Measure for Semantics and Semantically Constrained Pose Optimization: A Review

Linlin Xia, Sida Li, Linna Yi, Heng Ruan, and Daochang Zhang

***Abstract*— Simultaneous localization and mapping (SLAM) in robotics is a fundamental problem. The use of visual odometry (VO) enhances scene recognition in the task of ego-localization within an unknown environment. Semantically meaningful information permits data association and dense mapping to be conducted based on entities representing landmarks rather than manually designed, low-level geometric clues and has inspired various feature descriptors for semantically ensembled SLAM applications. This article illuminates the insights into the measure for semantics and the semantically constrained pose optimization. The concept of semantic extractor and the matched framework are initially presented. As the latest advances in computer vision and the learning-based deep feature acquisition are closely related, the semantic extractor is especially described in a deep learning paradigm. The methodologies pertinent to our explorations for object association and semantics-fused constraining that is amenable for use in a least-squares framework are summarized in a systematic way. By a collection of problem formulations and principle analyses, our review exhibits a fairly unique perspective in semantic SLAM. We further discuss the challenges of semantic uncertainty and explicitly introduce the term "semantic reasoning." Some technology outlooks regarding semantic reasoning are simultaneously given. We argue that for intelligent tasks of robots such as object grasping, dynamic obstacle avoidance, and object-target navigation, semantic reasoning might guide the complex scene understanding under the framework of semantic SLAM directly to a solution.**

***Index Terms*— Camera pose optimization, deep learning, object association, semantic extractor, semantic reasoning.**

I. Introduction

TO THIS day, the critical role of simultaneous localization and mapping (SLAM) in the development of autonomous mobile robots has been broadly discussed. As the major challenges in SLAM including the optimized state estimation, data association, loop closure detection, vivid map representation and rendering, and dynamic scenario applications have been making breakthroughs, the potential for lightweight, high-performance SLAM systems that can assist people in self-driving, virtual or augmented reality, digit city, and industrial automation intensely garners attention.

Manuscript received 23 January 2024; revised 26 May 2024; accepted 21 June 2024. Date of publication 15 July 2024; date of current version 24 July 2024. This work was supported by the Natural Science Foundation of Jilin Province, China, under Grant 20220101240JC. The Associate Editor coordinating the review process was Dr. Jing Yuan. *(Corresponding author: Linlin Xia.)*

Linlin Xia, Sida Li, Linna Yi, and Heng Ruan are with the School of Automation Engineering, Northeast Electric Power University, Jilin 132012, China (e-mail: xiall521@neepu.edu.cn).

Daochang Zhang is with the School of Science, Northeast Electric Power University, Jilin 132012, China (e-mail: daochangzhang@126.com).

Digital Object Identifier 10.1109/TIM.2024.3428639

Visual SLAM (V-SLAM) contributes to image appearance-based mapping and localization [1]. In general environments, V-SLAM systems perform satisfactorily. However, they are still vulnerable to significant changes in visual appearances and viewpoints. Also, they may suffer from multiple applying challenges, like limited illumination, featureless frames, and tiny pixel parallax for faraway features, thereby making algorithms less robustness-prone. Introducing semantics is an answer to this challenging issue. Semantic SLAM builds on the heritage of well-developed V-SLAM frameworks but aggregates the semantic information that leads to intuitive visualization and improved robustness [2]. The existing, fabulous deep learning achievements break through the dilemma of "hand-crafted feature extraction by a visual odometry (VO)." The use of deep neural networks facilitates the generation of high-level semantic information and the construction of a specialized knowledge base, apparently exploiting opportunities for semantic SLAM. Apart from this, we stress that we cannot ignore the flaws of conventional V-SLAM approaches. For one thing, cases of texture-less or texture-free scenes, intense illumination variations, and motion blur in images need the feature extraction and the following autonomous localization more robust; for another, dynamic objects should not be purely regarded as outliers and taken away from the camera pose estimator. After all, most V-SLAM pipelines are static scene assumption-dependent [3], [4]. Technically, for the evolution of SLAM, it is now in the era of robust perception. The new requirements cover robust performance, advanced understanding, resource awareness, and task-driven awareness [5]. All these push the advances in semantic SLAM.

One focus of current research activities is on leveraging more semantics-based constraints for threads of VO [6], back-end optimization [7], and loop closure detection [8]. V-SLAM systems track low-level geometric clues especially key points in images. Compared to these low-level geometric clues, object features are more stable and can provide richer semantic information. Reliable pose estimation is most readily understood by extracting the semantic primitives of concerned objects in the environment. Except for the increasing accuracy, the goals of such semantically ensembled localization have been identically aimed at improved robustness. Encouragingly, however, semantics-based recognition has a natural sort of advantage to eliminate dynamic, undesirable landmarks under complex environments.

Another major concern consists in mapping. Semantic maps provide a clear incentive for levels of intelligence in robots. For the above-mentioned high-grade human–computer interaction tasks, V-SLAM systems seek to realize dense 3-D reconstruction of concerned scenes by using ideas of

truncated signed distance function (TSDF) or by using a compact surfel-based model. Compared to such pixel-origin dense mapping, objects directly provide additional semantic constraints to the final consistent and dense 3-D maps. The combination of semantically ensembled strategies means an object-level enhancement to the power of pixel-level data association. When the geometrically expressed environments are given semantic labels, maps offer expressiveness. Additionally, it is important to appreciate that semantic SLAM exhibits higher applicabilities to more environments and distinguishes itself by not adding any extra sensors to systems.

This article illuminates the insights into the measure for semantics and the semantically constrained camera pose optimization. As powerful and eminently practical tools, certain deep learning ideas are highlighted. To the best of authors' knowledge, this is the first work that explicitly forms classifications for semantically meaningful descriptors and object-level constraining strategies in literature. Finding correspondences of the same semantic objects in consecutive frames is essential for semantics-aided localization, but there are currently no surveys that give a detailed overview on the topic "measure for semantics." This article fills this gap and helps researchers launch their efforts at efficient matching of semantic objects. Meanwhile, the existing semantic SLAM studies merely explore the object-level constraining strategies via different modules (such as VO, back-end, and loop closure detection). In contrast, this article organizes the strategies in terms of the correspondences between observed objects. By a collection of problem formulations and principle analyses, our review exhibits a fairly unique perspective in semantic SLAM.

The remainder of the article is organized as follows. The second section primarily describes a semantic extractor, which explains how the semantics are integrated into the existing V-SLAM frameworks. The methodologies pertinent to our explorations for object-level data association and semantically constrained pose optimization are summarized in paralleled Sections III and IV. The major contributions are critically reviewed, and this underpins our vision for future semantic reasoning. Section V further discusses the challenges of semantic uncertainty and the technology outlooks of semantic reasoning from a macroscopic view and attempts to find answers. Section VI draws conclusion.

II. Semantically Ensembled Localization and Mapping

Modern V-SLAM that carries semantic extractors can fulfill requirements for high-level perception and understanding [9]. Semantic observation is the core work. In fact, conventional V-SLAM approaches limit our measurements to primitive picture components such as points, line segments, and surfaces. This is primarily due to the fact that those geometrical primitives are processed by detectors and can be used as meaningful descriptions. However, object-level image understanding needs semantic observation [10]. Taking a specific example, for the extracted point features, via detecting and segmenting the commonalities of pixels, the points with a certain category help provide more constraints for the state extractor.

A. Semantic Extractors

A semantic SLAM system is constructed of two essential components: a semantic extractor and a modern V-SLAM framework. Inspired by deep learning techniques, various research groups have developed mechanisms for object detection, semantic segmentation, and instance segmentation. Those were successfully employed in semantic extractors.

1) Object Detection: The function of object detection is to acquire information related to the objects and spatial relationships and identify the category of each object by drawing the bounding box. Before deep neural network-based applications upsurge (AlexNet first ignited in 2012), conventional manual feature extraction approaches had to balance the computational complexity and efficiency. The desire for varieties of extracted features has motivated researchers to seek hand-crafted feature reinforcements [11]. Histogram of oriented gradient (HOG) feature descriptor and its extended version deformable part model (DPM) are the most representative tools in this age. After 2012, more and more object detection tasks employ the convolutional neural network (CNN) technique of region-based proposal extraction. Region-CNN series [12], [13] develops and enables superior region of interest (RIO) classification and object localization results. YOLO series [14] aims to reduce the computational requirements of anchor-based approaches. Their work has been widely applied to actual projects, such as civil and infrastructure engineering. Lately, anchor-free pipelines break the paradigm of "proposal extraction to nonmaximum suppression to candidate detection" and transform it into the solution of corner (or center) prediction. The state-of-the-art YOLOv8 [15] directly predicts the center of an object instead of the offset from a known anchor box.

Additionally, inspired by the "transformer" ideas, DEtection TRansformer [16] inferences the relations of the objects and the global image context, and directly outputs the matched categories and boundaries. This novel end-to-end object detection network distinguishes itself with bounding boxes or points-independent design.

2) Semantic Segmentation: Semantic segmentation is to assign each image pixel a specific label and is ascribed to pixel-level inference. Almost all the deep neural networks for semantic segmentation inherit the model from the fully convolutional network (FCN). In FCN, the fully connected layers are transformed into convolutional layers and this enables a classification network to output a heatmap [17]. Analogously, the typical FCN-based methods have major limitations. They suffer from the hierarchically lost spatial details and predefined fixed-size receptive fields. For the former, high-level semantic features may be lost due to the pooling (downsampling) process. The encoder–decoder networks [18], [19] have been developed to extract the features and recover the spatial size of the features obtained through upsampling. In particular, a class of lower resolution (corresponds to deep depth) and higher resolution (corresponds to shallow depth) fused solutions are proposed to enrich the spatial details [20]. For the latter, an object larger or smaller than the receiving field is likely to be encountered. Methods that use dilated convolutions to enlarge receptive fields [21], that efficiently capture local (pixel by pixel) and global (label by label) dependencies (image context) within a single model [22], and that adopt atrous spatial pyramid pooling (ASPP) to aggregate multiscale contextual information without losing resolution [23] have been well-developed.

