Probabilistic Data Association via Mixture Models for Robust Semantic SLAM

Kevin J. Doherty[1], David P. Baxter[1,*], Edward Schneeweiss[1,2,*], and John J. Leonard[1]

Abstract— Modern robotic systems sense the environment *geometrically*, through sensors like cameras, lidar, and sonar, as well as *semantically*, often through visual models learned from data, such as object detectors. We aim to develop robots that can use all of these sources of information for reliable navigation, but each is corrupted by noise. Rather than assume that object detection will eventually achieve near perfect performance across the lifetime of a robot, in this work we represent and cope with the semantic and geometric uncertainty inherent in object detection methods. Specifically, we model data association ambiguity, which is typically non-Gaussian, in a way that is amenable to solution within the common nonlinear Gaussian formulation of simultaneous localization and mapping (SLAM). We do so by eliminating data association variables from the inference process through max-marginalization, preserving standard Gaussian posterior assumptions. The result is a max-mixture-type model that accounts for multiple data association hypotheses. We provide experimental results on indoor and outdoor semantic navigation tasks with noisy odometry and object detection and find that the ability of the proposed approach to represent multiple hypotheses, including the "null" hypothesis, gives substantial robustness advantages in comparison to alternative semantic SLAM approaches.
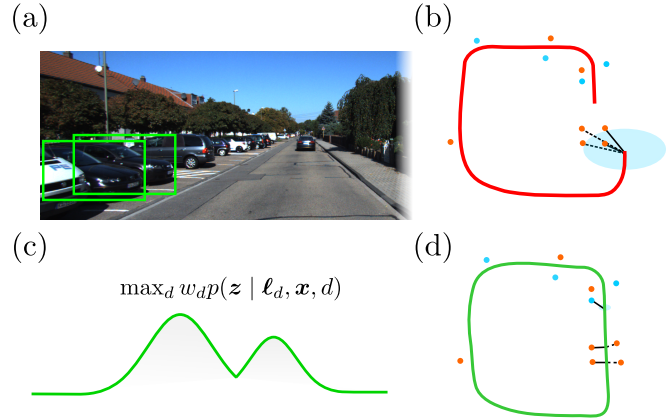
Fig. 1: Illustrative overview of the proposed approach. (a) Several object detections are made. (b) Robot pose ambiguity (blue covariance ellipse) results in multiple candidate landmark associations, the most likely of which may be *incorrect*. (c) All candidate associations for a measurement are represented as a "max-mixture" after data association variables are marginalized out. (d) Future evidence allows for the association to "switch" to the correct hypothesis, fixing the loop closure.

I. INTRODUCTION

The ability to build and use a map of environmental landmarks to navigate is one of the greatest strengths of the landmark-based *simultaneous localization and mapping* (SLAM) paradigm, but hinges critically on reliable recognition of previously mapped landmarks, i.e. data association. Consistent data association over long periods of time is vital if we aim to achieve robust robot navigation in the operational limit as "time goes to infinity." Unfortunately, long-term data association is substantially more challenging than short-term data association, since uncertainty in robot pose may grow large enough that many previously observed landmarks may be reasonable, albeit erroneous, candidates for a loop closure. For this reason, any mechanism by which we can associate landmarks *uniquely* is of interest.

Recent advances in the capabilities and reliability of deep neural networks for object detection and feature extraction have motivated the use of semantics to distinguish landmarks, taken to be objects, in the environment, and infer over time

the correct semantic class of each landmark. However, no detection system can be expected to have perfect performance over the lifetime of a robot. Rather than build navigation systems that depend on perfect detection and classification, we aim to develop methods that can take into consideration the error characterization of perception systems like neural network-based object detectors.

Robustness to misclassification and pose uncertainty requires the abilities to represent and resolve ambiguity in data association, and to reject incorrect loop closures. Traditional methods represent the problem of finding the correct set of *hypotheses* as a tree-search problem, where each node of the tree represents an association decision. Mitigating the complexity of search requires careful pruning of plausible hypotheses. Rather than explicitly search over association hypotheses, in this work we marginalize out data association variables at each point in time, allowing associations to arise implicitly from the inference of pose and landmark values.

The main contribution of this work is an *approximate max-marginalization* procedure for data associations which provides theoretical grounding for a "max-mixture"-type factor [1] within the context of semantic SLAM (shown in Figure 1); thus taking steps toward a unifying perspective on previous work in "robust SLAM" and recent work on data association for semantic SLAM, e.g. [2]. Approximate max-marginalization eliminates data association variables in a way

that preserves standard Gaussian distribution assumptions in SLAM in what otherwise becomes a non-Gaussian inference problem [3]. Our approach makes use of error characterization of an object detector, taking into consideration uncertainty to fuse detection information with geometric information from other sensors, like stereo cameras. Lastly, the proposed method incorporates loop closure rejection via incorporation of a *null-hypothesis* association, which we experimentally find to be critical in providing robustness to odometry noise and misclassification in the semantic SLAM problem.

