Research Article

INTERNATIONAL JOURNAL OF
ADVANCED ROBOTIC SYSTEMS

International Journal of Advanced Robotic Systems
May-June 2020: 1–17
© The Author(s) 2020
DOI: 10.1177/1729881420919185
journals.sagepub.com/home/arx
SAGE

A survey of image semantics-based visual simultaneous localization and mapping: Application-oriented solutions to autonomous navigation of mobile robots

Linlin Xia[1], Jiashuo Cui[1], Ran Shen[1], Xun Xu[2], Yiping Gao[1] and Xinying Li[1]

Abstract

As one of the typical application-oriented solutions to robot autonomous navigation, visual simultaneous localization and mapping is essentially restricted to simplex environmental understanding based on geometric features of images. By contrast, the semantic simultaneous localization and mapping that is characterized by high-level environmental perception has apparently opened the door to apply image semantics to efficiently estimate poses, detect loop closures, build 3D maps, and so on. This article presents a detailed review of recent advances in semantic simultaneous localization and mapping, which mainly covers the treatments in terms of perception, robustness, and accuracy. Specifically, the concept of "semantic extractor" and the framework of "modern visual simultaneous localization and mapping" are initially presented. As the challenges associated with perception, robustness, and accuracy are being stated, we further discuss some open problems from a macroscopic view and attempt to find answers. We argue that multiscaled map representation, object simultaneous localization and mapping system, and deep neural network-based simultaneous localization and mapping pipeline design could be effective solutions to image semantics-fused visual simultaneous localization and mapping.

Keywords

Semantic SLAM, visual SLAM, autonomous navigation, mobile robot, robustness, environmental perception

Date received: 5 October 2019; accepted: 23 February 2020

Topic: Robot Sensors and Sensor Networks
Topic Editor: Henry Leung
Associate Editor: Chunbo Xiu

Introduction

Autonomous robots are capable of performing specific tasks independently without any human interventions. As one of the principal attributes of autonomous robots, autonomous motion depends largely upon accurate ego-motion estimation and high-level surrounding environment perception. However, in cases where the artificial landmarks are unknown or the robots themselves are in GPS-denied environments, estimating ego-motion or perceiving scenes encounter great difficulties.

[1] School of Automation Engineering, Northeast Electric Power University, Jilin, China
[2] Institute for Superconducting and Electronic Materials, University of Wollongong, Wollongong, Australia

Corresponding author:
Linlin Xia, School of Automation Engineering, Northeast Electric Power University, Jilin 132012, China.
Email: xiall521@neepu.edu.cn

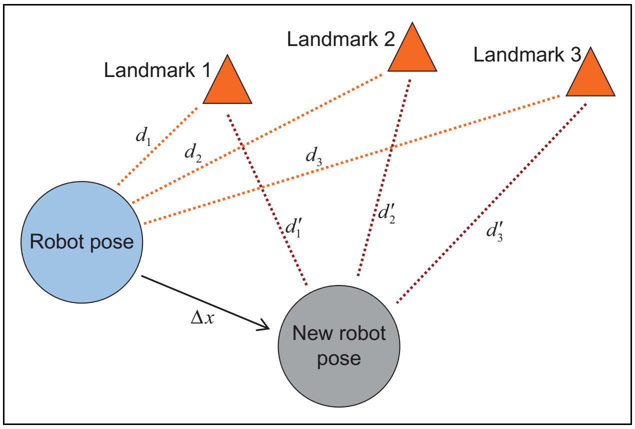

Figure 1. Diagrammatic representation of SLAM problem. SLAM: simultaneous localization and mapping.

The term "SLAM" stands for simultaneous localization and mapping (proposed by Smith and Cheeseman[1] in 1986), being recognized as an eminent tool for mobile robot ego-localization at an unknown location within an unknown environment.[2] Technically, the mobile robot incrementally builds a globally consistent map of concerned environment while simultaneously determines its location within this map. From the point view of mathematics, SLAM process can be abstracted as a concurrency estimation problem, which mainly covers the robot pose estimation and location estimation of available landmarks. The diagrammatic representation of SLAM problem is shown in Figure 1. For a long time, SLAM problem is basically solved via a series of range sensors,[3] like light detection and ranging, infrared radiation, or sound navigation and ranging within small-scale static environments (forms of range sensors conform to their individual physical principles). However, range sensor-based SLAM may have to face major challenges in dynamic, complex, and large-scale environments.

The SLAM that is implemented by means of external cameras (as the only external sensors) is termed as visual SLAM (V-SLAM). The significant advantage of V-SLAM over other typical SLAM frameworks (like range sensor-based SLAM) is its adaptability to the practical applications owing to richer image textures and simpler sensor configurations. Moreover, the development and maturation of computer vision (CV) allow V-SLAM to have access to graphical and visual supports. It is important to appreciate that solutions by CV have addressed some major difficulties in V-SLAM areas, such as detection, description and matching of image features, loop closure detection and 3D map reconstruction, and so on. Currently, with many open-source algorithms, the architecture of a V-SLAM system has been well-established. However, we must admit V-SLAM is vulnerable when either the motion of the robot or the environment is too challenging (e.g. fast robot dynamics, highly variable environments, severe illumination variations, highly limited visibility, or complex texture-less scenes).

Cadena et al.[4] firstly divided the timeline of SLAM into three periods and further summarized the individual achievements, as shown in Figure 2. Technically, they state that we are now entering the third stage of SLAM, videlicet, a stage of robust perception: the realization of robust performance, high-level understanding, resource awareness, and task-driven perception represent the themes in this age. The researchers of SLAM have worked on methods for solving high-level perception and understanding. Their efforts have been directed at semantics owing to their superiorities in aspects including improved robustness, intuitive visualization, and efficient human–robot–environment interaction. The studies that are associated with either semantic-based robustness/accuracy enhancements or semantically mapping are termed semantic SLAM. As V-SLAM could perform localization and mapping within a joint formulation, naturally, the above two processes of semantic SLAM could also be simultaneously evaluated by one estimator.

Table 1 lists the main surveys on SLAM from 2006 to present. As indicated, there have been few review articles that cover semantic SLAM (only Cadena et al.[4] mention the semantic concept-based mapping). Along the principal line of SLAM evolution, we attempt to conduct a broad review on current semantic SLAM area and to further illustrate some open problems and our insights into future research.

The outline of the remainder of this survey is as follows. The second section primarily presents a detailed description of semantic extractors, fundamental architecture of a modern V-SLAM system, and mainstream open-source algorithms. Special attention is then paid to the distinguished natures of a semantic SLAM. The perception, robustness, and accuracy problems that are, respectively, related to human–robot–environment interaction, environment adaptation, and reliable navigation are elaborated in paralleled third, fourth, and fifth sections. The sixth section focuses on the challenge discussions about semantic SLAM, seeking answers to these essential concerns. The seventh section draws conclusions.

The components of a semantic SLAM system

A semantic SLAM system is constructed of two essential components: a semantic extractor and a modern V-SLAM framework. Specifically, the semantic information is mainly extracted and derived from two processes. They are object detection and semantic segmentation.

Semantic extractor

Object detection is characterized by lightweight applicability, which not only can be applied to classify objects on the so-called object-level but also can be used to determine 2D

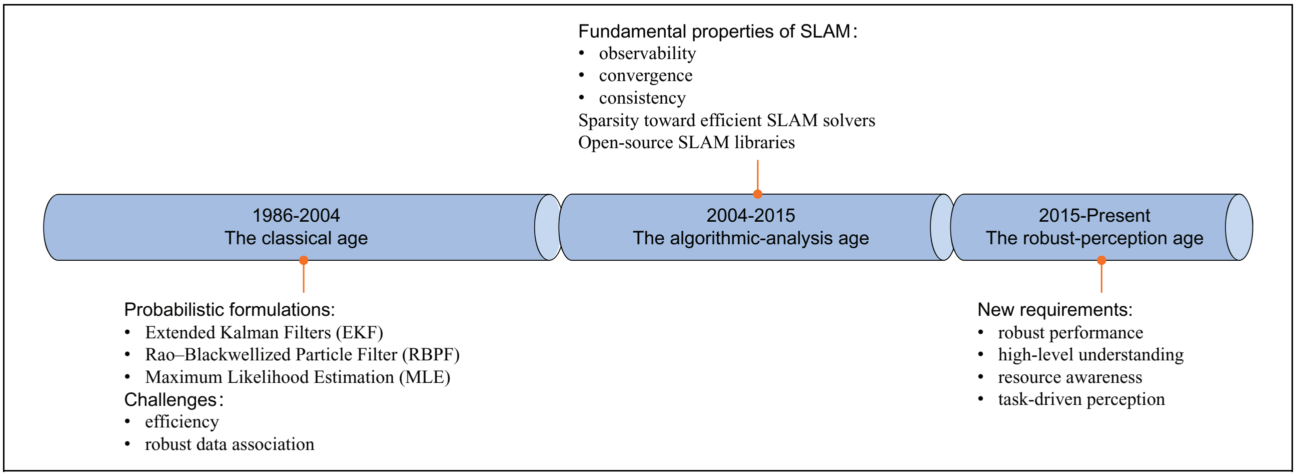
Fundamental properties of SLAM:
• observability
• convergence
• consistency
Sparsity toward efficient SLAM solvers
Open-source SLAM libraries
1986-2004
The classical age
2004-2015
The algorithmic-analysis age
2015-Present
The robust-perception age
Probabilistic formulations:
• Extended Kalman Filters (EKF)
• Rao–Blackwellized Particle Filter (RBPF)
• Maximum Likelihood Estimation (MLE)
Challenges:
• efficiency
• robust data association
New requirements:
• robust performance
• high-level understanding
• resource awareness
• task-driven perception

Figure 2. The development of SLAM. SLAM: simultaneous localization and mapping.

Table 1. Summary of SLAM-related review articles.

