ANNUAL REVIEWS

Annual Review of Control, Robotics, and Autonomous Systems

Advances in Inference and Representation for Simultaneous Localization and Mapping

David M. Rosen,[1] Kevin J. Doherty,[2] Antonio Terán Espinoza,[2] and John J. Leonard[2]

[1]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; email: dmrosen@mit.edu

[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; email: kdoherty@mit.edu, teran@mit.edu, jleonard@mit.edu

Annu. Rev. Control Robot. Auton. Syst. 2021. 4:215–42

First published as a Review in Advance on January 6, 2021

The *Annual Review of Control, Robotics, and Autonomous Systems* is online at control.annualreviews.org

https://doi.org/10.1146/annurev-control-072720-082553

Keywords

simultaneous localization and mapping, SLAM, robust estimation, certifiable perception, semantic SLAM, active SLAM

Abstract

Simultaneous localization and mapping (SLAM) is the process of constructing a global model of an environment from local observations of it; this is a foundational capability for mobile robots, supporting such core functions as planning, navigation, and control. This article reviews recent progress in SLAM, focusing on advances in the expressive capacity of the environmental models used in SLAM systems (representation) and the performance of the algorithms used to estimate these models from data (inference). A prominent theme of recent SLAM research is the pursuit of environmental representations (including learned representations) that go beyond the classical attributes of geometry and appearance to model properties such as hierarchical organization, affordance, dynamics, and semantics; these advances equip autonomous agents with a more comprehensive understanding of the world, enabling more versatile and intelligent operation. A second major theme is a revitalized interest in the mathematical properties of the SLAM estimation problem itself (including its computational and information-theoretic performance limits); this work has led to the development of

novel classes of certifiable and robust inference methods that dramatically improve the reliability of SLAM systems in real-world operation. We survey these advances with an emphasis on their ramifications for achieving robust, long-duration autonomy, and conclude with a discussion of open challenges and a perspective on future research directions.

1. INTRODUCTION

Simultaneous localization and mapping (SLAM) is the problem (and procedure) of constructing a globally consistent model of an environment from local observations of it. This is an essential capability for autonomous mobile robots, supporting such basic functions as planning, navigation, and control (1). In consequence, SLAM has been the focus of an intense and sustained research effort over the previous three decades (2–5). While this work has led to remarkable progress—including the widespread availability of complete high-quality, open-source SLAM solutions (6–8)—there remain numerous fundamental challenges in the development of SLAM systems capable of supporting truly robust, intelligent, long-duration autonomy (5). In this article, we survey recent progress and open challenges in SLAM, with a particular focus on two crucial areas for achieving this objective: inference and representation.

Inference addresses the algorithmic aspects of estimating a model of the environment from raw sensor data. Historically, computational tractability and empirical evaluation have been the primary drivers of algorithmic SLAM research. This empirical orientation has enabled tremendous progress in reducing SLAM to a practical technology, including the development of the sparse graphical optimization-based framework that forms the basis of current state-of-the-art techniques (4, 5, 9). At the same time, however, the historical emphasis on experimental evaluation (which is restricted to measuring performance empirically, under a particular set of operating conditions) leaves many fundamental theoretical and algorithmic properties of the SLAM estimation problem unresolved. These include, for example, such elementary aspects as the estimation accuracy that a SLAM system can achieve, what specific features of a given problem determine these performance limits, and under what conditions it is possible, even in principle, to efficiently compute a satisfactory SLAM estimate in practice (10).[1] A major theme of recent work has been a renewed focus on these fundamental theoretical and algorithmic challenges. In particular, we describe recent advances in characterizing the information-theoretic limits of SLAM, the development of the first practical inference algorithms that enjoy formal performance guarantees, and robust extensions.

Representation concerns the environmental attributes that can be captured in a SLAM system's model. Historically, these have been grounded in simple geometric primitives such as points, lines, and planes. A second major theme of current research, motivated in part by recent progress in machine learning for perception, is the development of richer representations (incorporating properties such as temporal dynamics, objects, affordances, and semantics) to enable higher-level reasoning and advanced autonomy, including environmental and human–robot interaction. Until recently, such efforts were largely limited by the necessity for a priori known object models, due to the difficulty of tasks such as object recognition and detection. Recent years have seen researchers revisiting seminal work such as Kuipers's spatial-semantic hierarchy (12), now supported by decades of progress in machine perception techniques that enable these ideas to be more fully

[1] Note that SLAM estimation problems are typically both high dimensional and nonconvex, which immediately raises the specter of computational complexity (11).

realized. We survey recent strides in the representational capabilities of SLAM systems, especially the modeling and representation of nonmetric information such as semantics and environmental topology; the ability to operate in changing environments; the interaction of SLAM with learning; and, broadly, the increasingly task-dependent nature of representations for SLAM.

1.1. Problem Formulation

Before proceeding, we provide a brief review of the SLAM problem and its mathematical formalization in order to ground our subsequent discussion. For a more comprehensive introduction, we encourage readers to consult excellent works by Thrun et al. (1), Stachniss et al. (4), and Dellaert & Kaess (9).

SLAM is fundamentally the problem of constructing a consistent global model from a collection of local observations. As real-world sensor observations are affected by measurement noise, we formalize this problem using the language of statistical estimation (1). Let $\{x_i\}_{i=1}^{n} \triangleq X \in \mathcal{X}$ denote a collection of latent states (the model) that we would like to estimate, and let $\tilde{Y} \triangleq \{\tilde{Y}_k\}_{k=1}^{m}$ denote a set of sensor measurements. We assume that each observation $\tilde{Y}_k$ is sampled from a probabilistic generative model according to

$$\tilde{Y}_k \sim p_k(\cdot|X_k) \quad \forall k \in [m], \qquad 1.$$

where $X_k \subseteq X$ denotes a subset of the states X comprising the complete model.

Equation 1 formalizes our notion of locality: Each observation $\tilde{Y}_k$ depends only on a (typically very small) subset X_k of the complete model X. This is a ubiquitous characteristic of SLAM problems and is a consequence of the principles of operation of real-world sensors. For example, laser scanners and cameras provide information only about those portions of an environment that are in a direct line of sight; similarly, GPS measurements depend only on the receiver's current position, not its previous or future locations.

The locality of the observation models in Equation 1 enables us to decompose the joint likelihood $p(\tilde{Y}|X)$ for the model X given the data $\tilde{Y}$ into a product of small conditional likelihoods:

$$p(\tilde{Y}|X) = \prod_{k=1}^{m} p_k(\tilde{Y}_k|X_k). \qquad 2.$$

This conditional factorization provides the mathematical basis for fusing the local observations $\tilde{Y}_k$ into the coherent global representation X we wish to obtain.[2]

It is often convenient to model the factorization shown in Equation 2 by means of probabilistic graphical models (13). The utility of this is twofold. First, the exploitation of the conditional independencies implied by Equation 2 is essential for achieving fast inference, and probabilistic graphical models make this independence structure directly accessible via their edge sets. Second, the graphical formalism provides a convenient modular modeling language for constructing the complex joint distributions in Equation 2 from simple constituent parts (9) (**Figure 1**).

The probabilistic graphical model formulation described in Equations 1 and 2 and **Figure 1** thus provides an elegant general abstraction for the problem of global estimation from local

[2]We remark that while Equations 1 and 2 as written describe the conditional likelihood $p(\tilde{Y}|X)$ for the data $\tilde{Y}$ given the model parameters X, our discussion straightforwardly extends to the joint distribution $p(\tilde{Y}, X) = p(\tilde{Y}|X)p(X)$ simply by appending the factor(s) describing the prior $p(X) = \Pi_l p_l(X_l)$ to the decomposition in Equation 2. Thus, Equation 2 suffices to describe both likelihood-based and fully Bayesian formulations of the SLAM problem.

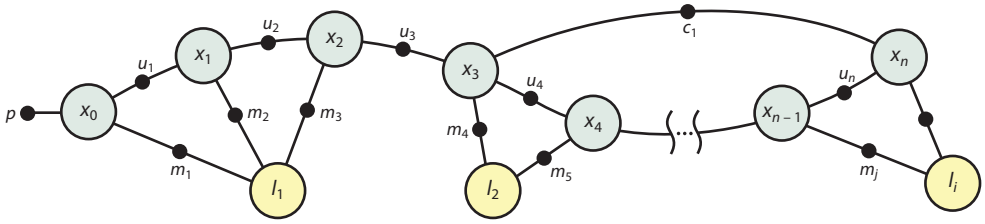

Figure 1

Factor graph model of the conditional factorization shown in Equation 2 for a simple pose-and-landmark SLAM problem (14). Here, variable nodes (corresponding to model parameters X) are shown as large circles, and factor nodes (corresponding to conditional densities p_k) are shown as small circles. Edges connect each conditional density $p_k(\tilde{Y}_k|X_k)$ to the subset of variables X_k upon which it depends (9). In this case, the variables consist of robot poses x and landmark positions l, and the factors are odometry measurements u, a prior p on the initial robot pose x_0, loop closure observations c, and landmark measurements m.

measurements. In particular, to instantiate a concrete estimation problem, it suffices to specify a representation for the model (i.e., the number and types of the states $X \in \mathcal{X}$ to be estimated) and the measurement models in Equation 1 that relate these quantities of interest to the available data $\tilde{Y}$. SLAM systems can thus be understood in terms of two fundamental properties:

- Representation: What quantities of interest X does the system model, and what are the generative models in Equation 1 that relate these parameters to the available data $\tilde{Y}$?
- Inference: What procedures are employed to perform inference over the probability distribution described by Equation 2?

1.2. Relation to Prior Surveys and Scope

The SLAM literature is vast, and a comprehensive summary of prior work is far beyond the scope of this article. Our focus is primarily on the advances that have occurred since 2016, when three previous surveys (5, 10, 15) were published. Readers are encouraged to consult the standard reference by Thrun et al. (1) for an elementary exposition of the SLAM problem, and the tutorials by Durrant-Whyte & Bailey (2, 3), Stachniss et al. (4), Dellaert & Kaess (9), and Grisetti et al. (16) for an overview of prior algorithmic progress in SLAM.

The survey by Cadena et al. (5) provides an extensive overview of the state of the art in SLAM as of 2016, including issues of robustness, scalability, higher-level representations, and active SLAM; we revisit many of these issues throughout this review, highlighting both recent progress and remaining open challenges. Huang & Dissanayake's (10) critique considered algorithmic and information-theoretic aspects of the SLAM estimation problem, including observability, convergence, accuracy, and consistency, and is closely aligned with our discussion of inference methods. Lowry et al. (15) addressed the specific problem of place recognition, which is closely related to issues of environmental representation and semantic mapping but aimed primarily at the problem of identifying loop closures.

By 2016, a consensus had emerged in the research community that a certain class of SLAM problems had become relatively well understood. For problems involving the estimation of simple geometric primitives (such as points, lines, planes, or camera calibrations) with well-characterized measurement models in Equation 1 (e.g., the projective mappings of vision sensors), maximum likelihood or maximum a posteriori estimation methods built atop factor graph representations of Equation 2 had become the method of choice (4, 5). A number of high-quality, open-source implementations of these methods became available, including iSAM (17), GTSAM (9, 18), and g^{2}o

(19), and the availability of benchmark data sets facilitated standardized measures of performance and steady progress. Dense and semidense visual SLAM methods, including ORB-SLAM (7) and ElasticFusion (8), demonstrated remarkable progress in camera pose estimation and 3D scene reconstruction for moderately sized scenes. Visual–inertial navigation, which seeks to estimate the trajectory of a moving sensor as accurately as possible (i.e., dead reckoning), had also seen substantial progress; Huang (20) provided a recent survey of progress in this area.