So far, semantic segmentation cannot differentiate objects (instances) of the same category. Instance segmentation is the answer to this.

3) Instance Segmentation: Instance segmentation is essentially a further refinement of the former pixel classification. It allows for mask extraction of semantically segmented objects and therefore faces challenges in both accuracy and efficiency. There are several problems that confront the designer, including how to retrieve the lost details in small object segmentation, how to deal with the geometric change and image occlusion, and even how to address the image degradation resulting from the illuminated or compressed source images. The proposal-based approach is the baseline technique in this domain. It leverages the bounding box to generate a fine instance mask. Mask R-CNN is a simple and flexible framework, as an extension of the Region-CNN series, it builds on Faster R-CNN but adds an extra branch for regression and mask prediction [24]. Via the mask head that works in parallel with the classification, Mask R-CNN essentially decouples the mask prediction and category prediction. The real-time YOLO series seeks to split the mask prediction process into two concurrent modules (branches). The final, linearly combined branch outputs that construct a mask are respectively derived from a set of prototype mask predictions and a vector of mask coefficient predictions for each proposal [25]. By contrast, Mask R-CNN is ease of generalization with regard to other related tasks, and YOLO series continuously upgrades and can simultaneously meet object detection and instance segmentation.

YOLOv8 [15] has aimed at the optimized instance segmentation at the beginning of its design. Except for the first 3×3 conv in the stem and C2f in the building block, the anchor-free design with a decoupled head makes it a current state of the art in instance segmentation.

B. Semantically Ensembled V-SLAM

The above-optimized image classification results are apparently a contributor to a live SLAM process. Based on ORB-SLAM2, Fig. 1 shows a modern semantically ensembled V-SLAM (blue blocks are newly added). Anatomically, it carries a semantic extractor that can predict the movable properties of objects. For the dynamic objects fused SLAM process, the reliable masks help enable the dynamic information to be excluded from the pose estimation. With simple static object-only cases, the masks or bounding boxes can either be included in the loop closure detection thread to increase the recall rate or be included as extra constraints in BA to improve the positioning accuracy.

1) Semantics-Aided Localization: In ego-localization tasks, the use of semantics is similar to the process of making robots think, viz. robots will be more or less attentive to specific locations on the images. The proposed Attention-SLAM [7] simulates human attention mechanism and focuses are selectively placed on saliencies in the map. Analogously, structure PLP-SLAM [26] concerns scale-consistent 3-D structures and OVD-SLAM [27] pays less attention to dynamic locations. Additionally, 3-D geometric primitives like ellipses, cylinders, and cubes could be utilized and directly involved in pose estimation. We give no more expatiation here as this article provides more insights into the pose precision.

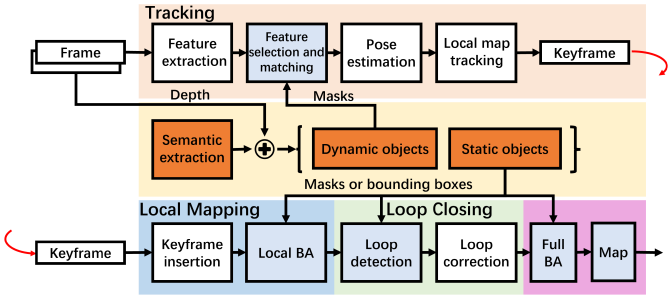

Fig. 1. Architecture of a modern semantically ensembled V-SLAM system. BA: bundle adjustment.

More theoretical supports and practical implementations of algorithms on semantics-aided localization can be found in Section IV.

2) Semantics-Aided Mapping: Dense maps are closer to the real world. The hypothesis of "highly static scenes" may suggest an undesired, continuous modeling of dynamic objects. In this domain, the flagship model DynaSLAM [28] successfully eliminates the potential dynamic objects via Mask R-CNN. In RGB-D mode, DS-SLAM [29] runs SegNet and motion consistency check in parallel. The constructed dense semantic 3-D octo-tree map adopts a log-odds score to filter out unstable voxels. Deep neural networks, such as SegNet and Mask R-CNN, are time-consuming. In the context of improving the adaptability of SLAM systems, lightweight pipelines CubeSLAM [30], QuadricSLAM [31], and EAO-SLAM [32] seek a real-time object-oriented mapping. Neural radiance fields (NeRFs) [33] is now pushing the prohibitive storage costs-denied mapping to a new level. The latest RO-MAP [34] and SNI-SLAM [35] can detect the object instances in real time and add them to the maps on-line. Lately, a novel idea of integrating text objects tightly into semantic mapping [36] catches and holds our concern. To better show the significant progress in semantic-aided mapping, we summarize the prominent online semantic map reconstruction schemes of the past five years in terms of model, method, map representation, and dataset, as listed in Table I.

III. Measure for Semantics

Object-level data association should fully use camera self-motion and 3-D geometric information to generate globally consistent semantic maps. Most existing algorithms solve for it by performing local low-level geometric feature matching, clustering, or statistical analysis to point clouds. We stress that the appearance descriptions of the concerned objects should by no means be ignored. Finding the right measures for semantics and efficiently matching these in consecutive image frames are two major concerns.

A. Measure for Object-Level Data Association

The process in which the front-end builds object-level landmarks and matches them over consecutive frames is termed "object-level data association." Multiple feature descriptors can be used to measure an object in terms of its geometric shape, appearance color, and spatial relationship. Motivated by the needs, fused descriptors could well apply to short-term association (e.g., feature matching) or long-term association (e.g., loop closure detection). We redirect the levels of

TABLE I
PROMINENT ONLINE SEMANTIC MAP RECONSTRUCTION SCHEMES

Model	Methods	Ref.	Year	Camera type	Map representation	Datasets
Cluster VO*	Multi-level probabilistic association mechanism + CRF	Huang *et al.* [37]	2020	Stereo	Point cloud	KITTI
Dyna-SLAM*	Instance semantic segmentation + ORB features	Bescos *et al.* [38]	2021	RGB-D/Stereo	Point cloud	KITTI
LIFT-SLAM	LIFT + Visual SFM	Bruno *et al.* [39]	2021	Mono	Point cloud	KITTI and EuRoc MAV
Sm-SLAM	Bag of words + YOLOv3	Qian *et al.* [8]	2022	RGB-D	Point cloud	TUM RGB-D
STDyn-SLAM*	SegNet + Optical flow + ORB features	Esparza *et al.* [40]	2022	RGB-D/Stereo	Octree	KITTI and EuRoc MAV
ObjectSDF*	MLP + SDF	Wu *et al.* [41]	2022	RGB-D	NIP	ToyDesk and ScanNet
SG-SLAM*	Epipolar constraints + Dynamic feature rejection strategy	Cheng *et al.* [42]	2023	RGB-D/Stereo	Octree	TUM RGB-D and Bonn RGB-D
SeMLaPS*	LPN + QPOS + SegConvNet	Wang *et al.* [43]	2023	RGB-D	3D occupancy map	ScanNet, SceneNN, and SMR
vMAP*	Depth guided sampling + Object mask prediction strategy	Kong *et al.* [44]	2023	RGB-D	NIP	TUM RGB-D, ScanNet, and Replica
RO-MAP*	EIF + Instance semantic segmentation + MLP	Han *et al.* [34]	2023	Mono	NIP	Cube-Diorama and Replica
SNI-SLAM	Hierarchical semantic representation + Fusion-based decoder	Zhu *et al.* [35]	2024	RGB-D	NIP	Replica, TUM RGB-D, and ScanNet
TextSLAM*	Locally planar characteristics + Deep learning-based semantic text feature extractor	Li *et al.* [36]	2024	RGB	3D text map	Real-world sequence

CRF: Conditional Random Field; ORB: Oriented FAST and Rotated BRIEF; LIFT: Learned Invariant Feature Transform; SFM: Structure from Motion; MLP: Multilayer Perceptron; SDF: signed distance function; NIP: Neural Implicit Representation; LPNs: Latent Prior Networks; QPOS: Quasi-Planar Over-Segmentation; SegConvNet: Segment-Convolutional Network; EIF: Extended Isolation Forest.The open-source models are marked with an asterisk. A click on the asterisk directs the readers to the open-source code.

descriptors: ① intersection over union (IOU), bag of words (BOW), HSV histogram, and embedding vector that concern partial information of an image; ② Global semantic understanding that perceives an entire image; ③ Object center metric, 3-D overlap ratio, and neighborhood topology that understand multiimages; and ④ hierarchical environmental representation that supports multiimage reasoning. Fig. 2 presents the diagrammatic interpretation of the first-level descriptors.

1) IOU, BOW, HSV Histogram, and Embedding Vector: As in Fig. 2, the positions of the car in these two consecutive frames change slightly, thus the IOU (overlap of two 2-D bounding boxes) is high. BOW converts the car's description into a vector based on the low-level key points (essentially, ORB feature points) and a pretrained visual vocabulary. It contributes to object appearance similarity measuring. The matched HSV color histogram concerns the car's mask area and is closer to human visual perception. The lower right embedding vector is the result of deep learning-based coding of the car's mask or bounding box. By toolbox FastReID [45], for instance, the distance of embedding vectors is calculated for object similarity checking.