Year	Topic	Reference
2006	Probabilistic approaches and data association	Durrant-Whyte and Bailey[5,6]
2008	Filter-based SLAM	Aulinas et al.[7]
2008	Visual SLAM	Neira et al.[8]
2010	Graph-based SLAM	Grisetti et al.[9]
2011	Examining and evaluating SLAM	Dissanayake et al.[10]
2011	Visual odometry	Scaramuzza et al.[11]
2012	BA	Strasdat et al.[12]
2015	Visual place recognition	Lowry and Sünderhauf[13]
2016	Multiple robot SLAM	Saeedi et al.[14]
2016	Fundamental properties	Huang and Dissanayake[15]
2016	Robust perception SLAM	Cadena et al.[4]
2017	Feature based, direct, and RGB-D SLAM	Taketomi et al.[16]
2017	Keyframe-based SLAM	Younes et al.[17]
2018	Dynamic SLAM	Saputra et al.[18]
2019	Event-based SLAM	Gallego et al.[19]

BA: bundle adjustment; SLAM: simultaneous localization and mapping.

positions of concerned objects. By contrast, semantic segmentation leads to pixel-level classification acquisition, that is, all pixels in an individual image have their own unique categories. Apparently, the latter exhibits more favorable precision owing to accurate boundaries. A section-by-section description follows.

Object detection. Object detection is recognized as an important branch of CV, whose development can be roughly divided into handcraft feature-based machine learning stage (2001–2013) and learning feature-based deep learning stage (2013 to present). The former is extremely dependent on handcraft features of images.[20–24] In fact, during that period, researchers were devoted to strength the representations of handcraft features by means of more diversified descriptor design. Moreover, due to the limited computational resources, they had to explore more efficient and practical calculation approaches. In spite of their struggle to balance the handcraft feature representations and calculation efficiency, object detection experiences unexpectedly complex design with poor robustness.

In recent years, due to the introduction of deep learning and graphics processing unit, object detection with high accuracy has made great progress in either theory or practice. Especially, deep neural network (DNN)-fused object detection has arrived at a preferred robustness and accuracy, whose pipeline can be approximately designed following the two stages below:

- Stage 1: To obtain 2D positions of objects.
- Stage 2: To classify objects.

Region convolutional neural network (R-CNN) series belong to typical two-stage networks, including R-CNN,[25] fast R-CNN,[26] faster R-CNN,[27] and the newest mask R-CNN.[28] R-CNN is not only the pioneering work of R-CNN series network, but also the earliest method adopted in CNN-based object detection tasks. In principle,

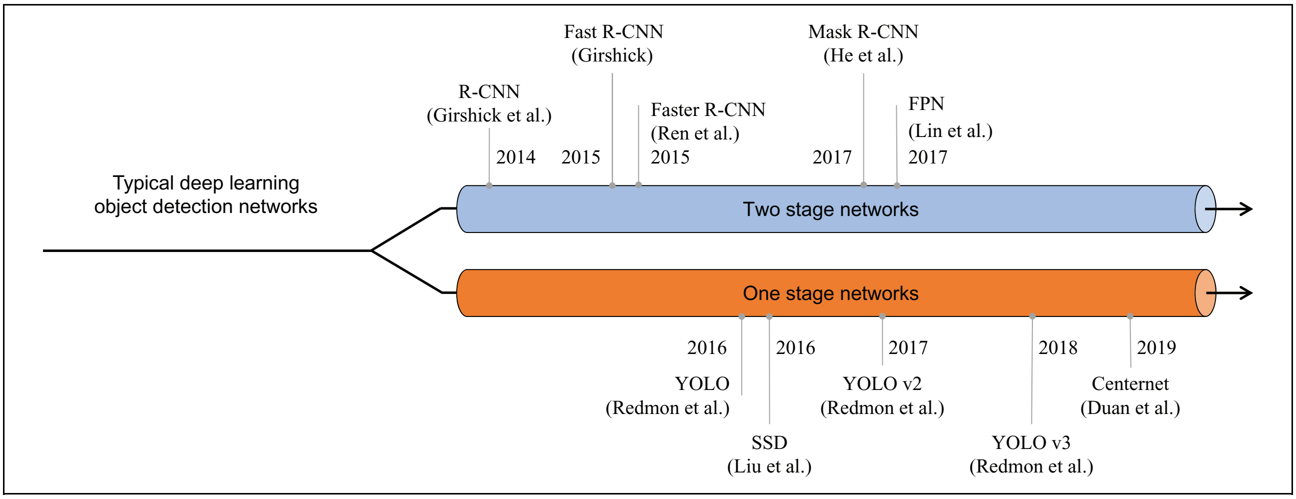
Typical deep learning object detection networks
Two stage networks
One stage networks
R-CNN (Girshick et al.) 2014
Fast R-CNN (Girshick) 2015
Faster R-CNN (Ren et al.) 2015
Mask R-CNN (He et al.) 2017
FPN (Lin et al.) 2017
2016 YOLO (Redmon et al.)
2016 SSD (Liu et al.)
2017 YOLO v2 (Redmon et al.)
2018 YOLO v3 (Redmon et al.)
2019 Centernet (Duan et al.)

Figure 3. The development of deep learning object detection networks.

R-CNN generates the region proposals via a selective search,[29] and the feature extraction and classification are, respectively, achieved via AlexNet[30] and support vector machines (SVMs).[31] Differing from which, fast R-CNN changes the order of generating region proposals and extracting image features and replaces SVMs with softmax. Faster R-CNN benefits from the generated object proposals for detection speed promoting via region proposal network, supplementary anchor, and sharing features. Quite clear, faster R-CNN would be faster, but it is still not fast enough for real-time SLAM tasks. By contrast, mask R-CNN sacrifices partial detection speed for more precise semantic segmentation purposes. As a consequence of which, it arrives at an instance-level result, that is, all pixels in each detected object have their own unique categories.

It is noteworthy that the latest type of object detection algorithms fulfills positioning and classification of objects simultaneously rather than deduce 2D positions of objects first. The representative Yolo series[32–34] (known as the most fast semantic extractor) employs $S \times S$ grids to replace region proposals, and the classification of these grids is consequently an ideal candidate for the final detection. Generally speaking, speed of Yolo series can be accepted by a real-time semantic SLAM system, but for higher accuracy, latest Centernet[35] provides a novel keypoint-based method. To clearly describe the development of object detection networks, a chronological overview is illustrated in Figure 3.

Semantic segmentation. In cases where the scenes with fantastic complexity are concerned, some care should be needed, and for guaranteed robust localization and mapping, the fine scene inference, videlicet, the deep association mining between numerous objects should be further considered. In comparison, object detection is suitable for coarse scene inferences,[36] and semantic segmentation is more general in that it applies to fine scene treatments. Analogously, the evolution of semantic segmentation has experienced "machine learning-based" to "deep learning-based" transform. Nowadays, the introduction of CNN has greatly upgraded the level of accuracy and efficiency for segmentation; thus, for cases where semantic SLAM systems are constructed, CNN-based solutions are generally to be preferred to the others.

Considering the practical applications of semantic segmentation in semantic SLAM systems, two things associated with networks (for semantic segmentation purposes) should be investigated. One is technical index (including accuracy and efficiency), one is applying condition (representing whether a network is valid for video segmentation or 3D image segmentation). The section is devoted to a description of deep learning-based semantic segmentation networks, mostly following the above lines of thought. The comparative performances of typical CNNs for semantic segmentation are listed in Table 2.

In general, almost all the deep learning-based networks for semantic segmentation inherit the model from fully convolutional network (FCN) (being recognized as landmark work by Long et al.[37]). As its name suggests, the authors modified all the most popular networks for classification (AlexNet, VGG-16, GoogleNet) to form the matched FCNs, so as to allow dense segmentation from arbitrary-sized image inputs. In addition, the encoding of CNNs enables the generations of different fine-grained semantic segmentation maps, and as the maps fuse in a skipping-connection-structure, a desired semantic segmentation result is achieved. However, FCNs themselves are not actually valid for both technical index and applying condition that a semantic SLAM requires (see Table 2 for reasons). The "SegNet" which is more concerned with decoding process appears available, so that a convolutional encoder–decoder structure is applied instead.[38] The contribution of DeepLab series networks[39–42] (including DeepLab-v1, DeepLab-v2, DeepLab-v3, DeepLab-v3+) consists in that they fully integrate the information of an

Table 2. The comparative performances of typical CNNs for semantic segmentation.

Name and reference	Architecture	Accuracy	Efficiency	Sequences	3D	Open source	Contribution(s)
FCN[37]	VGG-16	*	*	×	×	✓	Forerunner
SegNet[38]	VGG-16 + decoder	***	**	×	×	✓	Encoder–decoder
DeepLab series[39–42]	VGG-16/ResNet-101	****	*	×	×	✓	CRF, atrous convolution
Enet[43]	Enet bottleneck	**	***	×	×	✓	Bottleeneck module
PointNet[44]	Own MLP-based	**	*	×	✓	✓	3D CNN
Clockwork Convnet[45]	FCN	**	**	✓	×	✓	Clockwork scheduling

FCN: fully convolutional network; MLP: multi-layer perceptron; CNN: convolutional neural network; CRF: conditional random field.
*, **, ***, **** mean the level of performance, the more * the better performance that the system exhibits.
√ and × mean whether the certain function is supported.
√: supported; ×: unsupported.

image on various scales (termed "global context of image") and that they efficiently address "ambiguous boundary" problems likely to be encountered in FCN or SegNet. Specifically, DeepLab-v1 inserts a probabilistic graphical model (like conditional random field (CRF)) into a CNN-based pipeline and further model the segmentation result as a probabilistic graph. This probabilistic graph surely considers the global context of an image (i.e. the interactions between all pixels, not adjacent pixels only are considered) and contributes to finer segmentation results, but it indispensably burdens the load of calculation. DeepLab-v1 pioneers the use of "atrous convolution" in CNN models, and it derives a wider range of receptive fields without any load of complexity. By contrast, DeepLab-v2's pioneering work in contextual information capture on various scales is the adoption of atrous spatial pyramid pooling. DeepLab-v3 and DeepLab-v3+ further make some small revisions.

