

MIT Open Access Articles

A Multi-Hypothesis Approach to Pose Ambiguity in Object-Based SLAM

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Fu, Jiahui, Huang, Qiangqiang, Doherty, Kevin, Wang, Yue and Leonard, John J. 2021. "A Multi-Hypothesis Approach to Pose Ambiguity in Object-Based SLAM."

As Published: 10.1109/iroso51168.2021.9635956

Publisher: IEEE

Persistent URL: <https://hdl.handle.net/1721.1/153759>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-ShareAlike



A Multi-Hypothesis Approach to Pose Ambiguity in Object-Based SLAM

Jiahui Fu, Qiangqiang Huang, Kevin Doherty, Yue Wang, and John J. Leonard

Abstract—In object-based Simultaneous Localization and Mapping (SLAM), 6D object poses offer a compact representation of landmark geometry useful for downstream planning and manipulation tasks. However, measurement ambiguity then arises as objects may possess complete or partial object shape symmetries (e.g., due to occlusion), making it difficult or impossible to generate a *single* consistent object pose estimate. One idea is to generate multiple pose candidates to counteract measurement ambiguity. In this paper, we develop a novel approach that enables an object-based SLAM system to reason about multiple pose hypotheses for an object, and synthesize this locally ambiguous information into a globally consistent robot and landmark pose estimation formulation. In particular, we (1) present a learned pose estimation network that provides multiple hypotheses about the 6D pose of an object; (2) by treating the output of our network as components of a mixture model, we incorporate pose predictions into a SLAM system, which, over successive observations, recovers a globally consistent set of robot and object (landmark) pose estimates. We evaluate our approach on the popular YCB-Video Dataset and a simulated video featuring YCB objects. Experiments demonstrate that our approach is effective in improving the robustness of object-based SLAM in the face of object pose ambiguity.¹

I. INTRODUCTION

Object-based Simultaneous Localization and Mapping (SLAM) incorporates higher-level object primitives into the localization and mapping process. Compared to traditional approaches utilizing low-level geometric features, it tends to improve its robustness to measurement noise and textureless scenes and bears the potential to provide a rich representation for downstream scene understanding, planning, and manipulation tasks [1]. By encoding abundant robot-landmark geometric constraints in a compact form, 6D object poses act as a desirable type of object measurements to be leveraged [2]. When perceiving the world in terms of objects, robots could then readily integrate these 6D object pose measurements into the joint recovery of robot and landmark poses.

However, when there exists environment ambiguity due to complete or partial object (landmark) shape symmetries (e.g., from occlusion), the frontend may be misled to make false positive pose measurements. For instance, the frontend may throw out a random measurement of a mug’s orientation when its handle is obscured. Such a pose measurement may incur considerable estimation errors in the backend if it is used to recover robot and landmark poses.

All authors are with the MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA. {jiahuifu, hqq, kdoherty, yuewangx, jleonard}@mit.edu

This work was supported by ONR MURI grant N00014-19-1-2571 and ONR grant N00014-18-1-2832.

¹Video: <https://youtu.be/E4sheabxWBI>

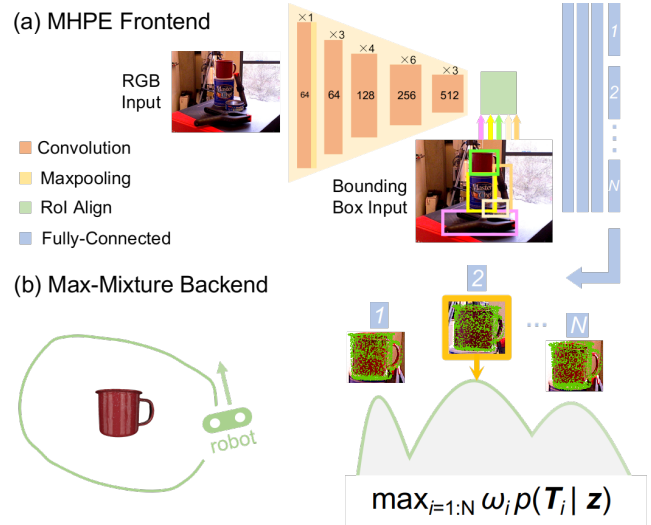


Fig. 1: Approach Overview. The system takes as input an RGB image along with its bounding box detection and outputs the optimized robot/landmark pose estimates. (a) The multi-hypothesis 6D object pose estimator (MHPE) in the frontend comprises three components: feature extraction, region-of-Interest (RoI) Align, and a multilayer perceptron. The output is N 6D pose hypotheses (T_i) for each object, coming from N separate branches in the last layer (blue), which are then sent to the max-mixture backend. (b) The max-mixture backend models the multi-modality embedded in the multi-hypothesis output. With MHPE covering the correct object pose (orange box, shown in green points matching the mug) in a big fraction of the observed frames, max-mixtures could gradually converge to the dominant mode, i.e., correct estimates of robot/landmark poses.

One way to counteract the effect of ambiguity is to generate multiple hypotheses each time, in which a consistent set of hypothesis could be covered. As a result, rather than conducting estimation based on one single frontend measurement at a time, as most current object-based SLAM systems do, we choose to enable the object-based SLAM system to be aware of the potential environment ambiguity through multiple hypotheses, i.e., generating and processing multiple possible object pose hypotheses for a single set of consistent estimates. In this way, by considering multiple hypotheses across many observations, an object-based SLAM system could then employ past evidence to *disambiguate* the true pose of the landmark (object) and recover a globally consistent set of robot and object pose estimates.