2) Global Semantic Understanding: Rather than leveraging pixel-level image details for diversity detection, global semantic understanding seeks to extract the semantics of an entire image (not limited to the semantics of individual objects). This makes it more valid for long-term data association. Unsupervised deep networks like generative adversarial networks (GANs) and BigBiGAN, deep encoder–decoder frameworks, and other ConvNets are powerful tools for the comprehension of an entire semantic image [46].

TABLE II
PERFORMANCE COMPARISON ON DIFFERENT BENCHMARK DATASETS

Model		Major contribution	Performance (maximum recall at 100% precision)				
			City Center	New College	KITTI-00	KITTI-05	KITTI-06
Supervised learning	NetVLAD* [47]	Pool convolutional features into a trainable VLAD pooling layer	86.34	85.17	96.72	87.90	95.50
	FILD++* [48]	Extract global convolutional features to recomment potential loop closures to be evaluated	90.01	82.37	94.92	**95.42**	**98.16**
	LoopNet [49]	Use attention mechanism to reduce the weight of moving objects	89.15	84.62	•	•	•
Unsupervised learning	CALC2.0* [50]	Construct a descriptor of both the appearance and semantic layout of an image	84.47	81.32	97.25	82.24	97.54
	PlaceNet [51]	Augment the encoder network PlaceNet with a semantic fusion layer	**92.50**	**90.66**	**98.50**	92.46	98.14

•: The corresponding model was trained on that dataset. The open-source models are marked with an asterisk. A click on the asterisk directs the readers to the open-source code.

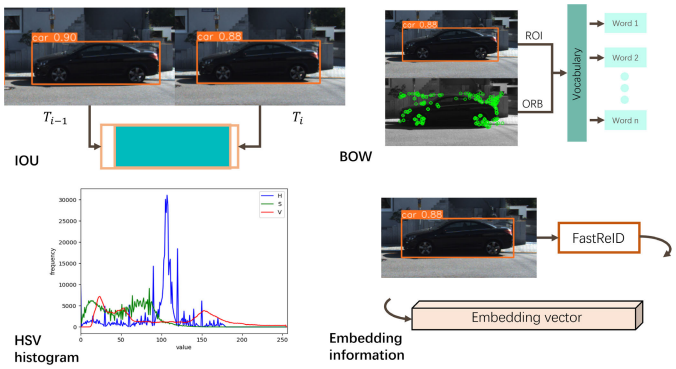

Fig. 2. Diagrammatic interpretation of first-level descriptors.

To improve the performance analyses of various networks in global semantic understanding, Table II demonstrates a summarized comparison on several publicly available benchmarking datasets. The networks are divided into two categories in terms of the machine learning mode (more specifically, supervised/unsupervised) and we should focus on the models that have experimented on the datasets. As benchmarking standards for loop closure detection algorithms, City Center dataset faces partial occlusion and unstable shadow features in some scenes, while New College dataset covers large regions that include challenging identical archways. KITTI-00, KITTI-05, and KITTI-06 (sequences that contain more loop closures in the KITTI vision suite) face significant viewpoint changes caused by different trajectories. As in Table II, the quantified performance is illustrated by the maximum recall at 100% precision. The optimal values are marked in bold. Metric recall is the proportion of correct loops retrieved from all actual loops, and clearly high recall returns the majority of all positive results.

Training deep CNN models on information-rich datasets facilitates high-level abstractions of input images. NetVLAD [47] introduces a differentiable VLAD layer to cluster and aggregate scene elements into a vector (as an image descriptor). It outperforms traditional nonlearned image representations and off-the-shelf CNN descriptors. FILD++ [48] constructs an incremental database using the global features to recommend potential loop closures. The architectural constructions appearing in the concerned benchmark datasets contribute to more representative deep feature extraction. As we show in Table II, FILD++ yields a very competitive performance compared to the state-of-the-art models. Mostly owing to the multiscale attention mechanism, LoopNet [49] specializes in solving loop closure detection in dynamic scenes and therefore performs satisfactorily in City Center (it features many dynamic objects such as cars and pedestrians in urban areas). Meanwhile, New College covers large regions with intense repeated structures and as a direct consequence of this, the ability of the attention mechanism in LoopNet to capture similarities is restrained.

To avoid wasting time using pixel-level supervision, more unsupervised models are employed in visual place recognition tasks. In CALC2.0 [50], a network comprising a semantic segmentator, variational autoencoder, and a triplet embedding network is built. Training the network in an unsupervised way, researchers construct a robust holistic-image descriptor describing both the visual appearance and semantic layout of an image. As shown in Table II, it achieves comparable results with NetVLAD. The novel PlaceNet [51] is a deep autoencoder network augmented with a semantic fusion layer for scene understanding, which generates semantic-aware deep features that are robust to dynamic scenes and scale invariant. As opposed to the state-of-the-art, PlaceNet shows optimal or suboptimal results on all datasets and demonstrates its ability

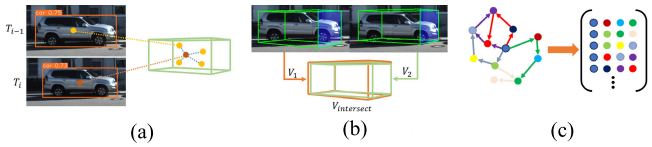

Fig. 3. Graphic interpretation of third-level descriptors. (a) Object center metric. (b) 3-D overlap ratio. (c) Neighborhood topology.

to perform robustly and consistently in various scenarios. To conclude, as an advanced plug-and-play model, PlaceNet is expected to be a great reference for global semantic understanding.

3) Object Center Metric, 3-D Overlap Ratio, and Neighborhood Topology: The third-level descriptors are identically illustrated by the centers, 3-D overlaps, and topologies in Fig. 3.

a) Object center metric: In principle, solving for object center metric consists of initially generating object landmarks incrementally and further narrowing the region of the center down with new observed angles. As in Fig. 3(a), a set of previous object centers $C = \{c_i\}$ restrict the range of the car's position, and it is readily understood that a center of the temporary point cloud c (in brown) in Cartesian space closes to more than one previous center (in yellow). The distance between c and the closest c_i in C is calculated as the object center metric, as shown as follows:

$$\text{center}(C, c) = \min(\|c_i - c\|) \quad \forall c_i \in C \tag{1}$$

where $\|\cdot\|$ denotes Euclidean distance.

b) 3-D overlap ratio: The purpose of the 3-D overlap ratio metric is to check the geometric consistency of objects in Cartesian space. As with IOU, the overlap ratio of two bounding boxes is calculated. In this higher level data association, 3-D bounding boxes that provide occupied space information are investigated. As in Fig. 3(b), for the two 3-D bounding boxes of the car, the max overlap ratio equals

$$\text{overlap}_{3D}(V_1, V_2) = \max\left(\frac{V_{\text{intersect}}}{V_1}, \frac{V_{\text{intersect}}}{V_2}\right) \tag{2}$$

where V_1 and V_2 are the volumes of the two 3-D bounding boxes. $V_{\text{intersect}}$ is the volume of the intersection.

c) Neighborhood topology: The neighborhood topology of objects explores the distribution of other objects around the detected object (the car, for instance) and their relative spatial relationships. Especially for multiobject environments, checking the topologies during the association and matching processes helps eliminate the ambiguities. The built topology matrix in terms of random walk descriptors is shown in Fig. 3(c). In essence, the associated objects form an undirected graph, where the vertice denotes the center position and the weights of edges are set as the spatial distances between centers of two linked objects. Such a topology matrix matching approach of finding correspondence objects based on semantic similarity is generally to be preferred to loop closure detection tasks.

4) Hierarchical Environment Representation: The main difficulty in perceiving the scene is the enormous number of possible candidate descriptions in which the system might be interested. The hierarchical environment representation is the advanced version of environment topology. Given a

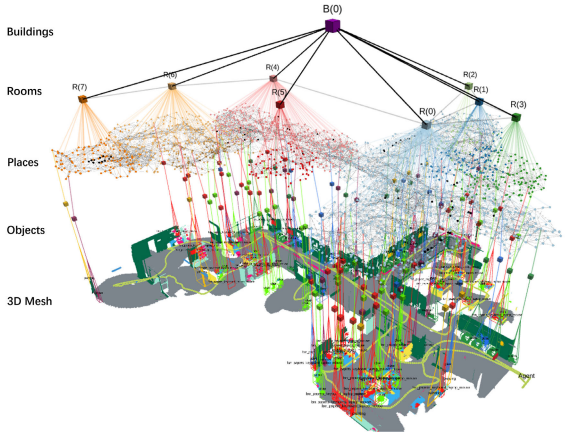

Fig. 4. 3-D scene graph created by Hydra [56]. Hydra: a suite of algorithms to build a 3-D scene graph in real-time.

room schema, a 3-D layered graph is generated through a combination of detecting some semantic entities like walls, bookshelves, and seats, and adding topological constraints to the identified entities by labeling or neural networks [52]. The bottom layer presents the individual semantic entities whose commonalities suggest a study in the context of a room schema. The top layer is a building node connected to all subgraphs of layer "Rooms" (see Fig. 4).

The nodes in Fig. 4 represent spatial concepts at multiple levels of abstraction (from a low-level geometry termed 3-D mesh to objects, places, rooms, and buildings) and edges represent relations between concepts. Long-term data association is obviously one of the important applications of this enhanced environmental map representation. When performing loop closure detection, we walk down the hierarchy of descriptors (from places to objects, to appearance descriptors). This optimizes the searching range for extremely time-consuming feature matching.