The remainder of this paper proceeds as follows. In Section II we discuss related works on the topics of data association, robust SLAM, and semantic SLAM. We describe the problem of semantic SLAM with unknown data association in Section III, where we outline our approach to data association at a high-level. In Section IV we describe in detail the max-marginalization procedure, data association weight computation, and define the "semantic max-mixture factor" used for optimization, including the representation of null-hypothesis data association for loop closure rejection. Finally, experimental results demonstrating the robustness of the proposed approach to odometry noise and misclassification during indoor navigation and results from the KITTI dataset [4] are provided in Section V.

II. Related Work

The proposed work intersects the topics of *data association*, *robust SLAM*, and *semantic SLAM*. Classical work on data association stems from target-tracking literature, where probabilistic data association (PDA) [5] and multi-hypothesis tracking (MHT) [6] were introduced. Subsequently these methods were applied to early filtering-based SLAM solutions making Gaussian noise assumptions [7], [8]. FastSLAM [9] later introduced a particle filtering-based approach to the non-Gaussian inference problem of data association; a data association sampler was introduced that serves as an alternative to explicit search over associations.

Later work in the area of *pose-graph optimization* focused on themes of multi-hypothesis SLAM and outlier rejection, including methods like switchable constraints [10], max-mixtures [1], junction tree inference [11], and robust estimation using convex relaxations [12]–[14]. These works consider mitigating the effects of perceptual aliasing, often in the context of laser scan matching or appearance-based loop closure (see, for example [15]), whereas in semantic SLAM we also want to locate and classify discrete objects. Nonetheless, as we demonstrate in this work, landmark-based semantic SLAM shares many similarities. Specifically, by taking an approximate Bayesian inference approach to the semantic SLAM problem we arrive at a specific case of the max-mixtures method [1] where component weights are directly computed as candidate association probabilities.

In the area of semantic SLAM, classifications from object detectors have been used to aid data association [2], [3], [16]–[18]. While many such works consider maximum-likelihood data association [16], [17], recent works have considered probabilistic data association making use of expectation-maximization [2], or alternate between sampling data associations and recomputing SLAM solutions [18]. In contrast to [2], we model data associations as a mixture, rather than averaging solutions for different associations. We also marginalize out poses and landmarks when computing data association probabilities, whereas in [2], point estimates of robot poses and landmarks are used to compute data association weights. In both [2] and [18], convergence requires iteratively recomputing data associations and performing factor graph optimization; we aim to avoid the complexity associated with this recomputation. In previous work [3], we addressed this problem for the general non-Gaussian SLAM case using nonparametric belief propagation. While that approach gives rich uncertainties representing data association ambiguity, in this work our goal is maximum *a posteriori* inference specifically in the nonlinear Gaussian case.

III. Semantic SLAM with Unknown Data Association

The semantic SLAM problem can be defined as the inference of vehicle poses $\mathbf{X} = \{\boldsymbol{x}_t \in \mathcal{X}\}_{t=1}^T$ and landmark states $\mathbf{L} = \{\boldsymbol{\ell}_j \in \mathcal{L}\}_{j=1}^M$, given a set of measurements $\{\boldsymbol{z}_t \in \mathcal{Z}\}_{t=1}^T$. This corresponds to the following maximum *a posteriori* (MAP) inference problem:

$$\hat{\mathbf{X}}, \hat{\mathbf{L}} = \underset{\mathbf{X}, \mathbf{L}}{\operatorname{argmax}} \, p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}). \tag{1}$$

We consider the vehicle state space $\mathcal{X} \triangleq \mathrm{SE}(3)$ throughout, while we consider landmark states containing both geometric and semantic components, i.e. $\mathcal{L} \triangleq \mathbb{R}^3 \times \mathcal{C}$ (for landmarks with only a positional component) or $\mathcal{L} \triangleq \mathrm{SE}(3) \times \mathcal{C}$ (for 6 degree-of-freedom landmark pose estimation), where $\mathcal{C} = \{1, \ldots, C\}$ is a fixed, *a priori* known set of C discrete semantic classes. When necessary, we denote the separate pose/positional components and semantic components of a landmark $\boldsymbol{\ell}_j$ as $\boldsymbol{\ell}_j^p$ and $\boldsymbol{\ell}_j^s$, respectively. We assume multiple measurements can be made at each point in time, such that $\boldsymbol{z}_{tk}$ is the k-th measurement made at time t. Lastly, we consider the measurement space $\mathcal{Z}$ as having jointly geometric and semantic components, e.g. range-bearing measurements with semantic class in $\mathcal{C}$ or 6-DoF pose measurements in $\mathrm{SE}(3)$ with semantic class in $\mathcal{C}$.