This idea motivates the design of our approach. Our objective is to enable an object-based SLAM system to reason about multiple hypotheses for an object pose. By producing and synthesizing these locally ambiguous information into a joint multi-hypothesis robot/landmark pose estimation setting, the system could solve for a globally consistent set of object and robot poses. Therefore, we instantiate our approach as follows (see Fig. 1):

- To receive ambiguity-aware measurements from the frontend, we develop a deep-learning-based multi-hypothesis 6D object pose estimator (MHPE) trained with known 3D object models to explore the potential pose hypothesis space.
- To solve the joint pose optimization with the multi-hypothesis formulation in the backend, we treat the multiple measurements as components of a mixture model and exploit nonlinear least square optimization to solve the max-mixture formulation [3] of pose hypotheses.

Experiments are conducted on the real YCB-Video Dataset and a synthetic video sequence featuring YCB objects in a virtual environment. Results demonstrate our approach’s effectiveness in improving the robustness of the object-based SLAM system in the face of ambiguous measurements.

Our paper proceeds as follows: we describe related work in Section II, elaborate on our pose estimation network design in Section III and max-mixtures formulation in Section IV, showcase experimental results in Section V, and conclude in Section VI.

II. RELATED WORK

A. 6D Object Pose Estimation

Various approaches have been developed to address the object pose estimation problem with different data modalities.

For RGB-based data inputs, traditional methods usually tackle the problem via feature matching between the query data and given object models [4]–[6], but are susceptible to low-texture regions or lighting changes. With the advancement in deep learning, PoseCNN [7] and SSD-6D [8], as two pioneering methods for estimating object pose in cluttered scenes, use neural networks to perform object detection and pose estimation at the same time. DOPE [9] uses only synthetic training data for robot manipulation tasks and the network shows good generalization to novel environments. Densefusion [10] fuses pixel-wise dense features from RGB and depth images and achieves strong pose estimation results.

For point-cloud-based inputs, ICP [11] is the most classical and widely used approach though prone to local minima due to non-convexity. [12]–[14] use probabilistic approaches to improve resilience to uncertain data. For its learning-based variants, SegICP [15] integrates semantic segmentation and pose estimation using neural networks, which achieves robust point cloud registration and per-pixel segmentation at the same time. The Deep Closest Point (DCP) [16] and Partial Registration Network (PRNet) [17] methods incorporate descriptor learning into the ICP pipeline, so that the model can learn matching priors from the data.

While all these above methods achieve good pose estimation accuracy, they either do not address the object shape ambiguity explicitly, or could not make a good compromise between estimation accuracy and processing speed, if serving as a SLAM frontend.

B. Uncertainty-Aware SLAM Backend

Uncertainty in SLAM involves many types of problems, such as loop closure and data association, where different backend methods have been developed. FastSLAM [18] tackles unknown data association in a particle-filter-based fashion that uses sampling to cover the overall probability space. Sünderhauf et al. [19], [20] develop methods to change the graph topology and achieve improved robustness to outliers with switchable constraints. The max-mixture method developed by Olson and Agrawal [3] represents discrete hypotheses as components of max-mixtures; this approach has recently been leveraged for unknown data association in the context of semantic SLAM by Doherty et al. [21].

C. Object-Based SLAM

SLAM++ [2] uses pixel-wise reprojection error and adds 6D object poses as camera-object pose constraints to the pose-graph optimization. But it needs 3D object models at runtime and is limited to few categories of pre-selected objects. QuadricSLAM [22] manages to estimate camera poses and quadric parameters together based on odometry and bounding box detections. CubeSLAM [23] conducts single-view 3D cuboid object detection and incorporates its refinement into the joint optimization of camera poses and objects. While these two works improve the camera pose estimation by including constraints from novel representations of the objects, they do not consider explicitly about the impact of shape ambiguity on these novel representations. PoseRBPF [24] employs a Rao-Blackwellized particle filter to track object poses across consecutive frames in the frontend, demonstrating robustness to object shape symmetry when a sufficient number of sampling particles are used, which, however, limits the computation speed.

With object (landmark) poses as measurements, our work here enhances the robustness of object-based SLAM in ambiguous environments by effectively generating multiple pose hypotheses from a learned pose estimation network and, different from what PoseRBPF does, then incorporates sequential observations into the backend factor graph for obtaining smoothed estimates of all robot/landmark poses.

III. FRONTEND: MULTI-HYPOTHESIS 6D OBJECT POSE ESTIMATOR (MHPE)

The MHPE frontend takes in an RGB image together with its object bounding boxes (could come from any off-the-shelf real-time object detectors), and outputs N hypotheses for the 6D pose of each object detected in the scene.

The object pose is represented by rotation $\mathbf{q} = (w, x, y, z)$ as a unit quaternion and 3D translation $\mathbf{t} = (t_x, t_y, t_z)$, which are defined with respect to the current camera frame. As $\mathbf{q} = (w, x, y, z)$ and $\mathbf{q} = (-w, -x, -y, -z)$ are equivalent, we enforce \mathbf{q} to have only non-negative real part w .

A. MHPE Network

The network processes the input data in three stages: feature extraction, region-of-interest alignment (RoI Align), and a multilayer perceptron (MLP) for final pose estimation (see Fig. 1).