B. Related Schemes

We summarize the major, recent studies that investigate object association via previously discussed feature descriptors, as listed in Table III. Quite clearly, the GAN-based methods that concern BOW or full graph semantics could well apply to place recognition and loop closure detection [46], [53]. For object-aware data association, the cross-level measuring schemes are more comprehensible. Liu et al. [55] perform a short-term object tracking by 2-D IOU to associate 2-D objects and further aggregate appearance metrics (3-D overlap ratio, object center) to seek corresponding 3-D objects in the map. HSV histograms are also included in their descriptor summation. Note here, that the efficient HSV appearance descriptor suffers from the objects that have similar appearance especially color, thus a solution by HSV histogram, directly, is not possible. Lin et al. [57] and Ji et al. [1], respectively, fulfill accurate matching of objects by fusing HSV histogram with neighborhood topology and embedding vector. Zins et al. [58] leverage object landmarks and 2-D IOU metrics in their camera relocalization tasks. Lately, high-level scene graph that describes the environment as a layered graph has gained more attention. Via aggregated BOW, topological maps of places, and hierarchical descriptors, Hughes et al. [56] manage to build 3-D scene graphs that can serve as an

TABLE III
DESCRIPTIONS OF OBJECT-LEVEL ASSOCIATION SCHEMES IN TERMS OF SEMANTICALLY MEANINGFUL DESCRIPTORS

Task	Ref.	Year	Level 1				Level 2	Level 3			Level 4
			IOU	BOW	HSV histogram	Embedding vector	Full graph semantics	Object center metric	3D overlap ratio	Neighborhood topology	Hierarchical graph
Indoor loop closure detection	Zhong *et al.* [46]	2021					✓				
Large-scale open scene-oriented loop closure detection	Yang *et al.* [53]	2021		✓							
Object-level data association	Qian *et al.* [54]	2021		✓							
Object-aware data association	Liu *et al.* [55]	2022	✓		✓			✓	✓		
Loop closure detection and optimization in 3D scene graphs	Hughes *et al.* [56]	2022		✓						✓	✓
Monocular semantic SLAM	Lin *et al.* [57]	2022			✓					✓	
Object-aided camera relocalization	Zins *et al.* [58]	2022	✓								
Object-level data association	Ji *et al.* [1]	2023	✓		✓	✓					
Marker-based loop closure detection	Touran *et al.* [59]	2023									✓
Clusters-based place recognition	Scucchia *et al.* [52]	2024									✓

advanced "mental model" for robots. A top–down loop closure detection that captures statistics across layers is simultaneously argued. Additionally, from aspects of adding more topological constraints to the detected semantic entities, fiducial markers and clusters are employed to generate practical, hierarchical graphs [52], [59].

In general, solving object association across frames and further finding correspondences of the same semantic objects in consecutive frames, as far as possible, provides a basis for semantically constrained pose optimization.

IV. SEMANTICALLY CONSTRAINED POSE OPTIMIZATION

In the back-end, SLAM problem can be modeled as a factor graph $G(V, E)$ where the vertices V represent the variables to be optimized such as robot poses and points in 3-D, and the edges E exert the constraints (factors) between two linked vertices.

Object-level semantic information establishes extra constraints for pose optimization. It is possible to parameterize object poses as optimizable variables or utilize object masks to constrain the reprojection process. In this section, the research progress of optimization strategies is summarized in terms of the correspondences between observed objects. If the object entities are independent of one another in the optimization process, they are referred to as independent semantic constraints. The associated semantic constraints are ones whose object-aware correspondences are explored and included in BA formulation.

A. Independent Semantic Constraining Strategy

1) Weighted Feature Points: Robust, reliable, and salient cues help reduce the uncertainty of semantically constrained pose estimation. Attention-SLAM [7] highlights the importance of feature points extracted from salient regions (see Fig. 5) and models a weighted BA process. As shown in Fig. 5(b), the white parts on the saliency mask indicate higher attention. The heatmap that combines the original image and the saliency mask leads to better visualization. WF-SLAM [60] defines feature point weights and initializes them with dynamic target detection results. PROB-SLAM [61] adjusts the optimization weights of feature points distributed on dynamic objects so that the dynamic SLAM system is more stable and less affected by moving objects. Following the idea of filtering the dynamic points that lead to significant pose drift, in pose optimization, OVD-SLAM [27] calculates an optimization weight for every map point. In particular, its pose refinement process also concerns nonrigid motion and static point recovery under low-dynamic environments. Generally, the weighted BA can be formulized as

$$e(x) = w\|x - \pi(\boldsymbol{R}X + \boldsymbol{t})\|^2 \tag{3}$$

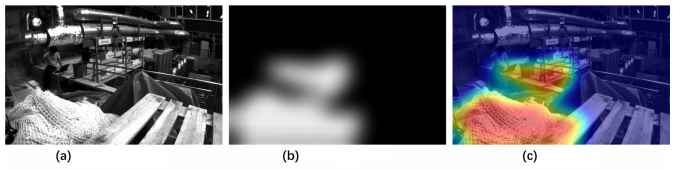

Fig. 5. Generated saliency mask and heatmap corresponding to the EuRoc MAV dataset. (a) Original image. (b) Saliency mask. (c) Heatmap.

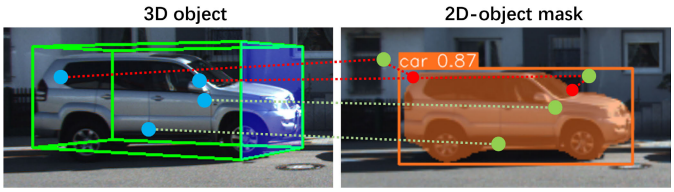

Fig. 6. Concerned points that lack semantic consistency.

where $e(x)$ is the defined reprojection error. w is the weight of the map point x (in the image coordinate). $\pi(\cdot)$ is the projection transform from the world coordinate system to the image coordinate system. $\boldsymbol{R}$ and $\boldsymbol{t}$ are rotation matrix and translation vector to be jointly optimized. X is the matched 3-D point of x (in the world coordinate).

This subset of SLAM is closely related to the studies that leverage semantic entities as constraints to solve dynamic problems. The research of probabilistic semantics simultaneously examines problems in semantics-induced uncertainty property. As previously discussed, dynamic objects should not be purely regarded as outliers and removed from the process. An advanced idea is to construct a motion estimation probability model and remove those unreliable features with high motion probability from the local map, as in Dyn-ORB-SLAM [62]. The uncertainty of deep learning models should be given enough attention as well [63]. The future work that tightly organizes deep learning (for semantic segmentation) and dynamic SLAM without incurring great runtime costs will receive extensive attention from researchers.

2) Semantically Constrained Points: Liu et al. [55] initiates the research of leveraging object masks to constrain the reprojection process. This proposal builds on the fact that semantic information remains consistent for the variance of viewpoints and observed scales. For the points that lack semantic consistency, more constraints are needed (as in Fig. 6, the red dotted lines show that some point landmarks of a 3-D object are not observed in the matched 2-D-object mask). A semantic residual e_{residual} denoted by the distance between the reprojection pixel and its nearest pixel belonging to the 2-D-object mask is defined. It follows that:

$$e_{\text{residual}} = \left\| z^s - \pi(p, \boldsymbol{T_C}) \right\|_{\Sigma}^2, \quad z^s = h(\pi(p, \boldsymbol{T_C}), m) \tag{4}$$

where z^s is the geometric semantic point observation. $\pi(p, \boldsymbol{T_C})$ is the projection function to project a 3-D point landmark p to the image using the camera pose $\boldsymbol{T_C}$ and the camera intrinsic parameters. Σ is the covariance matrix which represents the uncertainty related to the feature extraction. $h(z, m)$ is a function to search the nearest pixel valued 255 in m to z.

In this semantic optimization-type SLAM community, the earlier eminent work that enables continuous tracking of points using semantics is VSO [64]. Differing from using the aided object masks, VSO utilizes multiple fixed points to optimize camera pose and successfully reduces drift through medium-term data associations. However, this remains equivalent to a collection of pixel's class labels and therefore carries no segmented instance information. As discussed in Section II-A, deep learning bridges the gap in segmentation between different instances—rather than modeling class-specific uncertainties, incorporating such semantic mask constraining points into a SLAM pipeline must give great impetus to the studies of medium-term object-level association.

3) Object Bounding Box Representations: Another focus of semantically constrained pose optimization is on building object-level landmarks in a regular geometric shape. The backbones CubeSLAM [30] and QuadricSLAM [31] constitute the sources from which EAO-SLAM [32] derives its object descriptors. Given scenes in the TUM RGB-D dataset, cubes and ellipsoids are leveraged in place of so-called instance-level models and category-specific models. The jointly optimized object $O = \{O_i\}$ and camera poses $\boldsymbol{T_C} = \{\boldsymbol{T_{C_j}}\}$ yields

$$\{O, \boldsymbol{T_C}\}^* = \underset{\{\theta_y, \boldsymbol{s}\}}{\arg\min} \sum (e(\boldsymbol{\theta}) + e(\boldsymbol{s})) + \underset{\{\boldsymbol{T_C}\}}{\arg\min} \sum e(\boldsymbol{p}) \tag{5}$$

where $\boldsymbol{\theta}$ is the orientation, θ_y is the initialized yaw, and $e(\boldsymbol{\theta})$ is the object pose error. $\boldsymbol{s}$ is the half of side length of 3-D object and scale error $e(\boldsymbol{s})$ is defined as the distance between the projected edges of a cube and their nearest parallel line segment detector (LSD) segments. $e(\boldsymbol{p})$ is the commonly used point reprojection error in typical V-SLAM frameworks.

