Intelligent Systems with Applications 28 (2025) 200591

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Review

Semantic SLAM: A comprehensive survey of methods and applications

Houssein Kanso [a], Abhilasha Singh [a], Etaf El Zarif [b], Nooruldeen Almohammed [a], Jinane Mounsef [a,*], Noel Maalouf [c], Bilal Arain [d]

[a] Department of Electrical Engineering and Computing, Rochester Institute of Technology, Dubai, United Arab Emirates
[b] Department of Electrical and Computer Engineering, American University of Beirut, Lebanon
[c] Department of Electrical and Computer Engineering, Lebanese American University, Beirut, Lebanon
[d] Department of Computer Engineering, University of Sharjah, Sharjah, United Arab Emirates

ARTICLE INFO

Keywords:
Semantic SLAM
Object-level SLAM
Visual SLAM
Semantic mapping
Dynamic environments
Autonomous systems

ABSTRACT

This paper surveys the different approaches in semantic Simultaneous Localization and Mapping (SLAM), exploring how the incorporation of semantic information has enhanced performance in both indoor and outdoor settings, while highlighting key advancements in the field. It also identifies existing gaps and proposes potential directions for future improvements to address these issues. We provide a detailed review of the fundamentals of semantic SLAM, illustrating how incorporating semantic data enhances scene understanding and mapping accuracy. The paper presents semantic SLAM methods and core techniques that contribute to improved robustness and precision in mapping. A comprehensive overview of commonly used datasets for evaluating semantic SLAM systems is provided, along with a discussion of performance metrics used to assess their efficiency and accuracy. To demonstrate the reliability of semantic SLAM methodologies, we reproduce selected results from existing studies offering insights into the reproducibility of these approaches. The paper also addresses key challenges such as real-time processing, dynamic scene adaptation, and scalability while highlighting future research directions. Unlike prior surveys, this paper uniquely combines (i) a systematic taxonomy of semantic SLAM approaches across different sensing modalities and environments, (ii) a comparative review of datasets and evaluation metrics, and (iii) a reproducibility study of selected methods. To our knowledge, this is the first survey that integrates methods, datasets, evaluation practices, and application insights into a single comprehensive review, thereby offering a unified reference for researchers and practitioners. In conclusion, this review underscores the vital role of semantic SLAM in driving advancements in autonomous systems and intelligent navigation by analyzing recent developments, validating findings, and highlighting future research directions.

1. Introduction

Simultaneous Localization and Mapping (SLAM) is a technique that allows a mobile robot or device to create a map of its environment while simultaneously determining its position within that map (Azzam, Taha, Huang, & Zweiri, 2020; Cadena et al., 2016). This dual capability is essential for autonomous systems to function effectively in unknown static or dynamic environments. In robotics, SLAM provides the essential ability for robots to navigate and interact with their surroundings without predefined maps. This capability is crucial for applications ranging from autonomous vehicles (Takleh, Bakar, Rahman, Hamzah, & Aziz, 2018) to indoor mobile robots (Fang et al., 2021; Tian et al., 2025; Yousif, Bab-Hadiashar, & Hoseinnezhad, 2015) and drones (Antonini, Guerra, Murali, Sayre-McCord, & Karaman, 2020; Fu et al., 2023; Motlagh, Lotfi, Taghirad, & Germi, 2019). In autonomous navigation, SLAM allows vehicles to move safely and efficiently through complex and changing environments, such as urban streets or warehouses, by continuously updating their maps.

The primary goal of SLAM is to solve the coupled problem of mapping an environment and localizing the robot within that environment. This involves two main tasks: localization, which determines the position and orientation of the device, and mapping, which builds a coherent map of the environment based on sensory inputs. These tasks must be performed simultaneously because the accuracy of the map depends on the precise location of the device, and vice versa. This interdependence makes SLAM a challenging problem in robotics and computer vision.

Traditional SLAM methods are typically categorized into two main types: direct methods and indirect (feature-based) methods. Direct

* Corresponding author.
E-mail address: jmbcad@rit.edu (J. Mounsef).

https://doi.org/10.1016/j.iswa.2025.200591
Received 1 July 2025; Received in revised form 3 September 2025; Accepted 5 October 2025
Available online 10 November 2025

methods operate directly on raw image or depth data, such as pixel intensities or geometric measurements. These approaches estimate camera motion by minimizing the photometric or geometric error between consecutive frames, often using dense or semi-dense regions of the image. This makes them effective in environments with little texture, though they tend to be more sensitive to lighting variations. Prominent examples include LSD-SLAM (Engel, Schöps, & Cremers, 2014) and DSO (Engel, Koltun, & Cremers, 2017). In contrast, indirect methods, also called feature-based methods, first extract salient features from the environment, such as corners or blobs, using detectors like ORB, SIFT, or SURF. These features are matched across frames to estimate motion and reconstruct the environment. To refine both camera poses and 3D feature positions, many indirect SLAM systems employ bundle adjustment (Wang, Ma, Ren and Lu, 2021), an optimization technique that minimizes the overall reprojection error across multiple keyframes. Indirect methods are generally more robust to illumination changes and partial occlusions but may perform poorly in low-texture or repetitive environments. Notable examples include Parallel Tracking and Mapping (PTAM) (Klein & Murray, 2007) and ORB-SLAM (Mur-Artal, Montiel, & Tardos, 2015). Therefore, the foundation of many SLAM systems lies in the distinction between working with raw image data and tracking discrete visual features. Semantic SLAM builds upon both types by incorporating higher-level understanding, such as object labels and contextual information, which enables more accurate and intelligent mapping in complex, dynamic environments.

Traditional SLAM techniques primarily focus on the geometric and characteristic aspects of the environment. However, semantic SLAM introduces a significant advancement by incorporating high-level semantic information, allowing a better understanding of the environment in the scene, as shown in Fig. 1. The inclusion of semantic information allows for enhanced mapping by recognizing and labeling objects in the environment, creating maps that include not just the geometry but also the identity and function of different elements, which is more useful for tasks that require contextual awareness. Improved localization is achieved as semantic information helps in more robust data association and loop closure detection, reducing drift, and improving the accuracy of the SLAM solution. Semantic SLAM enables robots to perform more complex tasks that require an understanding of the environment, such as object manipulation, human–robot interaction, and autonomous driving in complex urban settings. It enhances performance in dynamic environments by effectively distinguishing between static and moving objects, enabling more adaptive, resilient navigation, and accurate mapping (Chen, Liu, Chen, Wang and Zhang, 2022; Li, Ye et al., 2025; Qian, Patath, Fu, & Xiao, 2021; Ran, Yuan, Zhang, Wu, & He, 2022).

The origins of SLAM can be traced back to the robotics and computer vision communities in the late 1980s and early 1990s. Initial approaches were primarily focused on probabilistic methods for basic localization and mapping. A seminal paper by Smith and Cheeseman in 1986 introduced the foundational concept of using probability theory to resolve uncertainties in robot navigation, which later became integral to SLAM (Smith & Cheeseman, 1986). The real turning point in SLAM research occurred in the late 1990s and early 2000s with the introduction of the EKF SLAM (Dissanayake, Newman, Clark, Durrant-Whyte, & Csorba, 2001). This method significantly improved the accuracy and reliability of the SLAM algorithms. As computational capabilities expanded, the mid-2000s saw the development of FastSLAM, a particle filter-based approach that addressed some of the scalability issues of EKF SLAM. FastSLAM provided a more efficient way to handle the SLAM problem by separating the mapping and the localization problems (Montemerlo, Thrun, Koller, & Wegbreit, 2002). A comprehensive tutorial on the state of SLAM detailed its challenges and emphasized the role of EKF in advancing the field (Durrant-Whyte & Bailey, 2006).

The introduction of Visual SLAM (vSLAM) marked a significant evolution in the field, replacing expensive laser sensors with cameras

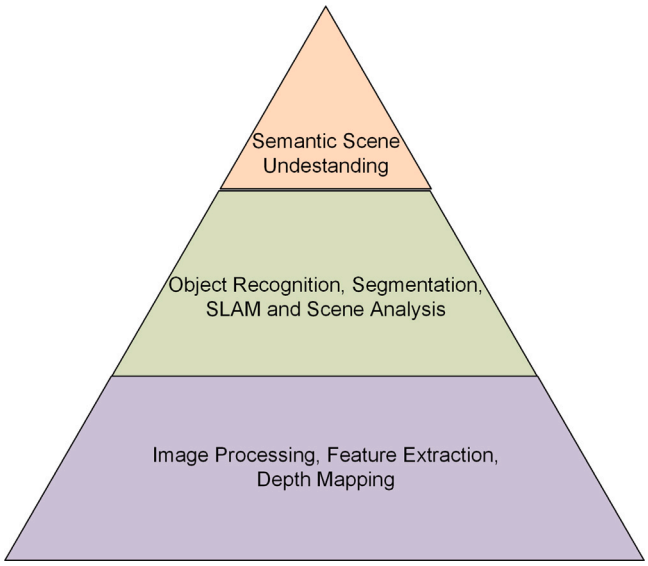

Fig. 1. Hierarchical structure of semantic SLAM systems.

to provide richer environmental data at lower cost (Davison, 2003). This camera-based approach gained further momentum with PTAM system, which demonstrated that a single small camera could effectively perform simultaneous tracking and mapping, making SLAM technology more accessible and versatile (Klein & Murray, 2007).

Traditional SLAM systems create geometric maps but lack understanding of object identities and scene semantics. Semantic SLAM addresses this limitation by combining spatial mapping with object recognition, enabling robots to build meaningful environmental representations. A foundational contribution was the introduction of probabilistic data association methods that reliably link semantic observations across multiple viewpoints (Bowman, Atanasov, Daniilidis, & Pappas, 2017). This work demonstrated how to effectively combine semantic labels with SLAM, providing the groundwork for future developments. SemanticFusion combined dense 3D mapping with convolutional neural networks to label and understand the environment dynamically. This approach showed that real-time semantic mapping was feasible and practical (McCormac, Handa, Davison, & Leutenegger, 2017). The evolution from CNN-based labeling to sophisticated inference methods, such as dynamic dense CRFs that maintain semantic consistency across temporal sequences, demonstrates the field's progression toward robust semantic understanding (You, Luo, Zhou, & Zhu, 2023).

The field of SLAM continues to evolve with advancements in deep learning and artificial intelligence. Semantic SLAM is increasingly focused on enhancing the interaction between autonomous systems and their environments (Chen, Wei, Lin and Lin, 2025), moving towards adaptive, intelligent systems. The detailed timeline of the evolution of SLAM and introduction of semantic SLAM can be visualized in Fig. 2. The integration of semantics into SLAM represents a pivotal advancement in robotics, offering the potential for more intuitive and intelligent machine behavior and significantly expanding the applications of robotic systems in complex environments.

Recently, interest in semantic SLAM and scene understanding has grown significantly, as evidenced by increasing publication numbers. Fig. 3 presents research trends from 2015 to 2025 for semantic SLAM, indoor scene understanding, and outdoor scene understanding, based on articles, proceedings, and book chapters from artificial intelligence, robotics, and engineering fields. The data reveals steady growth in semantic SLAM and indoor scene understanding research, while outdoor scene understanding shows particularly rapid growth in recent years, with continued expansion expected. Fig. 4 illustrates the distri-

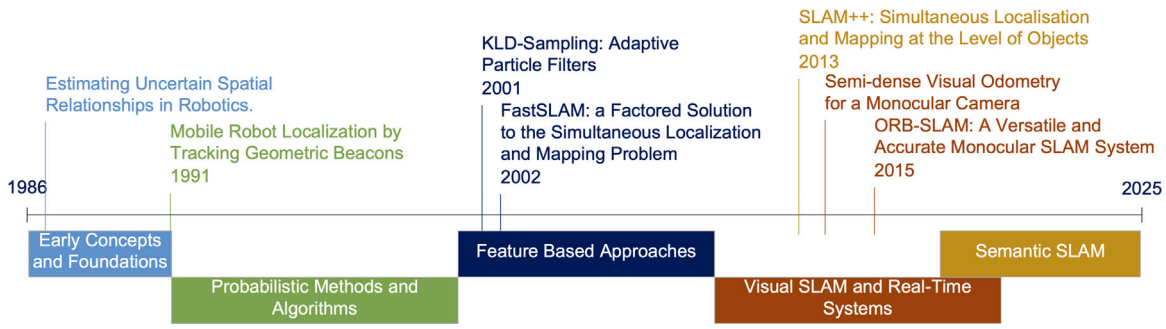

Fig. 2. History and evolution of SLAM over the years from 1986 to 2025.

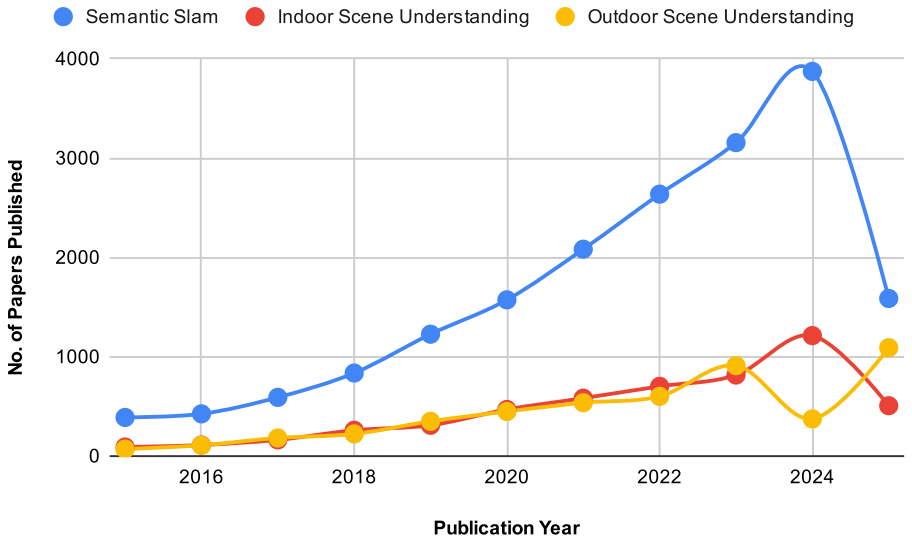

Fig. 3. Illustration of the number of papers published between 2015 to 2025 in the areas of semantic SLAM, indoor scene understanding, and outdoor scene understanding, highlighting the evolving research focus and growing interest in these fields over time.
Source: Digital Science (2024).

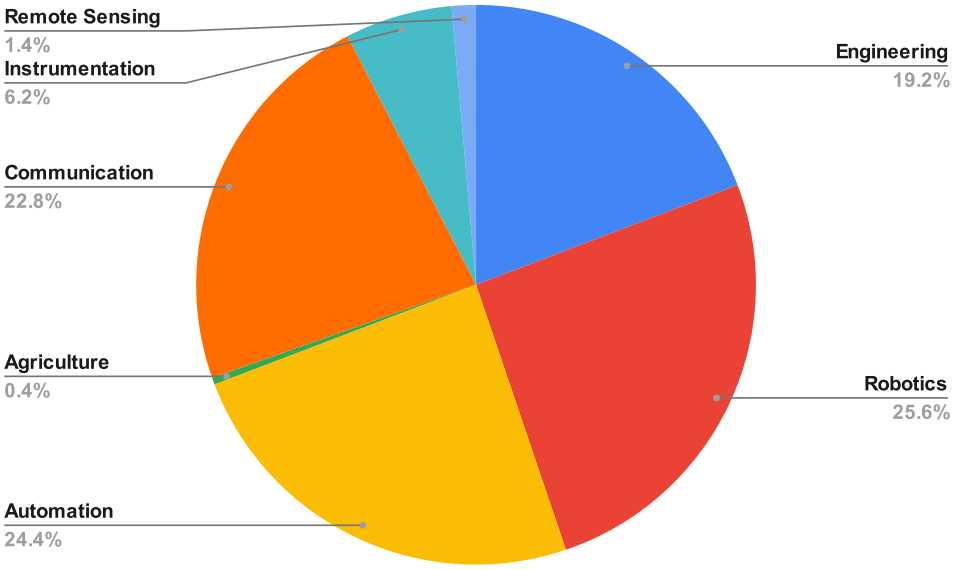

Fig. 4. Pie chart illustrating the distribution of research areas in semantic SLAM, where 24.4% of the research is in automation and 25.6% in remote sensing.

bution of semantic SLAM research across domains: robotics leads with 25.6% of publications, followed by automation (24.4%), communication (22.8%), and engineering (19.2%), while agriculture represents only 0.4% of the research output.

These trends reflect the growing interest and advancements in integrating semantic information and scene understanding in complex environments. Although significant progress has been made in geometric and feature-based SLAM from 2015 to 2025, research specifically

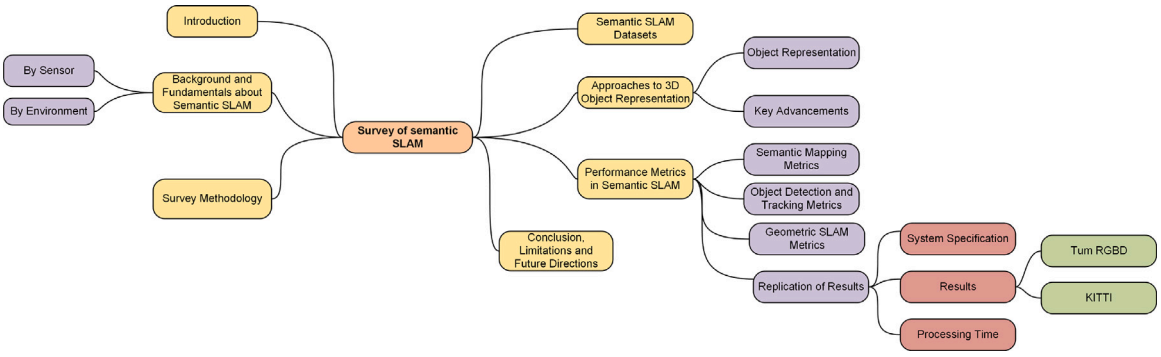

Fig. 5. Schematic diagram of the overall paper structure discussing various sections and their sub-sections.

dedicated to semantic SLAM remains limited. This presents a valuable opportunity for further exploration and development in this emerging field. Additionally, there is a limited number of review papers on semantic SLAM, highlighting the need for more comprehensive surveys and studies. Taking into account these gaps, the major contributions of this survey paper are listed below:

- Comprehensive coverage: We provide the most up-to-date and systematic review of semantic SLAM methods, filling the gap left by earlier surveys that focused only on geometric SLAM or isolated aspects of semantics.
- Dataset and metrics analysis: Unlike prior reviews, we analyze commonly used datasets and performance metrics in semantic SLAM, highlighting their strengths, limitations, and suitability for different scenarios.
- Taxonomy and categorization: We propose a structured taxonomy that organizes semantic SLAM approaches by sensor type, environmental context, and object representation, offering a clearer perspective than existing fragmented overviews.
- Reproducibility and benchmarking: We reproduce selected results from prior studies, which have not been addressed in previous surveys, to provide insights into reproducibility and reliability of semantic SLAM methods.
- Future outlook: We identify unresolved challenges, such as scalability, robustness in dynamic environments, and real-time semantic integration, and suggest research directions that go beyond the scope of earlier surveys.