We use ResNet-34 [25] without the last average pooling layer to get pixel level feature embeddings. The feature maps, as well as corresponding bounding boxes, are fed into the RoI Align layer. First introduced by MaskRCNN [26], this layer performs bilinear interpolation on the image feature embeddings and produces feature maps of the same shape for each bounding box region. The last MLP component is divided into two parts, one for rotation estimation and the other for translation. For N pose hypotheses for each object, the output size of the last fully connected layer is $4N$ for rotation (w, x, y, z) and $3N$ for translation (t_x, t_y, t_z). We apply Softplus to ensure non-negativity for “ w ” values. The final unit quaternion is obtained by normalizing the four outputs from the rotation branch.

B. Learning Objective for MHPE Outputs

The MHPE network aims to minimize the average distance (ADD loss [27]) between the two object point sets, transformed by the predicted and groundtruth poses, respectively. The ADD loss is defined as:

$$l_{ADD} = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \|(\hat{\mathbf{R}}\mathbf{x} + \hat{\mathbf{t}}) - (\mathbf{R}\mathbf{x} + \mathbf{t})\| \quad (1)$$

where $\mathbf{R} \in \text{SO}(3)$, $\mathbf{t} \in \mathbb{R}^3$, $[\mathbf{R}|\mathbf{t}]$ and $[\hat{\mathbf{R}}|\hat{\mathbf{t}}]$ are the groundtruth and estimated object poses, and \mathbf{x} is the 3D point in the object point sets \mathcal{X} from the known 3D object model. This metric performs well for the common single-hypothesis network, but needs adaption for the MHPE case, as different treatments of the individual loss in the hypothesis set may lead to distinct optimization directions.

Here, our pose estimation network is designed in a “mixture-of-expert” fashion [28], where the N hypotheses are made independently based on a shared set of feature maps. It is thus ideal to have each output branch acquire some specialty in certain output domains, hence helping improve the estimation accuracy and shed extra light on possible causes of pose ambiguity.

In this spirit, we choose to apply (1) in a winner-takes-all fashion, which is formulated as:

$$\begin{aligned} \hat{\mathbf{P}}_i &= \langle \hat{\mathbf{P}}_i^{(1)}, \dots, \hat{\mathbf{P}}_i^{(N)} \rangle \\ L_i(\hat{\mathbf{P}}_i) &= \underset{j \in [1, N]}{\operatorname{argmin}} l_{ADD}(\hat{\mathbf{P}}_i^{(j)}) \end{aligned} \quad (2)$$

where $\hat{\mathbf{P}}_i^{(j)} = [\mathbf{R}(\mathbf{q}_i^{(j)})|\mathbf{t}_i^{(j)}]$ is the j^{th} pose estimate in the output ensemble for the i^{th} input and \mathbf{R} is expressed as the rotation matrix from quaternion \mathbf{q} . The defined loss function, L_i for input i , only penalizes the most accurate estimation, regardless of “how bad” the rest of the hypotheses are. This treatment, compared to averaging, prevents other worse estimations from vanishing and adds some randomness, hence diversity, to the multi-branch estimator. Those branches with

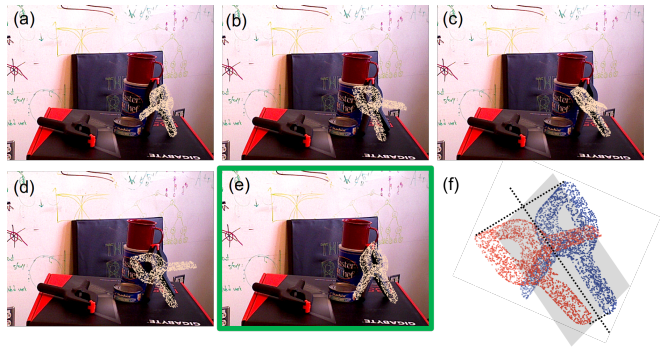


Fig. 2: Hypothesis diversity from the MHPE frontend on the YCB object clamp: Hypotheses (a) & (d) and (b) & (e) are two quasi-mirror images; hypothesis (c) is spurious, and the green-boxed hypothesis (e) gives a relatively good estimate. (f) illustrates that the 3D point clouds of the mirror image pair could incur similar distance loss with respect to the grey mirror plane.

an initialization too far away from the minimum could thus slowly evolve to different domains of the output space, hedging their bets while refraining from paying the price of making the wrong tentative decision. Fig. 2 shows one of our diverse pose estimation of a symmetrical clamp from the YCB-Video Dataset, illustrating the benefit of encouraging multiple hypotheses in different domains.

IV. BACKEND: MAX-MIXTURE-BASED MULTI-HYPOTHESIS MODELING AND OPTIMIZATION

Compared to its common single-hypothesis counterpart, the MHPE frontend can better tackle the ambiguity within the current observation via giving multiple pose candidates. Nevertheless, this leads to exponential computation complexity for picking a pose hypothesis combination that will bring about a sequence of optimal pose estimates.

To efficiently solve the optimization problem under the multi-hypothesis formulation, we here adapt max-mixtures to model multi-modal measurements and implicitly seek out consistent landmark pose hypotheses via optimization. Previously, the max-mixture method has been leveraged to solve data association problems in SLAM [3], [21] for its effectiveness in addressing large number of outliers in the least squares SLAM formulation. Thus by relying on max-mixtures’ ability to distinguish the better-behaved hypothesis, the backend is able to recover a globally consistent set of robot/landmark poses out of an exponentially increasing number of ambiguous measurement combinations (see Fig. 3).