We believe that Segnet and DeepLab (with no CRF) meet the technical index demands of building semantic SLAM systems. To take some specific examples, let us refer to some research.[46,47] Yu et al.[46] successfully constructed a dynamic scene-oriented SLAM system using SegNet. Li et al.[47] effectively solved the online monocular semantic SLAM construction by means of DeepLab-v2 (with no CRF). If heavy emphasis is placed upon the fine-grained semantic maps rather than upon the efficient mapping, DeepLab series networks (with CRF) are considered to be ideal tools.[48] On the contrary, if high efficiency mapping is strongly required, certain networks should be evaluated and be further applied. Enet[43] is reminiscent of specially designed network for the purposes of real-time semantic segmentation, but whose accuracy in semantic segmentation is relatively poor.

When it comes to issues of "applying conditions" of semantic segmentation processes, let us review two candidate networks: PointNet[44] and Clockwork Convnet.[45] The former is valid for direct segmenting of unstructured 3D point clouds, and the latter is concerned with time clues of a video or image sequences (image context established on the temporal scale). These two represent the leading favorable tools even though they do not seem to have significant advantages in either accuracy or efficiency. But we still hold the opinion that, with the rapid advance of computers, the relevant studies with respect to PointNet and Clockwork Convnet would be of practical significance.

Modern V-SLAM system

The architecture of a modern V-SLAM system. A modern V-SLAM typically includes:

- Sensor data acquisition: Acquiring images or a video via cameras.
- Visual odometry (VO): Preliminarily estimating the robot pose and landmark position via adjacent frames in an image sequence.
- State estimation: Globally estimating the state by means of the fused results that VO and loop closure detection provide.
- Relocalization: Relocating when tracking fails or map is reloaded.
- Loop closure detection: Determining whether the robot is located at the previous position.
- Mapping: Mapping according to the requirements of tasks.

Concerning the flow direction of sensor data and task level, a V-SLAM system generally contains two parts: the front end and the back end, whose schematic interpretation is given in Figure 4. As indicated, the VO and loop closure detection module simultaneously receive the inputs that certain sensors supply. Here, the function of VO is to provide preliminary robot pose estimation and the function of loop closure detection module is to provide scene similarity. The derived robot poses and scene similarity constitute the sources from which the robot globally optimizes the poses and landmarks and further plots the motion trajectories and environmental maps. Mathematically, the front-end task and the back-end task can be separately abstracted as "data association" problem and "state estimation" problem.

- The front end: Data association

The process that the front end tracks the same features (feature points or representative pixels) on different frames of one image sequence is referred to as "data association."

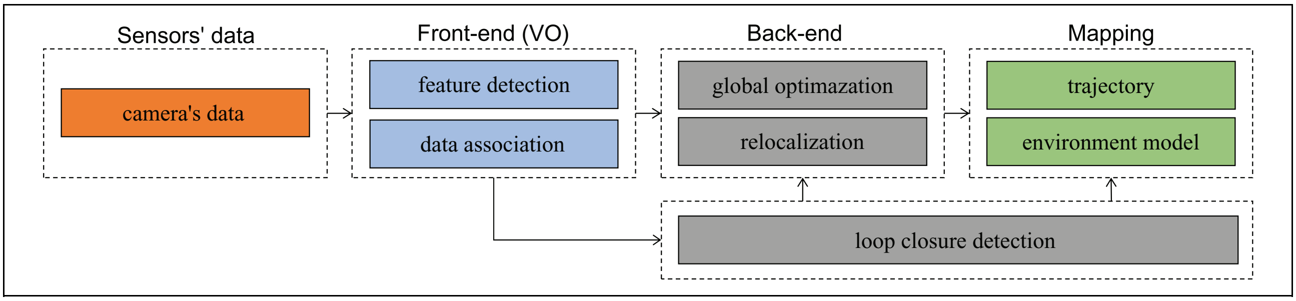

Figure 4. The architecture of a modern V-SLAM system. V-SLAM: visual simultaneous localization and mapping.

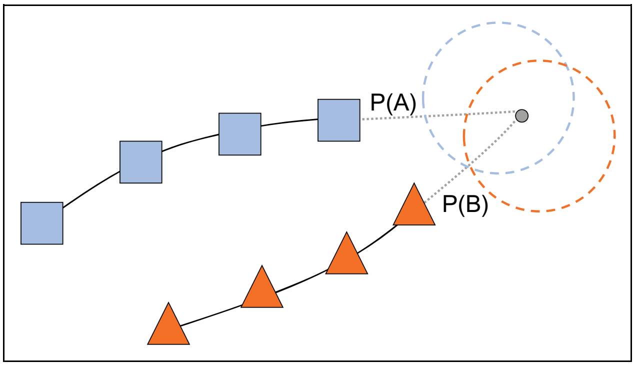

Figure 5. The diagrammatic interpretation of probabilistic data association.

Generally, early V-SLAM systems deal with "data association" via feature matching. Obviously, the insufficient description of local image features causes faulty data association with a high probability, which then leads to incorrect pose and landmark estimation. Some research that focus on eliminating the errors in data association (e.g. random sample consensus RANSAC) are proposed, but the not-yet essentially solved problems make it still unsatisfied. Later researchers begin to evaluate "data association" in probability ideas (i.e. making a soft decision to assign new features into tracking sequence). Probabilistic data association fully takes into account the uncertainty in feature assignments and minimizes erroneous associations. This is illustrated by the features in Figure 5.

Concerning the expression of data association in SLAM problems, Bowman et al.[49] were the advocates of expression $D \overset{\Delta}{=} Z\{(l_k, x_k)\}_{k=1}^{K}$, which indicates that observation Z_k (subscript k means kth) is dependent on x_k (camera pose) and l_k (landmark position). The maximum likelihood estimation (MLE) method is sequentially invited to solve for D.

- The back end: State estimation

Lu et al.[50] and Gutmann et al.[51] define SLAM as a maximum a posteriori estimation problem, which aims to estimate variable X (including robot poses and landmark positions) from a set of observations ($Z = \{z_1, ..., z_k\}$ have noises).

$$X^* = \arg\max_{X} P(X|Z) = \arg\max P(Z|X)P(X) \qquad (1)$$

Equation (1) conforms to the Bayesian theorem. Let $P(Z|X)$ denote the likelihood of the state Z (given the variable X) and let $P(X)$ denote the prior probability of variable X, so that the posterior probability can be expressed as $P(Z|X)P(X)$. The problem generalizes to determine an assignment variable X^* that maximizes $P(X|Z)$ and further determine variable X.

One of the most significant SLAM results is proposed by Davison et al.,[52] who pioneered the updating of the states of the camera and the landmark points by an extended Kalman filter (EKF). Differing from which, the representative bundle adjustment (BA)-based nonlinear optimization addressed the maximum posterior probability estimate problem by having the fused global constraints of the state variables be optimal rather than the pure iterations of EKF. By contrast, EKF-based SLAM has superior efficiency than optimization-based SLAM when dealing with small-scaled scene applications, but for the large-scale scene SLAM purposes, filter-based solutions appear insufficient superiorities due to the huge covariance matrix.

Honestly, the present V-SLAM frameworks involve a large quantity of image features, which restricts the conventional EKF-based solutions in SLAM tasks; special attention is therefore placed upon BA-based nonlinear optimization approaches. The BA ideas can be traced back to their use in the early 21st century. It is about solving structure from motion problem related to 3D reconstruction. Inspired by which, early SLAM researchers realized that BA would be probably helpful to high-precision state estimation, but they immediately found V-SLAM was actually an incremental process; the accumulated computing load made it not feasible to directly apply BA to a V-SLAM that emphasizes real-time requirements. The applicability demands of BA-based solutions were the original inspiration for the exploration of attributes of a V-SLAM; one of the major advances lies in that researchers exploited the sparsity of normal equations. They proved that the dependencies between state variables can be naturally represented in terms of a factor graph. This allows BA to have access to use a faster linear solver or an incremental solver, guaranteeing its adoption to a real-time required V-SLAM

Table 3. Open-source V-SLAM systems.

Name and reference	Year	Camera	Front end	Back end	Mapping	Relocalization	Loop closure detection
Mono-SLAM[52]	2007	Monocular	Feature based	Filter based	Sparse	×	×
PTAM[53]	2007	Monocular	Feature based	Optimization	Sparse	✓	×
KinectFusion[54]	2011	RGB-D	ICP	Optimization	Dense	×	×
Kintinuous[55]	2012	RGB-D	ICP and direct	Optimization	Dense	✓	✓
RGBD-SLAM v2[56]	2013	RGB-D	Feature based	Optimization	Dense	✓	✓
LSD-SLAM[57]	2014	Monocular	Direct	Optimization	Semi dense	✓	✓
SVO[58]	2014	Monocular	Direct	—	—	×	×
RTAB-MAP[59,60]	2014	RGB-D	Feature based	Optimization	Dense	✓	✓
ElasticFusion[61]	2016	RGB-D	ICP	Optimization	Dense	✓	✓
ORB-SLAM[62,63]	2015	All types	Feature based	Optimization	Sparse	✓	✓
DSO[64]	2017	Monocular	Direct	Optimization	Semi dense	×	×
BundleFusion[65]	2017	RGB-D	Feature based	Optimization	Dense	×	×
ProSLAM[66]	2018	Stereo	Feature based	Optimization	Sparse	✓	✓
OpenVSLAM[67]	2019	All types	Feature based	Optimization	Sparse	✓	✓

V-SLAM: visual simultaneous localization and mapping; PTAM: parallel tracking and mapping; SVO: semi-direct monocular visual odometry; ORB-SLAM: oriented FAST and rotated BRIEF SLAM; RTAB-MAP: real-time appearance-based mapping; ICP: iterative closest point; DSO: direct sparse odometry. √ and × mean whether the certain function is supported, √: supported; ×: unsupported.

system. The current optimization libraries (e.g. g2o, Ceres) make it easy to build solvers and process thousands of variables in one single second, which, therefore, makes BA-based graph optimization method to be the mainstream tool for the back-end state estimation.