The remainder of this paper is organized as follows (see Fig. 5): Section 2 describes the survey methodology. Section 3 presents the fundamentals of semantic SLAM, covering sensor types, environments, and approaches to 3D object representation. Section 4 reviews commonly used datasets for SLAM evaluation. Section 5 presents advancements in semantic SLAM methods. Section 6 highlights various applications of semantic SLAM while 7 discusses the practical challenges of semantic SLAM. Section 8 presents performance metrics for qualitative and quantitative assessment, including reproducing results from open-source implementations. Section 9 explores future research directions, and Section 10 presents our conclusions.

2. Survey methodology

Semantic SLAM has emerged as a pivotal area of research, yet several key challenges and gaps remain unaddressed in the current literature. While numerous surveys provide overviews of general SLAM technologies, there is a notable lack of comprehensive reviews specifically dedicated to semantic SLAM (Hughes et al., 2024; Xia, Li, Yi, Ruan, & Zhang, 2024). Compared to other forms of SLAM, semantic SLAM has been underrepresented in the literature. To the best of our knowledge, there are limited comprehensive surveys available on semantic SLAM. Most available surveys deal with different aspects of SLAM technologies, such as visual SLAM (Li, Wang and Gu, 2018; Pu, Luo, Wang, Huang, & Liu, 2023; Sahili et al., 2023; Wang, Tian, Chen, Xu and Ding, 2024; Zhang, Wang, & Su, 2021), which focus on visual data for mapping and navigation. This gap highlights the need for a dedicated survey that consolidates existing research on semantic SLAM, evaluating methodologies, effectiveness, and applications. Our paper addresses this void by performing a systematic review of the available literature on semantic SLAM and proposing future research directions, considering the advantages and limitations of current approaches. It also gives a global view of semantic SLAM, exploring how semantic data can be integrated into SLAM frameworks and its far-reaching impact on enhancing robotic perception and autonomous navigation. This section outlines the survey methodology adopted in this study. We define the inclusion and exclusion criteria to ensure that only the most relevant and high-quality articles are considered, thereby maintaining the rigor and reliability of our systematic review.

2.1. Search strategy

We begin by reviewing several key articles, majorly focused on semantic SLAM and scene understanding. To ensure comprehensive coverage of the relevant literature, we create a carefully structured search query, combining several terms associated with semantic SLAM and scene understanding. One such query used is: semantic SLAM, semantic SLAM with scene understanding, indoor scene understanding, and outdoor scene understanding. To enhance the search results, we additionally used author keywords like semantic SLAM for scene understanding-related papers.

To ensure a comprehensive and unbiased review, we adhered to a strict survey methodology with clearly defined inclusion and exclusion criteria as illustrated in Fig. 6. Articles were selected based on the following focused criteria:

- Paper Inclusion Criteria
  - Review articles, proceedings, and journals
  - Articles written in English
  - Articles published between 2015 and 2025
  - Articles focusing on semantic SLAM, indoor scene understanding, and outdoor scene understanding
  - Articles with SCI, SCIE, and Conference Proceedings citation index
  - Articles focusing on research areas like robotics, computer science, and engineering
- Paper Exclusion Criteria
  - Each article will be counted only once, even if it appears in multiple digital libraries
  - Articles that are not peer-reviewed
  - Studies that do not have experimental results

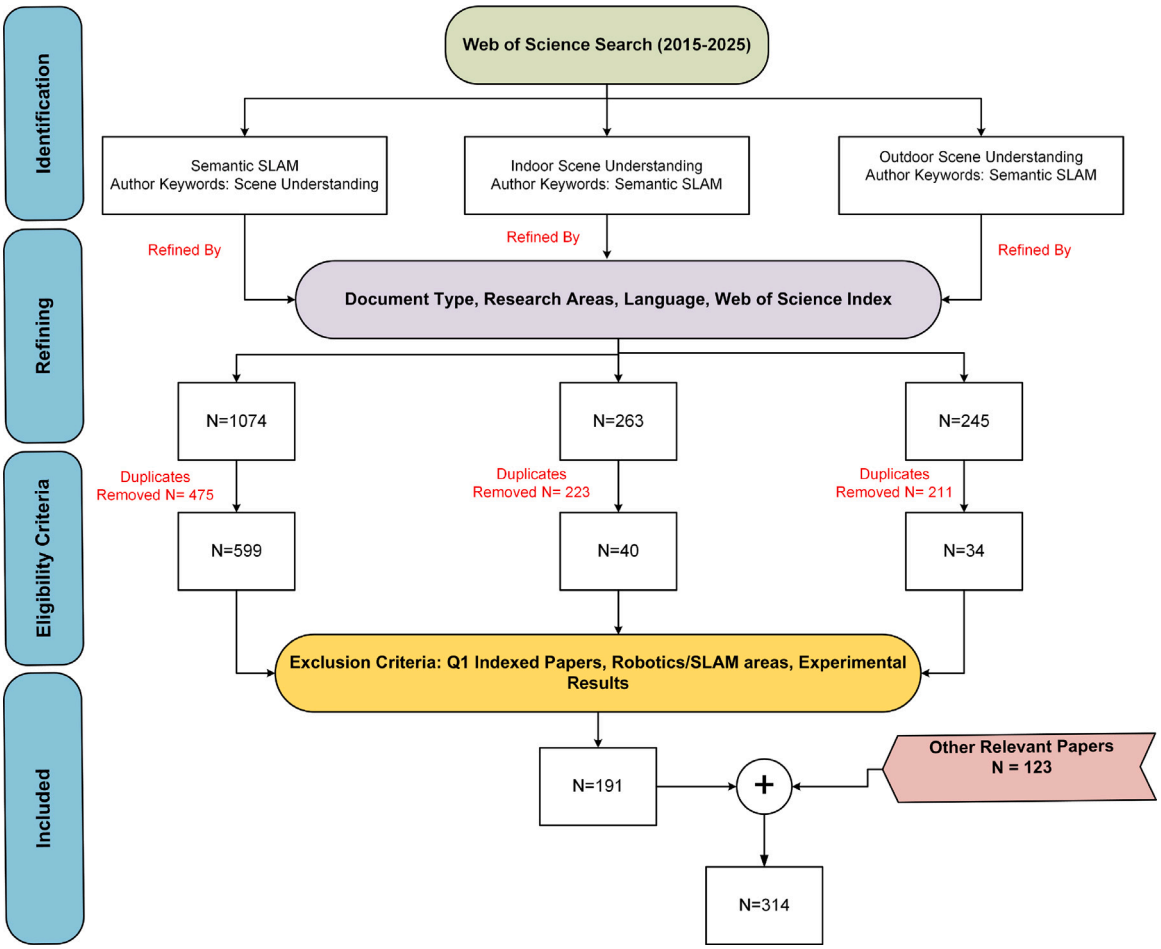

Fig. 6. Flowchart illustrating the systematic paper selection process employed in this survey, where duplicates were removed, followed by a screening of titles and abstracts for relevance. Articles were assessed against strict inclusion and exclusion criteria, focusing on semantic SLAM.

- Papers not directly related to the primary focus of semantic SLAM and robotics
- Duplicate studies from other sources not listed
- Papers not listed in Q1, with exceptions for significant contributions

2.2. Paper selection results

To ensure a thorough and methodical exploration of the research landscape, a systematic literature review was conducted based on the above-mentioned criteria, and the final results of the selected papers is illustrated as bar plot in Fig. 7.

Fig. 7 presents the annual distribution of the 191 Web of Science articles included in our survey, spanning from 2015 to 2025. The data reveal a clear upward trend in the number of publications related to semantic SLAM over the past decade. Initial contributions were relatively sparse between 2015 and 2018, with fewer than 10 papers published each year. However, from 2019 onward, there has been a noticeable increase, reflecting growing academic interest and technological advancement in the field. A significant rise began in 2020, with the number of publications more than doubling compared to previous years. This growth continued steadily, peaking in 2024 with 41 articles, followed closely by 37 publications in 2025. The sharp increase from 2020 onward suggests that semantic SLAM has become a rapidly evolving and highly active research area, likely driven by advancements in deep learning, 3D perception, and autonomous systems. This trend underscores the increasing relevance and applicability of semantic SLAM, motivating the need for a structured and comprehensive survey such as the one presented in this paper.

To enhance the breadth of our review, we also analyzed the related work sections of the initially selected articles. This allowed us to identify 123 additional papers relevant to SLAM techniques and scene understanding, bringing the total number of surveyed articles to 314. Additionally, the distribution of publication sources for the articles included in our survey is illustrated in Fig. 8. A significant majority of the reviewed works (84%) were published in peer-reviewed journals, reflecting the maturity and credibility of the research in the semantic SLAM domain. Conference papers accounted for 11%, while arXiv preprints contributed 3% and books comprised 2%. Although the proportion of arXiv articles is relatively small, those included in our study were carefully selected based on their high citation counts and demonstrated impact in the field, highlighting their relevance. After outlining the strict criteria we used to select articles, the next step is to provide the reader with the essential background of Semantic SLAM. This foundation will help contextualize the works we review later on.

3. Background and fundamentals about semantic SLAM

Here, we provide the necessary background and fundamental concepts of Semantic SLAM. This section serves to familiarize readers with the core principles, terminologies, and evolution of SLAM, establishing the foundation for the more advanced discussions that follow. Semantic SLAM augments traditional mapping and localization with high-level

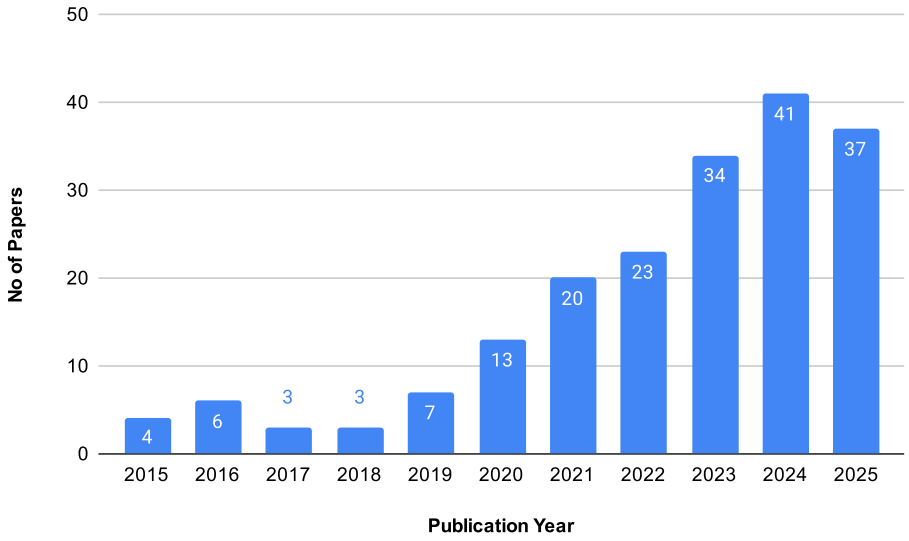

Fig. 7. Bar graph illustrating the number of publications per year in the Web of Science database from 2015 to 2025.

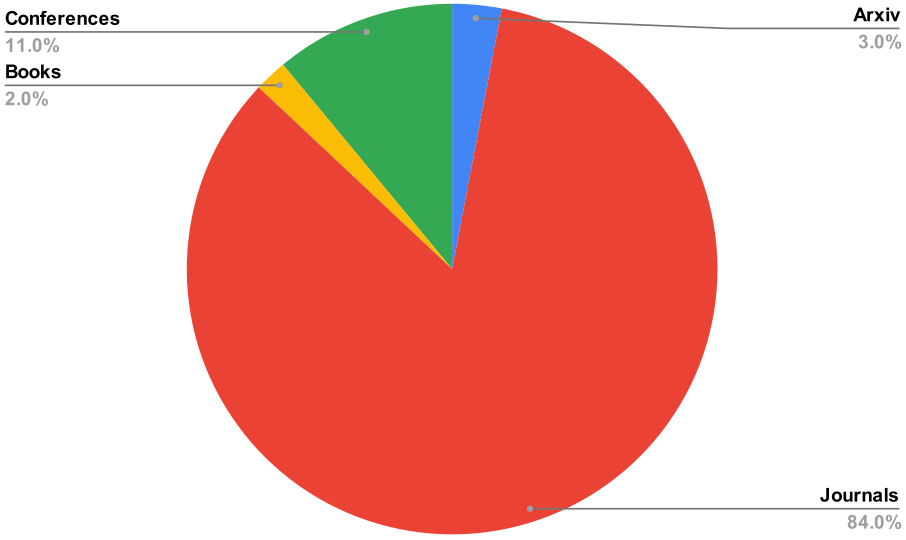

Fig. 8. A pie chart illustrating the distribution of the papers reviewed in our survey, showing that 84% of them were journal articles.

semantics (Cadena et al., 2016). By adding labels to 3D structure and layout, it provides contextual meaning for robots and humans. Geometry captures points, edges, and depth from sensors; semantics adds understanding beyond shape, improving handling of moving obstacles. A key benefit is real-time fusion of semantics with geometry to assign object identities, whereas classic SLAM yields accurate but semantically empty maps. The result is a precise geometric map enriched with labels (e.g., wall, chair, vehicle) that supports applications like autonomous driving, navigation, and environmental analysis (Atanasov, Zhu, Daniilidis, & Pappas, 2016; Azzam, Alkendi, Taha, Huang, & Zweiri, 2021; Bowman et al., 2017; Choe, Seong, & Kim, 2022; Ran, Yuan, Zhang, He et al., 2021). Modern ML and vision classify scene elements to enable intelligent interaction (Grinvald et al., 2019; Mukherjee, Das, Ghosh, Chowdhury, & Saha, 2021; Zhang et al., 2019). Robustness has improved via spatial-layout constraints and cross-view consistency (Ji et al., 2023); object-level methods refine associations and poses to keep identities consistent (Chen, Liu et al., 2022). Sparse GP regression further scales dense metric-semantic mapping for multi-robot use (Zobeidi, Koppel, & Atanasov, 2022).

Semantic extraction uses advanced deep learning to identify objects and support mapping. Beyond dense metric maps, text-based unsupervised segmentation builds topological semantic maps for assistive navigation without extensive labels (Sun, Ma, Zhou, & Cao, 2023). Together, these pipelines yield maps with both coordinates and context, enabling better robotic decision-making (Kostavelis & Gasteratos, 2017; McCormac et al., 2017; Tian, Liu, Ri, Liu, & Sun, 2019; Zhou, Yue et al., 2023). Understanding component interactions in SLAM clarifies semantic SLAM's role in autonomy (Cornejo-Lupa, Ticona-Herrera, Cardinale, & Barrios-Aranibar, 2020; Kostavelis & Gasteratos, 2015). Figs. 10 and 9 detail the workflow, highlighting how semantic labels enrich geometry to support more sophisticated navigation–environment interactions (see Fig. 9).

Adding semantics to SLAM enables object and feature recognition, so systems map the environment and interpret object purpose and context. Work is closing the gap to higher-level cognition, allowing mapping with real-time semantic recognition and categorization. The following sections survey emerging integration strategies that drive semantic SLAM forward.

Reviews often organize semantic SLAM by algorithmic style (feature-based, direct, dense), semantic integration (object- vs. scene-level), or application (indoor AR, driving, robotics). Many also contrast vision and LiDAR, or static scenes with semantic overlay versus dynamic-object SLAM. Here we adopt a more systematic view based on sensor modality and operating environment.

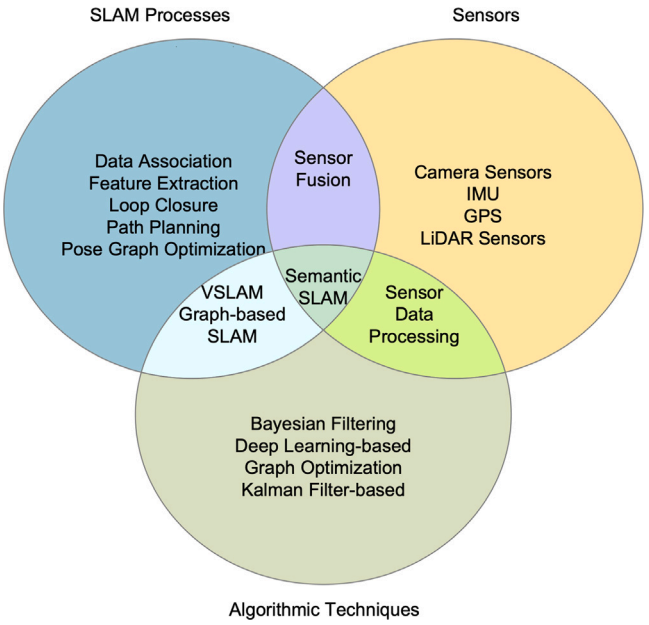

Fig. 9. Integration of processes, sensors, and algorithms in SLAM systems.

Organization by sensor type (monocular/stereo/RGB-D) classifying by sensor (monocular, stereo, RGB-D, LiDAR, or fused) aligns methods with their hardware capabilities and limits. Sensor choice directly affects map representation and tracking accuracy (Chen, Xiao et al., 2025). Distinguishing single-sensor from fusion setups (e.g., visual–inertial, vision–LiDAR) reflects common practice for robustness. Grouping by sensor enables fair comparisons under the same information regime (e.g., stereo-based dynamic SLAMs).

Separating indoor/outdoor (and structured/unstructured) is also useful. Semantics and motion statistics differ: indoors feature walls, furniture, and people; outdoors include vehicles, pedestrians, and open spaces. Conditions like lighting, texture, and weather strongly affect performance. Example priors differ (planar ground outdoors; ceiling planes indoors). This split enables environment-wide comparisons (e.g., indoor RGB-D vs. outdoor stereo+LiDAR).

Putting these axes together gives a taxonomy such as: Monocular (with/without IMU) – Indoor, Stereo – Outdoor, RGB-D – Indoor, LiDAR – Outdoor, and their sensor-fusion variants. The scheme of categorizing by the differences in sensors or scenarios of use helps understand trade-offs: e.g., monocular SLAMs forego depth sensors but can use priors (scale) and are commonly applicable in feature-rich scenes, whereas LiDAR systems are superior outdoors, but have no semantics in RGB. Organizing a survey by sensor and environment narrows the field: it clusters methods with similar operating conditions together, and strengths/weaknesses comparisons have meaning. As another illustration, a comparison between two RGB-D indoor SLAMs can emphasize their semantic fusion form, but a comparison of an RGB-D indoor approach with a stereo-LiDAR outdoor approach would confuse these totally different conditions.