Yet despite this remarkable progress, numerous open challenges in SLAM remain, especially regarding the representational richness and algorithmic reliability necessary to achieve persistent, intelligent, long-duration autonomy. Section 2 addresses the SLAM inference problem, including fundamental computational and information-theoretic limits, certifiably correct estimation methods, and robust and scalable solvers. Section 3 addresses the issue of representation for SLAM, seeking to bring us beyond elementary geometry to consider objects and a hierarchy of spatial relations, revisiting Kuipers's seminal work on semantic hierarchies for spatial AI. Section 4 closes the review with a discussion of open issues and prospects for future research.

2. ADVANCES IN ESTIMATION AND INFERENCE FOR SLAM

In this section, we survey recent theoretical and algorithmic advances in inference for SLAM. While historically, computational speed and empirical evaluation have been the primary metrics for assessing progress, a major theme of recent work has been a renewed focus on more deeply understanding the theoretical properties of the SLAM estimation problem in Equation 2, especially its geometric, algebraic, and graph-theoretic structure. These insights have illuminated many fundamental but previously only poorly understood aspects of the problem (including limits on achievable accuracy and computational cost), as well as enabled the development of novel classes of inference algorithms, including the first practical algorithms with formal performance guarantees for nonconvex SLAM estimation problems.

2.1. Computational Hardness and the Problem of Nonconvexity

The fundamental algorithmic challenge of SLAM is that the model in Equation 2 is a high-dimensional distribution over a nonconvex state space $\mathcal{X}$, and therefore performing inference within this model is computationally hard in general (13). Early research in SLAM explored a variety of approaches for performing tractable approximate inference (using, e.g., extended Kalman filters or Monte Carlo sampling) (1); however, by 2016 the community had settled on maximum likelihood estimation [or more generally M-estimation (21)] as the de facto method of choice (5, 9). In brief, this approach recovers a point estimate $\hat{X}_{\mathrm{MLE}} \in \mathcal{X}$ of the latent state X as the minimizer of an optimization problem of the form

$$\hat{X}_{\mathrm{MLE}}(\tilde{Y}) \triangleq \operatorname*{argmin}_{X \in \mathcal{X}} \sum_{k=1}^{m} l_k(X; \tilde{Y}_k), \qquad 3.$$

where each summand $l_k(X_k; \tilde{Y}_k)$ is the negative log-likelihood of the corresponding factor $p_k(\tilde{Y}_k|X_k)$ in the model in Equation 2, or a robust generalization thereof (21).

The maximum likelihood formulation in Equation 3 enjoys several attractive properties. From a theoretical standpoint, maximum likelihood estimation provides strong performance guarantees on the statistical properties of the resulting estimator [including asymptotic consistency and normality under relatively mild conditions (22)]. Computationally, the formulation of the estimation in Equation 3 as a sparse optimization problem admits the application of sparsity-exploiting first- or second-order smooth optimization methods (23) to efficiently recover critical points of

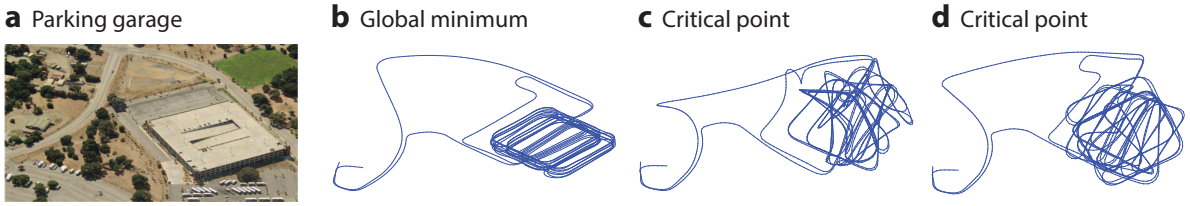

Figure 2

Examples of suboptimal estimates in pose-graph SLAM. Several estimates are shown for the trajectory of a robot as it enters and explores a multilevel parking garage, obtained as critical points of the maximum likelihood estimation in Equation 3. (*a*) The parking garage. (*b*) The true (globally optimal) maximum likelihood estimate $\hat{X}_{\mathrm{MLE}}$ computed using SE-Sync (26, 27). (*c*,*d*) Suboptimal critical points $\hat{X}$ obtained using local search. Panel *a* adapted from Reference 16 with permission from IEEE; panels *b*–*d* adapted from Reference 26.

the loss function. This computational efficiency is essential in enabling real-time robotics applications, where both computational and temporal resources on mobile platforms may be very limited. And indeed, current state-of-the-art algorithms and software libraries based on the formulation in Equation 3 are now capable of processing SLAM problems involving tens to hundreds of thousands of states on a single processor in real time (14, 18, 19, 24, 25).

However, the use of fast local optimization comes at the expense of reliability: Local search techniques can only guarantee convergence to a critical point $\hat{X}$ of the loss function, rather than the global minimizer $\hat{X}_{\mathrm{MLE}}$ required in Equation 3. Moreover, it is not difficult to find even fairly simple examples where suboptimal critical points are such poor solutions as to be effectively unusable as SLAM estimates (**Figure 2**). To address this potential pitfall, several strategies for initializing local search have been proposed in the literature, with the aim of favoring convergence to the true (global) minimizer (28–32). While these heuristics are often effective in practice, they do not provide any guarantees on the quality of the estimates $\hat{X}$ that are ultimately recovered.

These algorithmic difficulties can actually be understood as particular consequences of a fundamental computational stumbling block (11): As a high-dimensional nonconvex optimization, the maximum likelihood formulation in Equation 3 is general enough to encompass many problems that are known to be NP-hard, including, in particular, the fundamental problem of rotation averaging (33, 34).[3] This implies that in fact there cannot exist an algorithm that is capable of efficiently computing the maximum likelihood estimator $\hat{X}_{\mathrm{MLE}}$ required in Equation 3 in general, unless $P = NP$ (11).

In light of these considerations, as of 2016 several fundamental aspects of the reliability of state-of-the-art SLAM inference methods remained poorly understood (5, 10):

- Algorithmic: Under what conditions do SLAM inference methods successfully recover the correct estimate $\hat{X}_{\mathrm{MLE}}$ in Equation 3? Given the fundamental computational hardness of Equation 3, is it even possible to design SLAM estimation algorithms that are both practical and reliable? And if so, under what circumstances is this achievable?
- Statistical: Assuming that it is possible to compute $\hat{X}_{\mathrm{MLE}}$ in Equation 3, what are its statistical properties (e.g., achievable accuracy)? And what features of a given instance of Equation 3 determine these properties?

[3]This also entails that any estimation problem in the form of Equation 3 that subsumes rotation averaging—including, for example, the fundamental problem of pose-graph SLAM (26)—is also NP-hard (11).

2.2. Certifiably Correct SLAM

One of the most exciting advances of the last few years has been the development of the first class of SLAM estimation methods that are provably capable of efficiently recovering optimal solutions of Equation 3 for nonconvex problems, at least in certain practically important cases. These novel approaches, referred to as certifiably correct methods, are based on employing convex relaxation (rather than smooth local optimization) to search for high-quality state estimates. While the idea of applying convex relaxation in SLAM is not new [indeed, many well-known initialization techniques are based on this strategy (28–32)], what distinguishes certifiably correct methods from prior work is that the relaxations they employ are exact[4] provided that the noise on the data $\tilde{Y}$ in Equation 3 is not too large. Certifiably correct methods thus directly tackle the fundamental problem of nonconvexity in Equation 3: They enable the efficient computation of globally optimal solutions via convex programming within a restricted (but still practically relevant) operational regime (35).

Certifiably correct SLAM algorithms originated in the study of pose-graph SLAM specifically (**Figure 1**). Carlone and colleagues (36–38) proposed to address the problem of nonconvexity by leveraging Lagrangian duality. To that end, they developed a quadratically constrained quadratic program formulation of the SLAM problem and observed that the corresponding Lagrangian dual, a semidefinite program (39), was frequently tight in practice. This observation provides a simple means of certifying the optimality of a (correct) candidate solution $\hat{X}$ of Equation 3 using Lagrangian duality. However, due to the high computational cost of standard (interior-point) semidefinite programming methods, this certification approach (36, 37) still depends on the local search in Equation 3 to compute the estimate $\hat{X}$ itself. Rosen and colleagues (26, 27) subsequently studied the dual of Carlone and colleagues' semidefinite relaxation (Shor's relaxation of the original pose-graph SLAM problem) and proved that for sufficiently small noise it admits a unique low-rank solution that provides the exact MLE $\hat{X}_{\mathrm{MLE}}$; they also developed a specialized low-rank Riemannian semidefinite optimization method for efficiently solving large-scale instances of this problem. The resulting algorithm, SE-Sync, was the first practical certifiably correct method to appear in the SLAM literature. Rosen and colleagues (26, 27, 40) further observed that the semidefinite relaxation and low-rank optimization methods employed in SE-Sync could be directly generalized to a broad class of estimation problems (specifically, those formulated as rational polynomial optimization problems) via moment relaxation (41), thereby providing a general approach for synthesizing certifiably correct estimators.[5]

A nascent but rapidly growing body of work has subsequently adapted this approach to produce certifiably correct estimation methods for a variety of machine perception tasks, including rotation averaging (43, 44), calibrated two-view registration (45), extrinsic sensor calibration (46), 3D registration (47, 48), image segmentation (49), shape reconstruction (50), and alternative formulations of pose-graph SLAM (51), including sharper specialized relaxations for the 2D case (52, 53). In recent works, Fan & Murphey (54) and Tian et al. (55) have also shown how to adapt the low-rank Riemannian optimization used by Rosen et al. (26) and Briales & Gonzalez-Jimenez (51) to run in a distributed setting, enabling the first distributed certifiably correct methods

[4]A convex relaxation is called exact if its minimizer provides an exact solution to the original problem from which it was derived. Note that this condition can be checked post hoc, simply by verifying that the minimizer of the relaxation satisfies the constraints of the original problem.

[5]Kahl & Henrion (42) also proposed the use of moment relaxation for globally optimal geometric reconstruction in computer vision, although they considered only low-dimensional ($\leq$11D) estimation problems due to the high computational cost of standard interior-point semidefinite optimization methods.

for pose-graph SLAM and rotation averaging (56). The further development of this class of estimation methods remains a very active area of research.

2.3. Robust Estimation

It is well known that maximum likelihood estimators are also frequently nonrobust, meaning that corrupting an arbitrarily small fraction of the data $\tilde{Y}$ can cause the estimator $\hat{X}_{\mathrm{MLE}}$ in Equation 3 to diverge from the true latent value of the parameter X (21). This is true in particular for MLEs formulated as nonlinear least squares problems, as is frequently the case in robotics and computer vision applications. In practical SLAM applications, corrupted (outlier) measurements of this sort frequently arise from erroneous data associations (often due to visual aliasing) and are a primary source of the brittleness in current state-of-the-art systems. The development of SLAM estimation methods that are robust to outlier contamination is thus crucial for achieving reliable, long-duration autonomy.