Open-source V-SLAM system. We would like to review some open-source algorithms of V-SLAM, since this is so essential. Generally, V-SLAM systems can be classified according to the camera types, including but not limited to monocular, stereo, and RGB-D cameras. For a detailed demonstration, Table 3 further summarizes their characteristics containing the descriptions of front end, back end, relocalization, loop closure detection, and so on. We insist that key factors for a V-SLAM assessment would always be whether it enables dense mapping and loop closure detection, whether it supports a number of sensors, and whether it possesses real-time performances. It is important to appreciate that, for simplifying the present semantic SLAM designs, lots of studies directly refer to the well-established V-SLAM frameworks.[47,48]

Human–robot–environment interaction: Perception

We argue that the perception defined in area of semantic SLAM should consist of two aspects: understanding of environment and understanding of human. This perception is referred to as human–robot–environment interaction. Undoubtedly, an environment model (defined as semantic map) will play roles in these two understanding processes. Technically, the more information rich the semantic map is, the higher the so-called semantic level is. Since semantic map increasingly reveals its superiority in complex and autonomous robot tasks (e.g., avoid muddy road while driving), semantically mapping has become a significant and ongoing subject in present semantic SLAM studies. We would like to summarize the present research work and further state our vision for semantic maps within such semantic SLAM frameworks. Table 4 summarizes some semantic mapping studies.

Semantic map

Semantic maps can be categorized into object level and pixel level in a broad sense. Previous studies[75–78] established an embryonic concept of object-level semantic map by inserting some preestablished 3D models of known objects into meaningless sparse point cloud maps. Quite different, research[79–84] attempted to construct superior pixel-level semantic maps via applying some traditional tools, like SVM (even though SVM is commonly used in addressing industrial problems of prediction,[85–87] classification,[88] or fault diagnosis[89]), CRF, and so on, since these tools are considered to be useful for object identification and scene segmentation. However, the limited means, in most cases, tend to an unsatisfactory classifying precision. Inspired by the advances in deep learning, there has been more research in the area of CNN-based object identification, detection, and segmentation.[90–92] The sufficient achievements subsequently provide a guarantee for constructing more accurate semantic maps with pixel level.[93]

Li and Belaroussi[47] present a blend of most advanced semantic segmentation strategy (DeepLab-v2) and V-SLAM framework (large-scale direct monocular, LSD-SLAM). It distinguishes itself by successfully constructing a semi-dense 3D semantic map via a multiple-view monocular camera (rather than acquire a dense 3D semantic map with an RGB-D camera, as the study of McCormac et al.[48] indicates). It should be stressed that the highlight of such a

Table 4. Summary of semantic mapping studies.

Year	Reference	Camera type	3D reconstruction	Semantic labeling	Map expression	Data set
2013	Valentin et al.[68]	RGB-D	Surface reconstruction	CRF	Triangulated mesh	Indoor: NYU Outdoor: KITTI
2013	Sengupta et al.[69]	Stereo	Surface reconstruction	CRF	Mesh	KITTI
2015	Vineet et al.[70]	Stereo	VO	Random Forest	Voxel	KITTI
2016	Zhao and Chen[71]	RGB-D	VO	SVM	Voxel	NYU v2
2016	Li and Belaroussi[47]	RGB-D	LSD-SLAM	Deeplab v2	Voxel	Indoor: NYU v2 Outdoor: KITTI
2017	McCormac et al.[48]	RGB-D	RGB-D SLAM	CNN with CRF	Surfel	NYU v2
2017	Yang et al.[72]	Stereo	ORB-SLAM	CNN with CRF	Grid	KITTI
2018	Runz et al.[73]	RGB-D	RGB-D SLAM	Mask R-CNN and geometric segmentation	Surfel	TUM
2019	Narita et al.[74]	RGB-D	RGB-D SLAM	PSPNET with CRF mask R-CNN with CRF	Voxel	ScanNet v2

CRF: conditional random field; R-CNN: region convolutional neural network; CNN: convolutional neural network; VO: visual odometry; SLAM: simultaneous localization and mapping; PSPNET: pyramid scene parsing network.

blend also consists in its inversion back to enhance the performances of a large wider range of 2D single-view semantic segmentation approaches. Apparently, SLAM essentially elevates the accuracy of semantic segmentation.

Open problems

Time-varying semantic map. The semantic map lays the groundwork to the high-level semantic understanding, while its applicability to long-term robust positioning is still unsatisfactory. An ideal solution is to build a time-varying semantic map; if it were not for this fact, a model about spatiotemporal relations between objects in concerned scenes would not be established, and the following spatial changes (viz. the motion) of objects would not be predicted. Thus, we believe the introduction of time-varying semantic maps helps for both long-term and dynamic localization. We also believe that, fundamental to the development of such maps are certain artificial intelligence (AI) ideas about spatial and temporal reasoning. As far as we know, the present semantic SLAM rarely covers such studies.

Panoptic semantic map. As already discussed, the CNN-based semantic segmentation leads to superior fine-grained results. Even though they seem to be subtle enough, for some certain purposes, the segmented regions are not quite tiny (e.g. different styles of cars cannot be distinguished), which somehow limits their understanding level for scene perception. One of the important contributions of instance segmentation network in SLAM area just consists in its idea of further subdividing objects within the same category; nevertheless, it appears to be not available for irregular backgrounds.

Panoptic segmentation fully integrates the advantages of these two-segmenting means; as a new direction in CV community, it is expected to generate fine-grained results with globally consistent labelings in an elegant manner. The panoptic semantics mapping, therefore, is recognized as powerful and eminent tool for fostering the intelligence of autonomous robot as well as the contextual knowledge of augmented reality. Panopticfusion was a pioneering study in panoptic semantics-based 3D reconstruction,[74] which, however, unfavorably neglected the useful exploration of semantics-based positioning ideas. Due to the fact that semantic positioning is frequently overlooked in practical applications, we are firmly convinced that the semantic SLAM framework which simultaneously focuses on mapping and localization is still being explored.

Environmental adaptation: Robustness

As previously mentioned, V-SLAM is now at a robust-perception age. In a sense, a primary concern of semantic SLAM would be the "robustness" enhancements. We will concentrate on this central issue in terms of feature selection mechanism and optimized data association. Before a detailed review, we firstly summarize the relevant researches in robustness enhancements, as summarized in Table 5. More about object SLAM will be presented in Discussions section.

Feature selection mechanism

The acquisition of prior semantics of feature points leads to enhanced robustness of VO. Since we have initially assessed whether these feature points are suitable for a specific task, thus the selected robust features will contribute to better robot ego-motion tracking. Much more interesting, feature selection strategy could be flexibly

Table 5. Summary of semantic SLAM research in robustness enhancements.

Method	Reference	Year	Main contribution
Feature selection	Reddy et al.[94]	2015	Tracking stationary features
	Murali et al.[95]	2017	Lifelong localization
	Li et al.[96]	2018	A lightweight system
	Yu et al.[46]	2018	Adapting to dynamic environment
	Ganti et al.[97]	2018	An information-theoretic method
	Liang et al.[98]	2019	Visual saliency map
Optimized data association	Bowman et al.[49]	2017	Probabilistic data association
	Lianos et al.[99]	2018	Medium-term association
CNN-based image features	Yi et al.[100]	2014	Learned invariant features
	DeTone et al.[101]	2018	Self-supervised interest features
Object SLAM	Salas-Moreno et al.[78]	2013	A pioneer study
	Nicholson et al.[103]	2018	Objects described by ellipsoid
	Yang and Scherer[102]	2019	Objects described by cube

CNN: convolutional neural network; SLAM: simultaneous localization and mapping.

changeable for purposes of various tasks. We will review the recent studies from the following aspects.

Interested region feature selection. Liang et al.[98] proposed a VO framework for feature selection on basis of a visual saliency map (defined by visual saliency to each pixel of a single image, the closer to the red color, the higher the degree of visual saliency) filtered by semantics segmentation results. In fact, it is this blend map (integrates visual saliency map and semantics segmentation map) that consequently drives the feature selection process. The robustness of VO is tested to be superior with such robust feature points (selected by this blend map). Please see the research of Liang et al.[98] for more details.

In research,[95] the feature points derived from the parking cars are no longer used for mapping due to the fact that temporary objects should not be maintained in environmental maps. Also, such maps with no temporary objects lead to better robustness in lifelong localization tasks.

Informative region feature selection. The accuracy of pose estimation cannot be highly improved via feature points in regions with low information entropy.[104] Tracking with such features will consequently increase the risks of faulty data associations. Ganti and Waslander[97] propose an information-theoretic feature selection method by inviting the uncertainty concept of semantic segmentation for the calculation of information entropy. This immediately reduces the numbers of features, thus significantly improves the system performances of real time and robustness without any appreciably compromising in accuracy.

Dynamic feature selection. The extracted feature points (from images) probably belong to moving objects (so-called dynamic feature points), which greatly decrease the robustness of V-SLAM systems. Fortunately, high-level semantics can efficiently perform the division of stationary and dynamic feature points (so-called motion segmentation), so that certain positive mechanism works in dynamic scenes within which V-SLAM systems possess enhanced robustness.

Reddy et al.[94] employed a multilayer dense CRF tool to segment images. The distinguishable stationary feature exhibits stillness, making it feasible to separately track the stationary feature points. Consequently, a robust VO adapts to a dynamic scene. SLAM toward dynamic environments[46] seeks to joint semantic segmentation and moving consistency check to eliminate ORB feature points that initially exist in a dynamic object, which not only outperforms ORB-SLAM2[63] regarding accuracy and robustness in a dynamic environment but also builds a dense semantic octo-tree map for further 3D representation. Moreover, a lightweight 3D box inference tool is put forward by Li and Qin;[96] in their studies, the conventional semantic segmentation is even no more necessary for real-time semantic reasoning.

Optimized data association

In V-SLAM frameworks, in terms of the update frequency, the data association could be divided into two categories: short-term association (e.g. feature matching) and long-term association (e.g. loop closure detection). This mechanism ensures a maximum of data association reliability. However, in cases where the loop closure detection fails (e.g. unmanned vehicles are driving on long straight roads), VO will irreversibly drift and this consequently leads to the divergence of navigation systems. A study of semantic SLAM proposes image semantics based on medium-term association mechanism.[99] From an experimental point of view, this mechanism largely reduces the VO translational drift in unmanned driving scenes. There are several problems that confront the advocate of such image semantics-based mechanism. Bowman et al.[49] found a defect of such semantics associations in application, that is, invalid data association of objects' semantics greatly affects the results of localization and mapping. They therefore proposed a so-called probabilistic data association

Table 6. Summary of semantic SLAM research in accuracy enhancements.