3.1. Sensors

Semantic SLAM systems can be categorized based on their primary sensing modalities, each offering distinct advantages for different applications and environments. Table 1 presents a comparison of various sensor-based approaches evaluated in this survey. The following subsections examine five major categories: monocular semantic SLAM, stereo semantic SLAM, RGB-D semantic SLAM, LiDAR semantic SLAM, multi-modal semantic SLAM, and incremental semantic SLAM.

3.1.1. Monocular semantic SLAM

Monocular semantic SLAM extends traditional SLAM by adding object recognition with a single camera, lowering hardware cost and computation for small-scale devices (Cadena et al., 2016). Its major drawback is scale ambiguity, and it degrades in low-texture scenes and with moving objects (Engel et al., 2014; Gao et al., 2024). Whereas classic SLAM maps geometric features (points, lines), the semantic variant augments maps with labels from recognized objects, enabling richer interaction with the environment. The pipeline estimates camera pose and a sparse geometric map, builds 3D object models from images, recognizes objects, and fuses their positions into the SLAM map with continual refinement. Two parallel threads operate: a monocular SLAM thread for pose and geometry, and an object-recognition thread that detects objects in view and estimates their poses; recognized objects are inserted and their 3D poses are updated as more frames arrive (Han & Yang, 2023; Sun, Yuan, & Zhang, 2021). Recent advances strengthen data association with ensemble methods (Wu et al., 2020) and deep shape priors for accurate object modeling (Wang, Runz and Agapito, 2021); complementary work uses spatiotemporal consistency and graph constraints to filter out incorrect detections (Zhang, Yuan, Ran, Tao, & Wu, 2023). Geometric contour-based alignment further improves tracking under varying lighting by matching projected object boundaries across viewpoints (Lin, Wang, Xu, Zhao, & Chen, 2023). Building on these, newer frameworks introduce outlier-robust modeling via isolation forests (iForest) and semantic topological mapping, enabling higher-level tasks such as object-driven exploration and manipulation (Wu et al., 2023).

Building on the general pipeline in Fig. 10, Fig. 11 present two tightly coupled parallel threads—monocular SLAM and object recognition. The monocular SLAM thread uses an EKF to estimate camera motion and map features, employing a 1-point RANSAC-EKF for resilient data association. Upon object recognition, SLAM-derived camera pose (position and orientation) is leveraged to augment the estimate. Detected objects are transformed from their own reference frame into the SLAM frame and inserted into the map, where their 3D poses are iteratively refined by the SLAM back-end.

In the object recognition thread, SURF features are extracted from the images, and their correspondences are computed against known object models using Nearest Neighbor Distance Ratio (NNDR). Afterward, RANSAC performs a geometric consistency check to identify valid transformations. For planar objects, a homography is estimated, solving the Perspective-n-Point problem to estimate the object's translation and orientation. The object's pose is then refined using inlier correspondences and incorporated into the SLAM system, enhancing the geometric map with semantic information and enabling real-time SLAM.

A key component of this process is the equation that transforms the object's position and orientation from the object frame to the camera frame, as shown in (1):

$$H^{O}_{C_{k-m}} = H^{F}_{C_{k-m}}\left(t^{F}_{C_{k-m}}, q^{F}_{C_{k-m}}\right) H^{O}_{F}\left(t^{F}_{O}, q^{F}_{O}\right), \tag{1}$$

where t^{F}_{O} and q^{F}_{O} represent the position and orientation of face F in the object frame O, while $t^{F}_{C_{k-m}}$ and $q^{F}_{C_{k-m}}$ denote the position and orientation of face F relative to the SLAM camera C_{k-m}. This transformation integrates the object's features into the SLAM map by converting the object coordinates into the camera's coordinate frame.

The SLAM state vector x_k is then updated to include new object points. The state vector includes the camera position x_{C_k}, existing map points $y_1, \ldots, y_n$, and newly detected object points y^{F}_{W}:

$$x_k = \begin{pmatrix} x_{C_k} \\ y_1 \\ \vdots \\ y_n \\ y^{F}_{W} \end{pmatrix}. \tag{2}$$

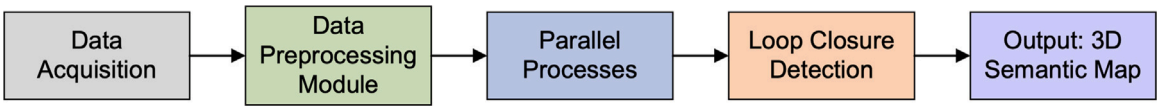

Fig. 10. General semantic SLAM framework.

Table 1
Advantages and disadvantages of different semantic SLAM sensor approaches.

Type	Advantages	Disadvantages
Monocular semantic SLAM	Low cost Lightweight High portability	Scale ambiguity Limited depth perception Vulnerable in dynamic scenes
Stereo semantic SLAM	Improved depth accuracy Enhanced environmental understanding Robustness	Higher computational cost Increased hardware complexity Higher cost
RGB-D semantic SLAM	Rich sensory data Ease of scene reconstruction Handles dynamic environments well	Limited range Sensitivity to lighting conditions Higher energy consumption
3D-Lidar based semantic SLAM	High accuracy Robust to lighting conditions Effective in large-scale environments	High cost Complexity Limited by weather conditions
Multi-modal semantic SLAM	Comprehensive understanding Robustness Improved accuracy	High computational cost Increased system complexity Costly
Incremental semantic SLAM	Continuous mapping Adaptability Resource efficiency	Complex algorithm design Drift Limited scalability

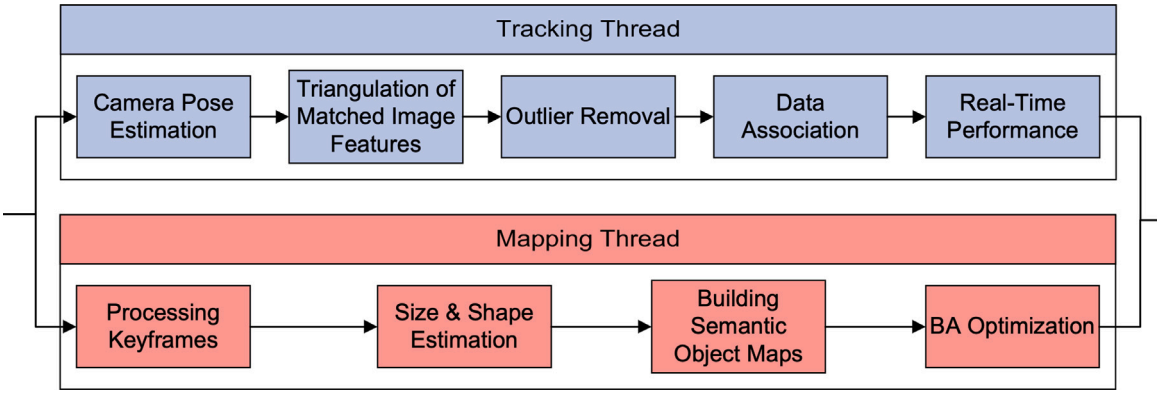

Fig. 11. Monocular semantic SLAM parallel threads.

By (2), object features enter the SLAM state via camera-frame coordinates, adding object points that EKF monocular SLAM can continually refine. Relocalization is also pivotal (Lee, Back, Hwang, & Chun, 2023a): when tracking fails—e.g., due to dynamics, dropped frames, or blur—the system realigns to a prebuilt map by matching current observations to known features, restoring robustness across varied settings (Chen et al., 2015; Civera, Galvez-Lopez, Riazuelo, Tardos, & Montiel, 2011; Li, Fu, Wang and Sun, 2025).

3.1.2. Stereo semantic SLAM

Stereo semantic SLAM is a leading paradigm that fuses stereo imagery with semantic cues for localization and mapping. It excels in dynamics, delivers stronger depth than monocular systems, and is more tolerant to illumination changes (Mur-Artal et al., 2015). The trade-off is higher compute and storage for stereo keyframes, which support optimization, loop closure, and relocalization—raising system cost (Meilland & Comport, 2013). To mitigate the "static-world" assumption of classic SLAM, stereo semantic methods employ self-instance segmentation and dynamic feature filtering (Bajpai, Burroughes, Shaukat, & Gao, 2016; Li, Song, Hao, Mao and Song, 2023), often building atop the well-established ORB-SLAM2 (Hu, Qi et al., 2025; Mur-Artal & Tardos, 2017; Zhai et al., 2024).

Concretely, stereo pairs are converted to depth, and parallel threads couple dynamic-feature filtering with ORB-SLAM2 to remain robust in motion-rich scenes. ORB features are extracted and their motion states inferred to separate static from dynamic points. Instance segmentation then contributes class-level semantics, partitioning frames into static and potentially dynamic regions; combining this with motion cues filters dynamic features. As in Fig. 12, four threads run concurrently: (1) pose tracking using static points, (2) local mapping, (3) loop detection/correction, and (4) dynamic-region recognition for filtering. This multi-threaded design preserves high-accuracy tracking and mapping despite moving objects (Bajpai et al., 2016; Li, Song et al., 2023).

Several key formulas and equations underpin stereo semantic SLAM. Absolute pose estimation and Essential matrix estimation are crucial for inferring the relative baseline motion between frames (Ai et al., 2023a), defined as shown in (3):

$$\mathbf{x}'^{T}\mathbf{F}\mathbf{x} = 0, \tag{3}$$

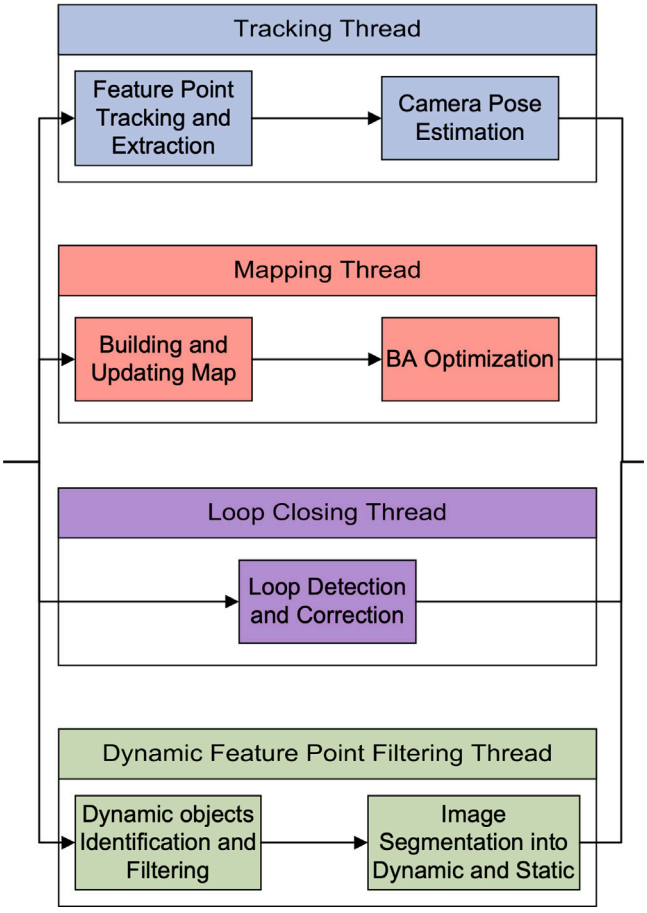

Fig. 12. Stereo semantic SLAM parallel threads.

where $\mathbf{x}$ and $\mathbf{x}'$ are corresponding points in stereo images. The polar constraint through which the motion probability p is calculated for each object, given a certain camera pose $\mathbf{P}$ and fundamental matrix $\mathbf{F}$, allows the differentiation between static and dynamic objects. The camera pose $\mathbf{T}$ is estimated using only static feature points with (4):

$$\mathbf{T} = \arg\min \sum \|\mathbf{x}_i - \mathbf{P}\mathbf{X}_i\|^2, \tag{4}$$

where $\mathbf{X}_i$ refers to the 3D point and $\mathbf{x}_i$ is its corresponding 2D point in the image (Tian, Yan and Li, 2023; Venator, Bruns, & Maier, 2020; Zhou & Wang, 2025).

3.1.3. RGB-D semantic SLAM

RGB-D semantic SLAM integrates color and depth information to enhance environmental understanding, achieve more accurate 3D mapping, and improve adaptability to scene changes. Sensors typically have a short effective range and experience reduced performance in bright sunlight. Additionally, they consume more power compared to simple cameras. Automatic SLAM RGB-D semantic is an advanced robotics method used to create smart maps by combining visual sensing with semantic information. Unlike traditional SLAM, which maps the environment based on shapes and appearances, RGB-D semantic SLAM labels these maps descriptively with terms such as "chair", "table", or "wall". This enhances the robot's ability to recognize and locate objects within the environment (Arth, Pirchheim, Ventura, Schmalstieg, & Lepetit, 2015; Ji, Wang, & Xie, 2021; Wang, Luo et al., 2025). Maintaining semantic coherence in dynamic scenes is tackled by combining spatiotemporal consistency with probabilistic propagation, enabling reliable separation of static and moving objects while preserving map integrity (Chen, Ling, Gao, Sun, & Jin, 2023). Online systems go further by coupling 2D/3D detection with semantic landmark association, providing real-time updates and constraints that correct pose drift during SLAM (Hempel & Al-Hamadi, 2022). For exploration, information-theoretic planners use Bayesian multiclass octrees with Shannon mutual information to choose viewpoints that reduce both geometric and semantic uncertainty (Asgharivaskasi & Atanasov, 2023). These ideas translate to factory floors via lightweight segmenters and dynamic keypoint classifiers tuned for industrial complexity (Gou et al., 2022).

Semantics improve a robot's task efficiency and interaction. Typical RGB-D semantic SLAM stacks three elements: (1) a dense SLAM core such as ElasticFusion (Memon, Iqbal, & Almakhles, 2024; Whelan, Salas-Moreno, Glocker, Davison, & Leutenegger, 2016); (2) CNN-based semantic segmentation over RGB; and (3) Bayesian fusion of labels into the 3D map. ElasticFusion maintains a surfel-based model resilient to revisits; surfels are adjusted to remain consistent with the real scene. A CNN performs pixel-wise labeling via max-unpooling and deconvolution to produce class probabilities. As depicted in Fig. 13, threads run in parallel: SLAM tracks pose and integrates depth to update surfels' geometry/color, while CNN predictions are registered via SLAM correspondences and fused probabilistically. Loop closure detects revisits and triggers global geometric optimization, changing surfel positions, normals, and semantic distributions. CRFs further regularize labels over surfels by enforcing spatial/appearance consistency, yielding rich, long-term consistent maps for advanced navigation (Li, Fan et al., 2023; Qin et al., 2022). Each surfel stores a class probability vector updated recursively, as in (5):

$$P(l_i \mid I_{1,\ldots,k}) = \frac{1}{Z} P(l_i \mid I_{1,\ldots,k-1}) P(O_{u(s,k)} = l_i \mid I_k), \tag{5}$$

where $P(l_i \mid I_{1,\ldots,k})$ represents the updated probability of surfel s belonging to class l_i, given images I_1 to I_k. $P(O_{u(s,k)} = l_i \mid I_k)$ represents the CNN's per-pixel probability output for the current image I_k, and Z is a normalizing constant.

The semantic predictions are further refined by incorporating the map's geometry using Conditional Random Fields. This method ensures that all labels remain consistent with the surrounding context. The energy required for this labeling in a fully connected graph can be mathematically expressed in (6):

$$E(x) = \sum_s \psi_u(x_s) + \sum_{s<s'} \psi_p(x_s, x_{s'}), \tag{6}$$

where the unary term $\psi_u(x_s)$ is defined as the negative logarithm of the surfel's internal probability distribution, which is illustrated in (7):

$$\psi_u(x_s) = -\log(P(L_s = x_s \mid I_{1,\ldots,k})). \tag{7}$$

The pairwise term $\psi_p(x_s, x_{s'})$ uses Gaussian edge potentials to enforce smooth predictions based on positional and appearance similarities, as handled by (8):

$$\psi_p(x_s, x_{s'}) = \mu(x_s, x_{s'}) \left(\sum_{m=1}^{K} w^{(m)} k^{(m)}(f_s, f_{s'}) \right), \tag{8}$$

where $k^{(m)}$ are Gaussian kernels applied to the feature vectors f_s of surfel s, and $\mu(x_s, x_{s'})$ is defined by the Potts model.

RGB-D semantic SLAM as implemented in the SemanticFusion framework combines mapping with object recognition. Injecting the ElasticFusion framework with CNNs and Bayesian updates can substantially boost the system. This integration allows for more accurate labeling in both 2D and 3D, enabling robots to become smarter and more efficient in their interactions and functions (Kostavelis & Gasteratos, 2017; Li, Gu, Dong, Dong and Han, 2020; Li, Hu et al., 2021; McCormac et al., 2017; Tian et al., 2019; Zhou, Yue et al., 2023).