Generally, estimating poses from a sequence of observations can be formulated as a maximum likelihood estimation (MLE) problem:

$$\hat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmax}} \prod_i \phi(\mathbf{z}_i | \mathbf{X}_i) \quad (3)$$

where \mathbf{X}_i denotes the latent variables involved in the i^{th} input, \mathbf{z}_i denotes the corresponding observation, and $\phi(\cdot)$ is

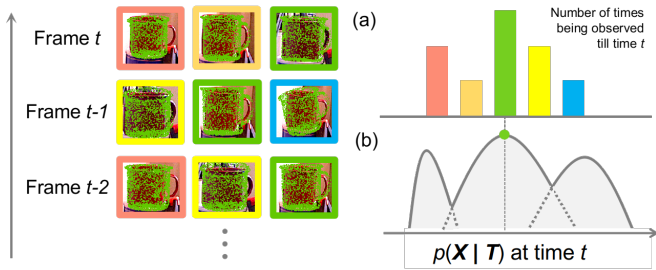


Fig. 3: Max-mixtures for processing multiple pose hypotheses. Left half: current pool of pose hypotheses (T_i) for the landmark “mug” with same hypothesis in the same outline color till time t . (a) Number of times each T_i has been observed so far, which is positively correlated to the posterior distribution of robot/landmark poses, $p(\mathbf{X}|\mathbf{T})$, where \mathbf{T} denotes all multi-hypothesis measurements. Hence the significance of *consistent inclusion* of the *correct pose* within the hypothesis set. Otherwise max-mixtures may converge to a sub-optimal hypothesis. (b) Schematic probability distribution of three pose hypotheses with respect to \mathbf{X} .

the likelihood function for this observation. The set of all latent variables, \mathbf{X} , is the camera and object pose sequence being jointly optimized in the backend. For odometry measurements between robot poses, the additive measurement noise is modeled as the canonical Gaussian noise. For the object pose observation with N hypotheses, a natural extension of the noise model is thus the sum of Gaussians or sum-mixtures:

$$\phi(\mathbf{z}_i | \mathbf{X}_i) = \sum_{j=1}^N \omega_{ij} \mathcal{N}(\log(\mathbf{z}_{ij} h(\mathbf{X}_i)^{-1})^\vee; \mathbf{0}, \Sigma) \quad (4)$$

in which each component corresponds to a hypothesis and follows the convention of the perturbation model for 6D poses in [29]. With an assumption of equally probable hypothesis, the weight ω_{ij} for component j is $\frac{1}{N}$. The measurement of interest here is the relative pose of the object with respect to the camera frame: ${}^c\mathbf{P} = h(\mathbf{X}_i) = h({}_o^w\mathbf{P}, {}_c^w\mathbf{P}) = {}_c^w\mathbf{P}_o^w\mathbf{P}$, which is a function of camera pose ${}_c^w\mathbf{P}$ and object pose ${}_o^w\mathbf{P}$ in the world frame.

Sum-mixtures falls outside the common nonlinear least squares (NLLS) optimization approaches for Gaussian noise models. For MLE estimation, we can instead consider the max-marginal [30], leading to the following max-mixture formulation:

$$\phi(\mathbf{z}_i | \mathbf{X}_i) = \max_{j=1:N} \omega_{ij} \mathcal{N}(\log(\mathbf{z}_{ij} h(\mathbf{X}_i)^{-1})^\vee; \mathbf{0}, \Sigma). \quad (5)$$

Max-mixtures switches to the “max” of all probability, retaining the better-behaved hypothesis while keeping the problem still within the realm of common NLLS optimization. Max-mixtures does not make a permanent choice out of multiple hypotheses. In an iteration of optimization, only one of the hypotheses actively contributes to the loss function and optimization steps via the “max” operator. Another hypothesis may turn active in the next iteration depending

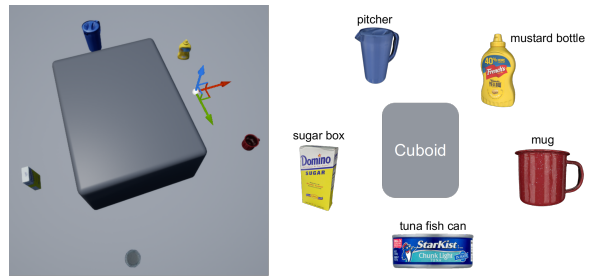


Fig. 4: Simulated video topview and YCB object layout. Some YCB objects are deliberately chosen to bear more or less *shape symmetries* (the mug and the tuna fish can), thus contributing to an *ambiguous environment setting* for ground-level observations. Note that objects are also scaled up to 50 times of its original size so as to match with the size of the observer, i.e., the AirSim car simulator, in the video.

on the updated estimate of the latent variable, \mathbf{X}_i . Hence, by optimizing (3) for the updated estimate of \mathbf{X} , we evaluate all the Gaussian components in (5) and identify the most likely one as the hypothesis contributing to the estimate in the next optimization step. Please refer to V-C for implementation details about optimization.

V. EXPERIMENTS AND RESULTS

In this section, in the goal of improving object-based SLAM results in the ambiguous environment setting, we would like to answer two questions: (1) Can MHPE predict a set of hypotheses that do contain the close-to-groundtruth object pose? (2) Can max-mixtures recover a globally consistent set of robot/landmark poses from MHPE measurements? To answer (1), we evaluate our approach on the YCB-Video benchmark in terms of distance loss and processing speed. For (2), we test our approach on a simulated video made with YCB objects for SLAM results, where the camera executes a much longer trajectory.

A. Datasets

YCB-Video Dataset. The YCB-Video Dataset [7] is built on 21 YCB objects [31] placed in various indoor scenes and consists of 92 video sequences with 133,827 images in all. Each sequence provides RGB-D images, object segmentation masks, and annotated object poses. Following [7], we divide the whole dataset into 80 videos for training and select 2,949 keyframes from the rest of 12 videos for testing. An extra set of 80,000 synthetic images created by [7] is also included in the training set.