Several approaches for addressing the problem of outlier contamination have previously appeared in the SLAM and computer vision literature (5). One line of work attempts to directly identify the set of inlier measurements by searching for the largest subset satisfying a notion of mutual consistency. A classical example of this class is random sample consensus (RANSAC) (57), which uses random sampling to search for sets of inlier measurements; while this approach works well for low-dimensional problems with a moderate proportion of outliers, the curse of dimensionality precludes its scaling to the high-dimensional problems typical in SLAM applications. In consequence, several more scalable implementations of consensus set search have been proposed specifically for use in SLAM. A prominent example is realizing, reversing, recovering (RRR) (58), which exploits the sequential structure of robotics applications to iteratively construct a consensus set: It integrates the measurements sequentially in small batches, checking the internal consistency of the resulting model each time; any batch that results in inconsistency is assumed to contain outliers and is discarded. While this approach can be very effective, it requires solving an (expensive) high-dimensional nonlinear estimation problem each time new data are added or removed. A more recent and computationally lightweight alternative is pairwise consistency maximization (PCM) (59); this method constructs a graph whose vertices correspond to the measurements $\tilde{Y}$ and whose edges connect mutually consistent observations, and then extracts the maximal clique as an estimate of the inlier set. Both RRR and PCM provide computationally tractable consensus set estimation methods for high-dimensional SLAM problems and are often very effective in practice; however, they provide no formal guarantees on the quality of the solutions they return.

An alternative class of approaches, originating in the work of Huber (60), is based on replacing the negative log-likelihood functions $-\log p_k(\tilde{Y}_k|X_k)$ that would appear in a standard maximum likelihood formulation of Equation 3 with robust losses $l_k(X;\tilde{Y}_k)$ that are less sensitive to the deleterious effects of outlier contamination; the resulting class of estimators thus achieves improved reliability at the cost of a (typically relatively minor) loss in statistical efficiency. Moreover, one can show that for suitable choices of the loss function, the resulting M-estimator is provably insensitive to contamination by a bounded fraction of outlier observations (21). Several robust M-estimation schemes have been proposed for use in SLAM, including the well-known switchable constraints (61), dynamic covariance scaling (62), and max-mixtures (63), all of which are straightforwardly implementable using the standard high-dimensional local optimization machinery already prevalent in these applications (14, 18, 19). Unfortunately, the shape required of a loss function in order to attain robustness against contamination tends to exaggerate the nonconvexity of the M-estimation in Equation 3, thus rendering these techniques more vulnerable to convergence to poor-quality critical points; this pitfall is further exacerbated by the fact that the methods typically employed

to initialize the local search (28, 30–32) are themselves no longer trustworthy when faced with potentially contaminated data.

In light of these considerations, a natural pathway to achieving practical robust estimation is attempting to combine the robust M-estimation in Equation 3 with the certifiably correct global optimization strategy outlined in Section 2.2. To that end, Yang & Carlone (64) recently described a general procedure that enables many geometric estimation problems involving the (robust) truncated squared-error loss to be reformulated as polynomial optimization problems[6] and demonstrated empirically that the semidefinite relaxations (41) of these problems are frequently tight, even when a large fraction of the data (in their experiments, greater than 50%) are outliers. While the resulting relaxations are too large to be solved directly using standard (interior-point) methods, they derive an efficient algorithm for solving the dual (sums-of-squares) problems, thereby enabling the certification of (correct) candidate solutions $\hat{X}$ for the original robust estimation problem in Equation 3.[7] Finally, they showed empirically that a local optimization strategy [based on graduated nonconvexity (65)] applied directly to Equation 3 very often succeeds in recovering certifiably optimal solutions, despite its nonconvexity. Taken together, these approaches provide an efficient and practically effective means of recovering certifiably optimal robust estimators $\hat{X}_{\mathrm{MLE}}$ for high-dimensional problems, even in the presence of substantial outlier contamination.

2.4. Information-Theoretic Limits of SLAM

Having considered the computational and algorithmic challenges of the M-estimation in Equation 3, we now address the statistical properties of the estimator $\hat{X}_{\mathrm{MLE}}$ itself. In particular, can we develop sharp limits and/or guarantees on the statistical performance of $\hat{X}_{\mathrm{MLE}}$ in SLAM?

As in the case of Sections 2.2 and 2.3, the fact that instances of Equation 3 are typically defined over high-dimensional nonconvex spaces can substantially complicate the analysis of $\hat{X}_{\mathrm{MLE}}$; nevertheless, recent work has identified certain spectral graph-theoretic properties (66) of the model (**Figure 1**) underlying an instance of SLAM as the key quantities controlling estimation performance, at least in certain practically important cases. Specifically, the connection Laplacian L [a generalization of the standard (scalar) Laplacian to graphs with matrix-valued data assigned to their edges (67)] and its spectral gap $\lambda(L)$ have emerged as objects of central importance. For rotation averaging, Boumal et al. (68) showed that the Cramér–Rao bound (a lower bound on the achievable covariance of any unbiased estimator) admits a simple expression in terms of the connection Laplacian L. Similarly, Khosoussi et al. (69) have recently derived an analogous result for 2D pose-graph SLAM. These results provide simple and sharp relations between the graphical structure of SLAM problems and the accuracy of SLAM estimates.

Interestingly, the analyses presented by Bandeira et al. (33), Rosen et al. (26), and Eriksson et al. (43) showed that the spectral gap also plays a central role in controlling the exactness of the semidefinite relaxations underlying the certifiably correct estimators for the rotation averaging and pose-graph SLAM problems; that is, the same quantity $\lambda(L)$ controls both the statistical and the computational hardness of these estimation problems. While much work remains to be done in this area, these early results are strongly indicative that spectral graph-theoretic tools will have an important role to play in designing reliable measurement networks for spatial perception.

[6]This reformulation involves introducing an auxiliary binary indicator variable, similarly to the approach of switchable constraints (61).

[7]This is analogous to the verification strategy employed by Carlone and colleagues (36, 37) for the specific case of pose-graph SLAM.

2.5. Open Questions and Future Research Directions

The theoretical and algorithmic advances described in this section lay the foundation for designing a new generation of principled, efficient, and provably reliable estimation algorithms for SLAM. However, much work remains to realize this potential in the form of standard technology that can be readily deployed by practitioners. In this section, we highlight three avenues for future work toward realizing this vision.

2.5.1. Efficient optimization methods for certifiably correct perception. The high computational cost of semidefinite programming remains a serious obstacle to the development of practical certifiably correct estimation methods. For example, all of the certifiably correct methods described in Section 2.2 either are restricted to small-scale problems that can be solved using standard (interior-point) semidefinite programming techniques or depend on specialized, purpose-built semidefinite optimization algorithms that are specifically tailored to each problem's structure. While semidefinite optimization remains a very active research area (70), the development of computational approaches that are reasonably general, easy to use, and well suited to the particular characteristics of machine perception applications (e.g., high dimensionality, ill conditioning, limited computational and temporal resources, and the need for high-precision solutions) remains an open problem. Reference 71 reports one initial step in this direction, but much work remains to be done.

Similarly, to date work on certifiable estimation has primarily addressed the offline (batch) setting. The development of efficient incremental semidefinite optimization methods—analogous to, e.g., iSAM (17, 18) and GTSAM (9)—would be extremely valuable for enabling certifiable estimation in real-time online perception tasks.

2.5.2. A priori performance guarantees. Certifiably correct perception methods approach the problem of solution certification in a post hoc, per-instance manner. This is sufficient to enable run-time verification and monitoring to confirm that a perception system is functioning as it should be. However, as designers and practitioners, we would also like to have formal results that clearly delineate in advance the circumstances under which such systems will succeed, as well as characterizations of their expected performance as statistical estimators. While there are some formal results that guarantee at least the existence of a noise regime within which certifiable perception methods will succeed (26, 43, 52, 72), at present there do not appear to be general, user-friendly theoretical tools for deriving sharp bounds on the size of this regime. Results of this type would be extremely useful in the design of measurement systems for machine perception applications, especially in safety- and life-critical applications (e.g., autonomous vehicles).

2.5.3. Beyond point estimation. The certifiably correct methods described in this section are all derived in the context of Equation 3, i.e., in the setting of point estimation; this is a reasonable approach whenever the underlying likelihood or posterior probability density is highly concentrated around a single mode. However, in practice it often occurs that the posterior is highly diffuse (due to a lack of sufficiently informative measurements) or even multimodal (as in the case, for example, of uncertain data association). In these cases, point estimation can dramatically underestimate the actual posterior uncertainty, even missing the existence of completely distinct but equally plausible solutions. Blindly trusting such a result can easily cause an erroneously overconfident belief in a completely wrong answer, potentially endangering the safety of the overall system.

Overcoming this challenge requires the development of inference methods that go beyond simple point estimation and attempt to explicitly characterize posterior uncertainty. This is

necessary both for introspection (i.e., to enable an autonomous agent to know what it doesn't know) and, by extension, for planning and active perception (to enable an autonomous agent to reason about how it could reduce its own uncertainty). The development of tractable estimation methods that can extract this richer information while scaling gracefully to high-dimensional problems is an important and fundamental open problem for future research in SLAM, although References 73–75 have proposed initial steps along these lines.

3. REPRESENTATION: BEYOND POINTS AND PLANES

This section surveys recent advances in map and robot state representation for SLAM. Our focus throughout is on the development of representations that go beyond geometry alone and on the corresponding problems that have arisen in this domain. A major throughline of recent work in representations for SLAM is the unification of semantic and geometric information, propelled by recent advances in machine learning. We explicitly discuss the challenges, opportunities, and major questions associated with the development of joint geometric and semantic representations; in so doing, we highlight three key research areas, considering the influence of semantics in each: navigation in dynamic and semistatic environments, abstraction and hierarchy, and learned representations. To contextualize progress in this area, we briefly review state-of-the-art geometric representations for SLAM.[8]

In considering the problem of constructing a global representation from local measurements, a fundamental question arises: What should the global representation be? More precisely, in designing a navigating robot, a choice of environmental representation must address which features (properties) of the world are relevant, what data and/or models are necessary to encode those features, and how that approach should be operationalized. Given that the answers to many of these questions depend on the specific task at hand, it is no surprise that numerous representations have been proposed, with no clear universally superior choice. Moreover, while the basic problem statement for SLAM has remained essentially unchanged for more than 30 years, the criteria for success have changed drastically. No longer do we expect SLAM methods to simply build geometric maps of static worlds and localize a robot. Modern SLAM methods often must also synthesize data from heterogeneous sensors to infer object categories, operate in dynamic and evolving environments, and support planning. These expectations mirror the increasingly stringent requirements in estimation for SLAM: We expect certifiability, robustness, and reliable operation with a variety of sensors, each possessing distinct noise characteristics.

Within the last decade, the expectation of SLAM methods to perform certain scene understanding tasks coupled with classical geometric estimation coincided with the recent successes (and accessibility) of machine learning methods for perception tasks [especially in computer vision (76, 77)], driven mainly by the availability of large-scale labeled data sets such as ImageNet (78). Such advances in learning, especially deep learning, motivated research in richer, semantic or object-level representations for SLAM (79). The expansion of SLAM to include semantic perception capabilities (most prominently using vision) to enable an embodied system to interact with the environment has recently been referred to as spatial AI (80–82).