Method	Reference	Year	Main contribution
Monocular scale initialization	Frost et al.[105]	2016	Reduces scale drift over long-range outdoor
	Sucar and Hayet[106]	2017	Reduces scale drift over small-scale indoor
Semantic and geometric joint optimization	Bowman et al.[49]	2017	First semantic and geometric joint optimization
	Lianos et al.[99]	2018	Medium-term data association
	Li et al.[96]	2018	A lightweight semantic inference method
Relocalization and loop closure detection	Stenborg et al.[107]	2018	Meeting seasonal change challenge
	Gawel et al.[108]	2018	Graph-based semantic relocalization method
End-to-end SLAM	Ummenhofer et al.[109]	2017	Inferring from a pair of images
	Wang et al.[110]	2018	Inferring from a video

SLAM: simultaneous localization and mapping.

mechanism to fully consider the uncertainty during the process of data association.

Open problems

Mainstream semantic SLAM methods improve the robustness of a VO via selecting features or optimizing data associations. However, with the full-scaled improvements of algorithms, the efforts for VO robustness enhancements by purely feature selecting or data association optimizing appear unsatisfactory. Recently, the CNN-based feature extractors appeared to be noticeable in the field of CV,[100] and they led to much more robust visual features that handcraft solutions never derive. Inspired by which, researchers in SLAM area are now making their attempts to reconstruct VO by so learned features,[101] so as to substantially improve VO robustness. Following this line of thought, we believe that the pursuit of enhanced feature stabilization and generalization ability for enhanced VO robustness would continue.

Reliable navigation: Accuracy

The accuracy of localization and mapping could suggest a reliability assessment of autonomous navigation systems. Generally speaking, if it were desired to elevate the accuracy enhancements, semantics could be included in nearly all the sessions of classic SLAM algorithm frameworks, such as initialization, back-end optimization, relocalization, loop closure detection, and so on. Before delivering a detailed discussion followed in this section, we would like to firstly summarize the relevant semantic SLAM research that devote to accuracy enhancements, as summarized in Table 6.

Monocular scale initialization

As a consequence of no absolute baseline length between images, the scales of monocular V-SLAM systems indispensably appear to be both ambiguous and drifting over time. Thus, a key problem in the development of monocular V-SLAM initialization would be how to rectify the scale ambiguity and drift. The highlight of both studies[105,106] consists in that they identically invite the concept of image semantics. As one form of image semantics, the size of object has been fully considered and the monocular scale initialization process is recognized to be more efficient with excellent concision. The experimental results based on public data sets also validate their effectiveness over a wide range of applications, that is, as small as small-object indoor scenes or as large as long-range outdoor scenes.

Semantic and geometric joint optimization

One of the most significant tightly coupled semantic and geometric joint optimization framework is proposed by Bowman et al.,[49] who pioneered the ideas of probabilistic data association models. If both continuous and discrete data are already involved in data association tasks, a solution by MLE method, directly, is not possible. For this, the authors skillfully broke their main problem down into subproblems, that is, they divided the so-called mixed association into two processes: discrete semantic association and continuous pose estimation. This two-step iterative computation problem could be easily solved by typical expectation maximization algorithm. Moreover, the principal importance of semantics that extracted by object detection is that they play roles in back-end optimization.

One of the ideas of incorporating the semantics (extracted by semantic segmentation) in SLAM back end is put forward by Linaos et al.[99] Given the fact that 2D object boundaries cannot precisely express boundaries of matched 3D objects, Linaos's theories are considered to be more valid in practical applications. The latest study[96] employs 2D object detection results to infer the bounding box of 3D objects. From an engineering perspective, this strategy can even be accepted by real-time semantic SLAM systems where the demands of accuracy could be moderately loose.

Relocalization and loop closure detection

Relocalization and loop closure detection usually employ identical techniques; they, however, tackle different problems. The purpose of relocalization is to restore the camera pose, while the function of loop closure detection is to derive geometrically consistent map. Regardless of how differently the individual techniques function, we are generally concerned with the identical theories. Therefore, this subsection is devoted to a description of semantics-based relocation algorithms, mostly following the application-oriented lines of thought.

The principal limitation of geometric localization lies in its long-term applicability to locating in changeable scenes (over time) within pre-built maps. However, the semantics-based solutions are the answers to this challenging issue. The evidence can be seen from a recent study,[107] where a semantics based cross-season localization algorithm is proposed. In principle, the geometric localization methods are dependent on similarities between image appearances, and this has apparently confronted the researchers that, even though the images are collected under identical positions, seasonal changes seem to be enough to make the concerned images unidentified, so that the matching relationship becomes unreliable. In this case, the semantics are certainly reminiscent, and one of the important contributions of research in cross-season localization has been the fact that topologies of semantic objects in a single image would be consistent over time. This cross-season localization method appears to be sufficiently reliable when applied to unmanned vehicles. A novel graph-based semantic relocalization idea was proposed by Gawel et al.,[108] in such a system, the keyframes with semantics are transformed into a large set of 3D graphs, and these 3D graphs are used to further match with the surrounding's map that is globally pre-built.

Apart from the seasonal changes, the introduction of semantics also helps to deal with the variation of larger viewpoint or illuminatio, or even partial structure changes of scenes caused by time. This relocalization and loop closure detection scheme produces a verification of accuracy enhancement of V-SLAM systems as an added benefit.

Open problems

Parts of semantic SLAM researchers pay their attention to the pipeline design of deep learning-based solutions, so as to build a trainable end-to-end SLAM system. Attempts have been made to estimate depth from a single image by means of CNNs in recent years.[111–113] Even if the feasibility has been testified, the difficulties caused by confining generalization ability of CNNs still remain as an inherently ill-posed problem. The efforts of researchers have been directed at exploiting some end-to-end pipelines to jointly estimate depth and camera motion from a pair of images.[109] In addition, Wang and Clark[110] provide an alternative solution and can be reference to further study, which directly infers poses and uncertainties from a video.

From their experiments, it has been learned that the hierarchical network design, together with careful parameter configuration and sufficient training, could result in the state-of-the-art accuracy on the given data sets. Meanwhile, opponents are still standing in the way of arguing the poor performance of pipeline-formed SLAM in practical applications; they emphasize the "interpretability" and "generalization capability" issues. For this, researchers are now working on deep learning modeled methods for better interpretability and multidimensional visualization.

Discussions

In the former sections, the issues associated with perception, robustness, and accuracy of semantic SLAM are currently referred to. Furthermore, among technical tools for SLAM performance enhancements, the matched open problems are posted. One of the major concerns of this survey is to present the feasible solutions to above open problems from a macroperspective. Therefore, this entire section is devoted to a macroscopic discussions. It is mainly related to multiscaled map expression, object SLAM, and weakly supervised and unsupervised learning SLAM.

Multiscaled map expression

We believe that the time-scaled maps contribute to the long-term autonomous location of robots. For a few years, the advocates of V-SLAM have ignored the existent problems in their research. For example, the spatiotemporal context (STC) in image sequences has been not taken into account in the process of mapping expression, which consequently makes it impossible to reconstruct the expected time-varying semantic maps. Lately, the research on recursive neural network (RNN) has helped to develop the ideas of STC in image sequences;[114] from our point of view, RNN could be identically invited for the mapping tasks of a V-SLAM that requires long-term locating with strong autonomy.

Together with time-varying map (contains the entire environmental information over a certain period of time), panoptic semantic map constitutes the main forms that may be taken in multiscaled expression. If it were desired to construct a panoptic semantic map within a V-SLAM framework, the keyframes need to be semantically segmented in a global perspective. As one source of the difficulty in CV community, several methods have been developed for segmenting foreground objects on pixel level; however, the problems of unifying labelings of foreground and background still remain. The rising panoptic segmentation network represents a solution to this class of problems.[93] It produces globally constraint labelings by fusing results derived from semantic segmentation and instance segmentation; a better understanding of the things being perceived, therefore, is achieved as expected.

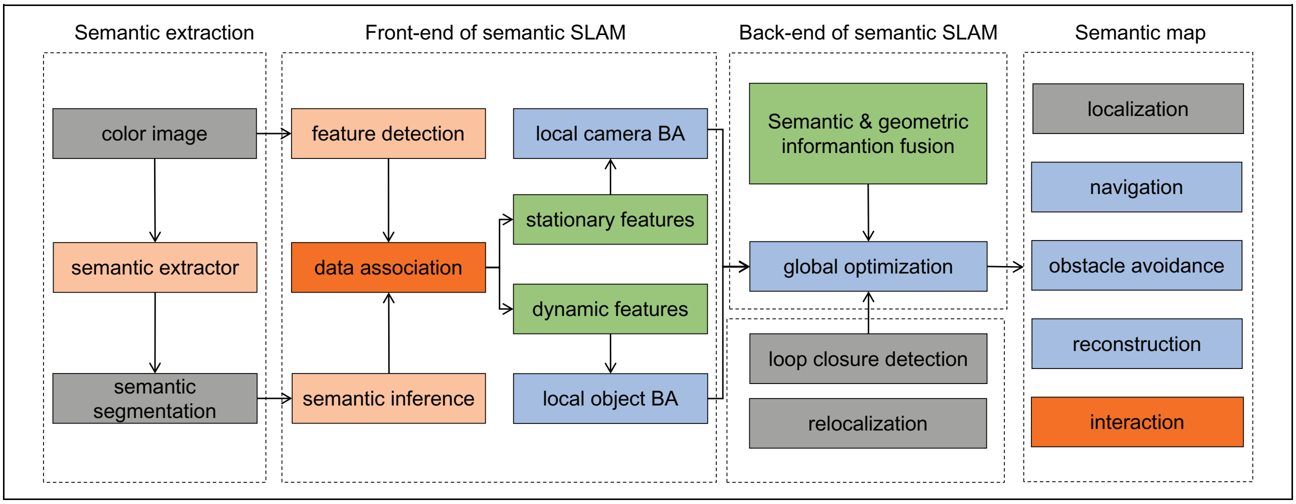

Figure 6. The architecture of a semantic SLAM system. SLAM: simultaneous localization and mapping.

