Semantic Visual Simultaneous Localization and Mapping: A Survey

Kaiqi Chen, Jianhua Zhang, Jialing Liu, Qiyi Tong, Ruyu Liu, Shengyong Chen

***Abstract*—Visual Simultaneous Localization and Mapping (vSLAM) has achieved great progress in the computer vision and robotics communities, and has been successfully used in many fields such as autonomous robot navigation and AR/VR. However, vSLAM cannot achieve good localization in dynamic and complex environments. Numerous publications have reported that, by combining with the semantic information with vSLAM, the semantic vSLAM systems have the capability of solving the above problems in recent years. Nevertheless, there is no comprehensive survey about semantic vSLAM. To fill the gap, this paper first reviews the development of semantic vSLAM, explicitly focusing on its strengths and differences. Secondly, we explore three main issues of semantic vSLAM: the extraction and association of semantic information, the application of semantic information, and the advantages of semantic vSLAM. Then, we collect and analyze the current state-of-the-art SLAM datasets which have been widely used in semantic vSLAM systems. Finally, we discuss future directions that will provide a blueprint for the future development of semantic vSLAM.**

***Index Terms*—Semantic Visual SLAM, Semantic Information, Localization, Mapping, SLAM Datasets.**

I. Introduction

SIMULTANEOUS localization and mapping (SLAM) is the foundation for robots to explore the environment autonomously. Humans can quickly locate themselves in unfamiliar and complex environments and reconstruct the environment through spatial perception. This ability can be enhanced with acquired training and plays a crucial role in human cognition and motor control development. Similarly, the mobile agents and robots can also estimate information about their motion and the environment, if they equip with different sensors and run the SLAM algorithm. SLAM has gradually become synonymous with robotics over the past few decades. In terms of theory and practice, SLAM has now been established a complete set of technical solutions. SLAM technology has been applied widely in many fields, such as Medical Service Robots [1], Autonomous Driving [2], Unmanned Aerial Vehicles (UAVs) [3], Augmented Reality (AR) [4] and Virtual Reality (VR).

There are various SLAM technology methods, but most are challenging to popularize due to high equipment costs or limited scenarios to be applied. SLAM based on vision sensors, which is also called visual SLAM (vSLAM), has recently become a popular research direction due to its low hardware cost, high accuracy in small scenes, and the ability to obtain rich environmental information. It has to be mentioned that the disadvantages of vSLAM are also very obvious. On the one hand, there are still many challenges in coping with lighting changes, dynamic object movements, and environments lacking textures. On the other hand, the system has a high computing load, and the constructed geometric maps are difficult to apply for path planning and navigation.

The rise of deep learning technologies in recent years has allowed researchers to solve the traditional SLAM problem. Based on deep learning techniques, the researchers extract feature points, descriptors, and semantic information and perform pose estimation. The integration of semantic information into traditional vSLAM improves the understanding of image features and builds highly accurate semantic maps, which are verified in earlier works [5]–[8]. Compared with traditional vSLAM, semantic vSLAM not only acquires the geometric structure information in the environment but also extracts semantic information about independent objects (e.g., position, orientation, and category). In localization, semantic vSLAM improves the accuracy and robustness of localization with the help of semantic constraints. In mapping, semantic information provides rich object information to build different types of semantic maps, such as pixel-level maps [9], and object-level maps [10], [11]. Therefore, semantic vSLAM can help robots improve the ability to accurately perceive and adapt to unknown complex environments and perform more complex tasks.

Several current survey papers have extensively discussed SLAM algorithms and systems. The early SLAM developments have been well-summarized in [12], [13]. Of course, there are also investigations involving specific domains, such as multi-robot collaborative SLAM [14], keyframe-based monocular SLAM [15], and SLAM for dynamic environments [16]. Meanwhile, there is a recent visual SLAM survey in [17], where the survey scopes from 2010 to 2016 and lacks a new frontier in semantics. Notably, [18] provides a detailed review of the tremendous progress which has been made in the SLAM community, and considers the future direction. While this survey contains a brief discussion of deep learning methods, it does not provide a comprehensive overview of this field, especially the explosion of research over the past seven years. Therefore, [19] introduces the current SLAM combined

This research was supported by the National Natural Science Foundation of China under Grant 62202137 and Zhejiang Provincial Natural Science Foundation of China under Grant LQ22F030004. (Corresponding author: Jianhua Zhang.)

K. Chen, J. Liu and Q. Tong are with the Institute of Computer Vision, College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: chenkaiqi96@outlook.com; liujialing98@hotmail.com; ahputqy@gmail.com).

R. Liu is with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China (e-mail: lry@hznu.edu.cn).

J. Zhang and S. Chen are with the Institute of Computer Vision, School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China (e-mail: zjh@ieee.org).

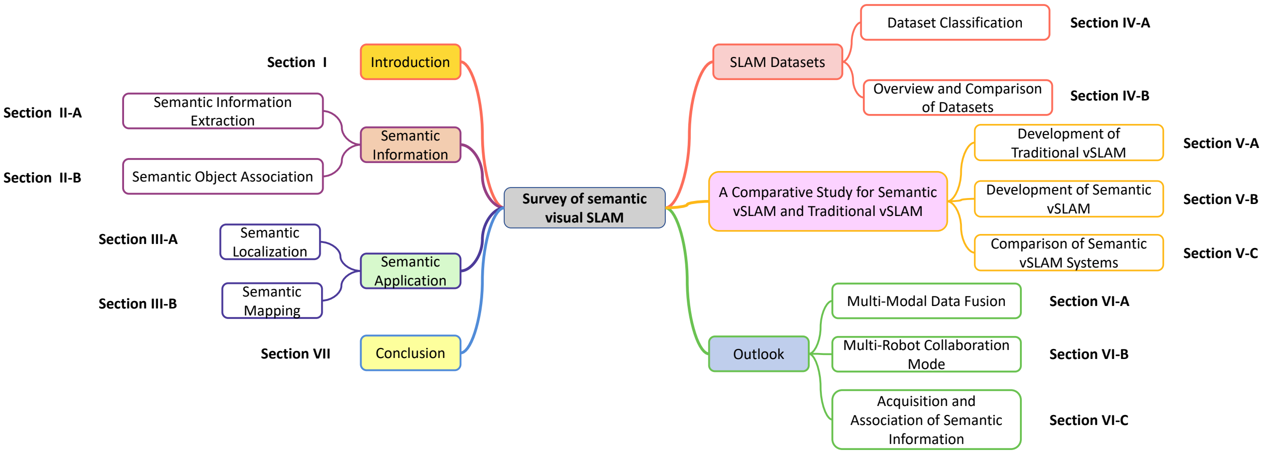
Survey of semantic visual SLAM
Section I
Introduction
Section II-A
Semantic Information Extraction
Semantic Information
Section II-B
Semantic Object Association
Section III-A
Semantic Localization
Semantic Application
Section III-B
Semantic Mapping
Section VII
Conclusion
SLAM Datasets
Dataset Classification
Section IV-A
Overview and Comparison of Datasets
Section IV-B
A Comparative Study for Semantic vSLAM and Traditional vSLAM
Development of Traditional vSLAM
Section V-A
Development of Semantic vSLAM
Section V-B
Comparison of Semantic vSLAM Systems
Section V-C
Outlook
Multi-Modal Data Fusion
Section VI-A
Multi-Robot Collaboration Mode
Section VI-B
Acquisition and Association of Semantic Information
Section VI-C

Fig. 1. The schematic diagram of the overall structure of the paper.

with deep learning methods. However, it does not summarize semantic data processing and association methods. [20] mentions the concept of semantic SLAM but does not emphasize the importance of semantics, especially the contribution made by the rapid development of semantic vSLAM in recent years. It is worth noting that there have been papers [17]–[20] detailing the work related to SLAM, but there are currently no surveys that systematically introduce the development history of semantic vSLAM. Therefore, this survey provides a detailed overview of existing works on semantic vSLAM, focusing on semantic information extraction, semantic applications, SLAM datasets, and comparison of semantic vSLAM with traditional vSLAM. To the best of our knowledge, this is the first survey paper that provides a comprehensive and extensive overview of semantic vSLAM.

The remainder of the paper is structured as shown in Fig.1. Section II introduces the extraction and association methods of semantic information in semantic vSLAM. Section III introduces the current application of semantic visual SLAM. Section IV introduces the existing SLAM datasets. Section V states the difference between traditional vSLAM and semantic vSLAM, listing and comparing recent semantic vSLAM works. Section VI introduces the current problems and future development directions of semantic vSLAM technology. A summary of semantic vSLAM is given in Section VII.

II. Semantic Information

In recent years, SLAM has begun to be combined with semantic information, which contains the position, orientation, colour, texture, shape, and specific attributes of objects in the environment. Compared with past SLAM methods, Semantic vSLAM not only can acquire geometric structure information in the environment during the mapping process, but also can recognize objects in the environment and acquire semantic information to adapt to complex environments and perform more intelligent tasks. Traditional vSLAM methods are often based on the assumption of a static environment, whereas semantic vSLAM can predict the movable properties of objects in dynamic environments. Similar object knowledge representations in semantic vSLAM can be shared, improving the operation and storage efficiency of SLAM systems by maintaining a shared knowledge base. Moreover, Semantic vSLAM can be applied to intelligent path planning and navigation, such as server robots selecting the optimal path for delivering supplies.