Due in large part to the recent progress in learning for machine perception and the historical success of geometric methods, the past decade has seen increased research interest in the

[8]For a review of progress in geometrically grounded SLAM representations, particularly those used in visual SLAM, we refer readers to a survey by Cadena et al. (5).

development of SLAM systems that estimate nonmetric properties of the environment (e.g., classifying static versus dynamic parts of the scene) and advances in representation learning methods (e.g., deep learning) that can be used to build richer maps. The coupling of map representation with the task (task-driven perception) is an emerging theme: We desire SLAM systems that will build environment representations that are useful for some task. Increasingly often, these tasks include active or interactive perception (83), intelligent exploration, manipulation, and learning as a task in itself.

3.1. Geometric Representations

Sparse, landmark-based (or feature-based) representations have been used since the earliest work on the SLAM problem more than 30 years ago. These representations seek to build and maintain a map of landmarks, i.e., salient, distinct environment features that can be reliably recognized. The problem of recognizing previously mapped features is known as data association. In the context of visual SLAM, the SIFT (84), SURF (85), and ORB (86) features are among the most commonly used feature descriptors. Visual SLAM methods leveraging sparse representations, such as ORB-SLAM (7, 87) and DSO and its variants (88–90), have had tremendous success at precisely localizing a camera in 3D environments. However, the maps built by these methods, while useful for localization, are not actionable. Geometrically, they consist of a sparsely distributed collection of points in 3D space, rather than an explicit characterization of free and occupied volumes and the boundaries (surfaces) that separate them. In consequence, they are not immediately convenient for planning collision-free paths or exposing potential routes for exploration.

In contrast, dense spatial representations attempt to build complete, albeit approximate, descriptions of surfaces or occupied space. These descriptions may take the form of occupancy grid maps (91); volumetric maps (in 3D) (92); meshes (93); dense point clouds, as in LSD-SLAM (6); or truncated signed distance function maps, as in ElasticFusion (8) and KinectFusion (94). Meshes and other forms of surface representations provide critical information for a system attempting to avoid colliding with its environment, whereas point-cloud-based methods do not. In these dense representations, the correspondence problem is addressed by computing the most probable location within the map from which a measurement could have been made. In many geometric maps, this is performed through a variant of iterative closest-point registration; alternatively, as with sparse representations, loop closures can be determined through place recognition using feature descriptors [e.g., as is done in the dense SLAM system Kintinuous (95)].

The principal limitations of solely geometric representations are as follows:

- Geometry alone cannot explain all potentially relevant sensory properties of the environment (e.g., color, tactile sensation, or weight). Thus, a comprehensive understanding of the environment requires reasoning capabilities above and beyond geometry.
- This fact suggests that geometry alone is insufficient for more sophisticated forms of sensor fusion that capture not only the physical appearance of objects but also interaction, sound, touch, and so on.
- Bare geometric representations do not naturally support human–robot interaction. For most tasks, humans do not specify, e.g., objects or locations in terms of numerical class labels or spatial coordinates. If robots aim to interact with the world alongside humans, then we need them to have at least some basic competency at interpreting higher-level, human-centric semantic descriptions.
- Though spatial abstractions for geometric data exist, they are not grounded in action. Any representation suited to reasoning about a task beyond localization must be actionable.

3.2. Unifying Semantics and Geometry

By 2015, the SLAM community had increasingly recognized the limitations of purely geometric perception and that concurrent advances in machine learning would enable richer, semantic representations (5, 96). Research on semantics in SLAM considers the development of more expressive map representations capable of incorporating objects and places and enabling higher-level autonomy. The ability to understand the world in terms of objects and places can provide robots with a number of benefits over traditional (dense or semidense) approaches, such as point clouds or octrees. Semantic maps can be encoded with much smaller memory and processing footprints and can provide robustness against the inevitable accumulation of small errors that can render purely geometric approaches brittle. Map representations based on human-understandable semantic primitives can also enable better ways for human operators to interact with autonomous systems, in more natural terms for the human. Thus far, there is no consensus on a mathematical formalism for the fusion of semantic information about the state of a robot and its environment with estimates of the local scene geometry. This section concerns the progress that has been made in establishing models for semantic information in SLAM systems.

The prevailing approach toward incorporating semantics into navigation systems is to treat the output of learned perception models (e.g., object detectors) as virtual sensor measurements. As in classical work on learned sensor models, this treatment essentially posits that some deterministic (though perhaps unknown) function relates the latent semantic category of an object with some other measurable physical properties, such as its appearance in a camera image, and that this function can be approximately learned from data. Historical efforts in robot mapping have sought to learn complex measurement models, e.g., those of sonar sensors (97). However, in this prior work there really is a (perhaps complex) relationship between the geometric structure of a scene and the measurements made in that environment; this is an immediate consequence of the physics of the sensing apparatus. In contrast, such a physically grounded relation need not hold between raw sensory data and semantics.

Semantics often arise through affordance rather than appearance. Such affordances may only become clear through interaction, and one cannot reason about interaction using single-image measurement models that currently dominate the navigation literature. Nonetheless, the recognition of objects from an a priori known set of classes—enabled by these approaches—is undoubtedly a useful capability for mobile robots. Recent work has focused on challenges arising within this relatively limited scope, such as characterizing the noise or uncertainty in the output of learned perception models, and the fact that these models are known to fail unpredictably even in nominal, nonadversarial operating conditions, causing drastic errors in systems that employ this information for navigation tasks. However, incorporating discrete measurements of object categories into the continuous geometric formulation of SLAM poses its own set of challenges for inference: Joint discrete–continuous estimation often leads to combinatorially large state spaces, making the determination of the most probable map difficult.

3.2.1. Joint inference of semantics and geometry. Given a model of object detections and classifications as the output of a (noisy) sensor, the problem of jointly estimating the latent semantic class and geometry of landmarks in the environment can be posed in terms of Equation 2 as

$$\hat{X}, \hat{L}, \hat{D} = \arg\max_{X,L,D} p(\tilde{Y} \mid X, L, D), \tag{4.}$$

where $\tilde{Y}$ denotes the full set of measurements (including semantic measurements); X the set of robot poses; L the set of environmental landmarks, which typically consist of some geometric

information (e.g., position, orientation, and size) coupled with a discrete semantic label from a known, fixed set of classes; and D the set of associations between measurements in Y and landmarks in L. A key observation is that discrete-valued categorical information about objects can be naturally combined with the already discrete inference problem of data association: Knowledge of an object's category can help distinguish it in clutter from other objects. This formulation unifies discrete models of semantic category, geometric estimation, and data association; however, in addition to being nonconvex and high-dimensional (as in the standard SLAM formulation), it now also involves combinatorial optimization. Moreover, in committing to the use of semantics for data association, one must cope with the errors of learned perception models.

Given this formulation, Bowman et al. (98) performed joint optimization via expectation maximization: First fix the data association probabilities and landmark classes and optimize the robot poses and landmark locations (with measurements weighted by the respective probabilities of their landmark correspondence), then fix the robot poses and landmark locations to compute new data association probabilities and landmark classes. This approach has the benefit of assigning soft associations to objects, gradually converging to a locally optimal solution to the problem in Equation 4. The data association probabilities can be computed via approximate matrix permanent methods (99).

An alternative approach to the combinatorial inference problem of Equation 4 is to reframe the optimization over discrete variables as one over only continuous-valued variables. In early work to this end, Sünderhauf et al. (96) optimized probabilities of semantic labels, which are defined over the $(K-1)$-dimensional unit simplex for a K-class semantic labeling problem. More recently, full (albeit approximate) posterior inference has been considered by marginalizing out the discrete variables, producing a mixture representation (100). A similar methodology has been applied in the context of maximum a posteriori inference to enable the use of continuous, gradient-based optimization methods to approximately solve Equation 4 (101). Finally, some have approached the combinatorial problem directly via multi-hypothesis tracking (75).

3.2.2. Semantic map representations. Many representations for semantic navigation consist of traditional geometric representations (such as truncated signed distance functions, occupancy grids, or meshes) in which each element is augmented with a semantic label. SemanticFusion (102) is one such dense representation; semantic octree-based occupancy maps have also been used (103), and Kimera (93) uses meshes.

Much work has also been done in the past several years in the area of object-level representations within SLAM. Since the early work on these representations (104–106), the community has largely shifted away from a priori known object models toward the use of learned perception models for object detection, recognition, and pose estimation. The simplest object-centric representations treat object landmarks as points in Euclidean space augmented with semantic labels. More recently, representations have been developed that permit the estimation of not only the class and position of objects but also their orientation and extent. These include the dual quadric formulation (107–109), which models objects as 3D ellipsoids, as well as CubeSLAM (110), which represents objects in terms of rectangular bounding volumes. Another recent consideration in work on object-centric representations has been the use of learned object descriptors; an example is the work of Sucar et al. (111), which captures the shape (via the occupied volume) of objects as well as the pose of the object.

A major theme among all of these representations is the use of some principally geometric representation augmented with semantic class. In subsequent sections, we discuss the incorporation of more complex semantic relationships into SLAM representations; the development of such models is an important underexplored area that we revisit explicitly in Section 3.6.

3.3. Beyond Static Worlds

The vast majority of SLAM systems have assumed that the world is static. The so-called static world assumption (that only the robot itself can change state) has enabled great progress in SLAM but is often violated in practice. In consequence, most practical implementations of SLAM implicitly regard dynamic objects as unmodeled disturbances, and rely on outlier rejection mechanisms [e.g., RANSAC (57)] to filter them out. More recently, methods that attempt to explicitly discriminate between dynamic and static objects in a visual scene have been developed in order to identify (static) areas of the scene that are more informative for localization (112). Particular interest in recent research is not restricted to the removal of dynamic elements from a scene, but also extends to modeling their dynamics over a variety of spatial and temporal scales.

In going beyond the static world assumption, a fundamental question arises: Did I move or did the world move? Often, without more information, a conclusion cannot be drawn. Prevailing methods that rely on the static world assumption break in scenarios where the entire scene moves but the robot stands still. A key representational element missing from such systems is the ability to reason about geometric ambiguity. Formally, this problem stems from a lack of observability. To achieve truly robust perception, a SLAM system must be able to reason about such ambiguous information. Alternatively, in light of making an error of judgment, such a system should be able to revise its decision, thereby fixing its error, and then use its corrected representation to resume state estimation. This can be experienced, for example, when one is stopped in traffic and a neighboring car begins to move: One may feel that it is oneself moving, when in reality it is the scene that moved. Priors grounded in semantics can help to address this observability problem: Knowledge that certain collections of geometric features correspond to a house rather than a car may suggest (though not ensure) that the landmark is stationary. The modeling of changing environments poses additional, more abstract challenges. If we allow environments or the objects within them to change over time, then the recognition of familiar places and identities of objects within becomes nontrivial. Like the ship of Theseus, whose pieces are gradually replaced until none of its original components exist, it is difficult or impossible in dynamic worlds to determine unambiguous object or scene identities. Consequently, solutions to SLAM in dynamic environments generally rely on some assumptions about the nature of the possible dynamics in order to make the problem tractable.

3.3.1. Dynamic environments. Operationally, highly dynamic environments consist of those in which any motion not due to the robot occurs on a timescale that makes it directly observable (pedestrians, cars, etc.). This is the most commonly studied form of dynamic SLAM. The direct observability of these motions reduces the problem of dynamic SLAM to a front-end classification and filtering problem: Dynamic components of the scene can be identified and modeled locally without becoming part of the global environmental representation [as in the work of Wang et al. (113)]. Recent methods such as those described in References 114–116 leverage semantics to inform dynamic SLAM systems, particularly in deciding which features correspond to objects that are likely to be moving. An emerging theme in this area is the estimation of the velocity of classified objects in the scene.