According to the analysis above, in semantic SLAM field, we are convinced of the promising advance of multi-scaled maps, which have same general characteristics in high-level human–robot–environment interaction and long-term autonomous location.

Object SLAM

From our point of view, DNNs are novel but unpractical ways in improving the robustness of a VO. In most cases, due to the overemphasized robustness of feature points, the overtrained DNN pipelines not only produce unexpected consumption of time but also exhibits unavailability in certain SLAM tasks under totally new scenes. A reliable object SLAM framework is illustrated in Figure 6, where the independent tracking for individual objects in a 3D scene is established. It enables the efficient feature selection and data association to be implemented in terms of 2D to 3D and single thread to multithread, so that practically improves the robustness and accuracy of a VO.

SLAM++[78] represents the earliest research in area of object SLAM. Due to the fact that the object data sets should be built beforehand, SLAM++ is still invalid for online tasks. Lately, the research on SLAM++ can be developed alternatively along two directions: one is represented by CubeSLAM[102] with an object description by cube, the other one is represented by QuadricSLAM[103] with an object description by ellipsoid.

We believe that object SLAM has broad prospects, and the point of the whole process is to directly track dynamic targets under 3D scenes. With the rapid advance of 3D object tracking (includes a 3D semantic estimator) in area of CV, there are reasons to believe that it simultaneously helps to construct an object SLAM system with more efficiency.

Weakly supervised and unsupervised learning SLAM

With the existing data sets, the end-to-end semantic SLAM pipeline generally leads to optimal localization accuracy, but the interpretability and generalization ability restricts its applicability to a wider range of applications. Take DNN as a specific example, the reduced generalization ability is often accompanied by overfitting due to over meticulous parameter configuration and training. The weakly supervised and unsupervised learning-based pipelines have been employed in the development of improved generalization ability of DNNs. However, the study is still in the preliminary stage. In fact, in end-to-end SLAM filed, unsupervised learning-based monocular depth estimation has been recognized as a main research direction;[115–117] meanwhile, interests of experts in machine learning are now focused upon the interpretability of DNNs. These clues make us believe that the advanced learning strategies would be powerful and practical tools for the semantic SLAM pipelines. It is important to appreciate that semantic SLAM pipelines can be easily integrated into deep reinforcement learning paradigm to construct a robot system with general intelligence.

Conclusions

For autonomous robot navigation tasks, a semantic SLAM that aims at better understanding and perceiving a message from the robot work volume has drawn an increasing attention. In this survey, we review the development of semantic SLAM concerning its perception, robustness, and accuracy and then discuss the open problems associated with the recent progress and challenges. Specifically, we attempt to seek possible solutions to these open problems from a macroscopic view and further state the suggestions in a constructive manner. We believe that SLAM frameworks are well-established and proven by practice, and semantic SLAM will distinguish itself by the eminent fusion of image semantics. The evolution of deep learning-based methods has apparently exploited the opportunity for researchers to use their powerful image processing capacities to estimate poses, detect loop closures, build 3D maps, and so on. From our point of view, deep learning and semantic SLAM are now inseparably related, and a blend of them must be experiencing a booming in the future studies.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by Natural Research Fund of Science and Technology Department, Jilin Province under Grant 20170101125JC, Science and Technology Program of Department of Education, Jilin Province under Grant JJKH20200117KJ, and Research Fund for Distinguished Young Scholars of Jilin City under Grant 20190104128.

ORCID iD

Linlin Xia https://orcid.org/0000-0002-5079-3788

References

1. Smith RC and Cheeseman P. On the representation and estimation of spatial uncertainty. *Int J Robot Res* 1986; 5(4): 56–68.
2. Gu ZP and Liu H. A survey of monocular simultaneous localization and mapping. *CAAI Trans Intell Syst* 2015; 10(4): 499–507.
3. Teng ZJ, Qu ZQ, Zhang LY, et al. Research on vehicle navigation BD/DR/MM integrated navigation positioning. *J North Electr Power Univ* 2017; 37(4): 98–101.
4. Cadena C, Carlone L, Carrillo H, et al. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. *IEEE Trans Robot* 2016; 32(6): 1309–1332.
5. Durrant-Whyte H and Bailey T. Simultaneous localization and mapping: part I. *IEEE Robot Autom Mag* 2006; 13(2): 99–110.
6. Bailey T and Durrant-Whyte H. Simultaneous localization and mapping (slam): part II. *IEEE Robot Autom Mag* 2006; 13(3): 108–117.
7. Aulinas J, Petillot YR, Salvi J, et al. The slam problem: a survey. *CCIA* 2008; 184(1): 363–371.
8. Neira J, Davison AJ, and Leonard JJ. Guest editorial special issue on visual slam. *IEEE Trans Robot* 2008; 24(5): 929–931.
9. Grisetti G, Kummerle R, Stachniss C, et al. A tutorial on graph-based slam. *IEEE Intell Transp Syst Mag* 2010; 2(4): 31–43.
10. Dissanayake G, Huang SD, Wang Z, et al. A review of recent developments in simultaneous localization and mapping. In: *2011 6th international conference on industrial and information systems (ICIIS 2011)*, Kandy, Sri Lanka, 16–19 August, 2011, pp. 477–482. Piscataway, NJ, USA: IEEE.
11. Scaramuzza D and Fraundorfer F. Tutorial: visual odometry. *IEEE Robot Autom Mag* 2011; 18(4): 80–92.
12. Strasdat H, Montiel JM, and Davison AJ. Visual slam: why filter? *Image Vision Comput* 2012; 30(2): 65–77.
13. Lowry S, Sünderhauf N, Newman P, et al. Visual place recognition: a survey. *IEEE Trans Robot* 2015; 32(1): 1–19.
14. Saeedi S, Trentini M, Seto M, et al. Multiple-robot simultaneous localization and mapping: A review. *J Field Robot* 2016; 33(1): 3–46.
15. Huang SD and Dissanayake G. A critique of current developments in simultaneous localization and mapping. *Int J Adv Robot Syst* 2016; 13(5): 1–13.
16. Taketomi T, Uchiyama H, and Ikeda S. Visual slam algorithms: a survey from 2010 to 2016. *IPSJ Trans Comput Vis Appl* 2017; 9(1): 16.
17. Younes G, Asmar D, Shammas E, et al. Keyframe-based monocular slam: design, survey, and future directions. *Robot Auton Syst* 2017; 98: 67–88.
18. Saputra MRU, Markham A, and Trigoni N. Visual slam and structure from motion in dynamic environments: a survey. *ACM Comput Surv (CSUR)* 2018; 51(2): 37.
19. Gallego G, Delbruck T, Orchard G, et al. Event-based vision: a survey. *arXiv preprint arXiv:190408405*, 2019.
20. Viola P and Jones M. Robust real-time object detection. *Int J Comput Vis* 2001; 4(34-47): 4.
21. Dalal N and Triggs B. Histograms of oriented gradients for human detection. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR 2005)* (eds C Schmid, S Soatto and C Tomasi), San Diego, CA, USA, 20–25 June, 2005, pp. 886–893. Los Alamitos, CA, USA: IEEE.
22. Felzenszwalb PF, Girshick RB, McAllester D, et al. Object detection with discriminatively trained part-based models. *IEEE Trans Patt Anal Mach Intell* 2009; 32(9): 1627–1645.
23. Girshick RB. *From rigid templates to grammars: object detection with structured models*. Chicago, IL, USA: University of Chicago, Division of the Physical Sciences, Department of Computer Science, 2012.
24. Lin TY, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: *30th IEEE conference on computer vision and pattern recognition (CVPR 2017)* (eds R Chellappa, Z Zhang and A Hoogs), Honolulu, HI, USA, 21–26 July, 2017, pp. 936–944. Piscataway, NJ, USA: IEEE.
25. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *27th IEEE conference on computer vision and pattern recognition (CVPR 2014)*, Columbus, OH, USA, 23–28 June, 2014, pp. 580–587. Piscataway, NJ, USA: IEEE.
26. Girshick R. Fast R-CNN. In: *15th IEEE international conference on computer vision (ICCV 2015)* (eds R Bajcsy, G Hager

and Y Ma), Santiago, Chile, 11–18 December, 2015, pp. 1440–1448. Piscataway, NJ, USA: IEEE.