3.1.4. 3D LiDAR-based semantic SLAM

3D LiDAR-based semantic SLAM incorporates semantic information into conventional SLAM procedures, providing accurate depth measurements and detailed environmental information. It performs well in both

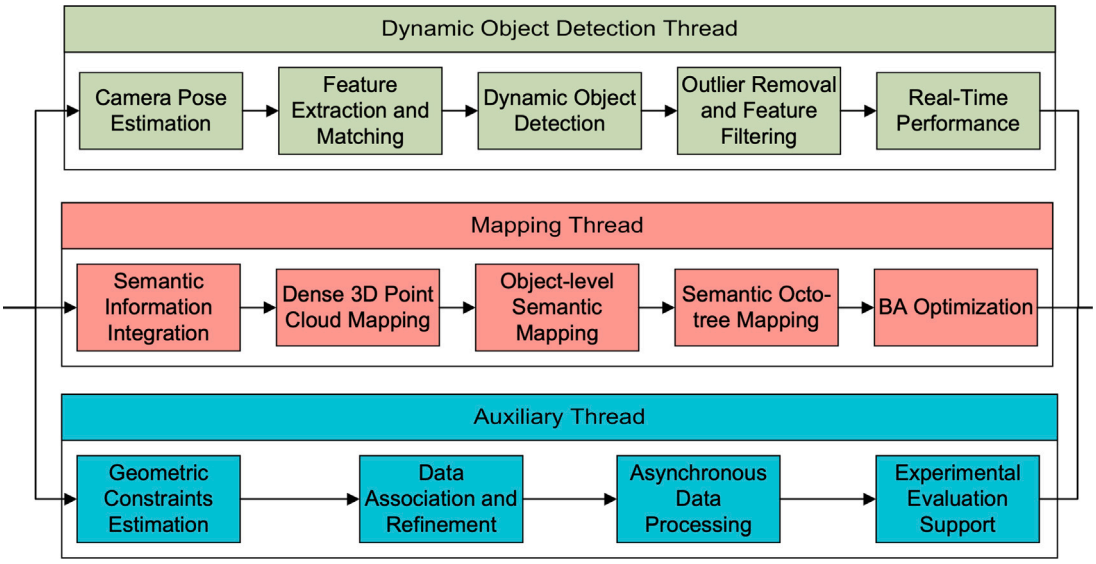

Fig. 13. RBG-D semantic SLAM parallel threads.

dark and bright environments (Gong et al., 2021; Liu, Mi and Chen, 2021; Yang, Chen, Liu, Zhang and Zhang, 2023). This approach has been successfully extended to complex indoor environments, where robot-assisted mobile scanning systems combine LiDAR-based SLAM with deep learning semantic segmentation to achieve comprehensive 3D reconstruction and automated point cloud labeling of building interiors (Hu, Gan, & Yin, 2023). Furthermore, the 3D LiDAR-based SLAM is also particularly well-suited for outdoor and large-area mapping applications (Behley et al., 2019; Li, Zhang, Li, Liu and Wang, 2020; Pu et al., 2025; Qiu, Zhuang, Yan, Hu, & Wang, 2019; Ruiz-Sarmiento, Galindo, & González-Jiménez, 2017; Tang, Huang et al., 2023; Zhu, Yuan, Zhang and Chen, 2025). Specialized applications have emerged for challenging natural environments, such as garden mapping systems that use semantic-based filtering to distinguish static structures from dynamic vegetation. This enables accurate static map construction through multi-frame and multi-resolution fusion techniques (Han et al., 2023). Recent advances have enhanced outdoor mapping through semantic-assisted topological map building, where online semantic segmentation enables adaptive node selection and robust place recognition in large-scale environments (He, Zhang, & Zhuang, 2022). However, LiDAR sensors are expensive and require sophisticated algorithms for data processing, with performance potentially hindered by weather conditions like rain or fog (Pak & Son, 2025a; Wei, Ni, Li, Hu, & Hu, 2024; Zhang & Singh, 2014). In 3D-LiDAR-based semantic SLAM, the general components of the 3D LiDAR-based semantic SLAM system, including tracking, mapping, and loop closing, are executed in parallel threads, which is illustrated in Fig. 14. These parallel processes work together to build and maintain a semantic map of the environment. These include (1) feature extraction, which identifies key features from raw LiDAR data crucial to understanding the environment's structure, and (2) feature matching and data association, which correlate these features against a map or between frames to determine the robot's position relative to its surroundings.

While these processes are running, semantic segmentation, powered by deep learning, classifies environmental components in the point cloud data. Object detection and tracking, which are parallel processes, identify and monitor objects over time. Pose estimation computes the exact position and orientation of the sensor, ensuring accurate localization. The system continuously updates and manages the map by incorporating new data, removing outdated information, and resolving inconsistencies. Optimization algorithms run in parallel, refining the map and trajectory quality by integrating new and historical data to minimize errors.

These processes are computationally intensive, making them essential for real-time data processing, which is critical in autonomous vehicles and robotic navigation in complex settings (Lou, Li, Zhang, & Wei, 2023). SLAM-generated maps are enriched with semantic information, significantly improving environmental understanding and navigation capabilities, particularly for applications such as autonomous driving. In semantic SLAM, point cloud data, conventionally denoted as $P = \{p_i\}$, where every p_i is a 3D point, is processed to include semantic labels l_i. These labels classify each point into various classes based on its characteristics through a classification model. Since these semantically rich point clouds may originate from different viewpoints, they are aligned and integrated into a coherent map using transformation matrices T. This process is evaluated for effectiveness using the mean Intersection-over-Union (mIoU), as defined in (9), as part of the semantic KITTI dataset:

$$\frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c + FN_c}, \tag{9}$$

Here, C is the number of classes under consideration, and for any class, c, TP_c, FP_c, and FN_c are the numbers of true positives, false positives, and false negatives, respectively. True positives are the points correctly classified as class c; false positives are those incorrectly labeled as class c despite belonging to another class; and false negatives are points that belong to class c but were mistakenly classified as a different class. This metric averages the IoU across all classes, giving a holistic view of the model's detection and classification accuracy. For example, in autonomous driving applications such as lane segmentation, it is crucial to have a precise and reliable understanding of the road environment, making these evaluations important for ensuring operational safety and effectiveness (Behley et al., 2019; Li, Zhang et al., 2020; Pugh, Chernak, & Jiddi, 2023; Qiu et al., 2019; Ruiz-Sarmiento et al., 2017; Tang, Huang et al., 2023).

3.1.5. Multi-modal semantic SLAM

Multi-modal semantic SLAM merges complementary sensors with semantics within a SLAM framework, enriching maps with object identity and attributes. Contemporary reviews emphasize the central role of deep learning in fusing heterogeneous sensors for robust autonomy (Tang, Zhao et al., 2023). Multiple modalities boost resilience to failures and noise but impose higher compute, tighter synchronization, and added hardware cost (Bresson, Alsayed, Yu, & Glaser, 2017; Rosen, Doherty, Terán Espinoza, & Leonard, 2021). Even so, the payoff is

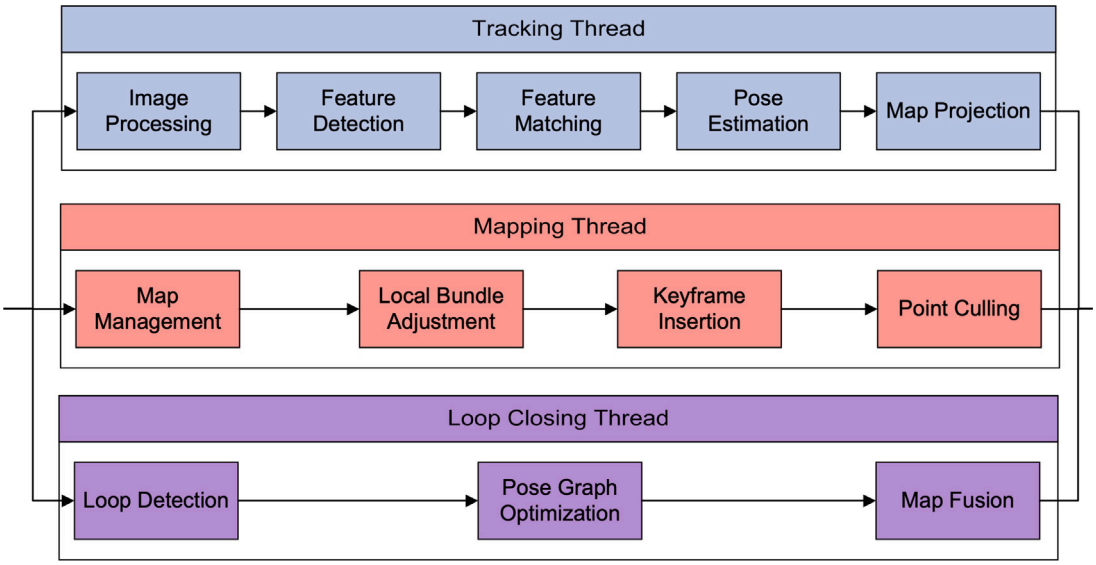

Fig. 14. 3-D LiDar-based semantic SLAM parallel threads.

stronger navigation and interaction in complex settings (Chen, Zhuang, & Wang, 2024; Chghaf, Rodriguez, & Ouardi, 2022; Xiao et al., 2025).

The system's performance is further improved by parallel processing, that are shown in Fig. 15, enabling efficient handling of large-scale maps and complicated environments. Parallel SLAM algorithms distribute computation and memory load across multiple processors using out-of-core techniques. The factor graph is subdivided into subgraphs, with local optimization for each segment while globally refining the whole graph. Selective updates are performed at the highest hierarchy level whenever new observations are made, concentrating computational resources on areas that change.

Additionally, some systems use a dual-thread strategy to separate the localization and mapping tasks. One thread optimizes and manages the local pose-feature graph, while another handles the pose-pose graph, with periodic synchronization. This separation ensures efficient task execution, which is important for real-time performance in robotic navigation and task execution in dynamic environments (Cadena et al., 2016). Mathematically, SLAM can be expressed as a MAP estimation problem, seeking to maximize the posterior probability of a map M and the robot's trajectory x, given measurements z and controls u, by maximizing (10):

$$(M^*, x^*) = \arg\max_{M,x} p(M, x | z_{1:t}, u_{1:t}). \tag{10}$$

Viewed probabilistically, SLAM addresses a high-dimensional Bayesian filtering problem: predict with the motion model, then correct with measurements (Hardegger, Roggen, Calatroni, & Tröster, 2016; Li, Jiang, Lei, Tang and Zhu, 2025; Yang, Zhang et al., 2023; Zhou, Mei, Liu and Bai, 2023). Extensions now capture epistemic uncertainty in semantics, informing belief-space planning that accounts for both pose and label uncertainty (Tchuiev & Indelman, 2023). For object-level data association, Dirichlet process mixtures cluster detections by class/position/size without fixing the number of objects (Wei, Chen, Chi, Wang and Sun, 2023). Confidence-aware fusion has likewise improved crowdsourced semantic maps, weighting pixel-level reliability across heterogeneous sources (Wijaya et al., 2022). Beyond VB inference, factor-graph methods jointly optimize coupled problems such as trajectory estimation and auto-calibration within a unified system (Liu et al., 2023).

SLAM problems can also be formulated in terms of factor graphs, where nodes represent states and edges represent observational and motion constraints. The objective in factor graphs is typically to minimize the sum of squared differences between predicted measurements and actual observations, as shown in (11):

$$\min_x \sum_i \| h_i(x_{i1}, x_{i2}, \dots, x_{ik}) - z_i \|^2_{\Sigma_i}, \tag{11}$$

where h_i represents the measurement functions that link the observed data z_i to the robot's states, accounting for the associated uncertainty Σ_i. In Multi-modal semantic SLAM, these models are further extended to include semantic tags and data from multiple sensors, enhancing the accuracy and richness of the robotic mapping and navigation.

3.1.6. *Incremental semantic SLAM*

Incremental semantic SLAM, a method that uses Visual–Inertial Odometry (VIO), ensures real-time map updates, adapts to dynamic environments, and reduces computational overhead by processing only new or changed data (Cao et al., 2025; Engel et al., 2017; Liu, Wu, Du, Zhang and Cong, 2024). While the advantages of incremental SLAM are clear, it faces issues such as accumulating inaccuracies over time, requiring sophisticated methods for effective implementation, and struggling with large-scale environments without regular global corrections (Bloesch, Omari, Hutter, & Siegwart, 2015). Note that recent frameworks have evolved to incorporate deep learning components that enhance performance in dynamic environments, with systems like SRVIO (Samadzadeh & Nickabadi, 2023). These methods achieve state-of-the-art results by intelligently fusing geometric constraints with semantic understanding to handle challenging scenarios that would cause traditional methods to fail.

The incremental SLAM process involves the continuous construction, updating, and refinement of a 3D mesh structure of the environment, relying on structural regularities for both mesh creation and state estimation. It is supported by parallel processes, which are crucial for simultaneously creating and updating a semantic map. These processes, that are shown in Fig. 16, cater to the extraction of semantic features from sensory data, detecting objects like doors and furniture using techniques such as deep learning. Meanwhile, key SLAM processes update both the map and the agent's location as new data is received.

Data association correlates new observations with the existing map, ensuring accuracy and consistency. State estimation uses filtering techniques, such as Kalman or particle filters, to compute variables like position and orientation. Additionally, 3D mesh generation from VIO keypoints creates a representation of the environment that is regularly updated and refined using structural constraints to ensure geometric accuracy. These processes—such as constraint enforcement, which applies structural regularities such as planarity, and optimization techniques to

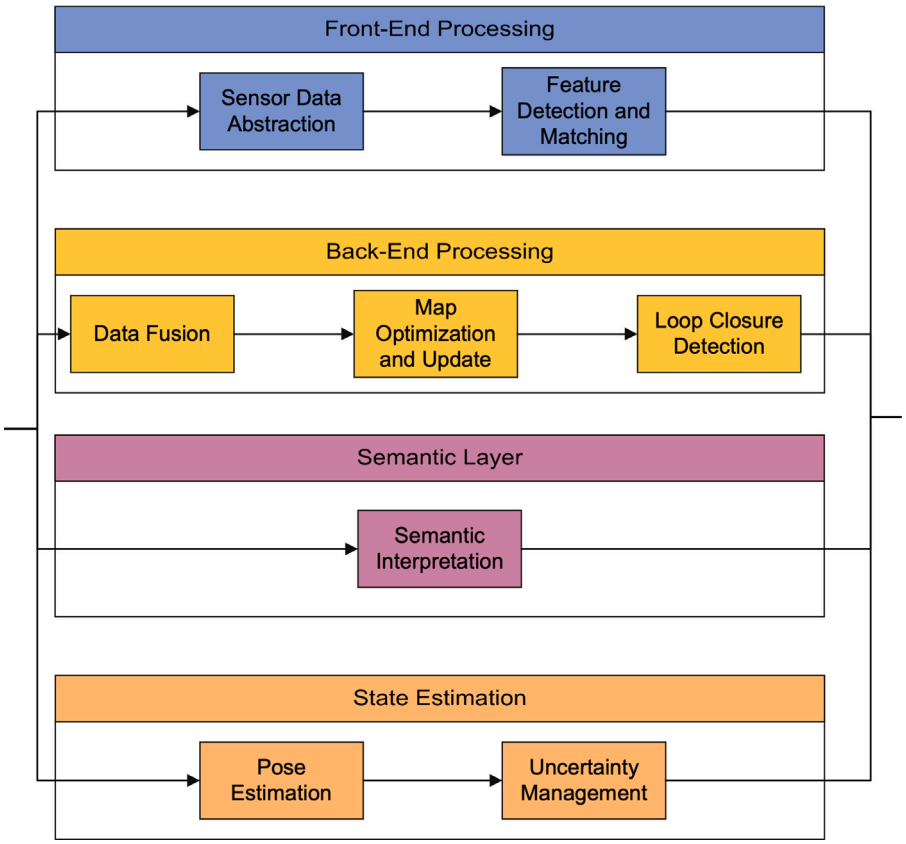

Fig. 15. Multi-modal semantic SLAM parallel threads.

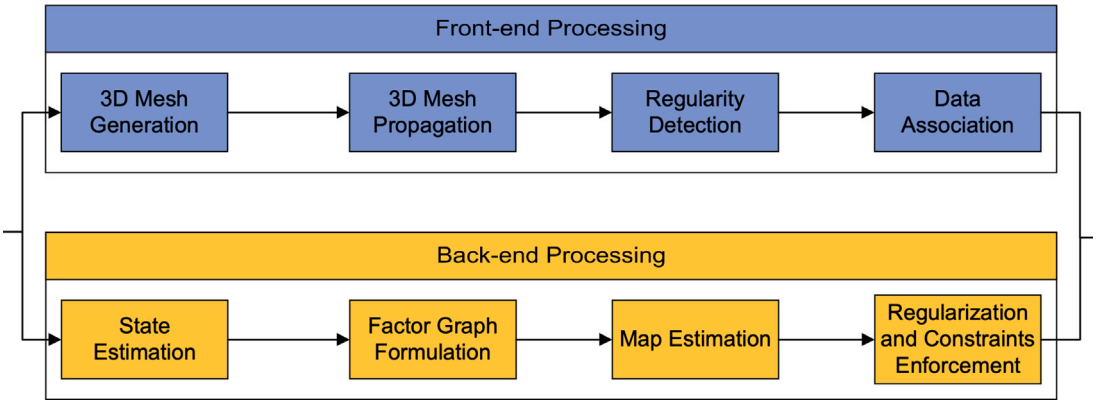

Fig. 16. Incremental semantic SLAM parallel threads.

minimize errors in map and trajectory estimates—run simultaneously. They interact to refine both the semantic map and state estimation, providing a robust and accurate environmental representation critical for autonomous navigation in complex environments (Tian et al., 2022).

Previous studies have focused on optimizing these processes by using detected structural regularities, such as planarity, to improve the accuracy and physical realism of the mesh. For every keyframe i, the variables are represented as shown in (12):

$$\xi = [R_i, p_i, v_i, b_i], \tag{12}$$

where R_i denotes the orientation of the IMU, p_i is the position, v_i indicates the velocity, and b_i represents the biases of the IMU.

The factor graph representing the back-end optimization problem includes the state X_t and the measurements Z_t, which are probabilistically interconnected as shown in (13):

$$p(X_t|Z_t) \propto p(X_t)p(Z_t|X_t) = \phi_0(x_0) \prod_{lc\in\Lambda_t \pi\in\Pi_t} A \prod_{(i,j)\in K_t} B, \tag{13}$$

where

$$A = \phi_R(\rho_{lc}, \pi_\pi)\delta(lc, \pi)$$

and

$$B = \phi_{\text{IMU}}(x_i, x_j).$$

ϕ_0 represents the prior on the initial state, ϕ_R are the regularity factors that link the points of interest ρ_{lc} with planes π_π, and ϕ_{IMU} incorporates the IMU data between keyframes.

The MAP estimation, which aims to find the most probable state configuration based on the measurements, is formulated in (14):

$$X_t^{\text{MAP}} = \arg\min_{X_t} \left\{ \|r_0\|^2_{\Sigma_0} + \sum_{\substack{lc\in A_t \\ \pi\in\Pi_t}} \delta(lc,\pi)\|r_R\|^2_{\Sigma_R} + \cdots \right\}. \tag{14}$$

This estimation aims to minimize the sum of squared residuals, where each term's distinct components address different aspects of the system's behavior and its interaction with the environment. The regularity constraints, particularly those related to co-planarity, are stated in (15):

$$r_R = n \cdot \rho_{lc} - d. \tag{15}$$

The constraint ensures that the landmark ρ_{lc} lies on the plane specified by the normal vector n and the distance d from the origin, enhancing the structural integrity of the generated mesh relative to real-world geometry (Rosinol, Sattler, Pollefeys and Carlone, 2019; Wen et al., 2020). This advanced approach not only improves the precision of state estimation but also produces a 3D mesh that is both accurate and representative of the actual environment, particularly in structured environments (Guo & Fan, 2022; Rosinol, Abate, Chang, & Carlone, 2020; Wu et al., 2025). Recent multi-robot extensions of these planar constraint-based methods show that high-quality metric-semantic reconstruction can be maintained across distributed systems (Tian et al., 2022).