More research work regarding the same is now improving. ClusterVO [37] simultaneously optimizes the poses of the camera and the motion of the surrounding rigid clusters/objects. In SQ-SLAM [65], superquadrics with shape parameters are used for object representation. Lee et al. [66] develop a pose ambiguity-aware object SLAM system in the presence of symmetric objects. The goal of uniformly representing much higher level primitives, like planes and quadrics is destined to stimulate the growth of this branch of semantic SLAM.

B. Associated Semantic Constraining Strategy

An associated semantic constraining occurs when object-aware correspondences between two or more objects (more precisely, *Object-Wall* or *Object-Ground*) are mined. Accepting certain environment assumptions, objects must form a geometric relationship with spatial structures. The spatial relationships between objects and structures, therefore, can be used as auxiliary constraints.

1) Object-Plane Occlusion Constraints: Yang and Scherer [67] propose a weaker but more general constraint that indicates objects should not be occluded by nearby planes in the camera view. This constraint is defined as the sum of 3-D corners' signed distance to the plane $\boldsymbol{\pi}$, viz.,

$$e_{\text{occlusion}} = \sum_{i=1:8} \max(0, -\boldsymbol{\pi} P_i) \tag{6}$$

where P_i is one of the eight cuboid corners. If the cuboid lies on the positive side of the plane, $e_{\text{occlusion}} = 0$.

2) Object-Plane Supporting/Tangency Constraints (for Quadrics): Hosseinzadeh et al. [68] hold the idea that almost all stable objects in the scene are supported by structural entities of the scene. They pioneered the study of online structure-aware SLAM that leverages a supporting/tangency constraint. A relationship is imposed between a quadric object and a structural infinite plane

$$e_{\text{tangency}} = \left\| \boldsymbol{\pi}_h^{\mathrm{T}} \boldsymbol{Q}^* \boldsymbol{\pi}_h \right\|_\sigma^2 \tag{7}$$

where $\boldsymbol{\pi}_h$ is the normalized homogeneous plane that supports the quadric $\boldsymbol{Q}^*$. σ is the tangent covariance of Mahalanobis norm $\|\cdot\|_\sigma$.

Analogously, SO-SLAM [69] includes a plane supporting constraint in its spatial constraint summation. Given that the X and Y axes of the object are orthogonal to the normal vector $\mathbf{n}$ of the supporting plane $\boldsymbol{\pi}_s$, the supporting relationship can be described as

$$e_{\text{tangency}} = \left\| \boldsymbol{\pi}_s^{\mathrm{T}} \boldsymbol{Q}^* \boldsymbol{\pi}_s \right\|_\sigma + \left\| \text{Rot}_X(\boldsymbol{Q}^*) \cdot \mathbf{n} \right\|_\vartheta + \left\| \text{Rot}_Y(\boldsymbol{Q}^*) \cdot \mathbf{n} \right\|_\vartheta \tag{8}$$

where $\text{Rot}_X(\boldsymbol{Q}^*)$ and $\text{Rot}_Y(\boldsymbol{Q}^*)$ are respectively the X- and Y-axis normal of the quadric $\boldsymbol{Q}^*$. σ is the tangent covariance from (7). ϑ is the rotation covariance.

3) Object-Plane Supporting/Tangency Constraints (for Cuboids): Zhou et al. [70] explore object-plane supporting/tangency constraints with chief reference to object surface planes. As research into SLAM optimization, they explicitly extend the idea of object-plane measurement in a unified BA framework. Given an object represented by a cuboid, they propose that the object surface planes that belong to the object itself must be close to the cuboid proposal. More precisely, if a plane $\boldsymbol{\pi}$ is associated with the object, it should have a similar angle and minimum distance to one of the cuboid surface planes. The following constraint (factor) is introduced

$$e_{\text{tangency}} = \min(q(\boldsymbol{\pi}) - q(\boldsymbol{\pi}_{oi})), \quad q(\boldsymbol{\pi}) = (\phi, \psi, d)^{\mathrm{T}} \tag{9}$$

where ϕ and ψ are the azimuth and elevation angle of the plane normal, respectively and restricted to $(-\boldsymbol{\pi}, \boldsymbol{\pi}]$ to avoid singularities. d is the distance from the Hessian form. $\boldsymbol{\pi}_{oi}$, $i \in \{1, 6\}$ denotes six surface planes of the concerned cuboid.

4) Plane–Plane Constraints (Manhattan Assumption): For a Manhattan world in which planes are mostly mutually orthogonal or parallel, imposing constraints on relative plane orientations is simply a matter of introducing a factor on the plane surface normals. Constraints between planes $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ with unit normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$, respectively, are implemented as

$$e_{||}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \left\| \left| \mathbf{n}_1^{\mathrm{T}} \mathbf{n}_2 \right| - 1 \right\|_\sigma^2, \quad \text{for parallel planes} \tag{10}$$

$$e_{\perp}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \left\| \mathbf{n}_1^{\mathrm{T}} \mathbf{n}_2 \right\|_\sigma^2, \quad \text{for perpendicular planes} \tag{11}$$

where σ is the tangent covariance from (7).

The theme of this branch of semantic SLAM studies lies in the introduction of object spatial constraints, as semantic priors, to improve the robustness and accuracy of object parameters and camera pose estimation. On the one hand, the studies offer excellent potential for enhanced, especially YOLO-based object detection designs. On the other hand, man-made environments contain many objects that are potentially used as landmarks in SLAM maps. People will have better access to the estimated, entities representing landmarks. Those are important clues for the desired lightweight semantic maps. For the plane landmarks discussed above, whose role is to encapsulate the high-level structure of regions. For cuboids in CubeSLAM works (parameterized in a 9 DoF form), the captured key properties in perceiving a scene include size, location, and orientation.

By far, the dug spatial constraints and semantic priors of objects are far from adequate. When referring to some higher level human–robot–environment interaction applications such as autonomous grasping of robots, even though performed almost unconsciously by humans, this grasping task, however, performed by robots requires semantics for modeling states of the world, and for describing the process of change from one world state to another. Semantic attributes related to the human world (such as room types, objects, and spatial layouts) are considered to be essential attributes for future robots.

To facilitate the comprehensiveness of this review study, we would like to present the performance assessment of the introduced SLAM algorithms explicitly. These algorithms are divided into four categories, each corresponding to a technical node, i.e., "weighted feature points," "semantically constrained points," "object bounding box representations," or "associated semantic constraining."

In fact, the datasets used to compare the algorithm performance are quite different since these algorithms have different focus. Given publicly available, unified experiments, Table IV demonstrates a comparison of a few introduced algorithms on the TUM RGB-D benchmark dataset by a metric absolute pose error (APE). APE is also called absolute trajectory error (ATE) in literature, which reflects the global consistency of SLAM systems. The root-mean-square error (RMSE) of APE is used as the main evaluation criterion. The minimum errors are marked in bold.

As in Table IV, OVD-SLAM [27], WF-SLAM [60], and Dyn-ORB-SLAM [62] are studies that leverage semantic entities as constraints to solve dynamic problems, so we see the data under the targeted dynamic environments that contain four high-dynamic and two low-dynamic sequences are given. For the high-dynamic sequences whose names start with "fr3/w/," they reflect the robustness of VSLAM and odometry algorithms to quickly moving dynamic objects in large parts of the visible scene. For the low-dynamic sequences, the robustness is evaluated while the dynamic objects move slowly. Among these three algorithms, OVD-SLAM restores static points on moving objects during its two-stage pose estimation and thus enables adaptability to low-dynamic environments. As was originally assumed, it achieves optimal or suboptimal results on sequences "fr3/s/half" and "fr3/s/xyz." WF-SLAM tightly couples semantic segmentation and geometric motion segmentation and builds a "pose-and-weights" joint optimization model, making itself well suited for dynamic target detection purposes. As we can see, it achieves optimal or suboptimal results on all dynamic sequences. For Dyn-ORB-SLAM, as it does not combine the probabilities of map points in BA optimization, it can still function as usual.

For the static sequences in Table IV, except for "fr3/teddy," typical office scenes are recorded. CubeSLAM [30] is

TABLE IV
ALGORITHM PERFORMANCE COMPARISON ON THE TUM RGB-D DATASET (IN TERMS OF RMSE OF APE, UNIT: cm)

Algorithm	Sequences											
	fr1/desk	fr1/desk2	fr1/xyz	fr2/desk	fr3/teddy	fr3/cabinet	fr3/w/xyz	fr3/w/half	fr3/w/static	fr3/w/rpy	fr3/s/xyz	fr3/s/half
OVD-SLAM [27]	-	-	-	-	-	-	1.35	**2.29**	**0.68**	3.49	0.90	**1.66**
WF-SLAM [60]	-	-	-	-	-	-	**1.27**	2.44	**0.68**	**2.52**	**0.87**	1.71
Dyn-ORB-SLAM [62]	-	-	-	-	-	-	2.36	5.11	0.78	10.46	1.19	1.87
Object-aware data association scheme [55]	1.60	**2.20**	1.00	1.10	**0.80**	-	-	-	-	-	-	-
CubeSLAM [30]	-	-	-	-	-	16.13	-	-	-	-	-	-
QuadricSLAM [31]	1.66	2.45	-	3.27	2.63	11.85	-	-	-	-	-	-
EAO-SLAM [32]	1.59	2.44	-	2.60	2.79	-	-	-	-	-	-	-
SO-SLAM [69]	-	-	-	-	-	**8.16**	-	-	-	-	-	-
Structure Aware SLAM [68]	**1.40**	-	**0.90**	**0.87**	-	-	-	-	-	-	-	-

-: The algorithm did not experiment on that sequence.