27. Ren SQ, He KM, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans Patt Anal Mach Intell* 2015; 39(6): 1137–1149.
28. He KM, Gkioxari G, Dollár P, et al. Mask R-CNN. In: *2017 IEEE international conference on computer vision (ICCV 2017)* (ed K Ikeuchi), Venice, Italy, 22–29 October, 2017, pp. 2980–2988. Los Alamitos, CA, USA: IEEE.
29. Hosang J, Benenson R, Dollar P, et al. What makes for effective detection proposals? *IEEE Trans Patt Anal Mach Intell* 2016; 38(4): 814–830.
30. Krizhevsky A, Sutskever I, and Hinton GE. Imagenet classification with deep convolutional neural networks. In: *26th annual conference on neural information processing systems 2012 (NIPS 2012)* (eds P Bartlett, FCN Pereira, CJC Burges, L Bottou and KQ Weinberger), Lake Tahoe, NV, USA, 3–6 December, 2012, pp. 1097–1105. Red Hook, NY, USA: Curran Associates.
31. Vapnik V and Lerner AY. Recognition of patterns with help of generalized portraits. *Avtomat i Telemekh* 1963; 24(6): 774–780.
32. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. In: *29th IEEE conference on computer vision and pattern recognition (CVPR 2016)* (eds R Bajcsy, L Fei-Fei and T Tinne), Las Vegas, NV, USA, 26 June–1 July, 2016, pp. 779–788. Piscataway, NJ, USA: IEEE.
33. Redmon J and Farhadi A. Yolo9000: better, faster, stronger. In: *30th IEEE conference on computer vision and pattern recognition (CVPR 2017)* (eds C Rama, Z Zhengyou and H Anthony), Honolulu, HI, USA, 21–26 July, 2017, pp. 6517–6525. Piscataway, NJ, USA: IEEE.
34. Redmon J and Farhadi A. Yolov3: an incremental improvement. *arXiv preprint arXiv:180402767*, 2018.
35. Duan KW, Bai S, Xie LX, et al. Centernet: keypoint triplets for object detection. In: *2019 IEEE/CVF international conference on computer vision (ICCV 2019)*, Seoul, Korea (South), 27 October–2 November, 2019, pp. 6568–6577. IEEE.
36. Hu JP, Li L, Xie Q, et al. A novel segmentation approach for glass insulators in aerial images. *J North Electr Power Univ* 2018; 38(2): 87–92.
37. Long J, Shelhamer E, and Darrell T. Fully convolutional networks for semantic segmentation. In: *IEEE conference on computer vision and pattern recognition (CVPR 2015)*, Boston, MA, USA, 7–12 June, 2015, pp. 3431–3440. Los Alamitos, CA, USA: IEEE Computer Society.
38. Badrinarayanan V, Kendall A, and Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Patt Anal Mach Intell* 2017; 39(12): 2481–2495.
39. Chen LC, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFS. *arXiv preprint arXiv:14127062*, 2014.
40. Chen LC, Papandreou G, Kokkinos I, et al. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. *IEEE Trans Patt Anal Mach Intell* 2017; 40(4): 834–848.
41. Chen LC, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:170605587*, 2017.
42. Chen LC, Zhu YK, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *15th European conference on computer vision (ECCV 2018)* (eds V Ferrari, M Hebert, C Sminchisescu and Y Weiss), Munich, Germany, 8–14 September, 2018, pp. 833–851. Cham, Switzerland: Springer.
43. Paszke A, Chaurasia A, Kim S, et al. ENet: a deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:160602147*, 2016.
44. Qi CR, Su H, Mo KC, et al. PointNet: Deep learning on point sets for 3D classification and segmentation. In: *30th IEEE conference on computer vision and pattern recognition (CVPR 2017)* (eds R Chellappa, Z Zhang and A Hoogs), Honolulu, HI, USA, 21–26 July, 2017, pp. 77–85. Piscataway, NJ, USA: IEEE.
45. Shelhamer E, Rakelly K, Hoffman J, et al. Clockwork convnets for video semantic segmentation. In: *14th European conference on computer vision (ECCV 2016)* (eds B Leibe, J Matas, N Sebe and M Welling), Amsterdam, the Netherlands, 11–14 October, 2016, pp. 852–868. Cham, Switzerland: Springer.
46. Yu C, Liu ZX, Liu XJ, et al. DS-SLAM: a semantic visual slam towards dynamic environments. In: *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS 2018)* (ed NR Gakkai), Madrid, Spain, 1–5 October, 2018, pp. 1168–1174. Piscataway, NJ, USA: IEEE.
47. Li XP and Belaroussi R. Semi-dense 3D semantic mapping from monocular slam. *arXiv preprint arXiv:161104144*, 2016.
48. McCormac J, Handa A, Davison A, et al. SemanticFusion: dense 3D semantic mapping with convolutional neural networks. In: *2017 IEEE international conference on robotics and automation (ICRA 2017)*, Singapore, Singapore, 29 May–3 June, 2017, pp. 4628–4635. Piscataway, NJ, USA: IEEE.
49. Bowman SL, Atanasov N, Daniilidis K, et al. Probabilistic data association for semantic SLAM. In: *2017 IEEE international conference on robotics and automation (ICRA 2017)*, Singapore, Singapore, 29 May–3 June, 2017, pp. 1722–1729. Piscataway, NJ, USA: IEEE.
50. Lu F and Milios E. Globally consistent range scan alignment for environment mapping. *Auton Robot* 1997; 4(4): 333–349.
51. Gutmann JS and Konolige K. Incremental mapping of large cyclic environments. In: *1999 IEEE international symposium on computational intelligence in robotics and automation (CIRA'99)*, Monterey, CA, USA, 8–9 November, 1999, pp. 318–325. Piscataway, NJ, USA: IEEE.
52. Davison AJ, Reid ID, Molton ND, et al. MonoSLAM: real-time single camera SLAM. *IEEE Trans Patt Anal Mach Intell* 2007; 29(6): 1052–1067.

53. Klein G and Murray D. Parallel tracking and mapping for small AR workspaces. In: *2007 6th IEEE and ACM international symposium on mixed and augmented reality (ISMAR)* (eds Nihon-Bācharu-Riariti-Gakkai), Nara, Japan, 13–16 November, 2007, pp. 225–234. Piscataway, NJ, USA: IEEE.
54. Newcombe RA, Izadi S, Hilliges O, et al. KinectFusion: real-time dense surface mapping and tracking. In: *2011 10th IEEE international symposium on mixed and augmented reality (ISMAR 2011)*, Basel, Switzerland, 26–29 October, 2011, pp. 127–136. Piscataway, NJ, USA: IEEE.
55. Whelan T, Kaess M, Fallon M, et al. Kintinuous: spatially extended kinectfusion. In: *3rd RSS workshop on RGB-D: advanced reasoning with depth cameras*, Sydney, Australia, 9–13 July, 2012, pp. 5724–5731.
56. Endres F, Hess J, Sturm J, et al. 3-D mapping with an RGB-D camera. *IEEE Trans Robot* 2013; 30(1): 177–187.
57. Engel J, Schöps T, and Cremers D. LSD-SLAM: large-scale direct monocular SLAM. In: *13th European conference on computer vision (ECCV 2014)* (ed D Fleet), Zurich, Switzerland, 6–12 September, 2014, pp. 834–849. Cham, Switzerland: Springer.
58. Forster C, Pizzoli M, and Scaramuzza D. SVO: fast semi-direct monocular visual odometry. In: *2014 IEEE international conference on robotics and automation (ICRA 2014)*, Hong Kong, China, 31 May–7 June, 2014, pp. 15–22. Piscataway, NJ, USA: IEEE.
59. Labbe M and Michaud F. Online global loop closure detection for large-scale multi-session graph-based SLAM. In: *2014 IEEE/RSJ international conference on intelligent robots and systems (IROS 2014)* (ed W Burgard), Chicago, IL, USA, 14–18 September, 2014, pp. 2661–2666. Piscataway, NJ, USA: IEEE.
60. Labbé M and Michaud F. RTAB-MAP as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *J Field Robot* 2019; 36(2): 416–446.
61. Whelan T, Salas-Moreno RF, Glocker B, et al. Elasticfusion: real-time dense slam and light source estimation. *Int J Robot Res* 2016; 35(14): 1697–1716.
62. Mur-Artal R, Montiel JMM, and Tardos JD. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans Robot* 2015; 31(5): 1147–1163.
63. Mur-Artal R and Tardós JD. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans Robot* 2017; 33(5): 1255–1262.
64. Engel J, Koltun V, and Cremers D. Direct sparse odometry. *IEEE Trans Patt Anal Mach Intell* 2017; 40(3): 611–625.
65. Dai A, Nießner M, Zollhöfer M, et al. Bundlefusion: real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM T Graphics (ToG)* 2017; 36(3): 24.
66. Schlegel D, Colosi M, and Grisetti G. ProSLAM: Graph SLAM from a programmer's perspective. In: *2018 IEEE international conference on robotics and automation (ICRA 2018)* (ed K Lynch), Brisbane, Queensland, Australia, 21–25 May, 2018, pp. 3833–3840. Piscataway, NJ, USA: IEEE.
67. Sumikura S, Shibuya M, and Sakurada K. OpenVSLAM: a versatile visual slam framework. In: *27th ACM international conference on multimedia (MM 2019)* (eds L Amsaleg, et al.), Nice, France, 21–25 October, 2019, pp. 2292–2295. New York, NY, USA: ACM.
68. Valentin JP, Sengupta S, Warrell J, et al. Mesh based semantic modelling for indoor and outdoor scenes. In: *26th IEEE conference on computer vision and pattern recognition (CVPR 2013)*, Portland, OR, USA, 23–28 June, 2013, pp. 2067–2074. Piscataway, NJ, USA: IEEE.
69. Sengupta S, Greveson E, Shahrokni A, et al. Urban 3D semantic modelling using stereo vision. In: *2013 IEEE international conference on robotics and automation (ICRA 2013)*, Karlsruhe, Germany, 6–10 May, 2013, pp. 580–585. Piscataway, NJ, USA: IEEE.
70. Vineet V, Miksik O, Lidegaard M, et al. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In: *2015 IEEE international conference on robotics and automation (ICRA 2015)*, Seattle, WA, USA, 26–30 May, 2015, pp. 75–82. Piscataway, NJ, USA: IEEE.
71. Zhao Z and Chen XP. Building 3D semantic maps for mobile robots using RGB-D camera. *Intell Ser Robot* 2016; 9(4): 297–309.
72. Yang SC, Huang YL, and Scherer S. Semantic 3D occupancy mapping through efficient high order CRFS. In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS 2017)*, Vancouver, BC, Canada, 24–28 September, 2017, pp. 590–597. Piscataway, NJ, USA: IEEE.
73. Runz M, Buffier M, and Agapito L. MaskFusion: real-time recognition, tracking and reconstruction of multiple moving objects. In: *17th IEEE international symposium on mixed and augmented reality (ISMAR 2018)*, Munich, Germany, 16–20 October, 2018, pp. 10–20. Los Alamitos, CA, USA: IEEE.
74. Narita G, Seno T, Ishikawa T, et al. Panopticfusion: online volumetric semantic mapping at the level of stuff and things. *arXiv preprint arXiv:190301177*, 2019.
75. Castle RO, Gawley DJ, Klein G, et al. Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras. In: *2007 IEEE international conference on robotics and automation (ICRA'07)*, Roma, Italy, 10–14 April, 2007, pp. 4102–4107. Piscataway, NJ, USA: IEEE.
76. Castle RO, Klein G, and Murray DW. Combining MonoSLAM with object recognition for scene augmentation using a wearable camera. *Image Vision Comput* 2010; 28(11): 1548–1556.
77. Civera J, Gálvez-López D, Riazuelo L, et al. Towards semantic SLAM using a monocular camera. In: *2011 IEEE/RSJ international conference on intelligent robots and systems: celebrating 50 years of robotics (IROS'11)* (ed Amato NM), San Francisco, CA, USA, 25–30 September, 2011, pp. 1277–1284. Piscataway, NJ, USA: IEEE.
78. Salas-Moreno RF, Newcombe RA, Strasdat H, et al. Slam++: simultaneous localisation and mapping at the level of objects. In: *26th IEEE conference on computer vision and pattern recognition (CVPR 2013)*, Portland, OR, USA, 23–28 June, 2013, pp. 1352–1359. Piscataway, NJ, USA: IEEE.