AirSim Simulated Video. Since the camera motion in the YCB-Video Dataset is relatively small compared to a typical SLAM-application scenario and its environment does not involve much observation ambiguity, we create a virtual environment of five YCB objects with some shape symmetries placed around a cuboid in Unreal Engine [32] (see Fig. 4). The AirSim [33] car simulator is adopted to simulate a robot-mounted camera circling around the objects, thus making observations in an ambiguous environment setting.

The observations provide RGB images, object segmentation masks, and automatically annotated 6D poses of the camera and the five objects.

B. Metrics

To facilitate comparison within the YCB-Video Dataset benchmark, we follow two metrics used in [7] and [24], i.e., ADD (introduced in III-B) and ADD-S. The ADD-S metric, i.e., average distance between the point in the estimation-transformed point set \mathcal{X}_1 and its closest point in the groundtruth-transformed point set \mathcal{X}_2 , is defined as:

$$l_{ADD-S} = \frac{1}{|\mathcal{X}_1|} \sum_{\mathbf{x}_1 \in \mathcal{X}_1} \min_{\mathbf{x}_2 \in \mathcal{X}_2} \|(\hat{\mathbf{R}}\mathbf{x}_1 + \hat{\mathbf{t}}) - (\mathbf{R}\mathbf{x}_2 + \mathbf{t})\| \quad (6)$$

where \mathcal{X}_1 and \mathcal{X}_2 are transformed from the object model point set.

The ADD metric reflects pose estimation error, while ADD-S focuses on shape similarity and is thus appropriate for evaluating symmetrical objects. Following [7], we set the distance threshold to 10cm and compute the area under the ADD and ADD-S curves (AUC) with higher AUC indicating higher accuracy.

C. Implementation Details

The MHPE network is implemented in PyTorch. We initialize ResNet-34 with weights pretrained on ImageNet. The final MLP module comprises of three fully connected layers of size 512, 512, and 256. The number of output hypotheses is set as $N = 5$ for a good compromise between sufficient hypothesis diversity and additional computation overhead. For training, we use a batch size of 64 and train 100 epochs with the Adam optimizer. The initial learning rate is 0.0001 with a decay rate of 0.1 at epoch 50. The backend max-mixture algorithm is implemented in C++ using the Robot Operating System [34] and the iSAM2 [35] implementation from the GTSAM library [36]. All training and testing are conducted on a laptop with an Intel Core i7-9750H CPU and an Nvidia GeForce RTX 2070 GPU.

D. Results on YCB-Video Dataset

We report our quantitative and qualitative object pose estimation results on the YCB-Video Dataset.

Quantitatively, we compare our MHPE results to those of PoseCNN and the more recent PoseRBPF [24], all with RGB images (see Table I) and object detection results from PoseCNN for the purpose of fair comparison. We choose the PoseRBPF-50-particle variant, which is of closer processing speed with ours and is therefore suitable for SLAM applications.

In particular, for MHPE results, we compare our hypothesis with the lowest ADD loss in each set (best hypothesis) against the single output from PoseCNN and PoseRBPF, as in our problem, it is the estimation error of the best hypothesis, rather than the average quality of the multi-hypothesis set, that indicates the existence of a consistent hypothesis mode for max-mixtures to switch to, and hence the possibility of recovering the true robot/landmark poses.

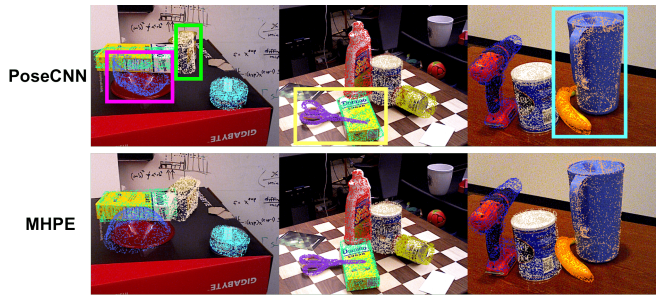


Fig. 5: Qualitative results of object pose estimation. Each object pose is annotated with the 2D image frame projection of a colored point cloud transformed by the pose estimate. Improved areas are marked by colored bounding boxes in the first row showing PoseCNN results.

Qualitatively, we display some sample visualizations of the object pose estimate in Fig. 5 for comparison between PoseCNN and the best hypothesis of our MHPE results. Since PoseRBPF does not provide multi-object pose estimation results on the YCB-Video Dataset, we do not include them in Fig. 5.

Overall Performance. Table I presents the ADD and ADD-S AUC values of all the 21 objects in the YCB-Video Dataset and the rough processing speed of each approach. We can conclude that with similar processing speed (ours: 23.0 fps & poseRBPF: 20.6 fps, excluding object detection as its results are used as system inputs), our best hypothesis from the MHPE frontend outperforms PoseCNN and PoseRBPF by 18.6%, 11.6% on ADD, and 9.4%, 11.1% on ADD-S, respectively. This demonstrates the effectiveness of our multi-hypothesis strategy in improving pose estimation accuracy and the reliability in providing valid hypothesis set for supporting the backend optimization of robot/landmark poses.

As shown by Fig. 5, when compared to PoseCNN, our best hypotheses from MHPE give more accurate pose estimations for the upside-down bowl in the first scene, the scissors and sugar box in the second, and the banana and the blue pitcher in the last scene.