The framework of semantic vSLAM can be roughly divided into the semantic information extraction module and the vSLAM module, as shown in Fig.2. Moreover, the key to semantic vSLAM methods is accurately identifying objects in the environment. For the process of semantic extraction, we can regard the process as identifying objects of interest in images and obtaining information about objects. The deep learning techniques that have emerged over the years are the most promising methods for semantic extraction. The research on semantic object extraction is gradually shifted from traditional machine vision algorithms to the direction of deep neural networks, such as CNN and R-CNN. Their semantic extraction accuracy and real-time performance can meet the requirements of SLAM. There are three commonly used methods for semantic vSLAM extraction of semantic information, namely object detection [21]–[28], semantic segmentation [29]–[32], and instance segmentation [33]. Furthermore, the processing of semantic object association is also crucial. We will describe the extraction and association of semantic information in the following.

A. Semantic Information Extraction

1) Object Detection: The object detection module in semantic vSLAM can help the SLAM system to acquire the objects from images. By combining it with SLAM, we can construct object-level semantic maps and improve the environment understanding. The current object detection methods used in semantic vSLAM are mainly divided into two categories: one-stage method and two-stage method.

Semantic vSLAM usually adopts SDD [21] or YOLO series (YOLOv1 [22],YOLOv2 [23],YOLOv3 [24],YOLOv4 [25]) as the one-stage object detection methods. SSD [21] is the first DNN-based real-time object detector that achieves above 70% mAP in PASCAL VOC datasets [34] with 40 FPS in

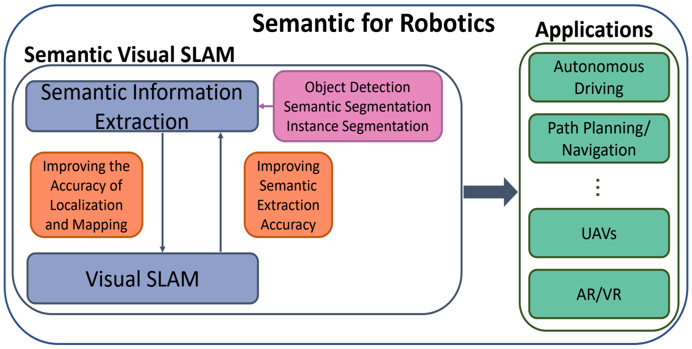

Fig. 2. The overall framework of semantic robotics. Semantic visual SLAM consists of semantic information extraction and visual SLAM modules, which influence each other. Semantic visual SLAM is widely used in autonomous driving, path planning, and navigation.

TitanX. SSD is also a one-stage object detection detector that balances speed and accuracy well. Hence, several semantic vSLAM works [35]–[38] deploy SSD as the detector module. [36]–[38] utilize SSD to detect static objects and do not consider dynamic objects, but Zhong et al. [35] reliaze that the most semantic vSLAM works perform poorly in dynamic and complex environments. To address these challenges, they detect potential moving objects (e.g., people, vehicles, animals) through the SSD detector and cull their regions to eliminate the influence of dynamic objects on pose estimation. Experimental results show that the system facilitates a robot to accomplish tasks reliably and efficiently in unknown and dynamic environments. However, it is inevitable to recognize non-dynamic regions as dynamic regions in violently removing potentially dynamic objects.

The YOLOv1 [22] detector is earlier than SSD, but its speed and detection performance are comparable to SSD, so it is also applied in semantic vSLAM for object detection [39]. Soon Redmon et al. propose YOLOv2 [23], which is a great improvement over the YOLOv1 and SSD in recognition type, detection accuracy, speed, and localization accuracy. Given its good performance enough for semantic vSLAM needs, some works [11], [40]–[42] try to use the YOLOv2 as the detector for semantic vSLAM. Bavle et al. [40] propose a particle filter localization approach based on semantic information from indoor environments. They fuse semantic information into a prior map, which assists the aerial robot with precise localization. In outdoor semantic vSLAM, [41], [42] use the YOLOv2 to detect outdoor vehicles and street road signs, respectively, obtaining a large number of object measurements for improving localization accuracy. Yang et al. [11] propose a single-image 3-D cuboid detection method suitable for indoor and outdoor scenes, which acquires 3D semantic objects by the YOLOv2 detector and vanishing point technique. YOLOv3 [24] draws on the residual network structure and multi-scale detection ideas to form a deeper network level, improving mean average precision (mAP) and small object detection. Therefore, a large number of semantic vSLAM works [43]–[48] use this detector to meet the accuracy of object detection and localization in dynamic environments. Among these methods, Nicholson et al. [43] specially design a sensor model for object detector based on YOLOv3. Thus, the challenge of detection of partially visible object is addressed, which substantially improves the accuracy of vSLAM. No matter the SSD or YOLO series, they all meet the most crucial requirement of semantic vSLAM, i.e., real-time. Therefore, they are widely used in many sematic vSLAM systems.

In addition to real-time performance, the detection accuracy also influences the performance of semantic vSLAM. Therefore, several works (e.g., [2], [49]) adopt the two-stage detectors for object detection, such as R-CNN [26], Fast R-CNN [27], Faster R-CNN [28]. Unlike the one-stage detectors, the two-stage detectors need to obtain the region proposal, classify the results and adjust the candidate bounding box positions. Due to the design idea of the two-stage detector, the real-time performance of the two-stage detector is usually slightly worse than that of the one-stage detector, but its detection accuracy is higher than that of the one-stage detector. Li et al. [2] recognize that Faster R-CNN performs well in detecting small objects, and consequently use the spatial-temporal consistency relationship between semantic information and sparse feature measurements for tracking static and dynamic objects. In order to make semantic vSLAM face more scenes, Iqbal et al. [49] propose the hybrid detector idea based on Fast R-CNN [28] and MobileNet [50], where the system can flexibly use different object detectors to cope with object detection in different environments. Moreover, Li et al. introduce text objects into the semantic map in [51], where they use the EAST text detector [52] and provide directional information for the detected text patches.

Although significant progress has been made in vSLAM and object detection in recent years, the object bounding boxes obtained by object detectors also contain foreground and other object information, which affects object reconstruction and global localization accuracy. Therefore, researchers try to use semantic segmentation or instance segmentation to obtain pixel-level objects.

2) Semantic Segmentation: Semantic segmentation is the cornerstone technology of image understanding, which can give the exact pixels corresponding to each type of object but cannot distinguish different individuals of the same type. It is pivotal in autonomous driving, UAVs, and wearable device applications. The current semantic segmentation methods used in semantic vSLAM are basically based on deep learning methods, such as U-Net [29], Bayesian SegNet [30], SegNet [31], PSPNet [32].

U-Net is one of the most commonly used segmentation models, which is simple, efficient, easy to build, and requires only a small datasets for training. Therefore, Qin et al. [53] classify image pixels into different categories based on the U-Net model, such as lanes, parking lines, speed bumps, and obstacles. Because these classes of objects have clear and stable features, they are used to improve localization accuracy. In addition, parking lines are also used for parking space detection and obstacles are used for path planning. SegNet is also often used for semantic segmentation tasks in outdoor environments [54]–[56], the advantages of which are better preservation of image edge information and higher running speed. [54]–[56] are based on the same framework [57], and each proposes a complete dynamic semantic vSLAM system. The difference is that [54] and [56] cull potential moving objects by default, while [55] combines semantic segmentation with the moving consistency checking method to filter out dynamic regions of the scene. Since feature points are removed from these regions, the robustness and accuracy of localization are improved in dynamic scenes. Compared with U-Net and SegNet, PSPNet considers the context relationship matching problem, which shows a good segmentation effect even in complex environments. Liu et al. [58] use the PSPNet to obtain semantic labels of sofas, cupboards, and desks in order to build high-precision semantic scene graphs.

Compared with object detection methods which output coarse detection bounding boxes, semantic segmentation methods can recognize objects at the pixel level, which dramatically helps semantic vSLAM to understand the environment. However, semantic segmentation cannot distinguish object instances from the same category, limiting the application scope.

3) Instance Segmentation: For detecting dynamic object instances, semantic vSLAM starts using instance segmentation methods, obtaining a pixel-wise semantic segmentation of the images. Instance segmentation is a further refinement of object detection to achieve pixel-level object separation. However, it cannot achieve the same real-time performance as object detection. The current common instance segmentation method used in semantic vSLAM (e.g., [59]–[62]) is Mask-RCNN [33], a powerful image-based instance-level segmentation algorithm that can segment eighty semantic object class labels. These works are suitable for dynamic environments because they fuse geometric information with Mask-RCNN to segment dynamic and static objects, obtaining pixel-wise semantic segmentation and instance label information. However, the real-time performance is greatly affected. Compared to traditional vSLAM methods, [59] is more robust to localization in dynamic scenes at the cost of removing dynamic objects. Moreover, [60]–[62] not only can track and reconstruct static backgrounds and objects in the scene in real-time, but also have robust tracking accuracy.

Although in current stage, three kinds of semantic extraction methods can meet the basic requirements of semantic vSLAM, there is still much room to be further improved with respect to recognition accuracy and operation speed for semantic extraction methods to be efficiently integrated into the semantic vSLAM systems, as in many complex environment, there are dynamic or occluded objects which will influence the performanc of object detection. To solve these challenges, vSLAM and semantic extraction methods will need to complement each other in the future, helping robots perform more intelligent tasks.

