programming for goal-directed manipulation in cluttered scenes. In *Robotics and Automation (ICRA), 2018 IEEE International Conference on*. IEEE, 2018.

[44] Z. Zhao and X. Chen. Semantic mapping for object category and structural class. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 724–729. IEEE, 2014.