3.3.2. Semistatic environments. Semistatic environments are those in which environmental change happens on a timescale that is observable only by repeated revisitation or observation of a given location. This can include, for example, furniture moving around or seasonal variation. The key challenge associated with modeling semistatic environments is that, operationally, learning and reasoning about these kinds of environmental changes requires some implementation of

environmental memory. This difficulty is in large part representational: What does it mean for a map to change over time? How can we build updatable maps, and how should we think about recording the evolution of environments over time once changes are permitted? Numerous approaches to this have been proposed. Rosen et al. (117), for example, developed a Bayesian framework for recursive estimation of the persistence of each feature in an environment. Krajník et al. (118) applied Fourier analysis to model the frequencies at which features in the environment may change. Bore et al. (119) made use of particle filters to learn the temporal dynamics of an arbitrary number of objects with a priori unknown feature correspondence. Zeng et al. (120) considered a semantic linking map representation that captures probabilistic spatial couplings between landmarks; this captures the intuition that in a semistatic environment, contextual semantic relationships between observed objects can be used to facilitate object search (e.g., to find cups, we might first look for tables or cabinets). Halodová et al. (121), Berrio et al. (122), and Pannen et al. (123) each considered the problem of managing changes to a map representation (though they did not explicitly attempt to track dynamic objects over time). The problem of change detection and map modification is especially pernicious for safety-critical autonomous vehicle systems that rely on high-resolution 3D maps for navigation.

3.3.3. Deformable environments. State estimation in deformable environments, or environments with deformable objects, remains a particularly challenging and underexplored research area in SLAM. A major milestone on this front was DynamicFusion (124), which demonstrated dense reconstruction of deformable objects using an RGB-D camera. More recent work has explored the fundamental theoretical problem of the observability of SLAM in deformable environments, in particular with application to mapping the interior of the human heart (125).

3.4. Abstraction and Hierarchy in Spatial Representations

Abstraction and hierarchy in spatial environment models have greatly improved the scalability, efficiency, and generalization of SLAM methods. Early work on the spatial-semantic hierarchy (12) outlined key ideas related to the construction and maintenance of a cognitive map of the environment. From a purely geometric standpoint, spatial partitioning algorithms enabled the scaling of dense representations of occupancy (92).

3.4.1. Topological models. Environment topology plays a major role in decision-making processes. For navigation, the (topological) decision to go left or right at a fork in the road can have a far more significant impact on the time to reach a destination than decisions about the specific (metric) motion plan. Topological maps describe an environment at this level of abstraction, i.e., at the level of connectivity. In the context of SLAM, topological maps provide a very compact representation that can be used for such coarse (but significant) navigation decisions, as well as localization.[9] In particular, topological feature graphs (126), whose vertices represent geometric features and whose edges represent obstacles, have recently been used to support information-theoretic exploration, where they provide a compact description of occupied space. Stein et al. (127) proposed a polygonal map representation from which topological navigation decisions can be obtained, enabling a broad class of learning-aided planning tasks. Neural topological SLAM (128) incrementally constructs a topological map (graph) in which vertices represent physical locations (identified with a set of features extracted from panoramic images via a neural network–based

[9]For a review of topological methods prior to 2016, we refer readers to a survey by Lowry et al. (15).

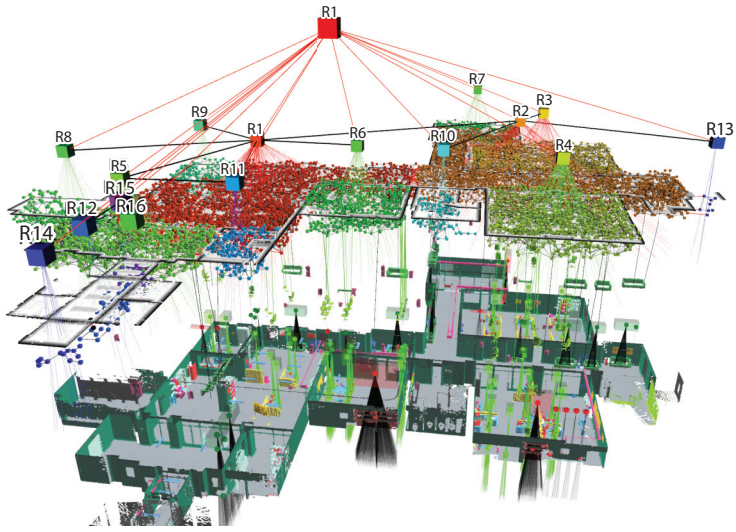

Figure 3

3D dynamic scene graphs (129). These are a recent application of scene graphs (previously common in the computer graphics community) to the SLAM problem and provide a substantial step toward linking scene understanding and spatial perception methods. Figure courtesy of A. Rosinol.

encoder) and edges represent connectivity/traversability between locations; this representation is used to support learning policies to navigate toward a goal, specified as an image taken at the target location.

3.4.2. Scene graphs. Representations that synthesize spatial and semantic information in hierarchies incorporating both metric and topological properties of the environment have been of very recent interest. Hierarchical models have the benefit of partitioning space at a variety of levels of abstraction, enabling efficient reasoning over large spatial scales. In particular, 3D scene graph models (129, 130) present a promising representational direction toward capturing object-level semantics, environment dynamics, and multiple spatial and semantic layers of abstraction (from the connectedness of unoccupied space to rooms and buildings and beyond). Scene graphs model the environment in terms of a directed graph where nodes can be entities such as objects or places and edges represent relationships between entities (depicted in **Figure 3**). The relationships modeled by a scene graph may be spatial or logical. Scene graph representations have also been used to learn physical descriptions of scenes in an unsupervised fashion from visual input alone (131), though this approach has yet to be applied directly in a robotics context.

3.5. Learning-Centered and Learned Representations

Many efforts thus far in semantic SLAM leverage learned perception models but opt for relatively straightforward environmental representations; these typically consist of some combination of classical geometric measurements and categorical output from a learned model, such as an object detector (e.g., 77). Such perception models are typically trained in an offline setting, using data that may not accurately reflect the conditions in which a robot is actually deployed. We may ask, then, whether it is possible to learn or refine perception models during robot navigation. In particular, knowledge of the scene geometry and robot motion during navigation motivates self-supervision

of learned perception models (e.g., 132, 133). We must address, in such settings, precisely what sorts of environment representations would permit these types of learning. Taking this a step further, we might consider learning parts of the representation itself, the logical extreme of which is end-to-end learning of navigation (as in 134).

3.5.1. SLAM for self-supervised learning. Work on SLAM-aware perception methods originated in the desire to leverage global spatial structure when recognizing objects in video streams (135). Since then, these capabilities have been extended to bootstrap supervision for a variety of learned perception models (132, 133, 136, 137). The central idea of these approaches is that a global (typically geometric) representation of the environment can be used as a supervisory signal for training a variety of models relating the motion of a camera and the scene geometry (e.g., visual odometry models). The ability to combine spatial and temporal information in order to improve learning is a unique feature of embodied spatial AI systems that undoubtedly warrants further attention.

3.5.2. Learning geometric representations. Coarse geometric models often fail to capture the precise volumes of objects, while dense representations require substantial memory to store a map. SceneCode (138) and NodeSLAM (111) incorporate learned intermediate representations for objects. From these compressed representations, the object geometry can be recovered. More generally, DeepFactors (139) incorporates compact representations of depth images toward the same goal.

Given the popularity of gradient-based methods for optimization (particularly backpropagation), there is great potential utility for differentiable SLAM representations. Recent work to this end (140) aims to enable backpropagation through traditional SLAM systems in order to seamlessly integrate with learning models such as neural networks.

3.5.3. Learning to navigate. Recently, several models have considered end-to-end learning for navigation. Broadly, these methods seek to learn functions mapping directly from sensor inputs (or a history of sensor measurements) to actions (e.g., 141, 142). A major consideration for these methods is how to structure the learning problem. Zhang et al. (141) used an agent with external memory representing an occupancy map. More recent work (128) has considered specifically structuring the learning problem to use a topological representation of the environment.

3.6. Open Questions and Future Research Directions

While the previous decade has brought great progress in the representational power of modern robotic mapping and localization systems, a number of fundamental issues remain open. In this section, we highlight three key avenues for future work: novel representations for sensor fusion, hierarchical abstractions for learning and attention, and the problem of identifying suitable concepts for grounding semantic models.

3.6.1. Novel representations for sensor fusion. While much progress has been made in recent years on the topic of semantic SLAM, most representations rely on the availability of high-signal-to-noise-ratio sensors that provide an abundance of accurate geometric measurements, such as RGB-D cameras and lidar. How can we do more with less? We would like to develop robots that process rich, multimodal sensory information from measurements that may individually only partially describe objects or scenes. Incorporating novel sensors such as event cameras [see, e.g., the recent survey by Gallego et al. (143)], light-field cameras, and tactile sensing together with more

traditional sensors such as cameras, lidar, inertial measurements, and sonar is a key area for future work. Similarly, the graphical optimization-based estimation frameworks that at present are commonly used in robotic state estimation provide a convenient and versatile language for describing sensor fusion problems, but they typically depend on having a relatively well-characterized model for each of the deployed sensors. Natural organisms, as well as machines, encounter physical changes to their sensors and configurations as time goes on: A camera may move slightly on an autonomous vehicle, lens distortion may occur, and so on. Systems that perform long-term sensor fusion must in some sense be adaptable. There is still ample room for fundamental contributions in many of these areas.

3.6.2. Hierarchical models for learning, navigation, and planning. Hierarchical, flexible models that abstract the minutiae of scene geometry in favor of higher-level concepts are needed for robust, scalable long-term mapping. However, the design of these abstractions raises several issues, including to what extent they should be learned or grounded in human-centric concepts, and how they can be structured to accommodate multiple spatial and temporal scales (to support large-scale operation).

3.6.2.1. To learn or not to learn? Thousands of years of human-made spatial-semantic abstractions have become embedded into our world. Given these prior abstractions, two natural questions arise: What should be learned from data, and what should be enforced as a prior? In particular, should a robot be explicitly programmed with human-centric abstractions, or should it attempt to learn its own abstractions through experience? The answers to these questions will naturally be task dependent, and the twin problems of developing both models of human semantics for robot use and online robot semantics (grounded in the actions a robot can perform) remain largely open.

3.6.2.2. Flexible representations and attention. What is the spatial representation for a robot that can get on a plane and fly from Boston to London? It may not matter explicitly where it is in Earth-centered coordinates while on an airplane: Only the local coordinates of the robot matter. Mixing topological constraints with metric constraints would be greatly beneficial in such situations. While this topic was initially explored by Sibley et al. (144), this is largely an open research area.

For the scalability of representations, another open question is what the robot should remember. Metric geometric models grow unboundedly in the space needed for navigation; however, it seems intuitively clear that not all metric information needs to be immediately accessible. Neither a human nor a robot needs to reason about the dense geometry of their home kitchen in order to navigate from one location to another while traveling in another country. Developing models that can pull relevant information from long-term memory into short-term memory for navigation tasks is another promising area for future work.