Sensor fusion involves integrating data from LiDAR, cameras, and Inertial Measurement Units (IMUs) to improve accuracy and robustness in SLAM. This integration uses the strengths of each sensor type to alleviate their respective weaknesses, providing more reliable 3D mapping in complex and dynamic environments. In sensor fusion, system state estimates —namely rotation, velocity, and position —are adjusted, each governed by specific formulas that describe how sensor data are fused over time. For rotation updates, the equation is given as:

$$\Delta R = \exp(\Omega \Delta t), \tag{16}$$

where Ω is the skew-symmetric matrix of the angular velocity $\omega(t)$, and Δt is the time interval. Eq. (16) is important for updating the orientation of the sensor platform based on angular velocity measurements, transforming these readings into a rotation matrix that reflects the changes in orientation over the time interval.

The velocity update, as shown in (17):

$$\Delta v = R(t_0) \cdot a(t) \Delta t, \tag{17}$$

uses the initial rotation matrix $R(t_0)$ and the linear acceleration $a(t)$ to compute changes in the sensor platform's velocity over time. The initial orientation at the start of the time period is accounted for, allowing the acceleration to be applied properly within the global reference frame, ensuring accurate velocity updates as the sensor platform moves.

Finally, the position is updated using (18):

$$\Delta p = \Delta v \Delta t + \frac{1}{2} a(t)(\Delta t)^2, \tag{18}$$

This equation integrates the change in velocity over time and adds the displacement due to constant acceleration, providing a full update of the platform's position. In dynamic environments, sequential updates of rotation, velocity, and position are important for maintaining SLAM accuracy. The system rapidly adapts to new sensor inputs, reducing errors due to sensor noise or processing delays. These updates not only improve real-time operational capabilities in SLAM but also contribute to the overall system's reliability and performance in increasingly complex and challenging conditions (Cai, Ou, & Qin, 2024; Yu, Xiang, & Su, 2022). Recent implementations have shown that incorporating semantic information during these incremental updates, such as through YOLOv4-based object detection, can further enhance localization accuracy. The combination of object detection with clustering algorithms provides semantic constraints that complement the geometric updates, resulting in more robust localization performance (Chai, Li, & Li, 2023). Beyond traditional object detection, visual–LiDAR fusion methods have demonstrated that salient obstacle detection can transform environmental features into reliable landmarks for mapping. Centroid and contour extraction from fused sensor data provides distinctive reference points for localization (Hu et al., 2022).

3.2. Environment

This section focuses on the application of semantic SLAM in both dynamic indoor and outdoor environments, highlighting the unique challenges and solutions for each context. In dynamic indoor environments, semantic SLAM must account for rapidly changing objects and obstacles, while outdoor environments introduce challenges such as varying terrain, weather conditions, and large-scale dynamic elements. By capturing semantic information, robots can achieve more robust localization and mapping in these complex and unpredictable settings.

3.2.1. Semantic SLAM for dynamic indoor environment

Visual semantic SLAM (VSLAM) applied in dynamic indoor environments seeks to enhance performance in indoor scenarios, where traditional methods usually fail in the presence of moving objects (Habibpour, Nemati, Meghdari, Taheri, & Nazari, 2024; Zhao, Zuo, & Hu, 2021). This work proposes a lightweight yet effective framework that integrates several parallel processes to improve the accuracy and robustness of the semantic SLAM system. First, the YOLOv7-tiny algorithm is applied to detect dynamic objects in the scene. Then, motion consistency detection and the Lucas–Kanade optical flow algorithm are employed to classify feature points as either static or dynamic. This approach aligns with recent developments in fast, semantic-aware motion detection. The fusion of depth information, feature flow, and semantic cues through probabilistic frameworks has proven effective for real-time dynamic object filtering (Singh, Wu, Do, & Lam, 2022). The dynamic points are filtered out, leaving only static points for mapping, which is crucial for maintaining the SLAM system's precision in environments with changing indoor scenes. In addition, ATY-SLAM incorporates an adaptive thresholding scheme for keyframe selection, further refining the process by considering environmental dynamics. These parallel processes enable ATY-SLAM to effectively address the challenges resulting from dynamic environments, overcoming the limitations of traditional SLAM systems that assume static settings (Deng, Hu, Wen, Zhang, & Jin, 2025; Qi, Hu, Xiang, Cai, & Zhao, 2023).

Delving deeper, the system initially incorporates the YOLOv7-tiny object detection model to recognize dynamic objects present in the scene. These resulting bounding boxes indicate areas where dynamic feature points are likely to appear. By removing these dynamic points, the system refines the feature set to focus on more quasi-static and stable points based on the model's prediction. This process removes dynamic feature points, leaving a more stable and cleaner set of points for further analysis. Specifically, all feature points detected in a frame F_k can be expressed using (19):

$$F_k = \{f_1, f_2, f_3, \ldots, f_n\}, \tag{19}$$

where dynamic points identified by YOLOv7-tiny are excluded from F_k, resulting in a refined set of static points denoted as P_k.

The method uses epipolar geometry to enhance motion consistency detection. The distance d of a feature point from the epipolar line is calculated to determine whether a point is dynamic or static, and is given by:

$$d = \frac{|P_2 F P_1|}{\sqrt{X^2 + Y^2}} \tag{20}$$

where P_1 and P_2 represent the corresponding feature points across frames, and F denotes the fundamental matrix. Feature points that exceed a distance threshold ϵ_{th} are classified as dynamic. Optical flow is employed to estimate the motion vectors u and v of the feature points, based on the brightness constancy assumption, calculated as:

$$\begin{bmatrix} I_x & I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -I_t \tag{21}$$

where I_x, I_y, and I_t represent the image gradients along the x-axis, y-axis, and the temporal derivative at the feature point, respectively.

Additionally, an adaptive thresholding method is used to improve keyframe selection, which is paramount for accurate mapping and localization. This method works by analyzing changes in model observations and assessing the angles and distances of matching points across frames. This adaptive threshold, T_{adaptive}, is carefully fine-tuned based on these ratios of matching points, angular deviations, and other factors, ensuring that only the most reliable frames are selected as keyframes (Qi et al., 2023).

3.2.2. Semantic SLAM for dynamic outdoor environment

The presence of dynamic objects in outdoor environments poses significant challenges for visual SLAM systems, requiring sophisticated approaches to maintain accurate localization and mapping. Early solutions, like DS-SLAM, integrate semantic segmentation with dynamic object handling using five parallel threads: tracking, semantic segmentation, local mapping, loop closing, and dense semantic map creation. By combining semantic segmentation networks with a method for motion consistency checking, DS-SLAM detects and discards moving dynamic objects, such as people, present in the scene. This methodology greatly enhances localization accuracy by reducing the effect of dynamic elements. A key component of DS-SLAM is the semantic octo-tree map, a dense mapping technique that incorporates semantic labels. These labels embed a deeper understanding of the environment, enabling higher-order tasks such as environment interaction and advanced path planning (Ran, Yuan, Zhang, Tang and He, 2021; Wu et al., 2024; Yu et al., 2018a). The evolution from DS-SLAM to systems like DE-SLAM represents a shift toward handling increasingly dynamic scenarios, with each iteration improving the robustness against moving objects (Xing, Zhu, & Dong, 2022). More recent frameworks have evolved to handle increasingly complex outdoor dynamic scenes through advanced deep learning models. These refined motion estimation algorithms achieve superior performance in challenging real-world conditions (Wen et al., 2023).

The robust computation of the fundamental matrix and the distance from the epipolar line are core algorithms in DS-SLAM, ensuring geometric consistency between frames and rejecting outliers caused by moving objects. The fundamental matrix F is used to compute the epipolar line $I = FP$, where P represents the homogeneous coordinates of points in the image. This calculation helps determine which points remain consistent across sequential frames. Additionally, the distance $D = \frac{|TPFP|}{\|X\|+\|Y\|}$ is used to assess point consistency, where $TFPP$ is the projection of the points and X and Y represent the spatial coordinates. Points with a distance exceeding a set threshold are classified as dynamically moving and are excluded from pose estimation.

In contrast, DS-SLAM maintains a semantic octo-tree map and updates the occupancy grid with log-odds scores. The log-odds score is given by $l = \log\left(\frac{p}{1-p}\right)$, where p represents the probability of occupancy. Its inverse, $p = \frac{e^l}{e^l+1}$, is used to update this probability over time. By accumulating evidence of occupancy through these log-odds scores, DS-SLAM ensures the map remains accurate and up-to-date, even in dynamic environments where objects may frequently move.

DynaSLAM extends ORB-SLAM2 by incorporating dynamic object detection through a combination of multi-view geometry and deep learning, while also reconstructing occluded areas of the scene using background inpainting. This capability is particularly important in applications requiring long-term autonomy and continuous learning of the environment. DynaSLAM runs through a structured, multi-threaded approach, where tasks such as tracking, semantic segmentation, local mapping, loop closure, and semantic map creation run in parallel. Real-time semantic segmentation is core to this system since it facilitates the detection of dynamic objects that could hamper the tracking and mapping processes. These advancements lead to reduced drift, more reliable trajectory estimation, high accuracy, and improved fidelity in mapping, particularly under challenging conditions like dynamic environments or varying lighting. Alternative approaches to dynamic SLAM have focused on feature weighting strategies, where detected semantic and geometric properties are used to assign reliability scores to features rather than completely removing them, offering a computationally lighter solution for real-time applications (Zhong, Hu, Huang, Bai, & Li, 2022). On the other hand, DS-SLAM takes a more comprehensive approach to using semantic data, integrating it more deeply into the system's operations (Bescos, Facil, Civera, & Neira, 2018; Li, Wang, Xu and Chen, 2021; Xie, Liu, & Zheng, 2021; Yang & Cai, 2024). The recent advancements on dynamic slam in both indoor and outdoor environments has been described in Table 2.

3.2.3. Approaches to 3D object representation

A key aspect of Semantic SLAM is how the environment and objects within it are represented in 3D, since the chosen representation directly influences the system's ability to perform accurate localization, mapping, and semantic understanding. Hence this section explains the approaches to 3D object representation in both indoor and outdoor environments. These object representations are broadly classified into two categories: Euclidean structured data and non-Euclidean structured data. Euclidean structured data encodes geometric information about the object's shape, size, and spatial relationships using Euclidean space coordinates. On the other hand, non-Euclidean structured data offers a more flexible and expressive way to encode geometric information, especially for objects with curvature, irregularities, or topological complexity.

While point clouds and meshes can be considered as both Euclidean and non-Euclidean data depending on the scale of observation, we categorize them as non-Euclidean due to their often infinite curvature, self-intersections, and variable dimensions. Analyzing such data on a broader scale helps understand the overall features of 3D objects, which is useful for tasks such as object recognition and correspondence. Fig. 17 represents the different classifications of object representations, with a detailed description of each category provided in Table 3. Once the basic principles are established, it becomes important to look at the datasets that drive this field, since they form the basis for training, validation, and comparison of Semantic SLAM methods.

4. Datasets

Datasets play a central role in enabling research progress. Here, we present the most widely used datasets in Semantic SLAM used for evaluating both traditional SLAM and semantic SLAM systems, highlighting their characteristics and the role they play in training, benchmarking, and validating algorithms. These datasets encompass real-world captures and simulated environments, supporting various sensor modalities including visual, RGB-D, and LiDAR systems. They span diverse scenarios, from static indoor scenes to dynamic outdoor environments, enabling comprehensive benchmarking of SLAM algorithms. The following subsections provide concise descriptions of key datasets and their characteristics.

Table 2
Dynamic semantic SLAM method comparison.

Reference	Dynamic object detection	Environment suitability	Strengths	Weaknesses
Gupta, Arbeláez, Girshick, and Malik (2015), Qi et al. (2025), Xu et al. (2019) and Yang, Ran, Wang, Lu, and Chen (2022)	Combines Mask R-CNN instance segmentation with residual-based motion filtering	Indoor	Dense object-level maps; tracks moving objects; reconstructs background	Low frame rate; COCO-dependent; needs GPU; not for large/outdoor scenes
Bescos, Campos, Tardós and Neira (2021) and Ying et al. (2023)	Classifies ORB features as static/dynamic; combines geometry and semantic cues	Outdoor	Real-time object-aware SLAM; full 6-DoF for camera/objects; good for driving	Sparse maps; ignores segmentation delay; needs stereo/RGB-D; accuracy depends on segmentation
Ge, Zhang, Wang, Coleman, and Kerr (2023) and Gonzalez, Marchand, Kacete, and Royan (2022)	Groups points by semantic class; motion modeled with mechanical constraints	Outdoor	Robust tracking for objects (e.g., cars); accurate camera pose with dynamic objects	No dense mapping; needs accurate segmentation/joint models; not for cluttered/indoor settings
Wang, Wu, Li and Yu (2024) and Judd and Gammell (2024)	Scene flow clustering and multilabel RANSAC; no semantics required	Indoor and Outdoor	Unsupervised; tracks multiple motions without semantics; occlusion-tolerant	High computation; not real-time; sparse maps; lacks object-level detail; needs stereo/depth sensors

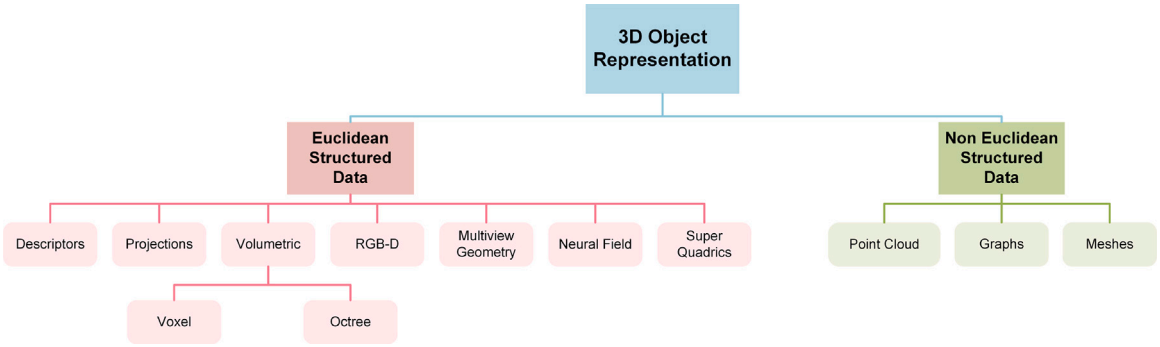

Fig. 17. A detailed illustration of different types of 3D object representations commonly used in semantic SLAM.

4.1. TUM RGB-D dataset

The TUM RGB-D dataset (Lin, Zhang et al., 2025; Sturm, Engelhard, Endres, Burgard, & Cremers, 2012) is a popular benchmark for evaluating SLAM and visual odometry systems. It provides RGB-D images captured using a Microsoft Kinect sensor, along with ground truth poses obtained through a motion capture system. The dataset encompasses various indoor environments, including both static and dynamic scenes, offering a diverse and challenging test bed for visual SLAM algorithms. In this paper, the dataset sequences are used to evaluate the results discussed in Sections 8, 8.4, with the list of sequences detailed in Table 4.

4.2. KITTI dataset

The KITTI dataset (Geiger, Lenz, & Urtasun, 2012) is a widely recognized benchmark for evaluating computer vision and SLAM algorithms, particularly in the context of autonomous driving. It contains high-resolution stereo and LiDAR data acquired from a vehicle navigating through various environments, such as urban, rural, and highway settings. The dataset provides ground truth poses obtained from a GPS/IMU system, enabling precise evaluation of SLAM performance in real-world outdoor environments. The KITTI-360 dataset extends this benchmark with longer sequences, providing more comprehensive evaluation capabilities for semantic SLAM and scene understanding tasks in complex urban environments (Liao, Xie, & Geiger, 2023). The KITTI dataset is used to replicate the results discussed in Sections 8, 8.4, with the specific sequences outlined in Table 5.

4.3. BONN dataset

The Bonn RGB-D Dynamic dataset (Palazzolo, Behley, Lottes, Giguère, & Stachniss, 2019a) is a key resource for advancing research in RGB-D SLAM. It consists of 24 dynamic sequences and 2 static sequences, capturing activities such as box manipulation and balloon play, designed to test SLAM algorithms in realistic, dynamic environments. Each sequence is accompanied by ground truth sensor poses obtained from an Optitrack Prime 13 motion capture system, along with a 3D point cloud of the static environment, recorded using a Leica BLK360 terrestrial laser scanner. The dataset is formatted similarly to the TUM RGB-D dataset, facilitating compatibility with existing evaluation tools.

4.4. A1 and Jackal

These publicly available datasets were part of the Kimera-Multi project (Tian, Chang et al., 2023; Wan & Luo, 2025). They were gathered using Unitree A1 quadrupedal robots and Clearpath Robotics Jackal wheeled robots, both equipped with RealSense D455 RGB-D cameras, IMUs, and Velodyne 3D LiDAR. The sequences contain RGB images, depth images, compressed grayscale images, wheel odometry, and LiDAR point clouds. The data was recorded in various locations across MIT's campus, including indoor and outdoor areas, underground tunnels, and an undergraduate dormitory. The diverse environments and revisited locations in these datasets make them particularly valuable for evaluating loop closure detection algorithms, where semantic features combined with traditional bag-of-words approaches have shown significant improvements in recognition accuracy (Sun, Wang, Ni, & Li, 2024).

Table 3
3D object representations for semantic SLAM in scene understanding.