When associations between measurements and landmarks are not known, they must be inferred. We take $\mathbf{D} = \{\boldsymbol{d}_t\}_{t=1}^T$ to be the set of associations of measurements $\boldsymbol{z}_t$ at all points in time to landmarks, such that $d_{tk} = j$ indicates that the k-th measurement taken at time t corresponds to landmark $\boldsymbol{\ell}_j$. The most common approach to SLAM with unknown data association is that of *maximum-likelihood*, in which the most probable set of data associations are computed and fixed, then used to solve for the most probable robot poses and landmark states. This approach can be brittle, as a single incorrect association can move the optimal solution for robot poses and landmark states far from their true values.

In order to mitigate the effects of data association errors one may consider *probabilistic data association*, in which

multiple associations for a measurement are given consideration commensurate with their probability. Generally this corresponds with marginalization of the data association variable, i.e.

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) = \mathbb{E}_{\mathbf{D}} \left[p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z}) \mid \mathbf{Z} \right] \qquad (2)$$
$$= \sum_{\mathbf{D}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z}) p(\mathbf{D} \mid \mathbf{Z}).$$

Common assumptions of additive Gaussian measurement noise make $p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z})$ Gaussian, such that the resulting belief $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$ is a sum of Gaussians, which generally falls outside the realm of traditional nonlinear least-squares optimization approaches to SLAM.

A recent approach to semantic SLAM [2] preserves the Gaussian nature of the problem by replacing the above expectation with one over $\log p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z})$, leading to an expectation-maximization algorithm for optimization. Convergence requires recomputation of the association weights $p(\mathbf{D} \mid \mathbf{Z})$, and prior to convergence the solution will lie somewhere *between* those obtained given fixed associations.

We propose an alternative solution to the MAP inference problem in which the "sum-marginal" above is replaced by the "max-marginal":

$$\hat{p}(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \triangleq \max_{\mathbf{D}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z}) p(\mathbf{D} \mid \mathbf{Z}). \qquad (3)$$

Each component of the max-marginal $\hat{p}(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$ is a weighted Gaussian, while the max operator simply acts to switch to the "best" data associations for any given point in the latent space of $\mathbf{X}$ and $\mathbf{L}$. The optimal solution for the max-marginal is identical to the MAP solution for the true posterior in (2) [19]. Exact computation of the true max-marginal is generally intractable due to the combinatorial number of plausible data associations, but as we will show, several reasonable approximations make the max-marginal a practical method for dealing with data association ambiguity.

IV. MAX-MIXTURE SEMANTIC SLAM

We consider a standard Gaussian SLAM framework with an odometry model $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ that is Gaussian with covariance Σ_t with respect to the relative transform from $\boldsymbol{x}_{t-1}$ to $\boldsymbol{x}_t$, denoted $g(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$ and Gaussian geometric measurement model $p(\boldsymbol{z}_{tk}^p \mid \boldsymbol{x}_t, \boldsymbol{\ell}_j)$ with covariance Γ, with respect to the nonlinear function $h(\boldsymbol{x}_t, \boldsymbol{\ell}_j)^1$. Semantic measurements with model $p(\boldsymbol{z}_{tk}^s \mid \boldsymbol{\ell}_j^s)$ are assumed independent of the pose from which a landmark was observed, as well as its position, and as in [2], [3] they are taken as samples from a categorical distribution with probability vector defined by a classifier confusion matrix (assumed to be known *a priori*). Lastly, we assume the geometric and semantic measurements factor as $p(\boldsymbol{z}_{tk} \mid \boldsymbol{x}_t, \boldsymbol{\ell}_j^p) = p(\boldsymbol{z}_{tk}^p \mid \boldsymbol{x}_t, \boldsymbol{\ell}_j) p(\boldsymbol{z}_{tk}^s \mid \boldsymbol{\ell}_j^s)$.

The unnormalized posterior $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$, can be written in the following general *factor graph* formulation:

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \propto \prod_i f_i(\mathbf{V}_i), \ \mathbf{V}_i \subseteq \{\mathbf{X}, \mathbf{L}\}, \qquad (4)$$

---

[1]The function h in this work is taken to be a relative transform in the case of full landmark pose measurements, or bearing, elevation, and range when only landmark position information is available.

(a) Association probabilities computed by marginalizing out poses and landmarks.

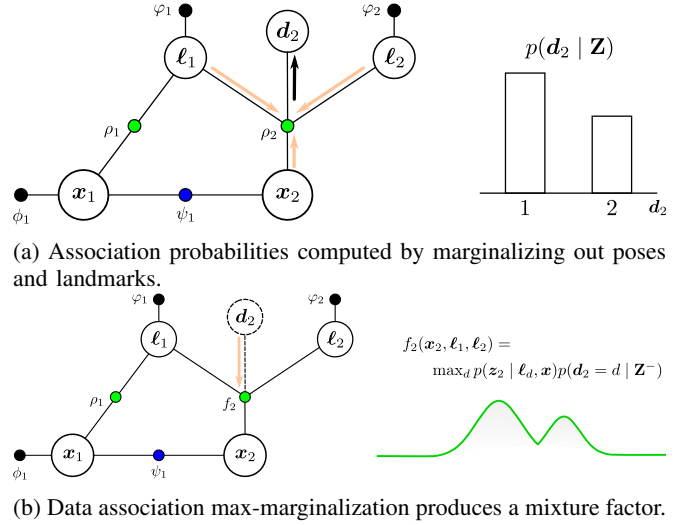(b) Data association max-marginalization produces a mixture factor.