3.6.3. Where do semantics come from? In comparison to work using semantics as a sensor, there has been comparatively little work on semantics as they arise from the compositions of other entities in a map or from the history of a robot's interactions with (and observations of) the environment. Groupings of geometric features (point clouds, lines, and planes) or local unoccupied regions (i.e., places) can each be associated with semantic phenomena in the environment. It may be necessary to build up a semantic model of the environment, rather than capture semantics in single measurements, in order to develop more intelligent robot systems. The following passage by Marr (145, p. 36) is relevant in this context:

> Finally, one has to come to terms with cold reality. Desirable as it may be to have vision deliver a completely invariant shape description from an image (whatever that may mean in detail), it is almost certainly impossible in only one step. We can only do what is possible and proceed from there towards what is desirable. Thus we arrive at the idea of a sequence of representations, starting with descriptions that could be obtained straight from an image but that are carefully designed to facilitate the subsequent recovery of gradually more objective, physical properties about an object shape.

For a number of reasons, scenes may contain semantic information not captured directly by individual sensor measurements. For example, the geometry of an object larger in extent than can be captured by a single image may be relevant, or the semantic relevance of an object may arise through interaction. More generally, some semantic properties of interest may have no physical grounding whatsoever; a natural example of such a property is ownership, which is completely divorced from an object's physical makeup. Finally, semantics are often contextual: Different properties or categorizations of objects may be important depending on the specific task at hand. There is a great need for flexible models that can describe these aspects of the world in order to build versatile robots that can robustly perform a variety of tasks.

4. DISCUSSION AND FUTURE PERSPECTIVES

Looking forward, we see three key challenges for guiding future research in SLAM:

- Long-term autonomy: Can we improve the robustness of SLAM systems to enable the reliable, persistent, and independent operation of robots in the real world?
- Lifelong map learning: Can we create systems that continually improve their mapping performance, despite (and perhaps even leveraging) the constant evolution of real-world environments, thereby enabling persistent deployment?
- SLAM and deep learning: Can we capitalize on the promise of recent breakthroughs in deep learning (146) and new semantic representations, while retaining the desirable properties of traditional model-based state estimation methods, including recently developed robust estimation algorithms?

In our view, long-term autonomy (i.e., the capacity of a robotic system to operate reliably, for extended periods of time, without human supervision or intervention) is an important measure of performance for evaluating future research in SLAM. Current state-of-the-art methods have largely solved the problems of spatially and temporally bounded deployment; what remains to address are the myriad infrequently encountered failure cases that arise in extended, real-world operation. By nature, these may be difficult to capture in laboratory settings via small-scale experiments or in the standard data sets that have traditionally been employed for empirical SLAM evaluation. We expect that the certifiable and robust SLAM inference methods described in Section 2 will have a particularly prominent role to play in addressing such long-tail failure modes, as their explicitly delineated operational assumptions and run-time verification enable potentially hazardous circumstances to be identified before they are encountered in operation.

Persistent operation will also require robots to evolve beyond the classical snapshot version of the SLAM problem and embrace a long-term existence via lifelong learning. The environment (and perhaps features of the robot itself) will change over time, and new inferential and representational approaches are required that are similarly adaptive. In particular, an important novel challenge in this regime is the need to explicitly account for, and actively manage, the uncertainty inherent in an ever-changing world. For long-term operation, inference and representation must be combined with planning and control to enable active, task-directed perception; these capabilities will provide autonomous agents the means to introspect (i.e., to monitor their own state of

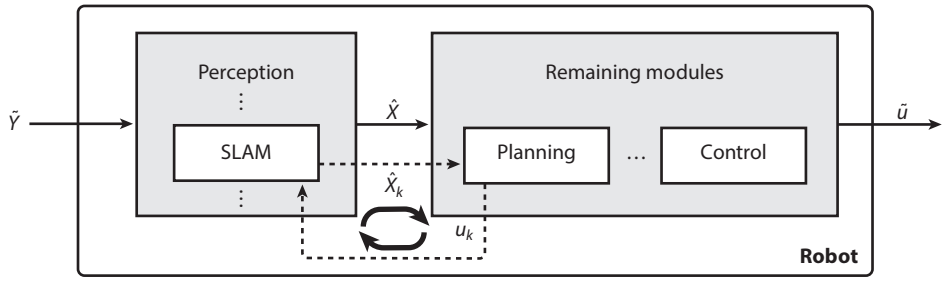

Figure 4

Block diagram of a robotic system, including perception and SLAM modules. Solid lines represent the basic flow of information through the perception subsystem: Sensor measurements $\tilde{Y}$ are received and processed, and a state estimate $\hat{X}$ is communicated to the rest of the system. Active perception (147) involves a tighter interconnection between planning and perception (*dashed loop*): Here, the planner proactively chooses its controls u_k to reduce the uncertainty in the predicted state estimate $\hat{X}_k$ that is expected after u_k is applied.

knowledge), equipping them to identify and seek out the information needed to reduce their own uncertainty (**Figure 4**). Recalling Bajcsy's (147) famous adage, intelligent agents do not just see, they look.

Regarding learning and adaptation, the integration of SLAM with deep learning specifically will be another key research area over the next decade. Indeed, end-to-end, data-driven machine learning techniques for SLAM are already starting to enter the literature (132, 136, 137, 140). There is ample opportunity to develop novel deep learning systems that are specifically adapted to the unique features of robotic perception; these include the rich, temporally coherent streams of sensory data available to robots, novel sensing modalities and data types beyond classical vision (e.g., direct 3D perception via lidar, event-based cameras, and tactile sensing), and the ability to close the loop around perception via active sensing. The community is also in need of larger and more varied data sets, tailored to the problem of SLAM, to more thoroughly investigate the potential of these approaches.

After 30 years of progress, the problem of constructing global representations from local measurements continues to inspire significant and fundamental advances in SLAM. Recent work has seen great strides in classical state estimation (Section 2), demonstrating both the theoretical and empirical effectiveness of certifiable and robust inference methods. However, these methods derive much of their power from being built atop well-characterized (typically geometric) models; such strong hypotheses may not always be realistic, especially in moving beyond short-term operation and representations grounded in simple geometric primitives. Conversely, it is the flexibility and representational power (Section 3) of learning systems that affords autonomous robots the capacity to build richer environmental models and adapt through experience. At present, these systems are often trained and deployed in a black-box end-to-end manner; this may make the learned representations difficult to interpret and hence difficult to integrate within larger-scale autonomous systems (see **Figure 4**). In the future, we envision SLAM systems that are built as a synthesis of these approaches, applying (narrowly scoped) learning in those parts of the system where it is required, but integrated within a classical model-based Bayesian state estimation framework (**Figure 1**) that enables us to take advantage of principled, highly developed (e.g., certifiable and robust) inference methods, introspection, and active perception. Such a synthesis could achieve the best of both worlds, equipping robotic agents with both the robustness and the adaptability necessary to achieve truly autonomous persistent operation.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by Office of Naval Research Multidisciplinary University Research Initiative grant N00014-19-1-2571 and Office of Naval Research grant N00014-18-1-2832.

LITERATURE CITED

1. Thrun S, Burgard W, Fox D. 2005. *Probabilistic Robotics*. Cambridge, MA: MIT Press
2. Durrant-Whyte H, Bailey T. 2006. Simultaneous localization and mapping: part I. *IEEE Robot. Autom. Mag*. 13(2):99–110
3. Bailey T, Durrant-Whyte H. 2006. Simultaneous localization and mapping (SLAM): part II. *IEEE Robot. Autom. Mag*. 13(3):108–17
4. Stachniss C, Leonard JJ, Thrun S. 2016. Simultaneous localization and mapping. In *Springer Handbook of Robotics*, ed. B Siciliano, O Khatib, pp. 1153–76. Cham, Switz.: Springer
5. Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, et al. 2016. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. *IEEE Trans. Robot*. 32:1309–32
6. Engel J, Schöps J, Cremers D. 2014. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *Computer Vision – ECCV 2014: 13th European Conference*, ed. D Fleet, T Pajdla, B Schiele, T Tuytelaars, pp. 834–49. Cham, Switz.: Springer
7. Mur-Artal R, Tardós JD. 2017. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans. Robot*. 33:1255–62
8. Whelan T, Leutenegger S, Salas-Moreno RF, Glocker B, Davison AJ. 2015. ElasticFusion: dense SLAM without a pose graph. In *Robotics: Science and Systems XI*, ed. LE Kavraki, D Hsu, J Buchli, pap. 1. N.p.: Robot. Sci. Syst. Found.
9. Dellaert F, Kaess M. 2017. Factor graphs for robot perception. *Found. Trends Robot*. 6:1–139
10. Huang S, Dissanayake G. 2016. A critique of current developments in simultaneous localization and mapping. *Int. J. Adv. Robot. Syst*. 13. **https://doi.org/10.1177/1729881416669482**
11. Sipser M. 2012. *Introduction to the Theory of Computation*. Boston: Cengage Learn.
12. Kuipers B. 2000. The spatial semantic hierarchy. *Artif. Intell*. 119:191–233
13. Koller D, Friedman N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press
14. Rosen DM, Kaess M, Leonard JJ. 2014. RISE: an incremental trust-region method for robust online sparse least-squares estimation. *IEEE Trans. Robot*. 30:1091–108
15. Lowry S, Sünderhauf N, Newman P, Leonard JJ, Cox D, et al. 2016. Visual place recognition: a survey. *IEEE Trans. Robot*. 32:1–19
16. Grisetti G, Kümmerle R, Stachniss C, Burgard W. 2010. A tutorial on graph-based SLAM. *IEEE Intell. Transp. Syst. Mag*. 2(4):31–43
17. Kaess M, Ranganathan A, Dellaert F. 2008. iSAM: incremental smoothing and mapping. *IEEE Trans. Robot*. 24:1365–78
18. Kaess M, Johannsson H, Roberts R, Ila V, Leonard JJ, Dellaert F. 2012. iSAM2: incremental smoothing and mapping using the Bayes tree. *Int. J. Robot. Res*. 31:216–35
19. Kümmerle R, Grisetti G, Strasdat H, Konolige K, Burgard W. 2011. g^2o: a general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pp. 3607–13. Piscataway, NJ: IEEE
20. Huang G. 2019. Visual-inertial navigation: a concise review. In *2019 International Conference on Robotics and Automation*, pp. 9572–82. Piscataway, NJ: IEEE
21. Huber P. 2004. *Robust Statistics*. Hoboken, NJ: Wiley