Reference	Data type	Characteristics	Advantages	Challenges
Chen, Shao, Zhang and Zhang (2022), Choudhary et al. (2017), Li, Zhou and Li (2024), Nie et al. (2020), Peng, Zhao and Wang (2024), Wen et al. (2021), Xie et al. (2022), Xu, Feng, Kamat, and Menassa (2020), Yang et al. (2020) and Zhu, Xiao and Fan (2025)	Descriptors	• Describe geometric or topological characteristics • Capture shape, surface, and texture information	• Object recognition • Shape similarity • Efficient 3D processing	• Deformable shape handling • Large-scale scalability
Gong et al. (2021), Huang et al. (2023), Jung et al. (2025), Liu, Mi et al. (2021), Sandstrom, Li, Van Gool, and Oswald (2023), Wang, Tian et al. (2025), Yang, Chen et al. (2023), You et al. (2022) and Ying and Li (2023)	Projections	• Convert 3D objects into 2D grids	• Retains key shape characteristics	• Information loss in dense tasks
Choi, Chao, Pantofaru, and Savarese (2015), Jin, Chen, Sun, and McLoone (2020), Mascaro, Teixeira, and Chli (2022), Popovic et al. (2021), Rosinol, Leonard, and Carlone (2023), Rosu, Quenzel, and Behnke (2020), Wang, Tian and Liu (2025) and Yan, Wang, He, Chang, and Zhuang (2020)	Volumetric (voxel/octree)	• Grid-based 3D space modeling	• Simple, structured encoding	• High memory cost • Poor resolution scalability
Cheng, Sun, Zhang, and Zhang (2023), Cheng, Wang, Mai, Min, and Meng (2021), Dang et al. (2019), Deng et al. (2020), Kuang, Yuan, and Liu (2022), Muthu et al. (2020), Yan et al. (2022) and Zhang, Zhang, Jin and Yi (2022)	RGBD	• Combines color and depth info (2.5D)	• Cost-effective, accurate pose and scene understanding	• Struggles with noisy/incomplete data
An, Pan, Li, and Wang (2022), He, Ding, and Lan (2024), Huang, Chen, Zhang, He, and Feng (2024), Islam, Ibrahim, Chin, Lim, and Abdullah (2024), Shi, Zha, Guo, Wang and Li (2020), Zheng et al. (2025) and Yang, Ye, Zhang, Wang, and Qiu (2024)	Multi-view geometry	• Combine multiple 2D images for 3D reconstruction	• Reduces noise and occlusion • Tolerant to lighting issues	• Sensitive to calibration errors • Not ideal for dynamic scenes
Bescos, Cadena and Neira (2021), Kong, Liu, Taher, and Davison (2023), Li, Guo et al. (2025) and Ruan, Zang, Zhang, and Huang (2023)	Neural field	• MLPs represent object surfaces	• Compact, watertight, coherent representation	• Complex temporal modeling • Requires large datasets
Han and Yang (2023), Peng, Xu et al. (2024), Tian et al. (2024), Tschopp, Nieto, Siegwart, and Cadena (2021) and Wei and Wang (2018)	Super quadrics (SQ)	• Compact 3D shape abstraction from point clouds	• Efficient representation with shape fidelity	• Training requires large datasets • Sensitive to temporal variance
Cho, Kim, Park, Sunwoo, and Jo (2020), Isele, Haas-Fickinger, and Zöllner (2021), Li, Fu, Sun, Li and Wang (2024), Li, Kong, Zhao, Huang, and Liu (2022), Pan, Hu, Cao, Kang, and Wang (2024), Vishnyakov et al. (2021) and Zhang, Huo, Huang, and Liu (2025)	Point cloud	• Unstructured 3D points without topology	• Flexible and detailed geometry	• Hard to model globally • Calibration sensitivity
Arshad and Kim (2024), Duan, Feng, and Wen (2022), Fernandez-Cortizas et al. (2024), Liu, Yuan and Kuang (2024), Qian, Fu, and Xiao (2022) and Zhang, Zhang, Liu, Naixue Xiong and Li (2024)	Graphs	• Nodes as vertices; edges encode relationships	• Scalable and expressive for both local/global tasks	• High complexity • Hard to visualize large graphs
Herb, Weiherer, Navab, and Tombari (2021), Rosu et al. (2020) and Wang, Zhang and Li (2020)	Meshes	• Polygons and vertices define surface geometry	• Preserves structure for segmentation and matching	• Irregular structure hampers DL integration • Sensitive to resolution and noise

Table 4
TUM RGB-D dataset sequences.

Sequence	Description	Image size	Frame rate
fr3_walking_xyz	Walking sequence with significant translational motion in x, y, z directions	640 × 480 pixels	30 Hz
fr3_walking_static	Static scene with minimal motion	640 × 480 pixels	30 Hz
fr3_walking_rpy	Walking sequence with rotational motion in roll, pitch, yaw	640 × 480 pixels	30 Hz
fr3_walking_half	Half walking sequence with moderate motion	640 × 480 pixels	30 Hz

Table 5
KITTI dataset sequences.

Sequence	Description	Image size	Frame rate
KITTI 00	Urban environment with moderate traffic	1242 × 375 pixels	10 Hz
KITTI 01	Highway environment with high-speed motion	1242 × 375 pixels	10 Hz
KITTI 02	Urban environment with dynamic objects	1242 × 375 pixels	10 Hz
KITTI 03	Rural environment with varying terrains	1242 × 375 pixels	10 Hz
KITTI 04	Urban environment with sharp turns and occlusions	1242 × 375 pixels	10 Hz

4.5. uHumans2

This dataset was created using the Unity simulator, where humans are simulated as realistic 3D models with standard graphic assets. It is used for 2D segmentation tasks and was developed as part of the Kimera project. For benchmarking purposes, the simulator provides ground truth poses for both humans and objects. The dataset contains visual–inertia data for various scenes, both with and without dynamic objects, covering environments such as offices, apartments, subways, and neighborhoods (Rosinol et al., 2021).

4.6. CarSim

These datasets consist of simulated urban outdoor scenes within the TESSE environment, designed similarly to the uHumans dataset. The simulated car is equipped with four monocular cameras positioned at the front, rear, left, and right. Additionally, the dataset includes ground truth pixel-wise semantic labels for precise analysis (Abate, Chang, Hughes, & Carlone, 2024; Zhi, Deng, Li, Zhang, & Hong, 2024). This multi-camera setup is particularly relevant for autonomous parking research, where similar sensor configurations have been successfully deployed in both outdoor valet parking scenarios (Abate et al., 2023) and indoor parking environments with semantic object detection capabilities (Shao, Zhang, Zhang, Shen, & Zhou, 2022).

4.7. openLORIS

This dataset is built for lifelong SLAM of service robots. The data is collected using two cameras: the RealSense D435i, which provides RGB-D images and IMU measurements, and the RealSense T265 tracking module, which captures stereo fisheye images and IMU measurements. The dataset includes five scenes, each containing 2 to 7 sequences taken at different times. Ground truth robot poses for each scene are provided by the Optitrack MCS and Hokuyo LiDAR systems (Shi, Li et al., 2020).

4.8. BeVIS (Indoor parking dataset)

The BeVIS dataset specifically targets the challenging domain of indoor parking environments. This comprehensive benchmark provides ground-truth trajectories for evaluating SLAM systems in parking structures, where GPS-denied conditions and repetitive structural patterns pose unique challenges. The dataset supports the evaluation of tightly-coupled semantic SLAM frameworks that integrate front-view cameras, inertial sensors, and surround-view systems for robust localization and detection of semantic objects in parking scenarios (Shao et al., 2023).

4.9. Scenesv2

The Scenes dataset v2 includes RGB and depth images from 14 scenes featuring various furniture items such as chairs, coffee tables, sofas, and tables, along with a selection of objects from the RGB-D Object dataset, including bowls, caps, cereal boxes, coffee mugs, and soda cans. Each scene has ground truth annotations and is represented as point cloud data, generated by aligning multiple video frames using Patch Volumes Mapping (Lai, Bo, & Fox, 2014; Shao, Liu, Lu, Li, & Akbar, 2025).

4.10. Freiburg cars

This dataset comprises RGB video sequences of 52 cars, captured with a camcorder in a full 360° rotation. Each video contains approximately 1500 to 3500 frames, which are uniformly downsampled to around 120 frames to accelerate the 3D reconstruction process (Sedaghat & Brox, 2015; Shao et al., 2025).

4.11. Redwood-OS chairs

This dataset features a large and diverse collection of RGB-D and reconstructed models, ranging from shoes, mugs, and toys to grand pianos, construction vehicles, and large outdoor sculptures. The data was captured using PrimeSense Carmine cameras with a resolution of 640 × 480 pixels and a frame rate of 30 Hz. Each scan consists of both color and depth images, with pixel values representing depth in millimeters (Choi, Zhou, Miller, & Koltun, 2016).

Further, to support both researchers and practitioners, we provide a summary of widely used open-source frameworks that extend traditional SLAM with semantic capabilities. Table 6 highlights prominent tools, their key features, and associated publications, offering practical resources for replicating results and advancing research in semantic SLAM.

The evolution from traditional SLAM datasets to semantically-annotated ones reflects the field's progression toward scene understanding. Modern datasets not only provide geometric ground truth but also semantic labels, instance segmentation, and dynamic object annotations. These diverse datasets are often used for benchmarking algorithms against various challenges, though the need for more specialized datasets targeting specific applications continues to grow. Finally, understanding these strengths and limitations of available datasets naturally leads us to explore how these resources have supported recent advances in Semantic SLAM, especially in enhancing scene understanding.

5. Advancements in semantic SLAM for scene understanding

Traditional SLAM techniques typically rely on geometric and probabilistic approaches, utilizing methods, such as feature-based tracking and EKFs, to estimate a robot's pose and map its environment. While these approaches are effective, they often face challenges in complex and dynamic environments due to limitations in feature extraction and sensitivity to sensor noise. Additionally, these approaches primarily rely on geometric features and lack a deeper semantic understanding of the environment. In contrast, semantic SLAM techniques integrate semantic information, such as object categories and semantic segmentation, into the mapping and localization process. By incorporating this layer of semantic understanding, semantic SLAM algorithms can generate more meaningful maps that not only represent spatial layouts but also provide insights into the environment's semantic content. This enables robots to make more informed decisions and interact more intelligently with their surroundings, opening up new possibilities for applications in areas like autonomous driving, robotics, and augmented reality.

Recent advances in semantic SLAM have focused on addressing the challenges of dynamic indoor and outdoor environments, which represent the most common and challenging real-world scenarios for

Table 6
Prominent open-source frameworks for semantic SLAM and their key contributions.

Framework	Key features	Applications	Reference
ORB-SLAM3 (with Semantic Extensions)	Multi-camera, stereo, and inertial SLAM; semantic object integration via Mask R-CNN	Robust semantic SLAM across diverse environments	Campos, Elvira, Rodr'iguez, Montiel, and Tardós (2020)
Kimera	Real-time metric-semantic mapping; 3D scene graphs; integrates visual–inertial odometry	Robot navigation, semantic scene understanding	Rosinol, Abate, Chang and Carlone (2019)
DROID-SLAM	End-to-end deep learning-based dense SLAM; robust to dynamics; lightweight	Visual odometry, dynamic scene tracking	Teed and Deng (2021)
SemanticFusion	Combines CNN-based semantic segmentation with ElasticFusion for dense maps	Indoor semantic mapping	McCormac, Handa, Davison, and Leutenegger (2016)
MaskFusion	Object-aware SLAM; fuses instance segmentation with 3D reconstruction	Augmented reality, dynamic object mapping	Rünz and Agapito (2018)
Co-Fusion	Multi-object segmentation and tracking in real-time; extends ElasticFusion	Dynamic SLAM with moving objects	Rünz and Agapito (2017)
Semantic voxblox	Incremental volumetric mapping with semantic fusion	Long-term mapping, mobile robotics	Palazzolo, Behley, Lottes, Giguère, and Stachniss (2019b)
PanopticFusion	Panoptic segmentation integrated into dense SLAM pipeline	Scene understanding, semantic mapping	Narita, Seno, Ishikawa, and Kaji (2019)
DS-SLAM	Dynamic Semantic SLAM using deep learning for segmentation and static/dynamic separation	Robust localization in dynamic scenes	Yu et al. (2018b)
OpenVSLAM (with semantics)	Versatile, modular SLAM with support for multiple camera models; extensible with semantics	Lightweight robotics, reproducible experiments	Sumikura, Shibuya, and Sakurada (2019)
MonoScene-SLAM (emerging)	Combines monocular SLAM with 3D scene completion and semantic priors	3D reconstruction from monocular cameras	Cao and de Charette (2021)

autonomous systems. These environments, characterized by frequent changes, moving objects, and diverse sensor conditions, provide ideal testbeds for evaluating the robustness and adaptability of different semantic SLAM methods. In this section, we present the recent advancements in Semantic SLAM techniques, with a particular focus on indoor and outdoor scene understanding. We emphasize how different methods incorporate semantic information to improve mapping and localization, reflecting the current trends and innovations in the field.

5.1. Key approaches in indoor scene understanding

The quality of the global map is important for accurate localization. To address this, Fan et al. proposed a novel semantic SLAM method that builds an accurate point cloud map while generating bounding boxes and masks using BlitzNet. The approach enables the creation of depth-stable points by accurately matching features in dynamic regions (Fan et al., 2020). Similarly, Han et al. provided a detailed review of indoor semantic mapping, covering aspects such as spatial mapping, semantic information acquisition, and map representation (Han, Li, Wang, & Zhou, 2021a). Chen et al. presented an extensive survey of semantic SLAM, detailing recent developments and analyzing the extraction and processing of semantic information using state-of-the-art datasets (Chen, Xiao et al., 2025).

Zhu et al. proposed a dense SLAM system called NICE-SLAM, which creates a hierarchical scene representation using local information. This representation, optimized with pre-trained geometric priors, enables detailed reconstruction of large indoor scenes while being more scalable, efficient, and robust (Zhu et al., 2022). Similarly, Wei et al. introduced DO-SLAM, a novel SLAM algorithm built upon ORB-SLAM2 and designed to enhance localization accuracy and system robustness in dynamic environments. By introducing outlier detection, this approach aims to mitigate the impact of dynamic objects on SLAM performance, improving both accuracy and reliability in challenging scenarios (Wei, Zhou, Duan, Liu and An, 2023). Additionally, Yu et al. presented DS-SLAM, which also extends ORB SLAM2 for highly dynamic environments. Their method uses five threads-tracking, semantic segmentation, local mapping, loop closing, and dense semantic map creation to improve the localization accuracy (Yu et al., 2018a). In a similar vein, Eslamian et al. proposed Det-SLAM, based on ORB SLAM3 and Detectron2, which identifies and eradicates dynamic spots to accomplish semantic SLAM in dynamic situations (Eslamian & Ahmadzadeh, 2022). Xu et al. further advanced this line of research with HMC-SLAM, a robust RGB-D SLAM system that leverages hierarchical multidimensional clustering to detect and filter dynamic features, significantly enhancing pose estimation in highly dynamic scenes (Xu, Zheng, Pan, & Yu, 2025). Moreover, Kim et al. developed SimVODIS, a unified framework that simultaneously performs visual odometry, object detection, and instance segmentation in a single self-supervised architecture that enables both geometric and semantic understanding for downstream SLAM or perception tasks in complex environments (Kim, Kim, & Kim, 2022). These semantic-based approaches have evolved to include adaptive fusion mechanisms that assign dynamic probabilities to detected objects, allowing the system to intelligently adjust its reliance on semantic versus geometric information based on scene complexity (Jiao, Wang, Li, Deng, & Xu, 2022). The computational efficiency of semantic integration can be further improved through selective frame processing. For example, the semantic segmentation can be applied only to keyframes rather than every frame, achieving real-time performance without sacrificing localization accuracy (Lee, Back, Hwang, & Chun, 2023b). Lightweight visual odometry systems have taken efficiency further by integrating adaptive geometric-semantic feature processing that dynamically balances computational load based on scene dynamics. This approach enables robust

performance even on resource-constrained platforms (Wei, Huang, Liu and Zhou, 2023). Beyond frame-level optimizations, deep learning approaches have also improved loop closure detection efficiency through weighted triplet loss functions that learn discriminative features for place recognition, reducing the computational burden of exhaustive frame matching (Dong et al., 2022).

Recent advancements in semantic SLAM for indoor environments, as highlighted by key research papers, underscore the integration of deep learning techniques like BlitzNet and Detectron2 to enhance semantic understanding and dynamic object detection within SLAM systems. Despite these strides, future research should prioritize further enhancing the robustness of SLAM systems in highly dynamic environments, optimizing them for real-time performance, and advancing the integration of multi-sensor data for improved accuracy and efficiency.

5.2. Key approaches in outdoor scene understanding

Recent approaches to outdoor semantic SLAM have emphasized robust handling of dynamic elements and environmental variability. Dynamic SLAM exemplifies this trend by integrating SegNet-based semantic segmentation with ORB-SLAM2, using spatial motion information to achieve a 39.5% accuracy improvement in challenging outdoor conditions (Wen et al., 2023). Lin et al. developed the DPL SLAM technique, which combines ORB SLAM3 with a line detector segment network for efficient pose estimation. They also incorporated CUDA-enabled YOLOv5 for object detection to extract semantic information and remove abnormal features. The authors claim that this novel algorithm excels by not relying on a single source of information, effectively handling both known and unknown dynamic objects (Lin, Zhang, Tian, Yu, & Lan, 2024). Similarly, Zhang et al. proposed a semantic-based visual SLAM technique using ORB SLAM3 with TensorRT-optimized YOLOX to detect humans and non-humans in both indoor and outdoor environments (Zhang & Li, 2023). RSO-SLAM by Qin et al. integrates instance segmentation and optical flow to enhance robustness and localization accuracy in dynamic scenarios. Using a "KMC kmeans + connectivity" algorithm and ORB SLAM2, it detects motion regions and effectively handles non-rigid objects and slow-moving targets. However, the system struggles when large moving objects dominate the field of view or when significant changes occur (Qin et al., 2024).

Li et al. proposed a VSLAM method based on ORB SLAM2 integrated with Deeplab v3+ that incorporates semantic information to eliminate the negative effects of dynamic objects on precise localization (Li, Song et al., 2023). Ai et al. developed a new stereo SLAM system that combines ORB SLAM2 with the deep learning model ENet to enhance the performance of camera pose and trajectory estimation. The author claims that this system is robust and practical, particularly in highly dynamic and complex urban environments (Ai et al., 2023b). Similarly, Esparza et al. proposed a stereo SLAM approach for both indoor and outdoor dynamic environments, using ORB SLAM2 with a neural network-based semantic segmentation and geometrical constraints to effectively eliminate dynamic objects (Esparza & Flores, 2022).

The synthesis of findings from papers on outdoor scene understanding highlights a notable shift towards integrating advanced semantic segmentation and deep learning techniques with traditional SLAM frameworks like ORB SLAM2 and ORB SLAM3. By incorporating methods such as Deeplab v3+, YOLOX, and ENet, these studies demonstrate significant enhancements in accurately localizing and mapping dynamic outdoor environments. This integration enables SLAM systems to effectively discern between stationary and moving objects, thus improving robustness in complex urban landscapes and varied outdoor conditions. However, challenges persist in scenarios involving large moving objects and rapid environmental changes, highlighting the ongoing need for further research. Overall, the fusion of semantic information with SLAM technologies promises advancements in autonomous navigation and spatial understanding, crucial for applications ranging from robotics to augmented reality in outdoor environments. A timeline diagram illustrating the most commonly implemented semantic SLAM systems for both indoor and outdoor scenes is shown in Fig. 18.