Fig. 2: Illustrative example of data association variable marginalization. In 2a, we marginalize out poses and landmarks to compute a distribution over candidate data associations (see Section IV-B). In 2b, we compute the max-marginal over data associations, resulting in a max-mixture factor (see Section IV-C).

where each factor f_i is in correspondence with one of the relevant (odometric or landmark) measurement models. From the measurement models, there is a clear partition of *geometric* information from the odometry and geometric landmark measurement models (which depend on both the robot poses and landmark locations) and the *semantic* information, which depends only on the class of the associated landmark. Since we do not know data associations, we instead infer them from data. At a high level, our approach is to apply variable elimination to data associations to produce an equivalent factor graph with data associations marginalized out. The proposed *max-mixture semantic SLAM* approach approximates optimization over the max-marginal in (3). In particular, we introduce a **proactive max-marginalization** procedure for **computing data association weights**. For associations to previous landmarks, as well as for the null-hypothesis case, the max-marginal over candidate associations is represented as a *factor* (in the factor graph SLAM framework) taking on the form of a "max-mixture" [1]. We term these factors, **semantic max-mixture factors**. The addition of **null-hypothesis data association** enables the rejection of incorrect loop closures. The resulting factor graph is amenable to optimization using standard nonlinear least-squares techniques, from which the optimal robot and landmark states can be recovered.

A. Proactive Max-Marginalization

Exact computation of the max-marginal over all possible data associations in Equation (3) is computationally expensive due to the combinatorial growth in the size of the set of possible data associations over time. Marginalizing out data associations *proactively*, i.e. as new measurements are made, and ignoring the influence of future measurements on association probabilities allows us to mitigate the complexity
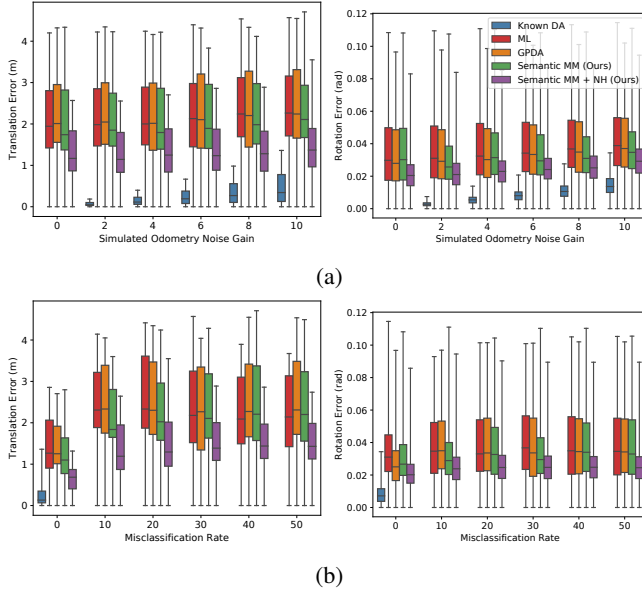
Fig. 3: Translation and rotation error as a function of (a) the simulated odometry noise and (b) misclassification rate parameters. The probabilistic approaches generally outperform maximum-likelihood, but the most significant advantages come from the incorporation of the null-hypothesis into the mixture-based framework, which outperforms all other methods across all misclassification rates and odometry noise. Legend is consistent throughout.

of full max-marginalization.

In particular, suppose we have some set of previous measurements $\mathbf{Z}^-$ and new measurements $\mathbf{Z}^+$. We aim to compute the max-marginal over associations to the new measurements, denoted $\mathbf{D}^+$. Formally, we have the following:

$$p(\mathbf{X}, \mathbf{L}, \mathbf{D}^+ \mid \mathbf{Z}^+, \mathbf{Z}^-) \propto \\ p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+)p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-)p(\mathbf{D}^+ \mid \mathbf{Z}^-), \quad (5)$$

from Bayes' rule, where we have used the conditional independences $p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+, \mathbf{Z}^-) = p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+)$, and $p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}^+, \mathbf{Z}^-) = p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-)$, since $\mathbf{D}^+$ consists of associations to only measurements outside of $\mathbf{Z}^-$. Applying max-marginalization to data associations, we obtain:

$$\hat{p}(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^+, \mathbf{Z}^-) = \\ p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-) \max_{\mathbf{D}^+} \left[p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+)p(\mathbf{D}^+ \mid \mathbf{Z}^-) \right]. \quad (6)$$

Here $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-)$ is the (potentially non-Gaussian) posterior distribution over poses and landmarks after sum-marginalization of data associations to the measurements $\mathbf{Z}^-$. For the purposes of optimization in the Gaussian case, we take this as the max-marginal $\hat{p}(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-)$.