B. Semantic Object Association

When a robot with sensors moves in an unknown environment, it will collect a series of data. In a vSLAM system, the image set $I_{1:T} = \{I_1, \cdots, I_T\}$ contains all images acquired from start time to current time T. Assuming that the current environment contains object labels $\mathcal{C} = \{1, ..., c\}$. For example, the common used objects in a semantic vSLAM system are doors, chairs, tables, people, and vehicles. x_t represents the camera pose of the image I_t, which includes the position and orientation. $\mathcal{X} \overset{\triangle}{=} \{x_t\}_{t=1}^{T}$ denote the set of camera trajectories at each time. Since the camera pose is incrementally estimated based on the last state rather than directly calculated, the result is susceptible to noise. The camera pose at time t can be expressed as follows:

$$x_t = f\left(x_{t-1}, u_t\right) + w_t, w_t \sim \mathcal{N}\left(0, R_t\right), \tag{1}$$

where u_t is the motion measurement of vSLAM at time t, and R_t is the covariance matrix of camera pose noise.

When the robot is at the pose x_t, it will observe landmark point measurements y_t and object measurements $\mathcal{L}_t \overset{\triangle}{=} \{L_{tm}\}_{m=1}^{M}$ through the camera. The corresponding camera measurement at time t is as follows:

$$z_t = h\left(x_t, y_t, \mathcal{L}_t\right) + v_z, v_z \sim \mathcal{N}\left(0, Q_z\right), \tag{2}$$

where Q_z is random measurement noise, $\mathcal{Z} \overset{\triangle}{=} \{z_t\}_{t=1}^{T}$ is the set of existing sensor measurements at the current time.

The SLAM system creates keyframes $\mathcal{F}_{1:D} \overset{\triangle}{=} \{F_d\}_{d=1}^{D}$ to reduce repeated compution. Assuming that there are some detectable objects in the current scene, each keyframe F_d can detect N_d object measurements through the object detection method. Usually, a keyframe F_d has multiple object measurements, which can be expressed as a set $\mathcal{L}_d \overset{\triangle}{=} \{L_{di}\}_{i=1}^{N_d}$, and

$$L_{di} = \left\{L_{di}^{c}, L_{di}^{s}, L_{di}^{b}\right\}, L_{di} \in \mathcal{L}_d, \tag{3}$$

where L_{di} represents the measurement information of the i object in the keyframe F_d, which usually consists of object category $L_{di}^{c} \in \mathcal{C}$, detection confidence L_{di}^{s}, and object detection bounding box L_{di}^{b} in works [36]–[38], [40], [41], [45], [48], [63], [64]. These pieces of information can be obtained from the semantic object extraction methods [21], [31], [33] in Section II-A.

The number of landmarks in the environment is much smaller than the number of object measurements. The reason is that the same landmarks can be observed in consecutive keyframes, and multiple object measurements are detected in each keyframe. Hence, the concept of object association S is introduced, which specifies that object measurements across keyframes are associated with the same landmark. It can be understood that each object measurement L_{di} is assigned a unique landmark O_k, i.e., $S_{di} = (O_k, L_{di})$. There are d_t keyframes at t time. The object association at current time t can be expressed as:

$$\mathbb{S}_t \overset{\triangle}{=} \left\{S_{ij}\right\}, i \in \{1, ..., d_t\} \quad and \ \ j \in \{1, ..., N_d\}, \tag{4}$$

As the robot moves around over time, the number of object measurements accumulates. Thus, object association is a dynamic process that varies with the continuously acquired object measurements. At t time, the object association is estimated in works [36]–[38], [63], [64]:

$$\mathbb{S}_t = arg\max_{\mathbb{S}_{t-1}} p\left\{\mathbb{S}_{t-1} | X, L, Z\right\}, X \subseteq \mathcal{X}, L \subseteq \mathcal{L}, Z \subseteq \mathcal{Z}, \tag{5}$$

where X, L, Z are subsets selected from all camera poses $\mathcal{X}$, object measurements $\mathcal{L}$ and sensor measurements $\mathcal{Z}$ by SLAM algorithms, respectively.

Once the object association is implemented, it is necessary to update the camera and object poses with the results of the object and camera optimization [36]–[38], [45], [63], [64].

$$X^{'}, L^{'} = arg \max_{X,L} \log P\{Z|X, L, \mathbb{S}_t\}. \tag{6}$$

The above equations describe the basic semantic vSLAM problems. From the above, the role of object association for semantic vSLAM is accurately associating semantic object measurements with object landmarks. The difficulty of object association is correctly associating new object measurements with existing 3D landmarks in the map when there are multiple objects of the same category, similar appearance, and proximity position in the current image. Object association helps the robot to obtain the number of real objects in the environment and integrate them into the semantic map, improving the perception of the environment. Furthermore, it can provide correct optimization constraints for updating the camera poses and object poses. Thus, Object association of semantic information is a worthwhile research problem in semantic vSLAM, some researchers consider semantic object association from a probabilistic perspective. For example, the probabilistic object association model by Bowman et al. [63] is a milestone work on robust object association in semantic vSLAM. They propose a probabilistic approach to model the object association process and to adopt the EM algorithm to find correspondences between the object measurements and landmarks. Furthermore, they fully consider the ambiguity of object association, which lays the mathematical foundation for the follow-up works of semantic vSLAM.

However, there are many indeterminable factors in real environment, such as a large amount of objects with the same category and close location, that greatly reduce the accuracy of object association. It also has a significant impact on the estimated camera poses and object poses, which leads to larger trajectory errors and lower accuracy of the map. To improve association accuracy, researchers try different schemes to solve object association in complex environments. [36] proposes the idea of a hierarchical topic model, which is based on hierarchical Dirichlet object association. They provide a rich basis for object association between object measurements and landmarks by using the information on the position, appearance, and category of object. Later, the hierarchical object association strategy [64] greatly reduces false associations between objects that are similar in location and appearance, which consists of short-term object association and global object association. Different from the work of [63], Doherty et al. [37], [38] propose a semantic SLAM approach for probabilistic object association based on approximate max-marginalization, which eliminates the ambiguity of object association variables during inference and retains the standard Gaussian posterior assumptions. Compared with other semantic vSLAM systems, this approach has significant robustness advantages. In addressing object association between consecutive keyframes, [49] uses a nonparametric statistical approach, which extracts the regions of object measurements and landmarks in the image and compares the depth similarity of the areas to determine whether they can be associated with each other.

Most probabilistic object association works focus on static objects, but these methods become ineffective in dynamic environments. Given the complexity of dynamic environments, [61] solves the problems of dynamic object association and occlusion by adopting a probabilistic EM framework. Moreover, they combine depth image and signed distance function methods to improve the accuracy of multi-object tracking. [65] adopts different strategies to associate dynamic and static objects. The association of static object measurement is based on the feature point matching mechanism in [66]. For dynamic object association, they consider the two main characteristics of the object at a constant velocity in a short time and feature point matching, then associate the object measurements with landmarks in the map by reprojection method.

Non-probabilistic object association methods are also very popular in object-level semantic vSLAM. For example, in [42], [44], [67], the Mahalanobis distance and Hungarian algorithm are used to associate new object measurements with landmarks, but these algorithms consume very large computational resources. [47] builds an integrated object association strategy which integrate parametric and nonparametric statistical tests, as well as IoU-based methods, and takes full advantage of the nature of different statistics. It is worth mentioning that Wang et al. [68] adopt different object association strategies for different sensor devices. One association strategy is performed in LiDAR mode by comparing the distance between the 3D bounding box and the reconstructed object. Another association strategy is in stereo or monocular camera mode by counting the number of matching feature points between the object measurements and the landmarks. If multiple object measurements in the current keyframe are associated with the same landmark, the closest object measurement is associated with that landmark, and the others are rejected. If object measurement is not associated with any existing landmark, then it will be initialized as a new landmark. [69] combines geometric and semantic information to propose a hybrid object association method, which enables drift-free tracking without an explicit relocalization module. However, the proposed method needs to improve real-time performance and tracking accuracy in the face of object switches and missed detection. Unlike the hierarchical object association strategy [64], the two-step strategy for object association proposed by [48] relies mainly on object category and appearance similarity to match landmarks, which is not applicable in outdoor environments.

Semantic extraction approaches and object association strategies in semantic vSLAM have received extensive attention in recent years. In addition, the current probabilistic object association strategy is one of the best approaches to improve the perception capability of SLAM systems, but its robustness and generality need further improvement. Furthermore, because of the complexity of real-world environments, semantic extraction and object association accuracy are highly susceptible to being affected.

III. Semantic Application

Semantic and SLAM technology are two parts that promote each other. Semantic information combined with localization and mapping can improve localization and scene understanding accuracy. In recent years, semantic vSLAM technology has driven the development of localization and mapping, which has significantly impacted research areas such as autonomous driving, mobile robots, and UAVs. This section will focus on two aspects of semantic localization and semantic mapping.

A. Semantic Localization

The purpose of localization is for the robot to obtain its orientation in an unknown environment, that is, to determine its position in the world coordinate system of that environment. Traditional vSLAM is susceptible to environmental factors, resulting in localization failure. Nevertheless, rich semantic information can be extracted in vSLAM, helping vehicles and robots perceive high-level information in the environment. Moreover, there are geometric constraints in the semantic information, which can effectively improve the localization accuracy of the system. Since there are obvious differences between indoor and outdoor environments, resulting in different localization difficulties. Therefore, we state semantic localization from two different environments.