The figure illustrates the evolution of techniques employed for constructing semantic maps and enhancing scene understanding from 2017 to 2024. Notably, the adoption of event cameras in semantic SLAM has surged in popularity from 2019 to 2024, indicating a promising avenue for future research. This trend suggests that event cameras hold significant potential for extracting semantic features in highly dynamic scenarios, paving the way for further advancements in the field. Furthermore, Table 7 provides a benchmark comparison of different methodologies and sensors used in semantic SLAM under various scenarios. The comparison highlights the evolution of semantic SLAM methods from traditional ORB-SLAM extensions to more advanced approaches based on semantic graphs, Gaussian splatting, and vision–language models. Earlier methods such as Blitz-SLAM, RDS-SLAM, and RS-SLAM focus mainly on indoor environments and RGB-D sensors, offering reliable dynamic handling but limited applicability outdoors. In contrast, recent frameworks like Kimera2, Dynamic-SLAM, and SG-SLAM extend capabilities to outdoor and highly dynamic environments through multi-sensor integration (LiDAR, IMU, stereo). Emerging approaches such as OpenGS-SLAM, SGS-SLAM, and Hier-SLAM++ demonstrate the integration of foundation models and 3D Gaussian splatting, enabling more generalizable and semantically rich representations. Another important trend is the gradual increase in open-source availability (e.g., RDS-SLAM, Kimera2, SG-SLAM), which facilitates reproducibility and benchmarking. Overall, the table underscores a shift from geometry-centric pipelines toward semantically enriched, multimodal, and open frameworks designed to handle real-world complexities more effectively.

5.3. Emerging trends in semantic SLAM

Recent research in semantic SLAM has introduced new techniques that significantly change how systems understand and build maps of the environment. One major development is the use of a method called 3D Gaussian Splatting (3DGS), which allows systems to create detailed and semantically rich 3D maps much faster and more accurately than before. For example, methods like SGS-SLAM (Li, Liu et al., 2024; Yang, Wang et al., 2025) and OpenGS-SLAM (Chen, Zhang, Zhao and Zhou, 2025; Guerrero-Font, Bonin-Font, Martin-Abadal, Gonzalez-Cid, & Oliver-Codina, 2021; Li, Gu, Liu, Long and Hu, 2018; Yang, Gao et al., 2025) use 3DGS to represent scenes using small 3D blobs (called Gaussians) instead of traditional pixels or point clouds. This leads to sharper object boundaries, faster map updates, and more precise object detection in the environment. Unlike older methods that use slow and blurry representations (like NeRF), these new techniques are more efficient and suitable for real-time use.

In addition, Hier-SLAM++ (Li, Hao, Stuckey, Reid and Rezatofighi, 2025; Zhang, Guo et al., 2024) goes a step further by combining 3DGS with powerful AI models such as SAM and CLIP, which have been trained on large datasets. This combination helps the system recognize and label a wide range of objects, even in unfamiliar environments, and works with both RGB-D and simple monocular camera inputs. These innovations make semantic SLAM systems smarter, faster, and more adaptable to real-world challenges.

Another emerging direction is the fusion of vision–language models with SLAM systems. For example, FindAnything (Abdelnasser et al., 2016; Laina et al., 2025) introduces an open-vocabulary SLAM framework that supports natural language queries during mapping, marking a shift toward generalizable, interactive scene understanding.

Additionally, SG-SLAM (Chen, Li et al., 2025; Dube et al., 2020; Wang, Lu et al., 2025) introduces a semantic-graph-enhanced LiDAR SLAM system. Instead of relying on point-wise labels, SG-SLAM constructs a robust object-level semantic graph, enhancing re-localization, loop closure, and global map consistency. It achieves real-time performance across challenging LiDAR datasets such as KITTI, MulRAN,

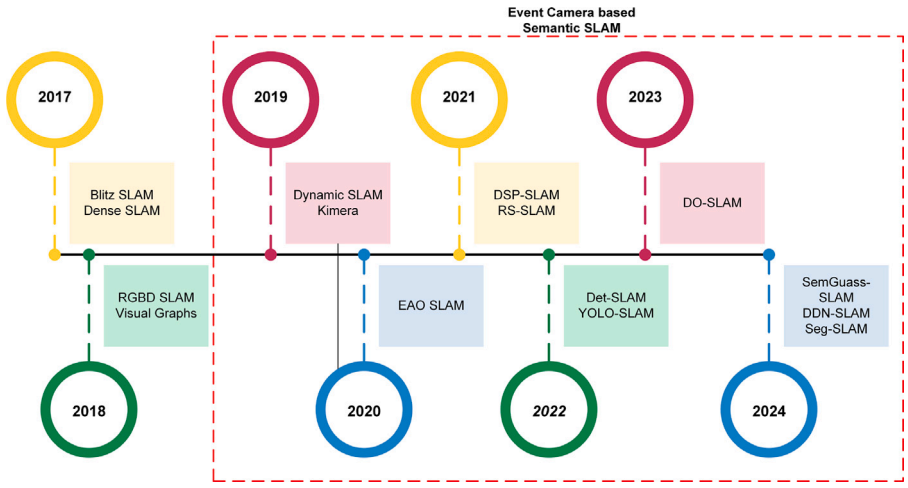

Fig. 18. Timeline diagram for the most commonly known semantic SLAM techniques.

and Apollo, outperforming both geometry-based and semantic SLAM baselines.

Collectively, these works demonstrate a move toward explicit, modular, and generalizable representations in semantic SLAM. They enable not only improved accuracy and robustness but also new capabilities such as zero-shot segmentation, open-set reasoning, and interactive manipulation, which are critical for next-generation applications in robotics, embodied AI, and autonomous systems. These developments illustrate how the field is progressing, but to truly assess their impact we must also consider the evaluation criteria used in the semantic SLAM. Therefore, the following section focuses on the performance metrics applied in Semantic SLAM research.

6. Applications of semantic SLAM

Semantic SLAM has advanced well beyond theoretical development, with real-world deployments across domains such as intelligent industry, smart cities, healthcare, and agriculture. This section provides an application-oriented perspective. We highlight representative use cases, examine practical deployment challenges, and connect algorithmic choices to application needs.

6.1. Intelligent/precision agriculture

Simultaneous Localization and Mapping (SLAM) is a fundamental technology for autonomous robot navigation in unknown environments, providing a joint estimate of robot poses and the 3D location of landmarks (Qadri & Kantor, 2021; Tiozzo Fasiolo, Scalera, Maset, & Gasparetto, 2023). Its applications in agriculture are expanding rapidly, driven by the need for precision agriculture, enhanced food security, and efficient resource management amidst global warming and a growing world population (Pak & Son, 2025b). SLAM enables robots to construct accurate 3D maps of agricultural fields, which is crucial for tasks such as plant phenotyping, crop counting, yield prediction, and intelligent irrigation.

However, agricultural environments present unique and significant challenges for traditional SLAM systems. These include highly unstructured and dynamic settings, varying illumination conditions (e.g., direct sunlight, shadows), lack of texture, repetitive patterns (e.g., rows of crops), wind-induced movement of plants, and the presence of dynamic objects like humans or other robots (Xie et al., 2024). To overcome these issues, robust agricultural SLAM systems increasingly rely on a combination of diverse sensor modalities and advanced data fusion techniques, particularly those incorporating semantic understanding.

6.1.1. Sensor modalities in agricultural SLAM

No single sensor is sufficient for robust SLAM in complex agricultural environments, necessitating a multi-sensor approach (Tiozzo Fasiolo et al., 2023). The most commonly employed and effective sensor modalities include:

LiDAR (Light Detection and Ranging): LiDAR is highly effective for acquiring 3D spatial information and reconstructing dense point clouds (Pak & Son, 2025b). It provides rapid access to precise surface information and is notably robust to varying external lighting conditions, a critical advantage in outdoor agricultural settings where cameras struggle. LiDAR's high measurement range, often up to 100 m, makes it suitable for both closed environments like greenhouses and open fields such as vineyards and orchards. 3D LiDAR, in particular, offers a larger field of view and directly generates dense point clouds, making it highly suitable for comprehensive map reconstruction. For instance, 3D LiDAR has been utilized for intelligent irrigation by leveraging its water-absorbing property to define water point clouds and segment surface water areas, enabling path creation for UAVs. Challenges, however, include potential degradation in rainy conditions due to reflections from raindrops, which can lead to noisy data. It may also perform poorly in environments with long corridors or low plant heights.

Visual Sensors (RGB, Stereo, RGB-D Cameras): Cameras are cost-effective and provide rich environmental information, including fine details that LiDAR might miss. RGB-D cameras, which provide precise depth information through physical measurements, are used for target detection and image segmentation. However, cameras are generally not robust enough for unstructured outdoor environments. They are highly sensitive to illumination changes, requiring constant exposure adjustments that can interfere with visual feature tracking, which often assumes constant brightness. Other challenges include lack of texture, repetitive patterns, and environmental dynamics like wind-blown crops, which can cause traditional visual SLAM algorithms (e.g., ORB-SLAM2) to fail or lose track (Lv et al., 2024b). Depth cameras are also sensitive to sunlight and have limited measurable depth outdoors.

IMU (Inertial Measurement Units): IMUs provide essential information about the robot's orientation (roll, pitch, yaw) and can fill temporal gaps between less frequent GNSS measurements due to their high update rates (Tiozzo Fasiolo et al., 2023). They are compact and have low power consumption, making them ideal for integration into robotic platforms. However, IMUs are prone to drift over time when used for dead reckoning, and their position estimates can become quickly inaccurate due to vibrations on rough terrain. For these reasons, IMUs are almost universally coupled with other sensors to provide robust orientation and mitigate drift.

Table 7
Benchmark comparison of core techniques and characteristics of semantic visual SLAM systems.

Reference	Method	Technique	Network	Sensors used	Public datasets	Indoor	Outdoor	Dynamic	Available
Fan, Zhang, Tang, Liu, and Han (2022)	Blitz SLAM	ORB SLAM2	BlitzNet, ResNet50	RGBD Camera	TUM RGBD	✔	✖	✔	✖
Lin et al. (2024)	DPL-SLAM	ORB SLAM3	YOLOv5-s	Intel D435i	TUM RGB-D, KITTI	✔	✔	✔	✖
Lv et al. (2024a)	MOLO-SLAM	ORB SLAM2	Mask-RCNN	LiDAR, Kinect, Realsense	TUM, KITTI	✔	✔	✔	✖
Qin et al. (2024)	RSO-SLAM	ORB SLAM2	YOLOv5-seg, Lite-FlowNet2	ZED2i Stereo	TUM, BONN, KITTI	✔	✔	✔	✖
Zhao et al. (2022)	KSF-SLAM	ORB SLAM2	SegNet	ZED stereo	TUM RGB-D, KITTI	✔	✔	✔	✖
Liu and Miura (2021b)	RDS-SLAM	ORB SLAM3	–	KinectV2	TUM RGBD	✔	✖	✔	✔
Ran, Yuan, Zhang, Tang et al. (2021) and Xiong et al. (2023)	RS-SLAM	ORB SLAM2	PSPNet	RGB-D	TUM	✔	✖	✔	✔
Abate et al. (2024), Zheng, Lin, and Yang (2024) and Zhang, Song et al. (2025)	Kimera2	Pose Graph	3D Dynamic Scene Graph	LiDAR, RGBD, IMU	A1, Jackal	✖	✔	✔	✔
Cheng et al. (2023)	SG-SLAM	ORB SLAM2	NCNN	RGBD Camera	TUM, BONN	✔	✖	✔	✔
Wu et al. (2020)	EAO-SLAM	ORB SLAM2	YOLOv3	RGBD Camera	TUM, Scenes V2	✔	✖	✔	✔
Li, Zou et al. (2023) and Luo, Rao, and Wu (2023)	FD-SLAM	ORB SLAM3	Fast-SCNN, Deepfillv2	RGBD Camera	TUM RGB-D	✔	✖	✔	✖
Wang, Runz et al. (2021)	DSP-SLAM	ORB SLAM2	–	LiDAR, Stereo, RGBD	KITTI3D, Redwood Chairs	✖	✔	✔	✔
Wen et al. (2023)	Dynamic SLAM	ORB SLAM2	SegNet	RGBD, IMU, LiDAR	KITTI	✖	✔	✔	✔
Cao et al. (2022) and Esparza and Flores (2022)	STDyn-SLAM	ORB SLAM2	SegNet + VGG16	ZED	KITTI	✖	✔	✔	✖
Yang, Gao et al. (2025)	OpenGS-SLAM	GS + Semantic Voting	SAM1.0, Mobile-SAMv2	RGBD Camera	Replica, TUM	✔	✖	✖	✔
Li, Liu et al. (2024)	SGS-SLAM	Semantic GS	CNN + Semantic Loss	RGBD Camera	ScanNet, TUM	✔	✖	✖	✔
Wang, Lu et al. (2025)	SG-SLAM	Semantic Graph	SegNet4D	LiDAR	KITTI, MulRAN	✖	✔	✔	✔
Li, Hao et al. (2025)	Hier-SLAM++	Hier GS + Semantic Loss	CLIP, SAM	RGB-D, Mono	Replica, TUM	✔	✖	✖	✖
Laina et al. (2025)	FindAny-thing	VL Semantic SLAM	CLIP, DINO, SAM	RGB Camera	Replica	✔	✔	✖	✖

GNSS (Global Navigation Satellite System)/RTK-GNSS: GNSS provides absolute position estimates, and with Real-Time Kinematic (RTK) corrections, it can achieve centimeter-level accuracy. This is vital for georeferencing collected data and ensuring global consistency of maps. However, its reliability is significantly reduced in densely vegetated areas due to signal blockage and multi-path reflection, a common issue in agricultural settings. RTK-GNSS also typically requires a reference station, adding to cost and setup complexity (Tiozzo Fasiolo et al., 2023).

Radar: Radar offers advantages as a robotic perception modality in adverse weather conditions, such as dust, fog, rain, and snow, where LiDAR and cameras may perform poorly. Radars also offer better

penetration in vegetation and can detect occluded targets. However, they typically build 2D images, limiting their direct use for 3D map reconstruction (Tiozzo Fasiolo et al., 2023).

Given the inherent limitations of individual sensors in dynamic and complex agricultural environments, multi-sensor fusion is critical for achieving robust SLAM performance. Modern agricultural SLAM heavily leverages advanced data fusion techniques, with a strong emphasis on integrating semantic information. In conclusion, robust SLAM in agriculture demands a combination of LiDAR, visual, IMU, and RTK-GNSS sensors, particularly in challenging environments. These modalities are most effectively integrated using tightly coupled data fusion techniques like factor graphs, critically enhanced by deep learning-based semantic segmentation and object detection to intelligently handle dynamic elements and extract meaningful, robust landmarks within complex agricultural scenes. Future research will likely focus on developing lighter deep learning architectures, improving real-time performance, and further integrating geometric and semantic information to enhance the system's ability to discern the motion state of targets and adapt to varying scales and conditions.

6.2. Intelligent industry and warehousing

In the logistics and warehousing domain, semantic SLAM has been successfully integrated into autonomous aerial inventory systems, enabling fast and reliable stock management without human intervention. Beul et al. (2018) developed a micro aerial vehicle (MAV) that performs fully autonomous stocktaking inside large warehouses, relying on a 3D LiDAR-based SLAM pipeline for localization and mapping, while integrating RFID readers and fiducial markers for item recognition. The system was evaluated in an operational warehouse, where it navigated narrow aisles, avoided static and dynamic obstacles such as forklifts, and maintained accurate pose estimation in highly self-similar environments. Experimental results showed robust navigation at velocities up to 2.1 m/s with a mean waypoint deviation of less than 10 cm, demonstrating industrial-grade accuracy and efficiency. This application highlights the practical relevance of semantic SLAM in Industry 4.0, where inventory automation reduces human workload, minimizes errors, and supports continuous real-time stock management. Importantly, it also illustrates how algorithmic advances in SLAM directly translate into measurable performance gains in intelligent industrial systems.

6.3. Autonomous driving

SLAM has become an indispensable technology in autonomous driving, enabling vehicles to achieve centimeter-level localization and robust perception in complex and dynamic road environments. In practice, autonomous vehicles must operate across diverse scenarios—highways, dense urban areas, tunnels, and adverse weather—where traditional GNSS or odometry-based localization methods often fail (Wang, Guo, Chen and Lu, 2025). SLAM systems leverage multi-sensor configurations, including LiDAR, cameras, and IMUs, to construct high-precision maps while simultaneously localizing the vehicle in real time. Semantic SLAM further augments this capability by integrating object-level understanding, allowing vehicles to detect and track dynamic agents such as cars and pedestrians, and to interpret traffic signs and lane markings as semantic landmarks. Case studies demonstrate that approaches like DynaSLAM and SG-SLAM significantly improve trajectory accuracy and robustness in urban driving scenes by filtering or modeling dynamic objects, while large-scale mapping frameworks now allow for the construction of kilometer-scale maps that can be continuously updated through crowd-sourced data. Despite these advances, deployment challenges remain: computational requirements are high, and maintaining robustness under variable lighting and weather conditions is still difficult. Nonetheless, SLAM-based localization and semantic scene understanding form the backbone of advanced driver assistance systems (ADAS) and are foundational for achieving safe and reliable Level 4+ autonomous driving in real-world intelligent transportation systems (Zheng, Wang, Rizos, Ding, & El-Mowafy, 2023).

By expanding the discussion of applications and deployment, this survey not only summarizes the state of semantic SLAM methods but also demonstrates their practical value across domains central to intelligent systems research and practice.

7. Practical challenges and deployment

While semantic SLAM has demonstrated remarkable progress in academic settings, the transition from laboratory prototypes to real-world deployment introduces a number of practical challenges. These challenges span hardware requirements, robustness under uncontrolled conditions, scalability in long-term operations, and the persistent gap between research systems and commercially viable products. Addressing these factors is critical to enable semantic SLAM to mature into a technology suitable for practitioners across domains such as intelligent industry, transportation, and agriculture.

7.1. Computational requirements

Semantic SLAM systems often combine geometric estimation with deep neural networks for object detection and segmentation, resulting in significant computational demand. Many approaches rely on GPU acceleration to achieve real-time performance, especially in dynamic environments where every frame must be processed at high frequency (Galagain, Poreba, & Goulette, 2025). On resource-constrained platforms, this creates a trade-off between accuracy and efficiency. While lightweight architectures such as MobileNet or Tiny-YOLO can partially mitigate these issues, they often sacrifice semantic richness. Designing architectures that balance accuracy, speed, and power consumption remains an open challenge, particularly for embedded and mobile robotic platforms.

7.2. Robustness in real-world environments

In real-world deployments, semantic SLAM must cope with factors rarely encountered in controlled laboratory experiments. These include dynamic objects such as humans, forklifts, or animals, changing illumination, sensor noise, and environmental variability. Esparza and Flores (2021) demonstrate that even state-of-the-art systems struggle when large moving objects dominate the field of view, while Han, Li, Wang, and Zhou (2021b) highlight the difficulties of maintaining semantic consistency in cluttered indoor scenes. Robustness therefore requires not only accurate semantic segmentation but also reliable filtering of outliers and adaptive fusion strategies that can adjust to environmental changes without degrading localization accuracy.

7.3. Scalability and long-term mapping

Scaling semantic SLAM to large and long-term deployments introduces further difficulties. Semantic maps must be continuously updated to remain relevant, yet doing so without introducing drift, redundancy, or inconsistency is non-trivial. For example, autonomous vehicles operating in urban environments require kilometer-scale semantic maps that must adapt to changing infrastructure, traffic patterns, and seasonal variations. Efficient map management, including storage, querying, and updating mechanisms, is thus essential for practical scalability (Galagain et al., 2025). Recent work has explored semantic graphs and incremental updates as solutions, but reliable large-scale deployment is still an active research challenge.