22. Ferguson T. 1996. *A Course in Large Sample Theory*. Boca Raton, FL: Chapman & Hall/CRC
23. Nocedal J, Wright S. 2006. *Numerical Optimization*. New York: Springer. 2nd ed.
24. Grisetti G, Stachniss C, Burgard W. 2009. Nonlinear constraint network optimization for efficient map learning. *IEEE Trans. Intell. Transp. Syst*. 10:428–39
25. Williams S, Indelman V, Kaess M, Roberts R, Leonard JJ, Dellaert F. 2014. Concurrent filtering and smoothing: a parallel architecture for real-time navigation and full smoothing. *Int. J. Robot. Res*. 33:1544–68
26. Rosen DM, Carlone L, Bandeira A, Leonard JJ. 2019. SE-Sync: a certifiably correct algorithm for synchronization over the special Euclidean group. *Int. J. Robot. Res*. 38:95–125
27. Rosen DM, Carlone L, Bandeira A, Leonard JJ. 2016. A certifiably correct algorithm for synchronization over the special Euclidean group. In *Algorithmic Foundations of Robotics XII: Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics*, ed. K Goldberg, P Abbeel, K Bekris, L Miller, pp. 64–79. Cham, Switz.: Springer
28. Martinec D, Pajdla T. 2007. Robust rotation and translation estimation in multiview reconstruction. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE. **http://doi.org/10.1109/CVPR.2007.383115**
29. Liu M, Huang S, Dissanayake G, Wang H. 2012. A convex optimization based approach for pose SLAM problems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1898–903. Piscataway, NJ: IEEE
30. Carlone L, Tron R, Daniilidis K, Dellaert F. 2015. Initialization techniques for 3D SLAM: a survey on rotation estimation and its use in pose graph optimization. In *2015 IEEE International Conference on Robotics and Automation*, pp. 4597–604. Piscataway, NJ: IEEE
31. Rosen DM, DuHadway C, Leonard JJ. 2015. A convex relaxation for approximate global optimization in simultaneous localization and mapping. In *2015 IEEE International Conference on Robotics and Automation*, pp. 5822–29. Piscataway, NJ: IEEE
32. Arrigoni F, Rossi B, Fusiello A. 2016. Spectral synchronization of multiple views in SE(3). *SIAM J. Imaging Sci*. 9:1963–90
33. Bandeira A, Boumal N, Singer A. 2016. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Math. Program*. 163:145–67
34. Hartley R, Trumpf J, Dai Y, Li H. 2013. Rotation averaging. *Int. J. Comput. Vis*. 103:267–305
35. Bandeira A. 2016. A note on probably certifiably correct algorithms. *C. R. Math*. 354:329–33
36. Carlone L, Dellaert F. 2015. Duality-based verification techniques for 2D SLAM. In *2015 IEEE International Conference on Robotics and Automation*, pp. 4589–96. Piscataway, NJ: IEEE
37. Carlone L, Rosen DM, Calafiore G, Leonard JJ, Dellaert F. 2015. Lagrangian duality in 3D SLAM: verification techniques and optimal solutions. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 125–32. Piscataway, NJ: IEEE
38. Carlone L, Calafiore GC, Tommolillo C, Dellaert F. 2016. Planar pose graph optimization: duality, optimal solutions, and verification. *IEEE Trans. Robot*. 32:545–65
39. Boyd S, Vandenberghe L. 2004. *Convex Optimization*. Cambridge, UK: Cambridge Univ. Press
40. Rosen DM. 2019. *Towards provably robust machine perception*. Paper presented at the RSS Pioneers Workshop, Robotics: Science and Systems XV, Freiburg im Breisgau, Ger., June 22–26
41. Lasserre J. 2010. *Moments, Positive Polynomials and Their Applications*. London: Imp. Coll. Press
42. Kahl F, Henrion D. 2007. Globally optimal estimates for geometric reconstruction problems. *Int. J. Comput. Vision* 74:3–15
43. Eriksson A, Olsson C, Kahl F, Chin TJ. 2018. Rotation averaging and strong duality. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 127–35. Piscataway, NJ: IEEE
44. Dellaert F, Rosen DM, Wu J, Mahony R, Carlone L. 2020. Shonan rotation averaging: global optimality by surfing $SO(p)^n$. In *Computer Vision – ECCV 2020: 16th European Conference*, Part VI, ed. A Vedaldi, H Bischof, T Brox, J-M Frahm, pp. 292–308. Cham, Switz.: Springer
45. Briales J, Kneip L, Gonzalez-Jimenez J. 2018. A certifiably globally optimal solution to the non-minimal relative pose problem. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 145–54. Piscataway, NJ: IEEE

46. Giamou M, Ma Z, Peretroukhin V, Kelly J. 2019. Certifiably globally optimal extrinsic calibration from per-sensor egomotion. *IEEE Robot. Autom. Lett*. 4:367–74
47. Briales J, Gonzalez-Jimenez J. 2017. Convex global 3D registration with Lagrangian duality. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5612–21. Piscataway, NJ: IEEE
48. Yang H, Shi J, Carlone L. 2020. TEASER: fast and certifiable point cloud registration. *IEEE Trans. Robot*. In press. **https://doi.org/10.1109/TRO.2020.3033695**
49. Hu S, Carlone L. 2019. Accelerated inference in Markov random fields via smooth Riemannian optimization. *IEEE Robot. Autom. Lett*. 4:1295–302
50. Yang H, Carlone L. 2020. In perfect shape: certifiably optimal 3D shape reconstruction from 2D landmarks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 618–27. Piscataway, NJ: IEEE
51. Briales J, Gonzalez-Jimenez J. 2017. Cartan-Sync: fast and global SE(*d*)-synchronization. *IEEE Robot. Autom. Lett*. 2:2127–34
52. Fan T, Wang H, Rubenstein M, Murphey T. 2019. Efficient and guaranteed planar pose graph optimization using the complex number representation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1904–11. Piscataway, NJ: IEEE
53. Mangelson J, Liu J, Eustice R, Vasudevan R. 2019. Guaranteed globally optimal planar pose graph and landmark SLAM via sparse-bounded sums-of-squares programming. In *2019 International Conference on Robotics and Automation*, pp. 9306–12. Piscataway, NJ: IEEE
54. Fan T, Murphey T. 2019. *Generalized proximal methods for pose-graph optimization*. Paper presented at the 19th International Symposium of Robotics Research, Hanoi, Vietnam, Oct. 6–10
55. Tian Y, Khosoussi K, How J. 2019. Block-coordinate minimization for large SDPs with block-diagonal constraints. arXiv:1903.00597 [math.OC]
56. Tian Y, Khosoussi K, Rosen DM, How JP. 2020. Distributed certifiably correct pose-graph optimization. arXiv:1911.03721 [math.OC]
57. Fischler M, Bolles R. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24:381–95
58. Latif Y, Cadena C, Neira J. 2013. Robust loop closing over time for pose graph SLAM. *Int. J. Robot. Res*. 32:1611–26
59. Mangelson J, Dominic D, Eustice R, Vasudevan R. 2018. Pairwise consistent measurement set maximization for robust multi-robot map merging. In *2018 IEEE International Conference on Robotics and Automation*, pp. 2916–23. Piscataway, NJ: IEEE
60. Huber P. 1964. Robust estimation of a location parameter. *Ann. Math. Stat*. 35:73–101
61. Sünderhauf N, Protzel P. 2012. Switchable constraints for robust pose graph SLAM. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1879–84. Piscataway, NJ: IEEE
62. Agarwal P, Tipaldi G, Spinello L, Stachniss C, Burgard W. 2014. Dynamic covariance scaling for robust map optimization. In *2013 IEEE International Conference on Robotics and Automation*, pp. 62–69. Piscataway, NJ: IEEE
63. Olson E, Agarwal P. 2013. Inference on networks of mixtures for robust robot mapping. *Int. J. Robot. Res*. 32:826–40
64. Yang H, Carlone L. 2020. One ring to rule them all: certifiably robust geometric perception with outliers. In *Advances in Neural Information Processing 33*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H Lin, pp. 18846–59. Red Hook, NY: Curran
65. Yang H, Antonante P, Tzoumas V, Carlone L. 2020. Graduated non-convexity for robust spatial perception: from non-minimal solvers to global outlier rejection. *IEEE Robot. Autom. Lett*. 5:1127–34
66. Chung F. 1997. *Spectral Graph Theory*. Providence, RI: Am. Math. Soc.
67. Singer A. 2011. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal*. 30:20–36
68. Boumal N, Singer A, Absil PA, Blondel V. 2014. Cramér-Rao bounds for synchronization of rotations. *Inform. Inference* 3:1–39
69. Khosoussi K, Giamou M, Sukhatme G, Huang S, Dissanayake G, How J. 2019. Reliable graphs for SLAM. *Int. J. Robot. Res*. 38:260–98

70. Majumdar A, Hall G, Ahmadi A. 2020. Recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics. *Annu. Rev. Control Robot. Auton. Syst.* 3:331–60
71. Rosen DM. 2020. *Scalable low-rank semidefinite programming for certifiably correct machine perception*. Paper presented at the 14th International Workshop on the Algorithmic Foundations of Robotics, Oulu, Finl.
72. Cifuentes D, Harris C, Sturmfels B. 2020. The geometry of SDP-exactness in quadratic optimization. *Math. Program.* 182:399–428
73. Fourie D, Leonard JJ, Kaess M. 2016. A nonparametric belief solution to the Bayes tree. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2189–96. Piscataway, NJ: IEEE
74. Hsiao M, Kaess M. 2019. MH-iSAM2: multi-hypothesis iSAM using Bayes tree and hypo-tree. In *2019 International Conference on Robotics and Automation*, pp. 1274–80. Piscataway, NJ: IEEE
75. Bernreiter L, Gawel A, Sommer H, Nieto J, Siegwart R, Lerma CC. 2019. Multiple hypothesis semantic mapping for robust data association. *IEEE Robot. Autom. Lett.* 4:3255–62
76. Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, ed. F Pereira, CJC Burges, L Bottou, KQ Weinberger, pp. 1097–105. Red Hook, NY: Curran
77. Redmon J, Divvala S, Girshick R, Farhadi A. 2016. You only look once: unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–88. Piscataway, NJ: IEEE
78. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–55. Piscataway, NJ: IEEE
79. Sünderhauf N, Brock O, Scheirer W, Hadsell R, Fox D, et al. 2018. The limits and potentials of deep learning for robotics. *Int. J. Robot. Res.* 37:405–20
80. Davison AJ. 2018. FutureMapping: the computational structure of spatial AI systems. arXiv:1803.11288 [cs.AI]
81. Davison AJ, Ortiz J. 2019. FutureMapping 2: Gaussian belief propagation for spatial AI. arXiv:1910.14139 [cs.AI]
82. Nicholson O. 2020. SLAMcore debuts full-stack spatial AI SDK in industry competition. *SLAMcore Blog*, Apr. 30. **https://blog.slamcore.com/sdk-debut**
83. Bohg J, Hausman K, Sankaran B, Brock O, Kragic D, et al. 2017. Interactive perception: leveraging action in perception and perception in action. *IEEE Trans. Robot.* 33:1273–91
84. Lowe DG. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60:91–110
85. Bay H, Ess A, Tuytelaars T, Van Gool L. 2008. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* 110:346–59
86. Rublee E, Rabaud V, Konolige K, Bradski G. 2011. ORB: an efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pp. 2564–71. Piscataway, NJ: IEEE
87. Mur-Artal R, Montiel JMM, Tardós JD. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans. Robot.* 31:1147–63
88. Engel J, Koltun V, Cremers D. 2017. Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.* 40:611–25
89. Wang R, Schworer M, Cremers D. 2017. Stereo DSO: large-scale direct sparse visual odometry with stereo cameras. In *2017 IEEE International Conference on Computer Vision*, pp. 3923–31. Piscataway, NJ: IEEE
90. Gao X, Wang R, Demmel N, Cremers D. 2018. LDSO: Direct Sparse Odometry with loop closure. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2198–204. Piscataway, NJ: IEEE
91. Moravec H, Elfes A. 1985. High resolution maps from wide angle sonar. In *1985 IEEE International Conference on Robotics and Automation*, Vol. 2, pp. 116–121. Piscataway, NJ: IEEE
92. Hornung A, Wurm KM, Bennewitz M, Stachniss C, Burgard W. 2013. OctoMap: an efficient probabilistic 3D mapping framework based on octrees. *Auton. Robots* 34:189–206