The consequences of this simple change are significant: evaluating the max operators in the above expression no longer requires examination of previous associations and can be done in linear time for the most recent measurement. The result is that we have arrived at an approximate max-product algorithm for SLAM with unknown data associations.

B. Data Association Weight Computation

Consider a single measurement of a landmark z, which in our case consists of the joint geometric and semantic measurement of the landmark. We assume the data association probability is proportional to the likelihood $p(z_{tk} \mid d_{tk}, \mathbf{Z}^-)$ with poses and landmarks marginalized out (see Figure 2a). From the factored measurement model assumption, the likelihood of the form $p(z_{tk} \mid d_{tk}, \mathbf{Z}^-)$ can be broken into the product of separate semantic and geometric likelihoods:

$$p(z_{tk} \mid d_{tk} = j, \mathbf{Z}^-) = \\ p(z_{tk}^s \mid d_{tk} = j, \mathbf{Z}^-)p(z_{tk}^p \mid d_{tk} = j, \mathbf{Z}^-). \quad (7)$$

Each term on the right-hand side can be expanded as follows into the summation over landmark classes:

$$p(z_{tk}^s \mid d_{tk} = j, \mathbf{Z}^-) = \\ \sum_c p(z_{tk}^s \mid \ell_j^s = c)p(\ell_j^s = c \mid \mathbf{Z}^-), \quad (8)$$

and integral over robot pose and landmark location:

$$p(z_{tk}^p \mid d_{tk} = j, \mathbf{Z}^-) = \\ \iint p(z_{tk}^p \mid d_{tk} = j, \boldsymbol{x}_t, \boldsymbol{\ell}_j)p(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-)d\boldsymbol{x}_t d\boldsymbol{\ell}_j. \quad (9)$$

With data associations marginalized out, the belief $p(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-)$ would be generally non-Gaussian. Consequently, we again make an approximation and use the single Gaussian component corresponding to the max-marginal $\hat{p}(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-)$ evaluated at the current estimate of $\boldsymbol{x}_t$ and $\boldsymbol{\ell}_j$,) which we denote $\hat{p}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{\ell}}_j \mid \mathbf{Z}^-)$:

$$p(z_{tk}^p \mid d_{tk} = j, \mathbf{Z}^-) \approx \\ \iint p(z_{tk}^p \mid d_{tk} = j, \boldsymbol{x}_t, \boldsymbol{\ell}_j)\hat{p}(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{\ell}}_j \mid \mathbf{Z}^-)d\boldsymbol{x}_t d\boldsymbol{\ell}_j. \quad (10)$$

Since all of the terms in the integral are now Gaussian, it can be simplified as follows, based on the method of [20]:

$$p(z_{tk}^p \mid d_{tk} = j, \mathbf{Z}^-) \approx \\ \frac{1}{\sqrt{|2\pi R_{tjk}|}}e^{-\frac{1}{2}\|h(\hat{\mathbf{x}})-z_{tk}^p\|^2_{R_{tjk}}}, \quad (11)$$

where $\hat{\mathbf{x}}$ is the mean of the joint distribution over $\boldsymbol{x}_t$ and $\boldsymbol{\ell}_j$. The covariance R_{tjk}, is defined as:

$$R_{tjk} \triangleq \left.\frac{\partial h}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}} \Sigma \left.\frac{\partial h}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}}^T + \Gamma, \quad (12)$$

where Σ is the block joint covariance matrix between pose $\boldsymbol{x}_t$ and candidate landmark $\boldsymbol{\ell}_j$, $\partial h / \partial \mathbf{x}$ is the Jacobian of the measurement function, and Γ is the covariance of the geometric measurement model. This result, combined with the expression in (8) gives the marginal likelihood in (7) that we normalize to compute data association probabilities.

(a) Max-Likelihood   (b) Gaussian PDA

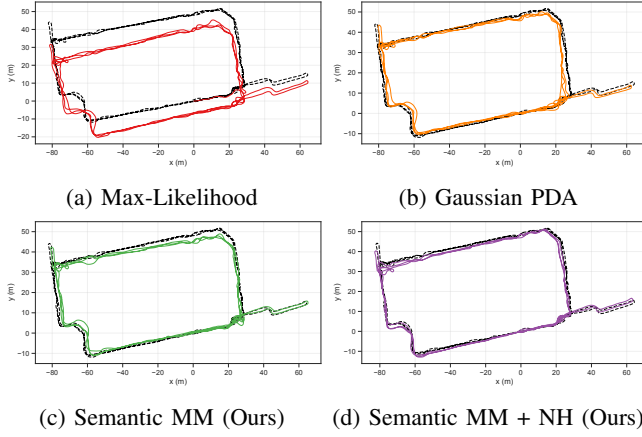(c) Semantic MM (Ours)   (d) Semantic MM + NH (Ours)

Fig. 4: Estimated trajectories for the MIT RACECAR dataset (origin aligned); reference trajectory in black. Association errors cause the maximum-likelihood approach to produce additional hallways.