evaluated on sequence "fr3/cabinet" (an office pedestal is recorded). This pedestal has little texture and structure, but has rectangular corners and flat surfaces. Most monocular SLAM algorithms fail on it, while CubeSLAM achieves a fairly good camera pose estimate. QuadricSLAM [31] is outperformed by EAO-SLAM [32] over the majority of sequences due to the limited number of semantic objects, while this quadrics-based algorithm shows a slightly improved pose estimate in the case of the "fr3/teddy." During the recording of "fr3/teddy," the Asus Xtion sensor was moved around a teddy bear in two rounds at different heights. The idea of building object-level landmarks in a regular geometric shape is indicated in (5). In this regard, it is worth mentioning that the teddy bear has soft fur but no regular structure. We notice that [55] yields a significant improvement of 69.58% compared to QuadricSLAM on the sequence "fr3/teddy." One object makes it difficult for SLAM to establish semantic constraints. However, the teddy bear has abundant textures (it wears a smooth shirt) which are repeatedly observed and exploited by this object-aware SLAM system. Additionally, as we explain earlier in Section IV-B, planar surfaces provide additional constraints and permit affordance constraints between objects and their supporting (or occlusion) planes. Consistent with this idea, it is seen that the SO-SLAM [69] and the structure-aware SLAM [68] that build on QuadricSLAM but encapsulate the high-level structure of regions greatly improve the localization performances. We believe the research frontier that imposes constraints between semantic entities will further facilitate pose optimization and lead to semantically meaningful maps.

V. DISCUSSION

In the former sections, the issues related to measuring semantics and intended pose optimization are investigated. The accurately judged object colors, sizes, spatial relationships, and measurements are critical to the estimated camera trajectory, whereas, we believe "semantic reasoning" underpins successful robot deployment in domains where the robot attempts to understand a previously unseen environment. This entire section will be centered around semantic reasoning, mostly following the "concept to property to discussion" lines of thought. A part-by-part description follows.

A. Concept of Semantic Reasoning

Inspired by the concept of deriving new knowledge from known facts in artificial intelligence (AI) technology of knowledge graph, we refer to semantic reasoning as a rule-based or logic-based deduction process structured by a semantic priors-to-constraints (or conditions) mining backbone. The dug words and symbols are easier to understand in the area of

question answering, while we believe the so-called semantically meaning information termed labels in SLAM has an identical nature to the concept of reasoning. A large set of rules can be found in the pyramid-like room schema, as in Fig. 4. They explicitly define for instance, that a kitchen is a room and it has stoves, which immediately allows the robot to use this knowledge to perform several types of inferences. The logic-based deduction refers in particular to the constraint mining that builds on spatial relationships of semantic entities, where the set supporting/tangency planes are semantic priors.

B. Challenges of Semantic Uncertainty

Research on semantic information mining and semantic reasoning has led to our exploration into the semantic uncertainty-solving problem. In essence, the uncertainty is introduced by the semantic detection networks. Reducing its impacts on SLAM performance is of great importance. The research sides that we are concerned with can be summarized as follows.

1) In Deep Learning Model Uncertainty Scale: This side of uncertainty solving represents to what extent we trust our learning models. The learning models are used to tackle the object-level data association and their implementations are complicated by the changes in viewpoints, illumination, and scene dynamics. In addition to the probabilistic formulation of SLAM (find [61] and [62] in technical node "weighted feature points,") to capture the uncertainty of deep models, we believe the idea of augmenting learning models into a Bayesian model [63] is constructive. Estimating Bayesian uncertainty contributes to both motion tracking and scene understanding. Given per-pixel uncertainty estimation, the factors termed belief metrics can also be used to guide the robot to determine if certain scene structures should be trusted.

2) In Outlier Elimination Scale: This side of uncertainty solving is equivalent to addressing the subsequent effects of the applied deep neural networks. For the object detection and semantic segmentation results, assume that the extracted bounding box is assigned an object label while it contains the background points or the extracted 2-D-object mask is recognized as a car while it contains the points that are not on the car body (as in Fig. 6), those points are outliers and they lead to wrong data association. To eliminate them, updating object-level landmarks with statistics (e.g., chi-square test and T-test) and clustering (e.g., DBSCAN [71]) algorithms is an option, and performing a semantic consistency check on the points is another. In addition to these, transforming the semantic binary object detection of boundaries into probability results is an effective means (as in [64]), which helps us select the maximum value of the probability values calculated in each bounding box.

3) In Wrong Detection Remedy Scale: This side of uncertainty solving is specifically oriented to the removal of object outliers induced by the wrong detection. In the area of object-level SLAM, most existing works assume that the detection is correct or artificially set some rules to eliminate the wrong detection. As in [32], the rule sets cover the number of overlaps, the defined object categories, the observation times of 3-D objects that have constructed objects. By contrast, building a unified framework to remove object outliers is more reasonable. A novel constraining mechanism that concerns spatiotemporal consistency has been discussed in [72]. We may conclude that the constraints that can robustly reflect the semantic spatial relationships among objects need to be mined. We will give a macroscopic discussion of this in Section V-C.

C. Technology Outlooks of Semantic Reasoning

This section is devoted to an exploration of semantic reasoning's technology outlooks. Special attention has been paid to 3-D scene graphs, object-aware correspondences, and motion reasoning.

1) High-Level Representation Reasoning of a Layered Graph: Hughes et al. [56] define 3-D scene graphs as an advanced "mental model" for robots. Building a rich, layered scene graph as in Fig. 4 in response to loop closures is quite difficult. From our point of view, a skillful method is to split the space utilizing certain object entities such as walls (suppose a robot enters a room through a doorway, it activates a room schema). Furthermore, a semantic attribute initialization of the skeletons of the environment can be executed based on region assumptions supported by those known object entities (e.g., a room with a refrigerator might suggest a kitchen). We present a simplified model with rooms boxed in by partitioning walls (see Fig. 7). The border effect is created by a network such as PlaneRecNet. Once the robot went right through the borders (walls), the new layered scene graphs were generated.

Walls are dominant planar surfaces in man-made environments. Essentially, this method is the exact dual of creating constraints. It leads to constraining the loop closure detection-targeted layered searching range. In addition, the derived semantics regarding regions will provide more priors for robot environment explorations that are previously thought to be built on hypothesize-and-test paradigms.

2) Potential Relationship Reasoning Between Semantic Entities: Generally speaking, mining semantic relationships is simply a matter of refining pose optimization. Understandably, the *Car-Road* supporting/tangency constraining (discussed in Section IV-B) simplifies a 6-D pose transformation to a 3-D concept. However, previous relationship reasoning efforts have mostly relied on a database of predefined relationships. *Point-Plane*, *Plane-Plane* (Manhattan), and *Object-Plane* constraints that are amenable for use in a least-squares framework have been well-described. We argue there would be more useful affordance constraints between entities to be found. The potential relationship reasoning can occur, for instance, when a function relation (kitchen hood installed over the cooking stove), a depth relation (collision between objects), or a scenario relation (office layout with desktops, and chairs) are recognized. All these clues among semantic entities increase the accuracy of correspondence to be established.

Additionally, introducing these useful affordance constraints can make the scene map representation remain compact. Given a fixed viewpoint, the depth of adjoining objects, as the fused semantic prior information, can also serve as the candidate for scene descriptions. We are firmly convinced that the potential relationship reasoning will be a powerful tool for the solution of semantic prior problems in a semantically constrained pose optimization context.

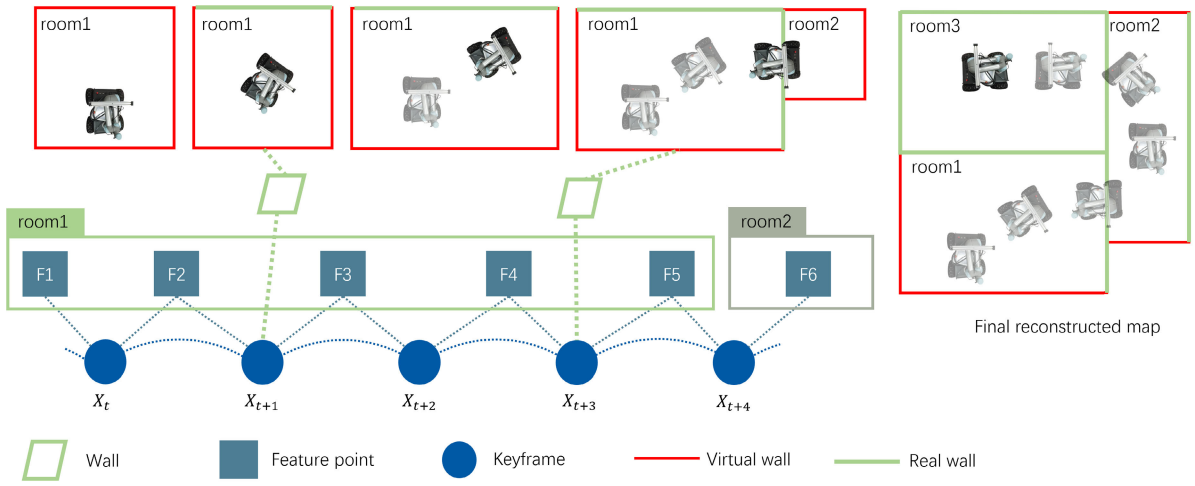

Fig. 7. Room division by partitioning walls.