Dealing with Ambiguity. From Table I, it is observed that the MHPE network performs better on most of the symmetrical objects (marked in bold). This proves the effectiveness of our object pose estimation strategy in dealing with shape ambiguity, as it is easier for symmetrical objects to have ambiguous poses, which may induce the network to generate random false positive hypotheses.

We further explain this with the “clamp” example in Fig. 2. The five pose hypotheses rendered here are MHPE’s estimates for the clamp pose, which include the correct estimate, (e), in green. Considering the diversity of the hypothesis set, which consists of seemingly inaccurate, and mirror-image pair hypotheses, a single-output network may throw out a random estimate as any of (a) & (d) or (b) & (e), since these two mirror-image pairs tend to incur similar ADD loss during training with respect to the grey mirror plane

TABLE I: Quantitative ADD and ADD-S AUC Results on YCB-Video Dataset. Symmetrical objects are in bold. Frame rate does not consider the effect of bounding box detection, which serves as part of pipeline inputs. Best ADD and ADD-S AUC values for each object are in bold (excluding ICP-refined results).

	RGB						RGB + Depth Image (for ICP)	
	PoseCNN		PoseRBPF 50 particles		MHPE Best Hypothesis		MHPE Best Hypothesis + ICP	
Frame Rate (fps)	5.9		20.6		32.3		2.1 (CPU-only)	
Objects	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S
002_master_chef_can	50.9	84.0	56.1	75.6	67.9	93.8	74.3	90.0
003_cracker_box	51.7	76.9	73.4	85.2	67.8	82.9	86.8	91.6
004_sugar_box	68.6	84.3	73.9	86.5	83.1	91.3	97.4	98.2
005_tomato_soup_can	66.0	80.9	71.1	82.0	79.5	92.2	38.1*	66.9*
006_mustard_bottle	79.9	90.2	80.0	90.1	81.6	90.8	98.0	98.7
007_tuna_fish_can	70.4	87.9	56.1	73.8	78.0	92.5	87.0	95.8
008_pudding_box	62.9	79.0	54.8	69.2	45.4	71.5	83.7	92.7
009_gelatin_box	75.2	87.1	83.1	89.7	76.1	87.8	97.2	98.4
010_potted_meat_can	59.6	78.5	47.0	61.3	69.1	85.5	66.4	78.4
011_banana	72.3	85.9	22.8	64.1	87.7	93.7	93.8	97.3
019_pitcher_base	52.5	76.8	74.0	87.5	76.8	88.8	96.9	98.1
021_bleach_cleanser	50.5	71.9	51.6	66.7	47.7	70.3	88.8	94.1
024_bowl	6.5	69.7	26.4	88.2	40.2	80.1	73.6	95.8
025_mug	57.7	78.0	67.3	83.7	40.6	72.8	69.3	91.1
035_power_drill	55.1	72.8	64.4	80.6	39.5	71.2	61.1	81.7
036_wood_block	31.8	65.8	0.0	0.0	64.6	85.5	85.2	93.0
037_scissors	35.8	56.2	20.6	30.9	64.5	88.9	63.5	88.9
040_large_marker	58.0	71.4	45.7	54.1	81.1	90.6	89.5	96.2
051_large_clamp	25.0	49.9	27.0	73.2	49.2	70.7	31.5	51.8
052_extra_large_clamp	15.8	47.0	50.4	68.7	8.6	47.4	9.7	31.3
061_foam_brick	40.4	87.8	75.8	88.4	75.1	92.6	90.5	97.2
All	53.7	75.9	57.1	74.8	63.7	83.1	72.6	85.2

(shown in (f)). Therefore, with each output branch evolving towards different domains of the solution space, MHPE can first produce a set of possible pose candidates and then let the backend to pick one based on the observations accumulated so far. The general poor performance on the extra large clamp may result from the given bounding box quality as it is hard for PoseCNN to distinguish between “large clamp” and “extra large clamp”. We also argue that for objects with poorer MHPE performance, these object pose hypotheses are consistently less accurate instead of randomly varying, and is hence less likely to cause big jumps in the backend robot pose estimation. Moreover, those jumps should be an even rarer occurrence considering the smoothing effect brought by the accumulation of observations in the backend.

ICP Post-Processing. We also provide AUC results with ICP post-processing, where the refined poses achieve significant accuracy improvements. While we are not comparing them to the non-post-processing PoseCNN and PoseRBPF results here, considering the great improvement in AUC values after ICP, we argue that our MHPE-predicted hypotheses are on average not far from the groundtruth and thus effectively prevent ICP from local-minimum failure. The only exception is for the “tomato soup can” (marked with asterisks) as in some frames, the lack of enough visible points

from heavy occlusion greatly impairs the registration quality.

E. SLAM Results for the Simulated Video

We apply our approach to the simulated video and test its effectiveness in the scenario with larger camera motion. For evaluation purposes, we compare the robot/landmark pose sequence optimized from our MHPE+max-mixtures approach against those obtained by two practical approaches when dealing with multiple hypotheses: averaging and random selection. Additionally, the random selection approach also resembles the behavior of the traditional single-hypothesis approach under ambiguous settings, as the frontend tends to throw out a random time-inconsistent measurement.

To better adapt to the simulated environment, in addition to the YCB-Video data, the MHPE network is trained on an extra of 18,523 images of a robot circumnavigating around each object at different ranges. The testing is then run on a 3,100-frame sequence in a similar environment, with the camera moving first in the inner and then the outer circle of the area enclosed by the five objects. Objects are detected in 1,611 frames of the sequence and the resulting pose estimations are used as the input to the SLAM backend.