C. Semantic Max-Mixture Factor

Assuming uniform priors on data associations, the distribution $p(d_{tk} \mid \mathbf{Z}^-)$ is proportional to the marginal likelihood in (7) and can simply be normalized over all assignments to d_{tk}. This results in a set $\mathcal{H} \subseteq \{1, \ldots, M\}$ of candidate landmark hypotheses, max-marginalization of which produces a max-mixture factor for a measurement $\boldsymbol{z}_{tk}$:

$$f(\boldsymbol{x}_t, \boldsymbol{\ell}_{\mathcal{H}}) = \max_{j \in \mathcal{H}} p(\boldsymbol{z}_{tk} \mid \boldsymbol{x}_t, \boldsymbol{\ell}_j) p(d_{tk} = j \mid \mathbf{Z}^-). \quad (13)$$

The max-marginalization step is visualized in Figure 2b, where we have eliminated the data association variable from the inference process. By augmenting the candidate set $\mathcal{H}$ to be $\mathcal{H} \cup \{\emptyset\}$, we allow a *null-hypothesis* data association to be made. In practice, we assume a probability for the null-hypothesis and normalize the remaining data associations such that the total probability of the augmented hypothesis set equals 1. The null-hypothesis component is assumed to be Gaussian with large standard deviation (e.g. 10^5).

Finally, we can recover maximum *a posteriori* landmark semantic class estimates (assuming uniform priors) as in [2] as follows:

$$\hat{\boldsymbol{\ell}}_j^s = \operatorname{argmax}_c \prod_t \sum_{\boldsymbol{d}_t} p(\boldsymbol{d}_t, \boldsymbol{\ell}_j^s = c \mid \mathbf{Z}), \quad (14)$$

which are recovered using the data association probabilities stored as the component weights of the max-mixture factors.

V. EXPERIMENTAL RESULTS

All computational experiments were implemented in C++ using the Robot Operating System (ROS) [21] and the implementation of iSAM2 [22] within the GTSAM [23] library for optimization and covariance recovery. We demonstrate our approach on 3D visual SLAM tasks using data collected during indoor navigation with an MIT RACECAR vehicle[2] equipped only with a ZED stereo camera [24], as well as with stereo image data from the KITTI dataset [4], [25].

[2]https://mit-racecar.github.io/

Experiments were run on a single core of a 2.2 GHz Intel i7 CPU. We use evo [26] for trajectory evaluation[3].

In both experiments, we compared two variants of the proposed method: semantic max-mixtures (MM) and semantic max-mixtures with null-hypothesis data association (MM+NH) to a known data association (Known DA) baseline, naïve maximum-likelihood (ML) data association (which makes a single association with the landmark maximizing Eq. (7)), as well as an expectation-maximization approach similar to that of [2], here referred to as Gaussian probabilistic data association (GPDA). We use a threshold on the marginal likelihood in (7) to identify new landmarks, as well as to produce the set of landmark candidates. This is similar to standard Mahalanobis distance-based thresholds, but also considers the semantic likelihood[4].

A. MIT RACECAR Dataset

We collected roughly 25 minutes of data during indoor navigation with the MIT RACECAR mobile robot platform over a roughly 1.08 km trajectory. We sampled AprilTag [27], [28] detection keyframes at a rate of 1 Hz resulting in 702 observations of 262 unique tags. Odometry was obtained using the ZED stereo camera visual odometry [24]. The use of AprilTags uniquely allows us to obtain a baseline "ground-truth" solution with known data associations. We artificially assigned semantic labels to each AprilTag by considering the true tag ID modulo C for a C-class semantic SLAM problem. In the experiments presented in this paper, we use $C = 2$ classes, as we found it to be one of the most challenging situations[5]. While AprilTags give generally accurate orientation information, we typically cannot expect this of neural network-based object detectors. For this reason, we set a large standard error on roll, pitch, and yaw of AprilTag detections to prevent orientation information from giving any substantial data association cues. Furthermore, this experimental setup allows us to apply classification error and simulate additional odometry noise in a repeatable way to observe the performance of each method with noise in classification and odometry.

In Figure 3, we provide box-plots summarizing statistical results of trajectory error on the MIT RACECAR dataset. Specifically, we considered robustness to simulated additional odometry noise and detector misclassification. We calibrated an initial odometry model, but in testing we add simulated Gaussian noise of x, y, and yaw[6], multiplied by a scale factor varying from 0 to 10. We simulate classification error for a misclassification rate α (from 0% to 50%) by sampling from a semantic measurement model with confusion matrix equal to $1 - \alpha$ on the diagonal and α on the off-

[3]We provide trajectory comparisons for all of the methods tested without landmarks visualized, but more detailed visualizations and videos can be found on the project page: https://github.com/MarineRoboticsGroup/mixtures_semantic_slam

[4]We use a χ^2 test with confidence 0.9 and null-hypothesis weight of 0.1.