3) Motion Reasoning in Terms of 4-D Spatio-Temporal Maps: The third question considers motion reasoning when the measurements of pure static landmarks cause performance degradation in dynamic scenes. In the dynamic object-aware SLAM domain, some efforts have been made to extract object semantics from the current frames, whilst, the linking of previously found semantics tends to be ignored. However, we do not share such a routine. We believe that rigid object entities increase the accuracy of motion inference. The motion constraints reflected by the motion consistency on rigid structures of dynamic object bodies, therefore, should be highlighted. This allows for a 4-D spatio-temporal map concept, implying that the 3-D structure of an articulated object remains consistent over time. As in [73], the related research work is making progress.

Dynamic worlds are all around. This makes us believe that semantics will imply more annotations such as motion intention. The semantic SLAM pipelines that concern motion reasoning are expected to be integrated into a deep learning paradigm and attract increasing attention in future studies.

VI. Conclusion

Semantically ensembled localization and mapping are now attracting more and more scientific attention. In this survey, we review the development of semantic SLAM concerning its framework, object-level data association, and semantics-constrained pose optimization. Specifically, a semantic extractor that allows deep learning-based object detection to be seamlessly integrated into a SLAM pipeline is in detail described. Inspired by the recent progress in 3-D layered scene graphs, we explicitly target semantic reasoning and extend this concept to potential relationship reasoning and motion reasoning. These technology outlooks and possible solutions are further discussed. We stress that our review exhibits a fairly unique perspective and the major concerns have been so summarized that they form explicit classifications for semantically meaningful descriptors and object-level constraining strategies. Additionally, we argue that semantic reasoning underpins successful robot deployment in high-level tasks, and it might guide the complex scene understanding under the framework of semantic SLAM directly to a solution.

References

[1] X. Ji, P. Liu, H. Niu, X. Chen, R. Ying, and F. Wen, "Object SLAM based on spatial layout and semantic consistency," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023, doi: 10.1109/TIM.2023.3316258.

[2] L. Xia, J. Cui, X. Li, D. Zhang, J. Zhang, and L. Yi, "A point-line-plane primitives fused localization and object-oriented semantic mapping in structural indoor scenes," *Meas. Sci. Technol.*, vol. 33, no. 9, Sep. 2022, Art. no. 095017, doi: 10.1088/1361-6501/ac784c.

[3] G. Singh, M. Wu, M. V. Do, and S.-K. Lam, "Fast semantic-aware motion state detection for visual SLAM in dynamic environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23014–23030, Dec. 2022, doi: 10.1109/TITS.2022.3213694.

[4] L. Xia, D. Meng, J. Zhang, D. Zhang, and Z. Hu, "Visual-inertial simultaneous localization and mapping: Dynamically fused point-line feature extraction and engineered robotic applications," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2022.3198724.

[5] C. Cadena et al., "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016, doi: 10.1109/TRO.2016.2624754.

[6] S. Shen, Y. Cai, W. Wang, and S. Scherer, "DytanVO: Joint refinement of visual odometry and motion segmentation in dynamic environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 4048–4055, doi: 10.1109/ICRA48891.2023.10161306.

[7] J. Li et al., "Attention-SLAM: A visual monocular SLAM learning from human gaze," *IEEE Sensors J.*, vol. 21, no. 5, pp. 6408–6420, Mar. 2020, doi: 10.1109/JSEN.2020.3038432.

[8] Z. Qian, J. Fu, and J. Xiao, "Towards accurate loop closure detection in semantic SLAM with 3D semantic covisibility graphs," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2455–2462, Apr. 2022, doi: 10.1109/LRA.2022.3145066.

[9] L. Xia, J. Cui, R. Shen, X. Xu, Y. Gao, and X. Li, "A survey of image semantics-based visual simultaneous localization and mapping: Application-oriented solutions to autonomous navigation of mobile robots," *Int. J. Adv. Robotic Syst.*, vol. 17, no. 3, May 2020, Art. no. 172988142091918, doi: 10.1177/1729881420919185.

[10] T. Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," *Sci. Robot.*, vol. 8, no. 79, Jun. 2023, Art. no. eadf6991, doi: 10.1126/scirobotics.adf6991.

[11] H. Zhou and G. Yu, "Research on pedestrian detection technology based on the SVM classifier trained by HOG and LTP features," *Future Gener. Comput. Syst.*, vol. 125, pp. 604–615, Oct. 2021, doi: 10.1016/j.future.2021.06.016.

[12] S. Beery, G. Wu, V. Rathod, R. Votel, and J. Huang, "Context R-CNN: Long term temporal context for per-camera object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13072–13082, doi: 10.1109/CVPR42600.2020.01309.

[13] P. Voigtlaender, J. Luiten, P. H. S. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6578–6588, doi: 10.1109/CVPR42600.2020.00661.

[14] Q. Qiu and D. Lau, "Real-time detection of cracks in tiled sidewalks using YOLO-based method applied to unmanned aerial vehicle (UAV) images," *Autom. Construct.*, vol. 147, Mar. 2023, Art. no. 104745, doi: 10.1016/j.autcon.2023.104745.

[15] H. Yi, B. Liu, B. Zhao, and E. Liu, "Small object detection algorithm based on improved YOLOv8 for remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1734–1747, 2024, doi: 10.1109/JSTARS.2023.3339235.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Switzerland: Springer, 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.

[17] N. H. T. Nguyen, S. Perry, D. Bone, H. T. Le, and T. T. Nguyen, "Two-stage convolutional neural network for road crack detection and segmentation," *Expert Syst. Appl.*, vol. 186, Dec. 2021, Art. no. 115718, doi: 10.1016/j.eswa.2021.115718.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[19] X. Weng, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Stage-aware feature alignment network for real-time semantic segmentation of street scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4444–4459, Oct. 2021, doi: 10.1109/TCSVT.2021.3121680.

[20] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "ExFuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 269–284, doi: 10.1007/978-3-030-01249-6_17.

[21] Y. Nakayama, H. Lu, Y. Li, and T. Kamiya, "WideSegNeXt: Semantic image segmentation using wide residual network and NeXt dilated unit," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11427–11434, May 2021, doi: 10.1109/JSEN.2020.3008908.

[22] F. Yuan, L. Zhang, X. Xia, Q. Huang, and X. Li, "A gated recurrent network with dual classification assistance for smoke semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4409–4422, 2021, doi: 10.1109/TIP.2021.3069318.

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818, doi: 10.1007/978-3-030-01234-2_49.

[24] S. Xu, S. Lan, and Q. Zhu, "MaskPlus: Improving mask generation for instance segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2019–2027, doi: 10.1109/WACV45572.2020.9093379.

[25] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++ better real-time instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1108–1121, Feb. 2022, doi: 10.1109/TPAMI.2020.3014297.

[26] F. Shu, J. Wang, A. Pagani, and D. Stricker, "Structure PLP-SLAM: Efficient sparse mapping and localization using point, line and plane for monocular, RGB-D and stereo cameras," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 2105–2112, doi: 10.1109/ICRA48891.2023.10160452.

[27] J. He, M. Li, Y. Wang, and H. Wang, "OVD-SLAM: An online visual SLAM for dynamic environments," *IEEE Sensors J.*, vol. 23, no. 12, pp. 13210–13219, Jun. 2023, doi: 10.1109/JSEN.2023.3270534.

[28] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4076–4083, Oct. 2018, doi: 10.1109/LRA.2018.2860039.

[29] C. Yu et al., "DS-SLAM: A semantic visual SLAM towards dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1168–1174, doi: 10.1109/IROS.2018.8593691.

[30] S. Yang and S. Scherer, "CubeSLAM: Monocular 3-D object SLAM," *IEEE Trans. Robot.*, vol. 35, no. 4, pp. 925–938, Aug. 2019, doi: 10.1109/TRO.2019.2909168.

[31] L. Nicholson, M. Milford, and N. Sünderhauf, "QuadricSLAM: Dual quadrics from object detections as landmarks in object-oriented SLAM," *IEEE Robot. Autom. Lett.*, vol. 4, no. 1, pp. 1–8, Jan. 2019, doi: 10.1109/LRA.2018.2866205.

[32] Y. Wu, Y. Zhang, D. Zhu, Y. Feng, S. Coleman, and D. Kerr, "EAO-SLAM: Monocular semi-dense object SLAM based on ensemble data association," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 4966–4973, doi: 10.1109/IROS45743.2020.9341757.

[33] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, p. 99, Dec. 2021, doi: 10.1145/3503250.

[34] X. Han, H. Liu, Y. Ding, and L. Yang, "RO-MAP: Real-time multi-object mapping with neural radiance fields," *IEEE Robot. Autom. Lett.*, vol. 8, no. 9, pp. 5950–5957, Sep. 2023, doi: 10.1109/LRA.2023.3302176.

[35] S. Zhu et al., "SNI-SLAM: Semantic neural implicit SLAM," 2023, *arXiv:2311.11016*.

[36] B. Li, D. Zou, Y. Huang, X. Niu, L. Pei, and W. Yu, "TextSLAM: Visual SLAM with semantic planar text features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 593–610, Jan. 2024, doi: 10.1109/TPAMI.2023.3324320.

[37] J. Huang, S. Yang, T. Mu, and S. Hu, "ClusterVO: Clustering moving instances and estimating visual odometry for self and surroundings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2165–2174, doi: 10.1109/CVPR42600.2020.00224.

[38] B. Bescos, C. Campos, J. D. Tardos, and J. Neira, "DynaSLAM II: Tightly-coupled multi-object tracking and SLAM," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 5191–5198, Jul. 2021, doi: 10.1109/LRA.2021.3068640.

[39] H. M. S. Bruno and E. L. Colombini, "LIFT-SLAM: A deep-learning feature-based monocular visual SLAM method," *Neurocomputing*, vol. 455, pp. 97–110, Sep. 2021, doi: 10.1016/j.neucom.2021.05.027.