93. Rosinol A, Abate M, Chang Y, Carlone L. 2020. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE International Conference on Robotics and Automation*, pp. 1689–96. Piscataway, NJ: IEEE
94. Newcombe RA, Davison AJ, Izadi S, Kohli P, Hilliges O, et al. 2011. KinectFusion: real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pp. 127–36. Piscataway, NJ: IEEE
95. Whelan T, McDonald JB, Kaess M, Fallon MF, Johannsson H, Leonard JJ. 2012. *Kintinuous: spatially extended KinectFusion*. Paper presented at RGB-D: Advanced Reasoning with Depth Cameras, Robotics: Science and Systems VIII, Sydney, July 9–13
96. Sünderhauf N, Dayoub F, McMahon S, Eich M, Upcroft B, Milford M. 2015. *SLAM – quo vadis? In support of object oriented and semantic SLAM*. Paper presented at The Problem of Mobile Sensors: Setting Future Goals and Indicators of Progress for SLAM, Robotics: Science and Systems XI, Rome, July 13–17
97. Moravec H, Blackwell M. 1992. Learning sensor models for evidence grids. In *1991 Annual Research Review*, pp. 8–15. Pittsburgh, PA: Carnegie Mellon Univ. Robot. Inst.
98. Bowman SL, Atanasov N, Daniilidis K, Pappas GJ. 2017. Probabilistic data association for semantic SLAM. In *2017 IEEE International Conference on Robotics and Automation*, pp. 1722–29. Piscataway, NJ: IEEE
99. Atanasov N, Zhu M, Daniilidis K, Pappas GJ. 2014. Semantic localization via the matrix permanent. In *Robotics: Science and Systems X*, ed. D Fox, LE Kavraki, H Kurniawati, pap. 43. N.p.: Robot. Sci. Syst. Found.
100. Doherty KJ, Fourie D, Leonard JJ. 2019. Multimodal semantic SLAM with probabilistic data association. In *2019 International Conference on Robotics and Automation*, pp. 2419–25. Piscataway, NJ: IEEE
101. Doherty KJ, Baxter D, Schneeweiss E, Leonard JJ. 2020. Probabilistic data association via mixture models for robust semantic SLAM. In *2020 IEEE International Conference on Robotics and Automation*, pp. 1098–104. Piscataway, NJ: IEEE
102. McCormac J, Handa A, Davison A, Leutenegger S. 2017. SemanticFusion: dense 3D semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation*, pp. 4628–35. Piscataway, NJ: IEEE
103. Sengupta S, Sturgess P. 2015. Semantic octree: unifying recognition, reconstruction and representation via an octree constrained higher order MRF. In *2015 IEEE International Conference on Robotics and Automation*, pp. 1874–79. Piscataway, NJ: IEEE
104. Salas-Moreno RF, Newcombe RA, Strasdat H, Kelly PH, Davison AJ. 2013. SLAM++: simultaneous localisation and mapping at the level of objects. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1352–59. Piscataway, NJ: IEEE
105. Civera J, Gálvez-López D, Riazuelo L, Tardós JD, Montiel J. 2011. Towards semantic SLAM using a monocular camera. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1277–84. Piscataway, NJ: IEEE
106. Castle RO, Gawley DJ, Klein G, Murray DW. 2007. Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras. In *2007 IEEE International Conference on Robotics and Automation*, pp. 4102–7. Piscataway, NJ: IEEE
107. Nicholson L, Milford M, Sünderhauf N. 2018. QuadricSLAM: dual quadrics from object detections as landmarks in object-oriented SLAM. *IEEE Robot. Autom. Lett.* 4:1–8
108. Sünderhauf N, Milford M. 2017. Dual quadrics from object detection bounding boxes as landmark representations in SLAM. arXiv:1708.00965 [cs.RO]
109. Ok K, Liu K, Frey K, How JP, Roy N. 2019. Robust object-based SLAM for high-speed autonomous navigation. In *2019 International Conference on Robotics and Automation*, pp. 669–75. Piscataway, NJ: IEEE
110. Yang S, Scherer S. 2019. CubeSLAM: monocular 3-D object SLAM. *IEEE Trans. Robot.* 35:925–38
111. Sucar E, Wada K, Davison A. 2020. Neural object descriptors for multi-view shape reconstruction. arXiv:2004.04485 [cs.CV]
112. Bescos B, Fácil JM, Civera J, Neira J. 2018. DynaSLAM: tracking, mapping, and inpainting in dynamic scenes. *IEEE Robot. Autom. Lett.* 3:4076–83
113. Wang CC, Thorpe C, Thrun S, Hebert M, Durrant-Whyte H. 2007. Simultaneous localization, mapping and moving object tracking. *Int. J. Robot. Res.* 26:889–916

114. Schorghuber M, Steininger D, Cabon Y, Humenberger M, Gelautz M. 2019. SLAMANTIC - leveraging semantics to improve VSLAM in dynamic environments. In *2019 IEEE/CVF International Conference on Computer Vision Workshop*, pp. 3759–68. Piscataway, NJ: IEEE
115. Zhang J, Henein M, Mahony R, Ila V. 2020. VDO-SLAM: a visual dynamic object-aware SLAM system. arXiv:2005.11052 [cs.RO]
116. Henein M, Zhang J, Mahony R, Ila V. 2020. Dynamic SLAM: the need for speed. arXiv:2002.08584 [cs.RO]
117. Rosen DM, Mason J, Leonard JJ. 2016. Towards lifelong feature-based mapping in semi-static environments. In *2016 IEEE International Conference on Robotics and Automation*, pp. 1063–70. Piscataway, NJ: IEEE
118. Krajník T, Fentanes JP, Santos JM, Duckett T. 2017. Fremen: frequency map enhancement for long-term mobile robot autonomy in changing environments. *IEEE Trans. Robot.* 33:964–77
119. Bore N, Ekekrantz J, Jensfelt P, Folkesson J. 2018. Detection and tracking of general movable objects in large three-dimensional maps. *IEEE Trans. Robot.* 35:231–47
120. Zeng Z, Röfer A, Jenkins OC. 2020. Semantic linking maps for active visual object search. In *2020 IEEE International Conference on Robotics and Automation*, pp. 1984–90. Piscataway, NJ: IEEE
121. Halodová L, Dvořráková E, Majer F, Vintr T, Mozos OM, et al. 2019. Predictive and adaptive maps for long-term visual navigation in changing environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 7033–39. Piscataway, NJ: IEEE
122. Berrio JS, Ward J, Worrall S, Nebot E. 2019. Updating the visibility of a feature-based map for long-term maintenance. In *2019 IEEE Intelligent Vehicles Symposium*, pp. 1173–79. Piscataway, NJ: IEEE
123. Pannen D, Liebner M, Hempel W, Burgard W. 2020. How to keep HD maps for automated driving up to date. In *2020 IEEE International Conference on Robotics and Automation*, pp. 2288–94. Piscataway, NJ: IEEE
124. Newcombe RA, Fox D, Seitz SM. 2015. DynamicFusion: reconstruction and tracking of non-rigid scenes in real-time. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 343–52. Piscataway, NJ: IEEE
125. Song J, Zhao L, Huang S, Dissanayake G. 2019. An observable time series based SLAM algorithm for deforming environment. arXiv:1906.08563 [cs.RO]
126. Mu B, Giamou M, Paull L, Agha-Mohammadi AA, Leonard JJ, How J. 2016. Information-based active SLAM via topological feature graphs. In *2016 IEEE 55th Conference on Decision and Control*, pp. 5583–90. Piscataway, NJ: IEEE
127. Stein GJ, Bradley C, Preston V, Roy N. 2020. Enabling topological planning with monocular vision. arXiv:2003.14368 [cs.RO]
128. Chaplot DS, Salakhutdinov R, Gupta A, Gupta S. 2020. Neural topological SLAM for visual navigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12872–881. Piscataway, NJ: IEEE
129. Rosinol A, Gupta A, Abate M, Shi J, Carlone L. 2020. 3D dynamic scene graphs: actionable spatial perception with places, objects, and humans. In *Robotics: Science and Systems XVI*, ed. M Toussaint, A Bicchi, T Hermans, pap. 79. N.p.: Robot. Sci. Syst. Found.
130. Armeni I, He ZY, Gwak J, Zamir AR, Fischer M, et al. 2019. 3D scene graph: a structure for unified semantics, 3D space, and camera. In *2019 IEEE/CVF International Conference on Computer Vision*, pp. 5663–72. Piscataway, NJ: IEEE
131. Bear DM, Fan C, Mrowca D, Li Y, Alter S, et al. 2020. Learning physical graph representations from visual scenes. arXiv:2006.12373 [cs.CV]
132. DeTone D, Malisiewicz T, Rabinovich A. 2017. Toward geometric deep SLAM. arXiv:1707.07410 [cs.CV]
133. DeTone D, Malisiewicz T, Rabinovich A. 2018. Self-improving visual odometry. arXiv:1812.03245 [cs.CV]
134. Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, et al. 2016. End to end learning for self-driving cars. arXiv:1604.07316 [cs.CV]
135. Pillai S, Leonard JJ. 2015. Monocular SLAM supported object recognition. In *Robotics: Science and Systems XI*, ed. LE Kavraki, D Hsu, J Buchli, pap. 34. N.p.: Robot. Sci. Syst. Found.

136. Pillai S, Leonard JJ. 2017. Towards visual ego-motion learning in robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5533–40. Piscataway, NJ: IEEE
137. Pillai S, Leonard JJ. 2017. *Self-supervised place recognition in mobile robots*. Paper presented at the Learning for Localization and Mapping Workshop, IEEE/RSJ International Conference on Intelligent Robots and Systems, Vancouver, Can., Sept. 24–28
138. Zhi S, Bloesch M, Leutenegger S, Davison AJ. 2019. SceneCode: monocular dense semantic reconstruction using learned encoded scene representations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11768–77. Piscataway, NJ: IEEE
139. Czarnowski J, Laidlow T, Clark R, Davison AJ. 2020. DeepFactors: real-time probabilistic dense monocular SLAM. *IEEE Robot. Autom. Lett.* 5:721–28
140. Jatavallabhula KM, Iyer G, Paull L. 2019. gradSLAM: dense SLAM meets automatic differentiation. arXiv:1910.10672 [cs.RO]
141. Zhang J, Tai L, Boedecker J, Burgard W, Liu M. 2017. Neural SLAM: learning to explore with external memory. arXiv:1706.09520 [cs.LG]
142. Mirowski P, Grimes M, Malinowski M, Hermann KM, Anderson K, et al. 2018. Learning to navigate in cities without a map. In *Advances in Neural Information Processing Systems 31*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, pp. 2419–30. Red Hook, NY: Curran
143. Gallego G, Delbruck T, Orchard G, Bartolozzi C, Taba B, et al. 2019. Event-based vision: a survey. arXiv:1904.08405 [cs.CV]
144. Sibley G, Mei C, Reid I, Newman P. 2010. Planes, trains and automobiles—autonomy for the modern robot. In *2010 IEEE International Conference on Robotics and Automation*, pp. 285–92. Piscataway, NJ: IEEE
145. Marr D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman
146. Assoc. Comput. Mach. 2019. *Fathers of the deep learning revolution receive ACM A.M. Turing Award*. Press Release, Mar. 27, Assoc. Comput. Mach., New York. **https://www.acm.org/media-center/2019/march/turing-award-2018**
147. Bajcsy R. 1988. Active perception. *Proc. IEEE* 76:966–1005