Long-term outdoor visual navigation must face challenges such as long-time operation, cross-weather, and significant light changes. Under these challenges, it is difficult to reliably match features between the image and the map, eventually resulting in poor localization accuracy or even complete failure of the localization algorithm. To solve these problems, some researchers try a localization algorithm [70] based on semantically segmented images and a semantic point feature map, which solves the problem of long-term visual localization. There are also established medium-term constraints based on the semantic information during tracking in [71], reducing the drift error of visual odometry. Facing the drastic viewpoint changes, the researchers [72] adopt semantic graph descriptor matching for global localization to achieve localization under multiple viewpoints.

The indoor robot localization problem is no less challenging than the outdoor one. vSLAM systems still rely on superficial image information to perceive the environment and lack cognitive-level capabilities. The robustness and reliability of SLAM have not yet reached practicality when entering complex indoor environments with dynamic or significant lighting changes. In improving the cognitive ability of robots in the environment, [43] comes up with an object-level semantic vSLAM system, which adopts dual quadrics representation of 3D landmarks for the first time, containing the size, position and orientation of the landmarks. Meanwhile, they derive a factor graph-based SLAM formulation that jointly estimates the dual quadric and camera pose parameters under the assumption of solving object association. Similarly, [11] is a monocular-based 3D object detection and mapping approach that improves camera pose and reduces monocular drift with the help of semantic object constraints. EAO-SALM [47], which borrows ideas from [43] and [11], is a framework for object pose estimation based on iForest. The framework contains an outliers-robust centroid, scale estimation algorithm, and an object pose initialization algorithm, which significantly facilitates the joint pose optimization. However, [43], [47] do not consider dynamic object factors, and ellipsoids or rectangles represent the objects created in the sparse maps without object details. Therefore, the researcher notices that semantic information can help distinguish between static and dynamic objects, improving robot localization accuracy and robustness in dynamic environments. Adopting semantic information to segment moving objects and filter out feature points associated with moving objects [35], [55], [73] is one of the frequently implemented approaches, improving the system localization in dynamic environments.

It can be seen that fusing semantic information is the fundamental way to raise robot localization performance. In improving localization, semantic works are often used in the stages of SLAM system initialization, back-end optimization, relocalization, and loop closing. Therefore, efficiently handling and utilizing semantic information is crucial to improving localization accuracy.

B. Semantic Mapping

Mapping is another goal of SLAM, which serves the localization in vSLAM. Usually, we hope the robot saves the map so that the robot does not need to build the map repeatedly in the next work, saving a lot of computational resources. In application, the maps constructed by vSLAM include sparse maps [57], [74], semi-dense maps [75], and dense maps [76], [77]. Compared with sparse maps, dense maps contain many 3D spatial points to describe the map, which is more suitable for localization, navigation, obstacle avoidance, and reconstruction.

However, the traditional vSLAM maps lack high-level environmental semantic information for human-computer interaction, making robots unable to perform complex tasks of intelligent obstacle avoidance, recognition, and interaction. In order to better solve map problems, it is more and more essential to establish an accurate and reliable 3D semantic map. Early semantic maps [5] often used a priori object CAD model database to construct 3D semantic maps, which can restore real scenes and save a large amount of space for storing dense point cloud maps. Nevertheless, the CAD models are limited to objects in a pre-defined database. In [9], [69], [78], [79], the researchers build static dense semantic maps, which integrates dense vSLAM with semantic segmentation labels. For dynamic environment reconstruction, [10], [35], [60] adopt instance-aware semantic segmentation to classify objects as background, moving objects, or potentially moving objects. However, they failed to achieve the real-time performance of the system. Considering the real-time problem of semantic vSLAM in mapping, some researchers [43], [48] try to construct sparse semantic maps. Represented by [11], [36], [44], [47], [51], [64], these methods are based on the ORB-SLAM2 framework and combine semantic objects with building sparse 3D semantic object maps in real-time.

It must be noted that semantic maps are more widely used in intelligent scenarios than traditional visual maps. However, it needs to face the challenges of heavy calculation, recognition of different types of objects, and map storage.

IV. SLAM Datasets

It is well known that most SLAM systems evaluate their algorithms on multiple public datasets to prove their effectiveness in some aspects due to the expensive equipment and the complexity of device operation. The most frequently used SLAM datasets include KITTI [80], TUM RGB-D [81], ICL-NUIM [82] and EuRoC MAV [83]. These datasets are collected from different environments, and suited for different vSLAM algorithms. Therefore, it is extremely important to find appropriate datasets to evaluate a vSLAM. Recently, Liu et al. [84] collated datasets commonly used in SLAM works in the past decade and provided a comprehensive introduction and analysis of them, which will facilitate the SLAM community to find suitable datasets. However, this survey does not provide a detailed introduction and analysis of the datasets suitable for semantic vSLAM. To fill this gap, we organize datasets suitable for semantic vSLAM, from which we evaluate and compare.

A. Datasets Classification

The categorization of datasets is typically based on sensor differences or applicable scenarios to help them understand and utilize existing SLAM datasets. Depending on the sensor, SLAM-related datasets can be divided into LiDAR, vision, and vision-LiDAR fusion datasets. The advantage of vision sensors is that they are inexpensive and ubiquitous vision devices, which can be mobile phones or cameras. Although these devices are not as powerful and accurate as radar devices, they are acknowledged to hold great potential in the SLAM community and are steadily moving forward. Furthermore, semantic-based vSLAM dramatically improves the performance of traditional vSLAM, which is attributed to obtaining rich environmental information in visual images through semantic extraction, helping robots to have a high-level understanding of unknown environments.

To help semantic vSLAM systems choose suitable datasets, we investigate the current open-source datasets and collect thirty-one datasets. Moreover, we provide a dataset selection guide for different vSLAM systems in section IV-B. For each dataset, we describe the dataset from eight dimensions, showing as much information about each dataset as possible, as shown in Table I. We expressly indicate whether the dataset contains semantic annotations. The annotation types are object frame, semantic segmentation, instance segmentation, and whether it is 2D or 3D. Moreover, we also indicate how many object categories are in the datasets because these semantic annotations are very helpful for examining semantic vSLAM. The details of each dimension of the dataset are shown as follows: • Name: Name of the dataset.

• Year: Year of publication of the dataset.

• Cited: the number of citations in the dataset when investigating the dataset (refer to Google Scholar for the number of citations).

• Sensors: The camera column contains the typical camera types: color, event, depth, and RGB-D. The LIDAR/RADAR column contains the standard 2D or 3D LIDAR (with beam numbers ranging from 4, 16, 32, and 64), while some datasets also have radar sensors. The IMU column is intended to express whether or not IMU is available in the data.

• Ground Truth: This dimension shows the acquisition of real ground localization information.

• Motion Pattern: The acquired moving platform is given in this dimension, indicating the different motion patterns.

• Environment: This dimension mainly introduces the sequence, length, or scene information of the dataset.

• Annotated information: This dimension mainly introduces whether the dataset contains semantic annotations, which can help semantic vSLAM to verify its performance and provide a large amount of annotation information for training detection models.

B. Overview and Comparison of Datasets

The results of the dataset collection are shown in Table I. This paper mainly displays thirty-one datasets that have been open-sourced in recent years, including six classics, highly cited datasets, and twenty-five datasets published in the last five years. Moreover, the paper also details four representative datasets [85]–[88] in the appendix.

Based on the results of the division in Table I, This paper provides the following recommended guidelines.

• Consider various sensor devices. Most of the collected datasets can be used for evaluation for semantic vSLAM, except for MulRan [89].

• Consider the challenges of the environment (light changes, weak textures, bad weather). Some SLAM systems try to illustrate the robustness of their systems in harsh environments, so researchers can choose datasets [90]–[93] for evaluation.

• Consider different scenarios. If researchers need a multi-scene dataset, then they can choose from urban datasets [87], [94], indoor datasets [95], [96], jungle datasets [97].

• Choose datasets with data annotation. If researchers are working on semantic vSLAM, they can choose the evaluation datasets [80], [85], [88], [93], [94], [97]–[103].

• Choose different motion patterns. For different application scenarios, it is necessary to select different motion patterns of capturing equipment, such as robots, cars, UAVs, USVs, handheld devices, and simulation devices.

Recently, the simulation device has been popular with the SLAM community because it does not require consideration of site constraints, and its equipment is less costly and time-consuming. For instance, TartanAir [85] adopts the Unreal Engine and collects the data using the AirSim plugin developed by Microsoft [104]. Compared with traditional datasets, the datasets collected in simulation contain all kinds of objects, motion diversity, and diverse scenarios.

V. A Comparative Study for Semantic vSLAM and Traditional vSLAM

The SLAM community has made tremendous progress in the last three decades, and we have witnessed the transition

TABLE I
OVERVIEW AND COMPARISON OF COMMON AND STATE-OF-THE-ART DATASETS.