Robot Pose Estimation. Quantitatively, in Fig. 6, we plot the running average of the estimation error for rotation and

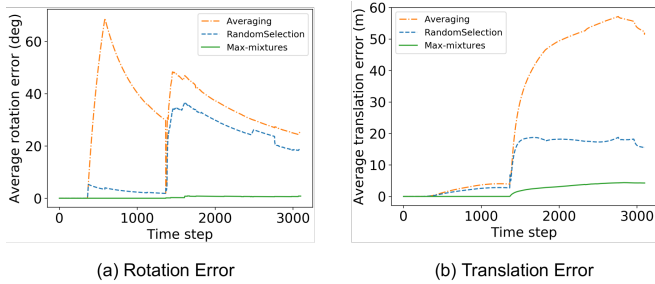


Fig. 6: Running average of the estimation error for rotation and translation across the course the robot motion. The groundtruth is extracted from the simulator.

translation, respectively. As shown in Fig. 6, we can conclude that generally max-mixtures performs the best while averaging does the worst. The discontinuities in (a) corresponds to the emergence of new objects. Considering the ambiguous measurement from the deliberately selected symmetrical objects, with a continuously lower rotation/translation error, our approach shows its robustness to the disturbance of the false-positive pose hypotheses and outperforms random selection, which to some extent, also indicates how a common, single-hypothesis SLAM system will behave. In Fig. 6 (b), the sudden climb in the translation error at about halfway through the course can be attributed to the reappearance of the mug in the video, and averaging obviously suffers from considerable drift due to ambiguous measurements from this loop closure.

Qualitatively, it can be observed from Fig. 7 that in terms of estimation accuracy, our approach outperforms both averaging and random selection by a large margin. We argue that by leveraging the accumulated observation from previous frames, our approach is able to recover consistent estimates from the ambiguous pose measurements, as opposed to its single-hypothesis-based counterparts.

This is especially useful for treating shape-ambiguous objects and loop closures. For example, the camera could obtain a better pose estimate of the mug by taking into account when last time the mug handle was visible, and this will, in turn, benefit loop closure detection. Furthermore, the result also implies that averaging over the candidate pool is the least effective approach, as the result of averaging over rotation matrices from the hypothesis ensemble can deviate from the actual rotation of any member hypothesis [38].

Landmark Pose Estimation. Here, we choose landmark pose estimation from MHPE as the frontend measurement to help recover robot poses. We compute the chordal distance between the groundtruth and the estimated landmark pose for comparison (see Fig. 8).

As shown in Fig. 8 (a), we could observe the pattern of a step-wise increase in error each time a new landmark is observed, which is then followed by a gradual error decrease as optimization proceeds. Compared to the other two approaches, our approach exhibits a much less steep surge between “stairs” as though with no prior knowledge about the object, it could employ past knowledge to help gradually

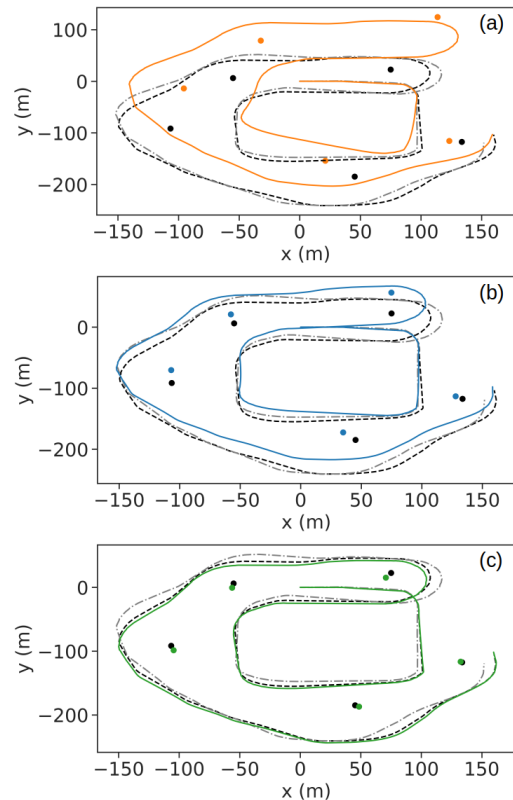


Fig. 7: Comparison of camera trajectories and object position estimates. We use *evo* [37] for trajectory evaluation. Objects are shown as dots. Groundtruth is shown in black. Odometry trajectories are shown by dash-dotted grey lines. (a) Averaging; (b) Random Selection; (c) Max-mixtures.

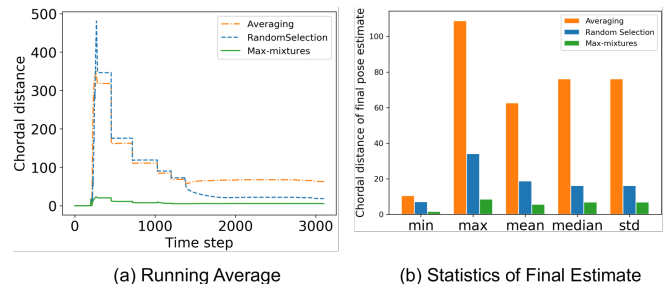


Fig. 8: Landmark pose estimation error.

switch to the better-behaved mode. Around time step 1,400 when the robot finishes its first circle and returns back to the mug, while averaging merges ambiguous measurements at the loop closure and hence renders the optimization off track (corresponds to trajectory in Fig. 7 (a)), max-mixtures manages to maintain a lower average error.

The above behaviors are also reflected in Fig. 8 (b), as for the final estimate of object poses, max-mixtures proves its superiority in all statistical descriptions, indicating a constantly higher estimation precision amongst all the three methods.