[40] D. Esparza and G. Flores, "The STDyn-SLAM: A stereo vision and semantic segmentation approach for VSLAM in dynamic outdoor environments," *IEEE Access*, vol. 10, pp. 18201–18209, 2022, doi: 10.1109/ACCESS.2022.3149885.

[41] Q. Wu et al., "Object-compositional neural implicit surfaces," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 197–213, doi: 10.1007/978-3-031-19812-0_12.

[42] S. Cheng, C. Sun, S. Zhang, and D. Zhang, "SG-SLAM: A real-time RGB-D visual SLAM toward dynamic scenes with semantic and geometric information," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023, doi: 10.1109/TIM.2022.3228006.

[43] J. Wang, J. Tarrio, L. Agapito, P. F. Alcantarilla, and A. Vakhitov, "SeMLaPS: Real-time semantic mapping with latent prior networks and quasi-planar segmentation," *IEEE Robot. Autom. Lett.*, vol. 8, no. 12, pp. 7954–7961, Oct. 2023, doi: 10.1109/LRA.2023.332264.

[44] X. Kong, S. Liu, M. Taher, and A. J. Davison, "VMAP: Vectorised object mapping for neural field SLAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 952–961, doi: 10.1109/cvpr52729.2023.00098.

[45] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "FastReID: A PyTorch toolbox for general instance re-identification," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 9664–9667, doi: 10.1145/3581783.3613460.

[46] Q. Zhong and X. Fang, "A BigBiGAN-based loop closure detection algorithm for indoor visual SLAM," *J. Electr. Comput. Eng.*, vol. 2021, pp. 1–10, Jul. 2021, doi: 10.1155/2021/9978022.

[47] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307, doi: 10.1109/TPAMI.2017.2711011.

[48] S. An, H. Zhu, D. Wei, K. A. Tsintotas, and A. Gasteratos, "Fast and incremental loop closure detection with deep features and proximity graphs," *J. Field Robot.*, vol. 39, no. 4, pp. 473–493, Jun. 2022, doi: 10.1002/rob.22060.

[49] H. Osman, N. Darwish, and A. Bayoumi, "LoopNet: Where to focus? Detecting loop closures in dynamic scenes," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2031–2038, Apr. 2022, doi: 10.1109/LRA.2022.3142901.

[50] N. Merrill and G. Huang, "CALC2.0: Combining appearance, semantic and geometric information for robust and efficient visual loop closure," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4554–4561, doi: 10.1109/IROS40897.2019.8968159.

[51] H. Osman, N. Darwish, and A. Bayoumi, "PlaceNet: A multi-scale semantic-aware model for visual loop closure detection," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105797, doi: 10.1016/j.engappai.2022.105797.

[52] M. Scucchia and D. Maltoni, "Region prediction for efficient robot localization on large maps," in *Proc. Int. Conf. Robot., Comput. Vis. Intell. Syst.*, 2024, pp. 244–259, doi: 10.1007/978-3-031-59057-3_16.

[53] H. Yang, T. Zhang, S. Jin, L. Chen, R. Sun, and L. Sun, "Visual loop closure detection based on binary generative adversarial network," *CAAI Trans. Intell. Syst.*, vol. 16, no. 4, pp. 673–682, 2021, doi: 10.11992/tis.202007007.

[54] Z. Qian, K. Patath, J. Fu, and J. Xiao, "Semantic SLAM with autonomous object-level data association," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Jun. 2021, pp. 11203–11209, doi: 10.1109/ICRA48506.2021.9561532.

[55] Y. Liu, C. Guo, and Y. Wang, "Object-aware data association for the semantically constrained visual SLAM," *Intell. Service Robot.*, vol. 16, no. 2, pp. 155–176, Apr. 2023, doi: 10.1007/s11370-023-00455-9.

[56] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," 2022, *arXiv:2201.13360*.

[57] S. Lin, J. Wang, H. Pei, H. Zhao, and Z. Chen, "Monocular semantic SLAM method based on object relation description," *J. Syst. Simul.*, vol. 34, no. 2, pp. 278–284, 2022, doi: 10.16182/j.issn1004731x.joss.20-0734.

[58] M. Zins, G. Simon, and M. Berger, "OA-SLAM: Leveraging objects for camera relocalization in visual SLAM," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2022, pp. 720–728, doi: 10.1109/ISMAR55827.2022.00090.

[59] A. Tourani, H. Bavle, J. L. Sanchez-Lopez, R. M. Salinas, and H. Voos, "Marker-based visual SLAM leveraging hierarchical representations," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2023, pp. 3461–3467, doi: 10.1109/iros55552.2023.10341891.

[60] Y. Zhong, S. Hu, G. Huang, L. Bai, and Q. Li, "WF-SLAM: A robust VSLAM for dynamic scenarios via weighted features," *IEEE Sensors J.*, vol. 22, no. 11, pp. 10818–10827, Jun. 2022, doi: 10.1109/JSEN.2022.3169340.

[61] X. Meng, B. Li, B. Li, B. Li, and B. Li, "PROB-SLAM: Real-time visual SLAM based on probabilistic graph optimization," in *Proc. 8th Int. Conf. Robot. Artif. Intell.*, Nov. 2022, pp. 39–45, doi: 10.1145/3573910.3573920.

[62] J. Jiao, C. Wang, N. Li, Z. Deng, and W. Xu, "An adaptive visual dynamic-SLAM method based on fusing the semantic information," *IEEE Sensors J.*, vol. 22, no. 18, pp. 17414–17420, Sep. 2022, doi: 10.1109/JSEN.2021.3051691.

[63] C. Chen, B. Wang, C. X. Lu, N. Trigoni, and A. Markham, "A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence," 2020, *arXiv:2006.12567*.

[64] K.-N. Lianos, J. L. Schonberger, M. Pollefeys, and T. Sattler, "VSO: Visual semantic odometry," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 234–250, doi: 10.1007/978-3-030-01225-0_15.

[65] X. Han and L. Yang, "SQ-SLAM: Monocular semantic SLAM based on superquadric object representation," *J. Intell. Robotic Syst.*, vol. 109, no. 2, p. 29, Oct. 2023, doi: 10.1007/s10846-023-01960-w.

[66] T. Lee, Y. Jang, and H. J. Kim, "Object-based SLAM utilizing unambiguous pose parameters considering general symmetry types," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 2120–2126, doi: 10.1109/ICRA48891.2023.10160309.

[67] S. Yang and S. Scherer, "Monocular object and plane SLAM in structured environments," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3145–3152, Oct. 2019, doi: 10.1109/LRA.2019.2924848.

[68] M. Hosseinzadeh, Y. Latif, T. Pham, N. Suenderhauf, and I. Reid, "Structure aware SLAM using quadrics and planes," in *Proc. Asian Conf. Comput. Vis.*, 2019, pp. 410–426, doi: 10.1007/978-3-030-20893-6_26.

[69] Z. Liao, Y. Hu, J. Zhang, X. Qi, X. Zhang, and W. Wang, "SO-SLAM: Semantic object SLAM with scale proportional and symmetrical texture constraints," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4008–4015, Apr. 2022, doi: 10.1109/LRA.2022.3148465.

[70] B. Zhou, M. Gilles, and Y. Meng, "Structure SLAM with points, planes and objects," *Adv. Robot.*, vol. 36, no. 20, pp. 1060–1075, Oct. 2022, doi: 10.1080/01691864.2022.2123253.

[71] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "DBSCAN: Past, present and future," in *Proc. 5th Int. Conf. Appl. Digit. Inf. Web Technol. (ICADIWT)*, Chennai, India, Feb. 2014, pp. 232–238, doi: 10.1109/ICADIWT.2014.6814687.

[72] J. Zhang, L. Yuan, T. Ran, Q. Tao, and Z. Wu, "Outlier elimination for monocular object SLAM based on spatiotemporal consistency constraints," *IEEE Sensors J.*, vol. 23, no. 8, pp. 8887–8898, Apr. 2023, doi: 10.1109/JSEN.2023.3252050.

[73] Y. Qiu, C. Wang, W. Wang, M. Henein, and S. Scherer, "AirDOS: Dynamic SLAM benefits from articulated objects," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 8047–8053, doi: 10.1109/ICRA46639.2022.9811667.

Linlin Xia received the Ph.D. degree from Harbin Engineering University, Harbin, China, in 2008.

She is currently a Professor with the School of Automation Engineering, Northeast Electric Power University, Jilin, China. Her research interests include multisource information fusion, navigation and positioning of robots, and advanced control.

Sida Li received the B.Eng. degree from Changchun University of Technology, Changchun, China, in 2022. He is currently pursuing the M.Eng. degree in control science and engineering with Northeast Electric Power University, Jilin, China.

His research interests include semantic SLAM and 3-D map reconstruction.

Linna Yi received the B.Eng. degree from Northeast Petroleum University, Daqing, China, in 2021. She is currently pursuing the M.Eng. degree in control science and engineering with Northeast Electric Power University, Jilin, China.

Her research interests include polarized 3-D reconstruction and VSLAM.

Heng Ruan received the B.Eng. degree from Northeast Electric Power University, Jilin, China, in 2022, where he is currently pursuing the M.Eng. degree in control science and engineering.

His research interests include semantic SLAM and polarized 3-D map reconstruction.

Daochang Zhang received the Ph.D. degree from Jilin University, Changchun, China, in 2015.

He is currently an Associate Professor with the School of Science, Northeast Electric Power University, Jilin, China. His research interests include matrix analysis, matrix algebra, and image processing.