Dataset	Year	Cited	Sensors			Ground Truth	Motion	Environment	Annotated information
			Camera	LiDAR/Radar	IMU	Pose	Pattern		
KITTI [80]	2012	8651	2*color (stereo) 2*gray (stereo)	1*Velodyne-64	N/A	RTK-GPS&INS	Car	Outdoors 39.2km	2D & 3D boxes 8 categories
TUM RGB-D [81]	2012	2670	1*RGB-D	N/A	N/A	MoCap	Handheld/ Wheeled Robot	Indoors 39 sequences	
ICL-NUIM [82]	2014	763	1*RGB-D	N/A	N/A	SLAM	Handheld	Indoors/8 sequences	
EuRoC MAV [83]	2016	989	2*gray (stereo)	N/A	Y	MoCap	UAV	Indoors/22 sequences	
TUM MonoVO [90]	2016	171	1*NA-gray 1*WA-gray	N/A	N/A	Loop Drift	Handheld	In-/outdoors diverse scenes/50 sequences	
Oxford RobotCar [105]	2017	908	3*color (stereo) 3*fisheye-color	2*Sick-2D 1*Sick-4	N/A	RTK-GPS&INS	Car	Outdoors/urban 100 sequences	
Complex Urban [91]	2019	101	2*color (stereo)	2*Velodyne-16 2*Sick-2D	Y	RTK-GPS+ FOG+SLAM	Car	Outdoors/urban diverse scenes	
ReFusion [106]	2019	39	1*RGB-D	N/A	N/A	MoCap /laser scanner	Handheld	Indoors 26 sequences	
RUGD [97]	2019	22	1*RGB	1*Velodyne-32	Y	GPS+IMU	Wheeled Robot	Outdoors/jungle 18 sequences	24 categories semantic segmentation
DISCOMAN [98]	2019	10	1*RGB-D 1*color (stereo)	1*Simulation	Y	Simulation	Robot	Indoors 200 sequences	semantic segmentation
H3D [94]	2019	110	3*color	1*Velodyne-64	Y	GNSS+IMU RTK(GSM)DGPS	Car	Outdoors/urban	3D boxes 8 categories
UZH-FPV [107]	2019	97	1*fisheye-RGB (stereo) 1*event	N/A	Y	Laser tracker	UAV	In-/outdoors 27 sequences	
ICL [95]	2019	10	RGB-D Mono	N/A	N/A	MoCap	MAV Handheld	Indoors 16 sequences	
EU Long-Term [108]	2020	38	2*stereo 2*fisheye-RGB	2*Velodyne-32 1*4-layer lidar 1*1-layer lidar 1*2D lidar	Y	RTK-GNSS/IMU	Car	Outdoors multi-season 63.4km	
Newer College [109]	2020	36	1*D435i(infrared)	1*Ouster-64	Y	6DOF ICP Localization	Handheld	Outdoors/2.2km 3D reconstruction	
TartanAir [85]	2020	52	2*color (stereo) 1*depth	1*Simulated-32	N/A	Simulation	Random	In-/outdoors 1037 sequences	semantic segmentation
CUHK-AHU [110]	2020	2	6*color	1*VLP-16	N/A	GPS+IMU	Car	Outdoors/34 sequences logistics/hill/urban	
IDDA [99]	2020	12	RDB-D	1*ARS 308-radar	N/A	Simulation	Car	Outdoors 105 scenarios	24 categories semantic segmentation
OpenLORIS-Scene [86]	2020	46	1*RGB-D 2*fisheye-RGB (stereo)	1*Hokuyo-2D 1*Robosense-16	Y	MoCap/ LiDAR SLAM	Wheeled Robot	Indoors/22 sequences temporal diversity	
MulRan [89]	2020	51	N/A	1*Ouster-64-3D 1*Navtech-CIR204-H	N/A	FOG+GPS+ICP	Car	Outdoors/urban 12 sequences	
Brno Urban [111]	2020	13	4*RGB 1*thermal camera	2*Velodyne-32	N/A	RTK-GNSS/INS	Car	Outdoors/urban 375.7km/67 sequences	
A*3D [100]	2020	47	2*color	1*Velodyne-64	N/A	/	Car	Outdoors 55-hours data	3D boxes 7 categories
Oxford Radar RobotCar [87]	2020	140	3*color (stereo) 3*fisheye-color	1*Radar 2*Sick-2D 2*Velodyne-32	N/A	GPS&INS+SLAM	Car	Outdoors/urban temporal diversity 32 sequences	
UrbanLoco [92]	2020	35	1*fisheye-color(HK)	1*Velodyne-32(HK) 1*Robosense-R32(SF)	Y	RTK-GNSS/INS(HK) RTK-GNSS(SF)	Car	Outdoors/urban diverse scenes	
Virtual kitti 2 [101]	2020	96	2*color (stereo) 1*depth	N/A	N/A	Simulation	Car	Outdoors/urban	2D & 3D boxes/14 categories semantic segmentation
TUM-VIE [96]	2021	7	2*uEye 2*GEN4-CD(event)	N/A	Y	MoCap	Handheld	Indoors 21 sequences	
TUK Campus [102]	2021	0	2*stereo 1*omnidirectional	1*Ouster-128 4*2D laser scanners	Y	GPS	Car	Outdoors campus	2D boxes/4 categories semantic segmentation
RADIATE [88]	2021	31	1*color (stereo)	1*Radar 1*Velodyne-32	Y	GPS+IMU	Car	Outdoors/22 sequences adverse weather	2D boxes 8 categories
Cirrus [103]	2021	4	1*RGB	2*Luminar Model H2	Y	GPS+IMU	Car	Outdoors 12 sequences	3D boxes 8 categories
VIODE [93]	2021	8	1*color (stereo)	N/A	Y	Simulation	UAV	In-/outdoors 12 sequences	instance segmentation
USVInland [112]	2021	1	1*color (stereo)	3*Radar 1*LS C16-16	Y	RTK-GNSS/INS	USV	Outdoors/canal 26km/33 sequences	

stage of SLAM technology in the industry. Cadena et al. [18] survey the development of SLAM over the past three decades and discuss that SLAM is entering a new era, the robust-perception age. Compared with the previous pure geometric vSLAM, the new-stage perceptual SLAM has a more robust performance and a higher level of environmental understanding, which is attributed to the application of image semantic information to SLAM for pose estimation, loop closing, and mapping. Therefore, this section will review the history of vSLAM development and introduce the semantic vSLAM development in recent years.

A. Development of Traditional vSLAM

The traditional vSLAM systems estimate the robot poses in unknown environments based on image information and build low-level maps, which use multi-view geometry principles. At present, the traditional vSLAM systems are mainly represented by filtering-based method [113], [114], keyframe-based BA method [66], [115], and direct tracking method [116], [117]. The filter-based vSLAM methods regard the system state at each moment as a Gaussian probability model and help the robot to predict the accurate poses according to the filter. Even with various noises, the filtering always predicts the real motion of the robot. For example, [113] chooses the extended Kalman filter (EKF). Since the visual SLAM pose estimation problem is not linear, the EKF cannot guarantee the global optimality of pose estimation. PTAM [115], as the first keyframe-based BA monocular vSLAM system, lays a foundation for subsequent keyframe-based BA vSLAM works.

ORB-SLAM [66] is based on PTAM architecture by adding the functions of map initialization and loop closing, and the optimization of keyframe selection and mapping. In localization, its localization error is much smaller than [113], [115]. Soon, the same researching team continuously improves ORB-SLAM and releases open-source vSLAM systems (i.e., ORB-SLAM2 [57], ORB-SLAM3 [74]). The localization accuracy of these systems is much higher than [76], [77], [118]. Direct tracking methods (i.e., DTAM [116], LSD-SLAM [117]) do not rely on the extraction and matching of feature points, but solve the camera motion by constructing the photometric error from the pixel gray values between the front and back frames. In the case of missing features and blurred images, these methods have better robustness than the previous two methods. However, the direct tracking methods are more sensitive to illumination changes and dynamic interference, so the positioning accuracy is generally inferior to [57], [66].

As introduced in Section III-B, traditional vSLAM represents the surrounding environment through point clouds, such as sparse maps, semi-dense maps, and dense maps. Since the point clouds in these maps do not correspond to objects in the environment, they are meaningless to the robot, as shown in Fig.3c. Therefore, researchers try to use geometric and a priori perception information to condense the features of 3D point clouds and understand them, which helps robots perceive high-level environmental details. Coinciding with the rise of semantic concepts, vSLAM systems combine with semantic information solutions greatly improve the ability of robots to perceive the unexplored environment. In Fig.3d, semantic information gives semantic labels to point clouds, helping build a semantic map of 3D landmark information. After years of development and verification, semantic information has improved the robustness of vSLAM to the environment and achieved more accurate loop closure.

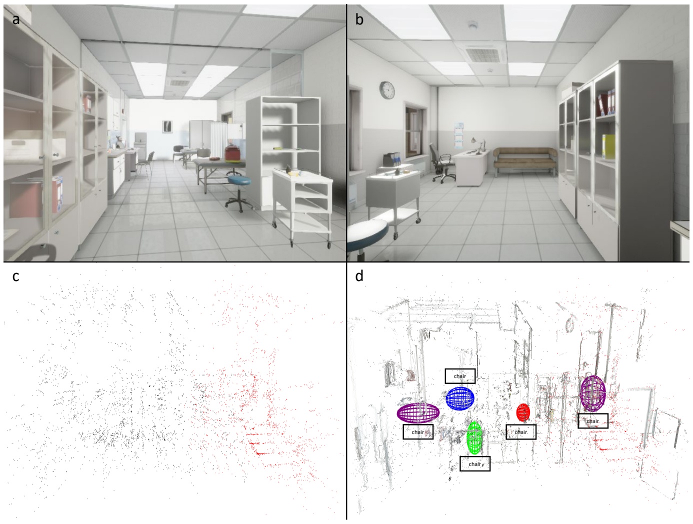

Fig. 3. (a)(b) Images of scenes from different perspectives. (c) 3D map based on point cloud representation of traditional vSLAM. (d)Environment reconstruction with semantic information.