[5]In general, with a "good" detector, the presence of many unique semantic classes among landmarks makes the data association problem easier.

[6]We take the baseline simulated noise model as $\sigma_x = 1.5$mm, $\sigma_y = 0.75$mm, and for yaw, $\sigma_\gamma = 0.000225$ rad in robot frame.

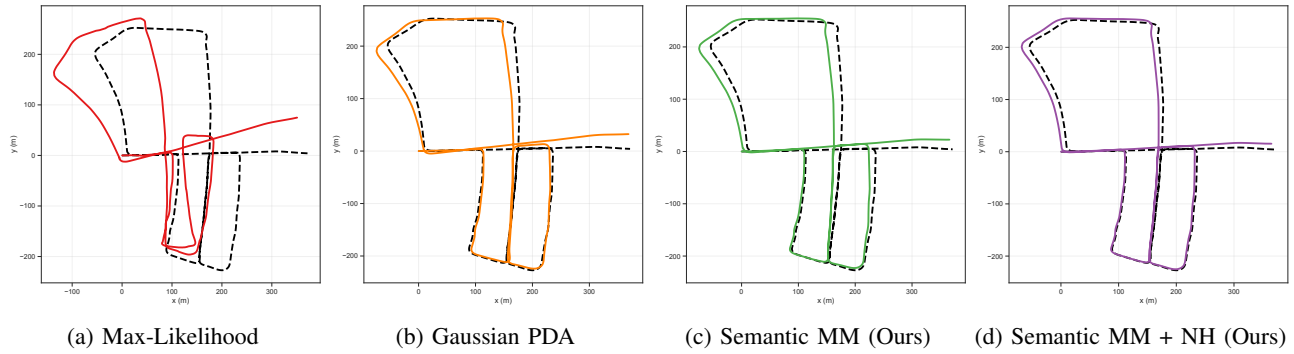| (a) Max-Likelihood | (b) Gaussian PDA | (c) Semantic MM (Ours) | (d) Semantic MM + NH (Ours) |

Fig. 5: Estimated trajectories for KITTI Sequence 5. False loop closures cause the maximum-likelihood data association method to fail catastrophically, while all of the probabilistic methods show improved robustness.

Method	Max Error	Mean Error	Median Error	RMSE
ML	126.46	52.78	59.66	62.35
GPDA	30.76	10.52	8.90	12.11
MM	23.23	9.31	8.31	11.37
MM + NH	**19.88**	**6.48**	**6.44**	**7.70**

TABLE I: Translation error (m) on KITTI Sequence 5.

Method	Max Error	Mean Error	Median Error	RMSE
ML	**0.42**	0.15	0.11	0.19
GPDA	0.51	0.06	0.052	0.069
MM	0.54	0.055	0.049	0.065
MM + NH	0.58	**0.043**	**0.037**	**0.053**

TABLE II: Rotation error (rad) on the KITTI Sequence 5.

diagonal. We find that while errors in all methods increase with added noise in odometry and misclassification, all of the probabilistic methods generally outperform maximum-likelihood. Most significantly, we find that the addition of the null-hypothesis to our approach drastically reduces error in all tests. Beyond the ability of the null-hypothesis method to reject bad loop closures, the addition of the null hypothesis may help prevent the max-mixtures approach from becoming "stuck" in a local optimum by decreasing the high cost associated with being "between" hypotheses; a region that must be crossed before hypothesis switching can take place. We contextualize these quantitative results with qualitative trajectories plotted in Figure 4 for a 2-class problem with 10% misclassification and 10% additional odometry error.

B. KITTI Dataset

We also evaluate our approach on stereo camera data from the KITTI dataset odometry sequence 5 [4]. In our experiments, we use the MobileNet-SSD object detector ([29]–[31]), from which detections were obtained at approximately 10 Hz. We threshold the confidence of the detector at 0.8, using detections of cars as landmarks. We use VISO2 stereo odometry for visual odometry [32]. We estimate the range and bearing to cars as the average range and bearing to all points tracked by VISO2 that project into the bounding box for a given car detection. Despite this very noisy landmark signal, we show qualitatively in Figure 5 that all of the probabilistic data association methods successfully recover reasonable trajectory estimates, while

the maximum-likelihood approach fails catastrophically due to incorrect loop closures. These results are corroborated by the translation and rotation errors, summarized in Tables I and II, respectively.

VI. CONCLUSION AND FUTURE WORK

We have proposed an approach to semantic SLAM with probabilistic data association based on approximate max-marginalization of data associations. This led to a "max-mixture"-type approach to factor graph SLAM, amenable to nonlinear least-squares optimization. We have shown how mixture component weights can be computed from joint semantic and geometric measurements, and null-hypothesis associations can be incorporated to reject bad loop closures. We evaluated the proposed approach on real stereo image data with known data associations using AprilTags [27] under a variety of simulated odometry and detection noise models, as well as on stereo image data from the KITTI dataset using noisy detections of cars as landmarks. We have shown that our approach is competitive with recent expectation-maximization methods for data association in semantic SLAM and drastically outperforms the common maximum-likelihood approach, particularly with use of null-hypothesis data association, while being similarly easy to implement within existing factor graph optimization frameworks. The particular benefits of the null-hypothesis in the proposed framework suggest that the ability to reject incorrect loop closures is a necessity for semantic SLAM systems relying on object detectors, and this work is one step toward unifying existing literature in robust SLAM with recent work on data association for object-level/semantic SLAM.