79. Xiong XH and Huber D. Using context to create semantic 3D models of indoor environments. In: *2010 21st British machine vision conference (BMVC 2010)* (eds F Labrosse, R Zwiggelaar, Y Liu and B Tiddeman), Aberystwyth, UK, 31 August–3 September, 2010, pp. 1–11. Cambridge, UK: British Machine Vision Association (BMVA).
80. Koppula HS, Anand A, Joachims T, et al. Semantic labeling of 3D point clouds for indoor scenes. In: *25th annual conference on neural information processing systems 2011 (NIPS 2011)* (ed JS Taylor), Granada, Spain, 12–14 December, 2011, pp. 244–252. Red Hook, NY, USA: Curran Associates.
81. Stückler J, Biresev N, and Behnke S. Semantic mapping using object-class segmentation of RGB-D images. In: *25th IEEE/RSJ International Conference on Robotics and Intelligent Systems (IROS 2012)*, Vilamoura, Portugal, 7–12 October, 2012, pp. 3005–3010. Piscataway, NJ, USA: IEEE.
82. Kostavelis I and Gasteratos A. Learning spatially semantic representations for cognitive robot navigation. *Robot Auton Syst* 2013; 61(12): 1460–1475.
83. Couprie C, Farabet C, Najman L, et al. Indoor semantic segmentation using depth information. *arXiv preprint arXiv: 13013572*, 2013.
84. Anand A, Koppula HS, Joachims T, et al. Contextually guided semantic labeling and search for three-dimensional point clouds. *Int J Robot Res* 2013; 32(1): 19–34.
85. Li GQ, Zhang Y, Zhang MJ, et al. The wind power real-time diction on the EEMD and SVM of the MRMR. *J North Electr Power Univ* 2017; 37(2): 39–44.
86. Yang M, Huang BY, Jiang B, et al. Real-time prediction for wind power based on Kalman filter and support vector machines. *J North Electr Power Univ* 2017; 37(2): 45–51.
87. Yang M, Chen XX, Zhang Q, et al. A review of short-term wind speed prediction based on support vector machine. *J North Electr Power Univ* 2017; 37(4): 1–7.
88. Ying H, Jiang LL, Li X, et al. Transient stability assessment in bulk power grid using v-nonparallel support vector machine. *J North Electr Power Univ* 2018; 38(5): 31–40.
89. Cui TX, Zhou XL, Liu WH, et al. Gear fault diagnosis based on Hilbert envelope spectrum and SVM. *J North Electr Power Univ* 2017; 37(6): 56–61.
90. Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*, Miami, FL, USA, 20–25 June, 2009, pp. 248–255. Piscataway, NJ, USA: IEEE.
91. He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition. In: *29th IEEE conference on computer vision and pattern recognition (CVPR 2016)* (eds Z Bajcsy, FF Li and T Tuytelaars), Las Vegas, NV, USA, 26 June–1 July, 2016, pp. 770–778. Piscataway, NJ, USA: IEEE.
92. Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks. In: *15th IEEE international conference on computer vision (ICCV 2015)* (eds R Bajcsy, G Hager and Y Ma), Santiago, Chile, 11–18 December, 2015, pp. 1529–1537. Piscataway, NJ, USA: IEEE.
93. Gupta S, Arbeláez P, Girshick R, et al. Indoor scene understanding with RGB-D images: bottom-up segmentation, object detection and semantic segmentation. *Int J Comput Vis* 2015; 112(2): 133–149.
94. Reddy ND, Singhal P, Chari V, et al. Dynamic body VSLAM with semantic constraints. In: *IEEE/RSJ international conference on intelligent robots and systems (IROS 2015)* (ed W Burgard), Hamburg, Germany, 28 September–2 October, 2015, pp. 1897–1904. Piscataway, NJ, USA: IEEE.
95. Murali V, Chiu HP, Samarasekera S, et al. Utilizing semantic visual landmarks for precise vehicle navigation. In: *20th IEEE international conference on intelligent transportation systems (ITSC 2017)* (eds E Rocklage, H Kraft, A Karatas and S Jorg), Yokohama, Japan, 16–19 October, 2017, pp. 1–8. Piscataway, NJ, USA: IEEE.
96. Li PL, Qin T, and Shen HJ. Stereo vision-based semantic 3D object and ego-motion tracking for autonomous driving. In: *15th European conference on computer vision (ECCV 2018)* (eds V Ferrari, M Hebert, C Sminchisescu and Y Weiss), Munich, Germany, 8–14 September, 2018, pp. 664–679. Cham, Switzerland: Springer.
97. Ganti P and Waslander SL. Visual SLAM with network uncertainty informed feature selection. *arXiv preprint arXiv:181111946*, 2018.
98. Liang HJ, Sanket NJ, Fermüller C, et al. SalientDSO: bringing attention to direct sparse odometry. *IEEE Trans Autom Sci Eng* 2019; 16(4): 1619–1626.
99. Lianos KN, Schonberger JL, Pollefeys M, et al. VSO: visual semantic odometry. In: *15th European conference on computer vision (ECCV 2018)* (eds V Ferrari, M Hebert, C Sminchisescu and Y Weiss), Munich, Germany, 8–14 September, 2018, pp. 246–263. Cham, Switzerland: Springer.
100. Yi KM, Trulls E, Lepetit V, et al. Lift: learned invariant feature transform. In: *21st ACM conference on computer and communications security (CCS 2014)* (ed GJ Ahn), Scottsdale, AZ, USA, 3–7 November, 2014, pp. 467–483. New York, NY, USA: ACM.
101. DeTone D, Malisiewicz T, and Rabinovich A. Superpoint: self-supervised interest point detection and description. In: *31st meeting of the IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW 2018)*, Salt Lake City, UT, USA, 18–22 June, 2018, pp. 337–349. Los Alamitos, CA, USA: IEEE.
102. Yang SC and Scherer S. CubeSLAM: monocular 3-D object SLAM. *IEEE Trans Robot* 2019; 35(4): 925–938.
103. Nicholson L, Milford M, and Sünderhauf N. QuadricSLAM: dual quadrics from object detections as landmarks in object-oriented SLAM. *IEEE Robot Autom Lett* 2018; 4(1): 1–8.
104. Xue L, Huang NT, Zhao SY, et al. Low redundancy feature selection using conditional mutual information for short-term load forecasting. *J North Electr Power Univ* 2019; 39(2): 30–38.
105. Frost DP, Kähler O, and Murray DW. Object-aware bundle adjustment for correcting monocular scale drift. In: *2016 IEEE international conference on robotics and automation*

(ICRA 2016), Stockholm, Sweden, 16–21 May, 2016, pp. 4770–4776. Piscataway, NJ, USA: IEEE.

106. Sucar E and Hayet JB. Probabilistic global scale estimation for monoSLAM based on generic object detection. In: *30th IEEE conference on computer vision and pattern recognition workshops (CVPRW 2017)*, Honolulu, HI, USA, 21–26 July, 2017, pp. 988–992. Los Alamitos, CA, USA: IEEE.
107. Stenborg E, Toft C, and Hammarstrand L. Long-term visual localization using semantically segmented images. In: *2018 IEEE international conference on robotics and automation (ICRA 2018)* (ed L Kevin), Brisbane, Queensland, Australia, 21–25 May, 2018, pp. 6484–6490. Piscataway, NJ, USA: IEEE.
108. Gawel A, Del Don C, Siegwart R, et al. X-view: graph-based semantic multi-view localization. *IEEE Robot Autom Lett* 2018; 3(3): 1687–1694.
109. Ummenhofer B, Zhou HZ, Uhrig J, et al. Demon: depth and motion network for learning monocular stereo. In: *30th IEEE conference on computer vision and pattern recognition (CVPR 2017)* (eds R Chellappa, Z Zhang and A Hoogs), Honolulu, HI, USA, 21–26 July, 2017, pp. 5622–5631. Piscataway, NJ, USA: IEEE.
110. Wang S, Clark R, Wen HK, et al. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. *Int J Robot Res* 2018; 37(4-5): 513–542.
111. Eigen D, Puhrsch C, and Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: *28th annual conference on neural information processing systems 2014 (NIPS 2014)* (ed M Welling), Montreal, Canada, 8–13 December, 2014, pp. 2366–2374. New York, NY, USA: Curran Associates.
112. Eigen D and Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: *15th IEEE international conference on computer vision (ICCV 2015)* (eds R Bajcsy, G Hager and Y Ma), Santiago, Chile, 11–18 December, 2015, pp. 2650–2658. Piscataway, NJ, USA: IEEE.
113. Liu F, Shen CH, Lin GS, et al. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans Patt Anal Mach Intell* 2015; 38(10): 2024–2039.
114. Sun B, Qiao C, Yang D, et al. Prediction method of thermal conductivity of nanofluids based on deep belief network. *J North Electr Power Univ* 2019; 39(1): 41–48.
115. Kuznietsov Y, Stuckler J, and Leibe B. Semi-supervised deep learning for monocular depth map prediction. In: *30th IEEE conference on computer vision and pattern recognition (CVPR 2017)* (eds R Chellappa, Z Zhang and A Hoogs), Honolulu, HI, USA, 21–26 July, 2017, pp. 2215–2223. Piscataway, NJ, USA: IEEE.
116. Zhan HY, Garg R, Weerasekera CS, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: *31st meeting of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2018)*, Salt Lake City, UT, USA, 18–22 June, 2018, pp. 340–349. Los Alamitos, CA, USA: IEEE.
117. Wang GM, Wang HS, Liu YL, et al. Unsupervised learning of monocular depth and ego-motion using multiple masks. In: *2019 international conference on robotics and automation (ICRA 2019)*, Montreal, Québec, Canada, 20–24 May, 2019, pp. 4724–4730. Piscataway, NJ, USA: IEEE.