B. Development of Semantic vSLAM

The early works of semantic vSLAM can be traced back to the early investigation by Salas-Moreno et al. [5]. They realize that traditional vSLAM systems operate at the low-level primitives which need to be processed harshly. Furthermore, building Maps by traditional methods is only noisy point clouds that do not appear like the maps shown in human vision. It is worth noting that the system is limited to a pre-established object database and has strict requirements on the location of the detected objects. However, it provides the necessary foundation for the subsequent development of semantic vSLAM. In recent years, the feature point-based vSLAM system has shown outstanding accuracy and robustness in localization, so researchers have tried to build a semantic vSLAM system based on the ORB-SLAM2 algorithm framework. For example, researchers introduce 3-D rectangular into the map to construct a lightweight semantic map in [11], [36], [64]. Other researchers [43], [45], [119] adopt semantic 3-D ellipsoids to build semantic maps because of their ability to compactly represent the size, location, and orientation of landmarks. Liao et al. incorporate three spatial structure constraints based on [43], and propose a monocular object SLAM algorithm for indoor environments with fully coupled three spatial structure constraints [39]. Soon after, EAO-SLAM [47] integrates the methods of [11], [43] and improves the object pose estimation based on the iForest method, making it possible to estimate the position, pose and scale of landmarks more accurately.

However, previous works on semantic vSLAM often assume that the scenes are static, which limits its application scenarios. These methods do not obtain robustness in localization and mapping when facing dynamic environments. To solve the challenge, some researchers propose dynamic outlier detection strategies to remove dynamic objects. For example, the motion consistency check algorithm based on semantic segmentation [55], the multiview geometry algorithm based on semantic segmentation [59], and the moving object detector [35]. [120] use prior semantic information to build an efficient online probabilistic model for monitoring dynamic outliers. In [121]–[123], they turn dynamic object regions into realistic static images, improving vision-based localization and mapping tasks in dynamic environments. Moreover, other researchers focus on object motion tracking and pose estimation in the literature [124]–[126]. [127] integrates their previous works [125], [126] and proposes a novel feature-based dynamic SLAM system that leverages semantic information to localize the robot, build the environment structure, and track motions of rigid objects. They rely on a denser object feature to ensure more robust tracking than [125], and their object tracking accuracy is much better than [126], due to their method can track occluded objects.

In summary, the development of semantic vSLAM has received much attention in recent years, but many solutions are limited to specific scenarios and face many challenges for practical applications. When facing with processing a large number of semantic object measurements in a short period, effective filtering and association of semantic information are still worth further research.

TABLE II
COMPARISON OF THE PROPERTIES OF SEMANTIC VISUAL SLAM SYSTEMS.

Method	Year	Input	Full Shape	Detailed Shape	Large Scene	Dynamic Scene	FPS	online
Co-Fusion [128]	2017	RGB-D	✓	✓		✓	/	✓
MaskFusion [60]	2018	RGB-D	✓	✓			30	✓
Fusion++ [79]	2018	RGB-D	✓	✓			4-8	✓
DS-SLAM [55]	2018	RGB-D	✓	✓		✓	/	✓
Detect-SLAM [35]	2018	RGB-D	✓	✓		✓	/	✓
DynaSLAM [59]	2018	Multiple	✓	✓	✓	✓	/	✓
DynSLAM [10]	2018	Stereo	✓	✓	✓	✓	2	
Li et al. [2]	2018	Stereo	✓	✓	✓	✓	5.8	✓
EM-Fusion [61]	2019	RGB-D	✓	✓		✓	/	✓
MID-Fusion [62]	2019	RGB-D	✓	✓		✓	2-3	✓
QuadricSLAM [43]	2019	RGB	✓		✓		/	✓
CubeSLAM [11]	2019	RGB	✓		✓		10-30	✓
Liu et al. [58]	2019	RGB-D	✓	✓		✓	/	✓
HDP-SLAM [36]	2019	RGB, RGB-D	✓		✓		/	✓
Deep-SLAM++ [129]	2019	RGB-D	✓	✓			/	✓
ClusterSLAM [124]	2019	Stereo			✓	✓	7.1	✓
DXSLAM [130]	2020	RGB-D			✓		21.6	✓
AVP-SLAM [53]	2020	RGB+IMU	✓		✓		15	✓
NodeSLAM [131]	2020	RGB-D	✓	✓			/	✓
EAO-SLAM [47]	2020	RGB	✓	✓			/	
TextSLAM [51]	2020	RGB	✓	✓	✓		/	✓
ClusterVO [132]	2020	Stereo	✓	✓	✓	✓	8	✓
VDO-SLAM [127]	2020	Stereo, RGB-D	✓	✓	✓	✓	5-8	✓
Empty Cities [122]	2021	RGB	✓		✓	✓	/	✓
DSP-SLAM [68]	2021	Multiple	✓	✓	✓		10-20	✓
DynaSLAM II [65]	2021	Stereo, RGB-D	✓	✓	✓	✓	10-12	✓
Qian et al. [48]	2021	RGB-D	✓				30	✓
Sharma et al. [69]	2021	RGB-D	✓	✓			/	✓
SO-SLAM [39]	2022	RGB	✓		✓		/	✓
Chen et al. [64]	2022	RGB, RGB-D	✓		✓		/	✓

C. Comparison of Semantic vSLAM Systems

To compare semantic vSLAM more graphically, we collect thirty semantic vSLAM systems from 2017 to 2022, as shown in Table II. For each semantic vSLAM system, we describe it in nine basic dimensions, which can reveal the advantages and disadvantages of the system. The details of each dimension are shown below. Method: name of the semantic vSLAM system. Year: the year in which the method was published. Input: type of sensor used by semantic vSLAM. Full Shape: can the object be reconstructed completely? Detailed Shape: Is it possible to know the detailed Shape of the reconstructed object? Large Scene: Can it be used for large scenes? Dynamic Scene: Can it be used for dynamic scenes? FPS: Semantic vSLAM run rate(/ means the run rate is unknown). Online: Can the system run online?

As shown in the table II, different semantic vSLAM systems have their characteristics. For example, The advantage of [59], [68] is that it is suitable for various types of sensors, meeting real-time needs. [2], [10], [62], [65], [127] show advantages in reconstructing 3D objects, making up for the shortcomings of [11], [43], [124], [130]. However, reconstructing objects consumes many computing resources, making it potentially inferior to other semantic vSLAM systems in real-time performance. [65], [124], [127], [132] are applicable for dynamic outdoor scenes, and the robustness of these systems is much better than indoor or static vSLAM. From the semantic vSLAM works between 2017 and 2022, the input data types of semantic vSLAM systems are becoming increasingly multi-modal, and these works have an increasing emphasis on object reconstruction. Furthermore, the application scenes of these works are gradually developing toward large-scale and dynamic environments. The real-time performance of semantic vSLAM can meet the requirements of current applications without reconstructing the appearance of objects.

Looking at the development of SLAM in recent years, semantic vSLAM has been considered the best approach to improve the perception capability of vSLAM systems. Traditional vSLAM systems mainly use low-level geometric features for matching and localizing, such as corner points, lines, and surface features. With the introduction of semantic information, the semantic vSLAM system can perceive advanced information about the environment, such as identifying pedestrians and detecting vehicles, which greatly enriches map information and improves the localization accuracy. Currently, semantic information can be used in all stages of the traditional SLAM algorithm framework, including initialization, backend optimization, relocalization, and loop closing. However, the contradiction between the current limited computational resources and the increasing demand for computational resources of algorithms greatly hinders the development of semantic vSLAM. For instance, in semantic information extraction, the systems need to obtain real-time semantic information and require timely filtering and associating of semantic information. In addition, it should be noted that semantic vSLAM is still in the development stage, and many hidden problems need to be solved. For example, wrong object association will make the systems more vulnerable in object-level SLAM.

VI. Outlook

A. Multi-Modal Data Fusion

Some semantic SLAM works use multi-modal sensors (e.g., RGB cameras, Depth cameras, LiDAR) for pose estimation and mapping in unknown environments. Multi-modal semantic SLAM systems can be more robust and accurate in complex and dynamic environments. Because these systems incorporate multi-modal semantic information, reducing the ambiguity of object associations. Moreover, these systems more accurately recognize dynamic objects to reduce localization drift caused by dynamic objects. Of course, the processed multi-modal environmental information can be used to construct dense semantic maps. However, in complex and highly dynamic environments, the semantic information acquired by these sensors alone is no longer sufficient for real needs. Therefore, future semantic SLAM works can try to fuse more sensors (e.g., Millimeter-wave radar, Infrared cameras, and Event cameras) and prior semantic maps (e.g., 2.5D maps). While multi-modal approaches can obtain richer semantic object information and help improve the ambiguity of object associations, they also bring challenges, such as calibration and synchronization of multiple sensors, real-time fusion, and the association of multi-modal semantic information.