In this work we used classifications from object detectors to disambiguate data associations, but alternative semantic descriptors can be modeled and incorporated similarly, for example using neural network-based feature matching techniques [33]–[35]. Additionally, our experience simulating semantic SLAM with AprilTags suggested that orientation can provide very useful cues for data association. While we focused on approaches that give limited geometric information about objects, methods like [36] that infer the full pose of objects may greatly improve accuracy and robustness of data association.

REFERENCES

[1] 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.

[2] 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*. IEEE, 2017, pp. 1722–1729.

[3] K. Doherty, D. Fourie, and J. J. Leonard, "Multimodal semantic SLAM with probabilistic data association," in *2019 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[4] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI Vision Benchmark Suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[5] D. Reid, "An algorithm for tracking multiple targets," *IEEE transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, 1979.

[6] Y. Bar-Shalom and E. Tse, "Tracking in a cluttered environment with probabilistic data association," *Automatica*, vol. 11, no. 5, pp. 451–460, 1975.

[7] I. J. Cox and J. J. Leonard, "Modeling a dynamic environment using a bayesian multiple hypothesis approach," *Artificial Intelligence*, vol. 66, no. 2, pp. 311–344, 1994.

[8] ——, "Probabilistic data association for dynamic world modeling: A multiple hypothesis approach," in *Advanced Robotics, 1991.'Robots in Unstructured Environments', 91 ICAR., Fifth International Conference on*. IEEE, 1991, pp. 1287–1294.

[9] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *Proc. of the AAAI National Conference on Artificial Intelligence, 2002*, 2002.

[10] N. Sünderhauf and P. Protzel, "Switchable constraints for robust pose graph SLAM," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 1879–1884.

[11] A. V. Segal and I. D. Reid, "Hybrid inference optimization for robust pose graph estimation," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 2675–2682.

[12] L. Carlone, A. Censi, and F. Dellaert, "Selecting good measurements via ℓ1 relaxation: A convex approach for robust estimation over graphs," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 2667–2674.

[13] L. Carlone and G. C. Calafiore, "Convex relaxations for pose graph optimization with outliers," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 1160–1167, 2018.

[14] P.-Y. Lajoie, S. Hu, G. Beltrame, and L. Carlone, "Modeling perceptual aliasing in slam via discrete–continuous graphical models," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1232–1239, 2019.

[15] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[16] S. Yang and S. Scherer, "CubeSLAM: Monocular 3D object detection and SLAM without prior models," *arXiv preprint arXiv:1806.00557*, 2018.

[17] 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*, 2013, pp. 1352–1359.

[18] B. Mu, S.-Y. Liu, L. Paull, J. Leonard, and J. P. How, "SLAM with objects using a nonparametric pose graph," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 4602–4609.

[19] D. Koller, N. Friedman, and F. Bach, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[20] M. Kaess and F. Dellaert, "Covariance recovery from a square root information matrix for data association," *Robotics and autonomous systems*, vol. 57, no. 12, pp. 1198–1210, 2009.

[21] M. Quigley, K. Conley, B. P. 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*, 2009.

[22] 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.

[23] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," Georgia Institute of Technology, Tech. Rep., 2012.

[24] "StereoLabs ZED camera." [Online]. Available: https://www.stereolabs.com/zed/

[25] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[26] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.

[27] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, oct 2016, pp. 4193–4198.

[28] D. Malyuta, C. Brommer, D. Hentzen, T. Stastny, R. Siegwart, and R. Brockers, "Long-duration fully autonomous operation of rotorcraft unmanned aerial systems for remote-sensing data acquisition," *Journal of Field Robotics*, p. arXiv:1908.06381, Aug. 2019. [Online]. Available: https://doi.org/10.1002/rob.21898

[29] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7310–7311.

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[31] 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*. Springer, 2016, pp. 21–37.

[32] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d Reconstruction in Real-time," in *IEEE Intelligent Vehicles Symposium*, Baden-Baden, Germany, June 2011.

[33] Y. Tian, B. Fan, and F. Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 661–669.

[34] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *European Conference on Computer Vision*. Springer, 2016, pp. 467–483.

[35] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3279–3286.

[36] L. Nicholson, M. Milford, and N. Sünderhauf, "QuadricSLAM: Constrained Dual Quadrics from Object Detections as Landmarks in Semantic SLAM," *IEEE Robotics and Automation Letters (RA-L)*, 2018.