VI. CONCLUSION

In this paper we develop a novel approach that allows the robot to reason about multiple hypotheses for an object's pose, and synthesize this locally ambiguous information into a globally consistent object and robot trajectory estimation formulation. We instantiate this approach by designing a deep-learning-based frontend for producing multiple object pose hypotheses and a max-mixture-based backend for selecting pose candidates and conducting joint robot/landmark pose optimization. Experiments on the real YCB-Video Dataset and the synthetic YCB object video demonstrate the effectiveness of our approach to performing object pose and robot trajectory estimation in ambiguous environment settings. Future work may involve providing extra clues to facilitate max-mixtures optimization, such as by regressing weights for each pose hypothesis from the MHPE frontend.

REFERENCES

- [1] D. M. Rosen, K. J. Doherty, A. T. Espinoza, and J. J. Leonard, "Advances in Inference and Representation for Simultaneous Localization and Mapping," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, 2021.
- [2] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1352–1359, 2013.
- [3] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 826–840, 2013.
- [4] A. Collet, M. Martinez, and S. S. Srinivasa, "The moped framework: Object recognition and pose estimation for manipulation," *The international journal of robotics research*, vol. 30, no. 10, pp. 1284–1306, 2011.
- [5] M. Zhu, K. G. Derpanis, Y. Yang, S. Brahmabhatt, M. Zhang, C. Phillips, M. Lecce, and K. Daniilidis, "Single image 3d object detection and pose estimation for grasping," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3936–3943, IEEE, 2014.
- [6] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3762–3769, 2014.
- [7] 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.
- [8] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [9] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *arXiv preprint arXiv:1809.10790*, 2018.
- [10] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3343–3352, 2019.
- [11] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 239–256, Feb. 1992.
- [12] D. Hähnel and W. Burgard, "Probabilistic matching for 3D scan registration," in *Proceedings of the VDI Conference*, 2002.
- [13] G. Agamennoni, S. Fontana, R. Siegwart, and D. Sorrenti, "Point clouds registration with probabilistic data association," in *Proceedings of The International Conference on Intelligent Robots and Systems (IROS)*, 2016.
- [14] T. Hinzmann, T. Stastny, G. Conte, P. Doherty, P. Rudol, M. Wzorek, E. Galceran, R. Siegwart, and I. Gilitschenski, "Collaborative 3D reconstruction using heterogeneous uavs: System and experiments," in *International Symposium on Experimental Robotics*, 2016.
- [15] J. M. Wong, V. Kee, T. Le, S. Wagner, G.-L. Mariottini, A. Schneider, L. Hamilton, R. Chipalkatty, M. Hebert, D. M. Johnson, and et al., "SegICP: Integrated deep semantic segmentation and pose estimation," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep 2017.
- [16] Y. Wang and J. M. Solomon, "Deep closest point: Learning representations for point cloud registration," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [17] Y. Wang and J. M. Solomon, "PRNet: Self-supervised learning for partial-to-partial registration," in *33rd Conference on Neural Information Processing Systems*, 2019.
- [18] M. Montemerlo and S. Thrun, "Simultaneous localization and mapping with unknown data association using FastSLAM," in *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, vol. 2, pp. 1985–1991 vol.2, 2003.
- [19] N. Sünderhauf and P. Protzel, "Towards a robust back-end for pose graph SLAM," in *2012 IEEE international conference on robotics and automation*, pp. 1254–1261, IEEE, 2012.
- [20] N. Sünderhauf and P. Protzel, "Switchable constraints for robust pose graph SLAM," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1879–1884, 2012.
- [21] K. J. Doherty, D. P. Baxter, E. Schneeweiss, and J. J. Leonard, "Probabilistic data association via mixture models for robust semantic SLAM," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1098–1104, IEEE, 2020.
- [22] L. Nicholson, M. Milford, and N. Sünderhauf, "QuadricSLAM: Dual quadrics from object detections as landmarks in object-oriented SLAM," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2018.
- [23] S. Yang and S. Scherer, "CubeSLAM: Monocular 3-D object SLAM," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.
- [24] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, "Poserbpf: A rao-blackwellized particle filter for 6d object pose tracking," *arXiv preprint arXiv:1905.09304*, 2019.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- [27] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Asian conference on computer vision*, pp. 548–562, Springer, 2012.
- [28] A. Guzman-Rivera, D. Batra, and P. Kohli, "Multiple choice learning: Learning to produce multiple structured outputs," in *Advances in Neural Information Processing Systems*, pp. 1799–1807, 2012.
- [29] T. D. Barfoot and P. T. Furgale, "Associating uncertainty with three-dimensional poses for use in estimation problems," *IEEE Transactions on Robotics*, vol. 30, no. 3, pp. 679–693, 2014.
- [30] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [31] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*, pp. 510–517, IEEE, 2015.
- [32] Epic Games, "Unreal engine."
- [33] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "AirSim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and service robotics*, pp. 621–635, Springer, 2018.
- [34] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, p. 5, Kobe, Japan, 2009.
- [35] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "isam2: Incremental smoothing and mapping using the bayes tree," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 216–235, 2012.
- [36] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," tech. rep., Georgia Institute of Technology, 2012.
- [37] M. Grupp, "evo: Python package for the evaluation of odometry and SLAM." <https://github.com/MichaelGrupp/evo>, 2017.
- [38] R. Hartley, J. Trunpf, Y. Dai, and H. Li, "Rotation averaging," *International journal of computer vision*, vol. 103, no. 3, pp. 267–305, 2013.