B. Multi-Robot Collaboration Mode

Multi-robot systems are one of the most important research directions in robotics. In the multi-robot cooperative SLAM system, mutual communication and coordination among robots can effectively utilize spatially distributed information resources and improve problem-solving efficiency. Moreover, the damage of a single robot in the system will not affect the operation of other robots, which have better fault tolerance and anti-interference than single-robot systems. In traditional SLAM research, there are two collaboration methods for multi-robot systems. One is that each client robot builds a local map individually, and the server receives and fuses all the local maps to build a globally consistent map. The other is a decentralized architecture. The premise of multi-robot collaboration is how to efficiently and accurately perform multi-robot global localization, but the appearance-based localization methods are difficult to achieve accurate localization under significant viewpoint differences and light changes. Recently, the fusion of semantic information (e.g., text information) helps the multi-robot system to be more robust, which is attributed to the appearance and context-based semantic localization methods that can perform global localization stably and accurately. In addition, multi-robots bring multi-view semantic information for semantic vSLAM. For example, in object association, observing objects from a multi-view increases the number of observations of the same object, which can effectively avoid the ambiguity problem of object association. But it also increases the computational cost simultaneously.

C. Acquisition and Association of Semantic Information

The acquisition and association of semantic information is still a problem worthy of research in semantic vSLAM systems. The current semantic information acquisition method is based on the deep learning model, and the generalization and accuracy of the model determine the accuracy of the semantic information. For example, when an object is occluded, it is easily ignored by object detection methods. As the number of object measurements accumulates, it becomes more difficult to associate object measurements to landmarks correctly. Current object association methods are often based on semantic information such as distance, orientation, and appearance. However, it is impossible to accurately associate objects by adopting the conventional methods when objects of the same category, close to each other, obscured objects or dynamic objects appear in the environment. Therefore, we need more in-depth research to mine potential semantic information constraints which can improve object association and global localization.

VII. Conclusion

This survey summarizes the recent developments in semantic information for robot vision perception, involving semantic information extraction, object association, localization, and mapping. To give the reader an overview of the current state of the field, we summarize some representative works in the survey. We introduce three types of deep learning model-based methods for obtaining semantic information: object detection, semantic segmentation, and instance segmentation. We also introduce the problem of semantic information association. We summarize the application of semantic information in vSLAM. Moreover, we collect and evaluate thirty open-source SLAM datasets. Finally, we present the differences between traditional and semantic vSLAM, listing thirty semantic vSLAM systems. Most of the references in the survey are from the last five years, and we also provide some views on the future development of semantic vSLAM.

References

[1] G.-Z. Yang, B. J. Nelson, R. R. Murphy, H. Choset, H. Christensen, S. H. Collins, P. Dario, K. Goldberg, K. Ikuta, N. Jacobstein *et al.*, "Combating covid-19—the role of robotics in managing public health and infectious diseases," p. eabb5589, 2020.

[2] P. Li, T. Qin *et al.*, "Stereo vision-based semantic 3d object and ego-motion tracking for autonomous driving," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 646–661.

[3] J. Qian, K. Chen, Q. Chen, Y. Yang, J. Zhang, and S. Chen, "Robust visual-lidar simultaneous localization and mapping system for uav," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[4] J. Liu, R. Liu, K. Chen, J. Zhang, and D. Guo, "Collaborative visual inertial slam for multiple smart phones," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 553–11 559.

[5] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1352–1359.

[6] X. Li and R. Belaroussi, "Semi-dense 3d semantic mapping from monocular slam," *arXiv preprint arXiv:1611.04144*, 2016.

[7] L. Ma, J. Stückler, C. Kerl, and D. Cremers, "Multi-view deep learning for consistent semantic mapping with rgb-d cameras," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 598–605.

[8] N. Sünderhauf, T. T. Pham, Y. Latif, M. Milford, and I. Reid, "Meaningful maps with object-oriented semantic mapping," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5079–5085.

[9] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and automation (ICRA)*. IEEE, 2017, pp. 4628–4635.

[10] I. A. Bârsan, P. Liu, M. Pollefeys, and A. Geiger, "Robust dense mapping for large-scale dynamic environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7510–7517.

[11] S. Yang and S. Scherer, "Cubeslam: Monocular 3-d object slam," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.

[12] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[13] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (slam): Part ii," *IEEE robotics & automation magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[14] S. Saeedi, M. Trentini, M. Seto, and H. Li, "Multiple-robot simultaneous localization and mapping: A review," *Journal of Field Robotics*, vol. 33, no. 1, pp. 3–46, 2016.

[15] G. Younes, D. Asmar, E. Shammas, and J. Zelek, "Keyframe-based monocular slam: design, survey, and future directions," *Robotics and Autonomous Systems*, vol. 98, pp. 67–88, 2017.

[16] M. R. U. Saputra, A. Markham, and N. Trigoni, "Visual slam and structure from motion in dynamic environments: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–36, 2018.

[17] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: A survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, pp. 1–11, 2017.

[18] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[19] 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," *arXiv preprint arXiv:2006.12567*, 2020.

[20] M. Sualeh and G.-W. Kim, "Simultaneous localization and mapping in the epoch of semantics: a survey," *International Journal of Control, Automation and Systems*, vol. 17, no. 3, pp. 729–742, 2019.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[23] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[24] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Computer Vision and Pattern Recognition*. Springer Berlin/Heidelberg, Germany, 2018, pp. 1804–2767.

[25] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[27] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[30] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[31] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[32] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[33] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[34] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[35] F. Zhong, S. Wang, Z. Zhang, and Y. Wang, "Detect-slam: Making object detection and slam mutually beneficial," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1001–1010.

[36] J. Zhang, M. Gui, Q. Wang, R. Liu, J. Xu, and S. Chen, "Hierarchical topic model based object association for semantic slam," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 11, pp. 3052–3062, 2019.

[37] K. Doherty, D. Fourie, and J. Leonard, "Multimodal semantic slam with probabilistic data association," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 2419–2425.

[38] K. J. Doherty, D. P. Baxter, E. Schneeweiss, and J. J. Leonard, "Probabilistic data association via mixture models for robust semantic slam," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1098–1104.

[39] 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 Robotics and Automation Letters*, vol. 7, no. 2, pp. 4008–4015, 2022.

[40] H. Bavle, S. Manthe, P. De La Puente, A. Rodriguez-Ramos, C. Sampedro, and P. Campoy, "Stereo visual odometry and semantics based localization of aerial robots in indoor environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1018–1023.

[41] E. Sucar and J.-B. Hayet, "Bayesian scale estimation for monocular slam based on generic object detection for correcting scale drift," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5152–5158.

[42] M. Jayasuriya, J. Arukgoda, R. Ranasinghe, and G. Dissanayake, "Localising pmds through cnn based perception of urban streets," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6454–6460.

[43] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2018.

[44] M. Hosseinzadeh, K. Li, Y. Latif, and I. Reid, "Real-time monocular object-model aware sparse slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7123–7129.

[45] K. Ok, K. Liu, K. Frey, J. P. How, and N. Roy, "Robust object-based slam for high-speed autonomous navigation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 669–675.

[46] M. Shan, Q. Feng, and N. Atanasov, "Orcvio: Object residual constrained visual-inertial odometry," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5104–5111.

[47] 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 *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4966–4973.

[48] Z. Qian, K. Patath, J. Fu, and J. Xiao, "Semantic slam with autonomous object-level data association," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 203–11 209.

[49] A. Iqbal and N. R. Gans, "Localization of classified objects in slam using nonparametric statistics and clustering," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 161–168.

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

[51] B. Li, D. Zou, D. Sartori, L. Pei, and W. Yu, "Textslam: Visual slam with planar text features," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2102–2108.

[52] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5551–5560.

[53] T. Qin, T. Chen, Y. Chen, and Q. Su, "Avp-slam: Semantic visual mapping and localization for autonomous vehicles in the parking lot," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5939–5945.

[54] V. Murali, H.-P. Chiu, S. Samarasekera, and R. T. Kumar, "Utilizing semantic visual landmarks for precise vehicle navigation," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–8.

[55] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "Ds-slam: A semantic visual slam towards dynamic environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1168–1174.

[56] P. Ganti and S. L. Waslander, "Network uncertainty informed semantic feature selection for visual slam," in *2019 16th Conference on Computer and Robot Vision (CRV)*. IEEE, 2019, pp. 121–128.

[57] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[58] Y. Liu, Y. Petillot, D. Lane, and S. Wang, "Global localization with object-level semantics and topology," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4909–4915.

[59] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "Dynaslam: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4076–4083, 2018.

[60] M. Runz, M. Buffier, and L. Agapito, "Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2018, pp. 10–20.

[61] M. Strecke and J. Stuckler, "Em-fusion: Dynamic object-level slam with probabilistic data association," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5865–5874.

[62] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger, "Mid-fusion: Octree-based object-level multi-instance dynamic slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5231–5237.

[63] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[64] K. Chen, J. Liu, Q. Chen, Z. Wang, and J. Zhang, "Accurate object association and pose updating for semantic slam," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[65] B. Bescos, C. Campos, J. D. Tardós, and J. Neira, "Dynaslam ii: Tightly-coupled multi-object tracking and slam," *IEEE robotics and automation letters*, vol. 6, no. 3, pp. 5191–5198, 2021.

[66] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[67] J. Li, D. Meger, and G. Dudek, "Semantic mapping for view-invariant relocalization," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7108–7115.

[68] J. Wang, M. Rünz, and L. Agapito, "Dsp-slam: Object oriented slam with deep shape priors," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 1362–1371.

[69] A. Sharma, W. Dong, and M. Kaess, "Compositional and scalable object slam," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 626–11 632.

[70] E. Stenborg, C. Toft, and L. Hammarstrand, "Long-term visual localization using semantically segmented images," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 6484–6490.

[71] K.-N. Lianos, J. L. Schonberger, M. Pollefeys, and T. Sattler, "Vso: Visual semantic odometry," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 234–250.

[72] A. Gawel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, "X-view: Graph-based semantic multi-view localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1687–1694, 2018.

[73] S. Li and D. Lee, "Rgb-d slam in dynamic environments using static point weighting," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 2263–2270, 2017.

[74] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[75] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 15–22.

[76] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard, and J. McDonald, "Real-time large-scale dense rgb-d slam with volumetric fusion," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 598–626, 2015.

[77] T. Whelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison, "Elasticfusion: Dense slam without a pose graph." Robotics: Science and Systems, 2015.

[78] K. Tateno, F. Tombari, I. Laina, and N. Navab, "Cnn-slam: Real-time dense monocular slam with learned depth prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6243–6252.

[79] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level slam," in *2018 international conference on 3D vision (3DV)*. IEEE, 2018, pp. 32–41.

[80] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[81] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 573–580.

[82] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," in *2014 IEEE international conference on Robotics and automation (ICRA)*. IEEE, 2014, pp. 1524–1531.

[83] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[84] Y. Liu, Y. Fu, F. Chen, B. Goossens, W. Tao, and H. Zhao, "Simultaneous localization and mapping related datasets: A comprehensive survey," *arXiv preprint arXiv:2102.04036*, 2021.

[85] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4909–4916.

[86] X. Shi, D. Li, P. Zhao, Q. Tian, Y. Tian, Q. Long, C. Zhu, J. Song, F. Qiao, L. Song *et al.*, "Are we ready for service robots? the openloris-scene datasets for lifelong slam," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 3139–3145.

[87] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6433–6438.

[88] M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "Radiate: A radar dataset for automotive perception in bad weather," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1–7.

[89] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "Mulran: Multimodal range dataset for urban place recognition," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6246–6253.

[90] J. Engel, V. Usenko, and D. Cremers, "A photometrically calibrated benchmark for monocular visual odometry," *arXiv preprint arXiv:1607.02555*, 2016.

[91] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban dataset with multi-level sensors from highly diverse urban environments," *The International Journal of Robotics Research*, vol. 38, no. 6, pp. 642–657, 2019.

[92] W. Wen, Y. Zhou, G. Zhang, S. Fahandezh-Saadi, X. Bai, W. Zhan, M. Tomizuka, and L.-T. Hsu, "Urbanloco: A full sensor suite dataset for mapping and localization in urban scenes," in *2020 IEEE International*

Conference on Robotics and Automation (ICRA). IEEE, 2020, pp. 2310–2316.

[93] K. Minoda, F. Schilling, V. Wüest, D. Floreano, and T. Yairi, "Viode: A simulated dataset to address the challenges of visual-inertial odometry in dynamic environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1343–1350, 2021.

[94] A. Patil, S. Malla, H. Gang, and Y.-T. Chen, "The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9552–9557.

[95] S. Saeedi, E. D. Carvalho, W. Li, D. Tzoumanikas, S. Leutenegger, P. H. Kelly, and A. J. Davison, "Characterizing visual localization and mapping datasets," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6699–6705.

[96] S. Klenk, J. Chui, N. Demmel, and D. Cremers, "Tum-vie: The tum stereo visual-inertial event dataset," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8601–8608.

[97] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5000–5007.

[98] P. Kirsanov, A. Gaskarov, F. Konokhov, K. Sofiiuk, A. Vorontsova, I. Slinko, D. Zhukov, S. Bykov, O. Barinova, and A. Konushin, "Discoman: Dataset of indoor scenes for odometry, mapping and navigation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2470–2477.

[99] E. Alberti, A. Tavera, C. Masone, and B. Caputo, "Idda: a large-scale multi-domain dataset for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5526–5533, 2020.

[100] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, and J. Lin, "A* 3d dataset: Towards autonomous driving in challenging environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2267–2273.

[101] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," *arXiv preprint arXiv:2001.10773*, 2020.

[102] H. E. Keen, Q. H. Jan, and K. Berns, "Drive on pedestrian walk. tuk campus dataset," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 3822–3828.

[103] Z. Wang, S. Ding, Y. Li, J. Fenn, S. Roychowdhury, A. Wallin, L. Martin, S. Ryvola, G. Sapiro, and Q. Qiu, "Cirrus: A long-range bi-pattern lidar dataset," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5744–5750.

[104] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and service robotics*. Springer, 2018, pp. 621–635.

[105] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[106] E. Palazzolo, J. Behley, P. Lottes, P. Giguere, and C. Stachniss, "Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7855–7862.

[107] J. Delmerico, T. Cieslewski, H. Rebecq, M. Faessler, and D. Scaramuzza, "Are we ready for autonomous drone racing? the uzh-fpv drone racing dataset," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6713–6719.

[108] Z. Yan, L. Sun, T. Krajník, and Y. Ruichek, "Eu long-term dataset with multiple sensors for autonomous driving," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 697–10 704.

[109] M. Ramezani, Y. Wang, M. Camurri, D. Wisth, M. Mattamala, and M. Fallon, "The newer college dataset: Handheld lidar, inertial and vision with ground truth," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4353–4360.

[110] W. Chen, Z. Liu, H. Zhao, S. Zhou, H. Li, and Y.-H. Liu, "Cuhk-ahu dataset: Promoting practical self-driving applications in the complex airport logistics, hill and urban environments," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4283–4288.

[111] A. Ligocki, A. Jelinek, and L. Zalud, "Brno urban dataset-the new data for self-driving agents and mapping tasks," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3284–3290.

[112] Y. Cheng, M. Jiang, J. Zhu, and Y. Liu, "Are we ready for unmanned surface vehicles in inland waterways? the usvinland multisensor dataset and benchmark," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3964–3970, 2021.

[113] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[114] A. I. Mourikis, S. I. Roumeliotis *et al.*, "A multi-state constraint kalman filter for vision-aided inertial navigation." in *ICRA*, vol. 2, 2007, p. 6.

[115] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM international symposium on mixed and augmented reality*. IEEE, 2007, pp. 225–234.

[116] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2320–2327.

[117] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[118] C. Kerl, J. Sturm, and D. Cremers, "Dense visual slam for rgb-d cameras," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 2100–2106.

[119] C. Rubino, M. Crocco, and A. Del Bue, "3d object localisation from multi-view image detections," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1281–1294, 2017.

[120] N. Brasch, A. Bozic, J. Lallemand, and F. Tombari, "Semantic monocular slam for highly dynamic environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 393–400.

[121] B. Bescos, J. Neira, R. Siegwart, and C. Cadena, "Empty cities: Image inpainting for a dynamic-object-invariant space," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5460–5466.

[122] B. Bescos, C. Cadena, and J. Neira, "Empty cities: A dynamic-object-invariant space for visual slam," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 433–451, 2020.

[123] B. Besic and A. Valada, "Dynamic object removal and spatio-temporal rgb-d inpainting via geometry-aware adversarial learning," *IEEE Transactions on Intelligent Vehicles*, 2022.

[124] J. Huang, S. Yang, Z. Zhao, Y.-K. Lai, and S.-M. Hu, "Clusterslam: A slam backend for simultaneous rigid body clustering and motion estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5875–5884.

[125] M. Henein, J. Zhang, R. Mahony, and V. Ila, "Dynamic slam: The need for speed," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2123–2129.

[126] J. Zhang, M. Henein, R. Mahony, and V. Ila, "Robust ego and object 6-dof motion estimation and tracking," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5017–5023.

[127] ——, "Vdo-slam: a visual dynamic object-aware slam system," *arXiv preprint arXiv:2005.11052*, 2020.

[128] M. Rünz and L. Agapito, "Co-fusion: Real-time segmentation, tracking and fusion of multiple objects," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4471–4478.

[129] L. Hu, W. Xu, K. Huang, and L. Kneip, "Deep-slam++: Object-level rgbd slam based on class-specific deep shape priors," *arXiv preprint arXiv:1907.09691*, 2019.

[130] D. Li, X. Shi, Q. Long, S. Liu, W. Yang, F. Wang, Q. Wei, and F. Qiao, "Dxslam: A robust and efficient visual slam system with deep features," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4958–4965.

[131] E. Sucar, K. Wada, and A. Davison, "Neural object descriptors for multi-view shape reconstruction," 2020.

[132] J. Huang, S. Yang, T.-J. Mu, and S.-M. Hu, "Clustervo: Clustering moving instances and estimating visual odometry for self and surroundings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2168–2177.