7.4. From research to commercial products

Finally, there exists a significant gap between academic prototypes and industrial-grade solutions. Academic research often evaluates systems on curated datasets under controlled conditions, whereas real deployment demands robustness, reliability, and maintainability. A recent review of autonomous forklifts (Fraifer et al., 2025) underscores that while many SLAM-based prototypes achieve promising results in warehouses, deployment in production settings faces hurdles related to safety certification, integration with enterprise systems, and cost of hardware. Bridging this gap requires not only algorithmic innovation but also attention to system engineering, reliability, and human–robot interaction in real industrial workflows.

In summary, although semantic SLAM has advanced significantly in terms of algorithms and benchmarks, addressing the challenges of computational efficiency, robustness, scalability, and deployment readiness remains key to translating research into practical impact. Overcoming these challenges will determine the technology's adoption across intelligent cities, industry, agriculture, and other domains.

8. Performance metrics used in semantic SLAM

This section briefly reviews the evaluation metrics and comparison methods used in both qualitative and quantitative assessments of SLAM and semantic SLAM algorithms. Typically, SLAM systems are evaluated based on accuracy, focusing on positional error, rotational error, and computation time. We classify evaluation metrics into semantic mapping, geometric SLAM, and tracking metrics, corresponding to the core functional components of semantic SLAM systems. Semantic mapping metrics evaluate how well the system understands and labels the environment, geometric SLAM metrics measure the accuracy of pose estimation and map reconstruction, and tracking metrics assess performance in following dynamic objects over time. This structure helps clearly separate the different aspects of system performance and aligns with how these metrics are commonly used in the literature. A detailed sub-classification of these metrics is illustrated in Fig. 19, and selected metrics from each category are explained as follows. Metrics such as Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) directly reflect the accuracy of localization and mapping, which form the foundation for reliable semantic interpretation of the environment. Similarly, metrics related to semantic mapping and object recognition contribute to evaluating how well SLAM systems capture and represent the meaningful structure of a scene. By linking geometric accuracy with semantic consistency, these measures provide a more complete assessment of a system's capability in advancing scene understanding.

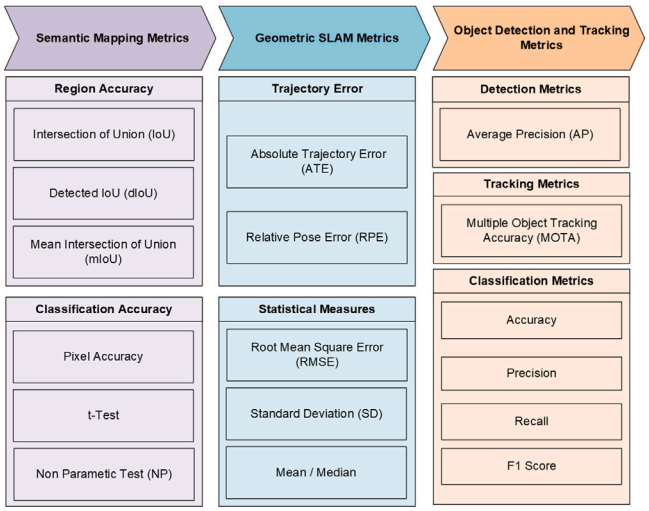

Fig. 19. Performance metrics for evaluation of semantic SLAM in scene understanding.

8.1. Tracking metrics

Accuracy (Karpyshev et al., 2022), precision (Hu, Wu et al., 2025; Liu, Lei et al., 2024; Shoukat et al., 2024; Tardioli et al., 2016; Vasilopoulos, Pavlakos, Schmeckpeper, Daniilidis, & Koditschek, 2022), and recall (Wang, Zheng, & Li, 2023) are key metrics for evaluating the performance of classification models in machine learning. Each metric provides insight into a different aspect of the model's effectiveness, with its relevance varying depending on the use case. Accuracy represents the ratio of correct predictions to the total number of predictions, whereas precision measures the proportion of true positives among all positive predictions, typically expressed as a percentage. In contrast, recall, also known as sensitivity, determines how often the model correctly identifies the true class within the dataset. The balance between precision and recall is especially important in visual loop closure detection. Semantic-aware models must minimize both false positives (requiring high precision) and false negatives (requiring high recall) to maintain map consistency in dynamic environments (Osman, Darwish, & Bayoumi, 2023). These metrics can be calculated using the equations provided from (22) to (24):

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total predictions from samples}} \tag{22}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{23}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{24}$$

F1 score measures the harmonic mean of precision and recall, as shown in Eq. (25) (Karpyshev et al., 2022):

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{25}$$

Additionally, Average Precision (AP) is a key performance indicator of detection metrics that reflects the trade-offs between precision and recall, with values ranging between 0 and 1. This metric is calculated using Eq. (26) (Xing et al., 2022):

$$\text{AP} = \int_{r=0}^{1} P(R) dR \tag{26}$$

where $P(R)$ represents precision as a function of recall R. MOTA is an ideal performance metric for tracking multiple objects, features, or landmarks over time, and it is defined by Eq. (27) (Chen et al., 2019; Li, Wang et al., 2018; Sahili et al., 2023):

$$\text{MOTA} = 1 - \frac{\sum_t FN_t + FP_t + IDS_t}{\sum_t GT_t} \tag{27}$$

where FN_t represents the number of false negatives, FP_t the number of false positives, IDS_t the number of identity switches at time t, and GT_t the ground truth.

8.2. Semantic mapping metrics

Semantic metrics evaluate the system's ability not only to map the environment and localize within it but also to understand and categorize the elements present. These metrics assess the performance of the system in integrating semantic information, such as identifying objects and their relationships, alongside traditional spatial mapping. Commonly used metrics include Intersection over Union (IoU) (He, Li, Wang, & Wang, 2023) and pixel accuracy (PA) (Han et al., 2021a). IoU is calculated based on the overlap between ground truth and predicted bounding boxes, as shown in Eq. (28):

$$\text{IoU}_c = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{28}$$

Pixel accuracy (PA) is a common metric used to evaluate the performance of semantic segmentation, defined as the ratio of correctly classified pixels to the total pixels in the image, as shown in (29):

$$PA = \frac{\sum_{j=1}^{k} n_{jj}}{\sum_{j=1}^{k} t_j} \tag{29}$$

where n_{jj} represents the total number of pixels that are both classified and labeled as class j, essentially corresponding to the number of true positives for class j. t_j refers to the total number of pixels labeled as class j.

Mean Intersection over Union (mIoU) is an extension of IoU, used for evaluating multiple classes or segments (Liu, Sun and Liu, 2021), whereas Distance Intersection over Union (DIoU) calculates the distance between the centroids of predicted and ground truth regions (Jiang, Guo, Jiang, Hu, & Zhu, 2021), with the formulas given in Eqs. (30) and (31). Additional metrics used to evaluate semantic SLAM include the t-test (Trejos, Rincón, Bolaños, Fallas, & Marín, 2022) and the Non-Parametric (NP) test (Wilcoxon, 1992). The t-test quantitatively assesses SLAM system performance by comparing the estimated trajectory with the ground truth, as shown in Eq. (32), whereas the NP test processes red and green point clouds and applies the Wilcoxon Rank-Sum test to verify the null hypothesis.

$$\text{mIoU} = \frac{1}{c} \sum_{c} \text{IoU}_c \tag{30}$$

where c is the total number of classes, and IoU_c represents the intersection of union for a specific class c.

$$\text{DIoU} = 1 - \text{IoU}_\text{c} + \frac{\rho^2\left(\mathbf{b}, \mathbf{b}_{gt}\right)}{l^2} \tag{31}$$

where $\rho(b, b_{gt})$ represents the Euclidean distance between the central points of the predicted bounding box b and the ground truth bounding box b_{gt}; l is the diagonal length of the smallest enclosing box that covers both bounding boxes.

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \tag{32}$$

where X represents the sample mean, μ is the population mean, s indicates the standard deviation, and n denotes the sample size.

8.3. Geometric SLAM metrics

Geometric SLAM metrics are crucial for evaluating both performance and reliability. They assess accuracy in positioning and mapping, robustness in handling diverse and dynamic environments, and consistency in trajectory and map generation. These metrics also measure computational efficiency, scalability to large environments, real-time performance, and user-centric factors like ease of integration and usability. Common error metrics include RMSE (Han & Xi, 2020; Zhu, Wang, Lu, & Jiang, 2024), ATE (Peng, Tong, Yang, Wang, & Zou, 2025; Zhang, Wang and Zhang, 2022), RPE (Guan et al., 2020; Peng, Ran, Yuan, Zhang and Xiao, 2024), Mean (Zhang & Li, 2023), Median (Chen, Xue, Fang, Pan, & Zheng, 2020; Lin, Su et al., 2025), and Standard Deviation (Han & Xi, 2020), all of which quantify the differences between estimated and actual positions and trajectories. These are commonly expressed in meters (m), centimeters (cm), or millimeters (mm), based on the camera trajectory measured in different use cases. For rotational RPE, the error is generally measured in degrees, percentages, or degrees per 100 m (Qin et al., 2024). By using these metrics, developers can enhance the reliability and efficiency of semantic SLAM systems for various applications.

8.3.1. Absolute Trajectory Error (ATE)

Absolute Trajectory Error (ATE) is a crucial metric that calculates the error between the estimated camera trajectory and the ground truth. This metric becomes particularly important when evaluating SLAM systems in dynamic environments, where methods like SA-LOAM (Li, Kong et al., 2021) and Pseudo-Anchors (Yang, He, Zhuang, Wang and Yang, 2023) demonstrate improved trajectory accuracy through semantic feature integration (Deng et al., 2019; Huang, Wang, Yun, Jiang, & Gong, 2025; Zhao, Zhang, Gu, Yang, & Huang, 2019):

$$\text{ATE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{p}_i^{\text{estimated}} - \mathbf{p}_i^{\text{ground truth}}\right\|^2} \tag{33}$$

where n represents the number of data points, and $\left\|\mathbf{p}_i^{\text{estimated}} - \mathbf{p}_i^{\text{ground truth}}\right\|^2$ is the squared Euclidean norm between the estimated and ground truth positions for the ith sample.

A lower ATE indicates more accurate localization and mapping, helping developers identify and improve discrepancies in their semantic SLAM algorithms.

8.3.2. Relative Pose Error (RPE)

The Relative Pose Error (RPE) is a key metric in evaluating the performance of visual SLAM by measuring the accuracy of relative motion, or pose change, between consecutive frames or time steps (Cheng et al., 2023). It helps in assessing how well the semantic SLAM system tracks the incremental movement of the camera or sensor. The translational and rotational components of the RPE (RPE_t and RPE_r) are shown in (34) and (35):

$$\text{RPE}_\text{t} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left\|\mathbf{t}_k^{\text{estimated}} - \mathbf{t}_k^{\text{ground truth}}\right\|^2} \tag{34}$$

$$\text{RPE}_\text{r} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left\|\log\left(\mathbf{q}_k^{\text{estimated}} \cdot \left(\mathbf{q}_k^{\text{ground truth}}\right)^{-1}\right)\right\|^2} \tag{35}$$

where n represents the number of data points, $\left\|\mathbf{t}_k^{\text{estimated}} - t_k^{\text{ground truth}}\right\|^2$ is the squared Euclidean norm of the predicted and ground truth translation vectors, and $\left\|\log\left(\mathbf{q}_k^{\text{estimated}} \cdot \left(\mathbf{q}_k^{\text{ground truth}}\right)^{-1}\right)\right\|^2$ is the logarithm of the relative rotation between the predicted and ground truth quaternion vectors for the kth sample.

8.3.3. Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) is a commonly used metric for quantifying the average magnitude of errors between estimated and true values, as shown in (36) (Liu & Miura, 2021a; Zhou, Tao, Fu, Zhu and Ma, 2023):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(e_i\right)^2} \tag{36}$$

where e_i represents the error between the estimated and ground truth values for the ith sample, and n is the total number of data points.

8.3.4. Statistical measures

In semantic SLAM systems, apart from other metrics, statistical measures such as mean, median, and standard deviation are also vital for evaluating system performance and reliability (Ahmed, Masood, Fremont, & Fantoni, 2023; Wu, Guo et al., 2022). The mean, or average, quantifies the central tendency of a dataset, making it key to assessing the overall performance of SLAM systems. The median, the middle value of an ordered dataset, provides a measure of central tendency that is less affected by outliers compared to the mean. Standard deviation indicates the amount of variation or dispersion within a dataset, showing how far the values deviate from the mean. By using these statistical metrics, developers and researchers can gain deeper insights into

Table 8
Benchmark comparison of performance metrics using the TUM RGBD datasets.

Reference	Accuracy (%)	ATE (m)	RPE_t (m)	RPE_r (deg)	IoU (%)
Fan et al. (2022)	✖	0.0159	0.0182	0.5785	✖
Wu, Guo et al. (2022)	✖	0.0546	0.0315	0.7417	✖
Cheng et al. (2023)	✖	0.0175	0.02196	0.5611	✖
Qian et al. (2021)	92.19	0.0429	✖	✖	✖
Wu et al. (2020)	✖	✖	✖	✖	81.75
Bavle, De La Puente, How, and Campoy (2020)	✖	0.0365	✖	✖	✖

Table 9
Benchmark comparison of performance metrics using the Bonn datasets.

Reference	ATE (m)	RPE_t (m)	RPE_r (deg)
He et al. (2023)	0.0245	0.1878	14.2961
Singh et al. (2022)	0.0620	0.0690	✖
Jiang, Xu, Li, Feng, and Zhang (2024)	0.1230	✖	✖
Wu, Guo et al. (2022)	0.0890	✖	✖
Cheng et al. (2023)	0.0644	✖	✖
Li, Guo et al. (2025)	0.0290	✖	✖

the performance, robustness, and reliability of SLAM systems, thereby facilitating better optimization and enhancement of algorithms and implementations (Wu, Zhao et al., 2022). The corresponding formulas are shown in (37) and (38):

$$\text{Mean Error} = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{p}_i^{\text{estimated}} - \mathbf{p}_i^{\text{ground truth}}\right\| \tag{37}$$

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2} \tag{38}$$

where n represents the number of data points, x_i is the ith data point, $\left\|\mathbf{p}_i^{\text{estimated}} - \mathbf{p}_i^{\text{ground truth}}\right\|$ is the Euclidean norm of the predicted and ground truth positions for the ith sample, and μ is the average value of the dataset.

In addition to discussing the various performance metrics used in semantic SLAM, a benchmark comparison of these metrics for both indoor and outdoor scenes is provided for the TUM RGBD, Bonn, and KITTI datasets in Tables 8–10, respectively.

On the indoor TUM RGB-D benchmark (Table 8), semantically informed pipelines achieve centimeter-level trajectory errors. Fan et al. (2022) report the lowest ATE (0.0159 m) and the lowest translational RPE (0.0182 m), with Cheng et al. (2023) close behind (ATE 0.0175 m; RPE_t 0.02196 m), while Wu, Guo et al. (2022) is notably higher (ATE 0.0546 m; RPE_t 0.0315 m). Rotational accuracy is consistently strong for the top entries ($RPE_r \approx 0.56°$–$0.74°$). Only Qian et al. (2021) report a classification Accuracy (92.19%) and only Wu et al. (2020) report IoU (81.75%), highlighting heterogeneous metric reporting across different works. Overall, these results indicate that semantic integration chiefly benefits geometric accuracy on indoor RGB-D data, whereas inconsistent disclosure of semantic metrics limits strict cross-paper comparison.

Table 9 highlights the variability of performance on the Bonn dataset, which is characterized by dynamic indoor scenes. He et al. (2023) achieve the lowest ATE and report rotational accuracy alongside moderate translational drift (ATE 0.0245 m; 14.3°; RPE_t = 0.1878 m). Singh et al. (2022) show competitive results with the lowest reported translational RPE, while (ATE 0.062 m; RPE_t = 0.069 m) (Li, Guo et al., 2025) also perform strongly (ATE 0.029 m). In contrast, Jiang et al. (2024) and Wu, Guo et al. (2022) present higher ATE values, reflecting less robustness to challenging dynamics. (ATE 0.123 m and 0.089 m respectively). Cheng et al. (2023) falls in the mid-range (0.0644 m). Overall, these results demonstrate that while several methods achieve sub-decimeter accuracy, robustness to dynamic variations remains inconsistent across approaches, underlining the need for more standardized reporting of translational and rotational errors.

Table 10 summarizes performance on the KITTI dataset, which is widely used for outdoor driving scenarios with high dynamic complexity. Qin et al. (2024) achieve strong relative pose performance, reporting the lowest translational and rotational RPE values, despite a higher ATE. (RPE_t = 0.0072 m; RPE_r = 0.002°; ATE 2.31 m). Lv et al. (2024a) show the largest ATE and translational RPE, indicating limited robustness in large-scale outdoor conditions. (ATE 3.53 m; RPE_t = 1.807 m). By contrast, Esparza and Flores (2022) and Wang, Li et al. (2020) achieve more balanced results, while Esparza and Flores demonstrate a particularly low translational RPE. (ATE 1.33 m and 1.45 m; RPE_t = 0.0233 m). Chen, Liu et al. (2022) report the only accuracy percentage but also the highest ATE, suggesting a trade-off between recognition accuracy and localization drift. (Accuracy 80.82%; ATE 4.61 m). Singh et al. (2022) focus on rotational performance. (RPE_r = 0.87°). Overall, these results emphasize that while certain methods excel in pose accuracy, achieving consistently low ATE across challenging outdoor environments remains difficult, underscoring the inherent complexity of large-scale dynamic driving datasets.

8.4. Replication of results from open-source papers

Replication of results from open-source papers plays a crucial role in validating the reliability and generalizability of Semantic SLAM methods, as it allows researchers to benchmark existing approaches under consistent conditions. We evaluated selected semantic SLAM systems algorithms on the TUM and KITTI datasets. The benchmark metrics, consistent with those described earlier, were selected to demonstrate algorithm performance across diverse indoor and outdoor environments, including static and dynamic settings.

8.4.1. System specifications

The experiments were conducted on a high-performance computer with the following specifications:

- Processor: AMD Ryzen 9 3950X 16-Core Processor with 32 threads, operating at a base clock speed of 2.2 GHz and a maximum clock speed of 3.5 GHz.
- GPU: NVIDIA GeForce RTX 2080 Ti with 11 GB of VRAM, supporting CUDA version 12.1.
- Operating System: Ubuntu 16.04.
- Robotics Framework: ROS Melodic.

We employed widely-used datasets to ensure reproducible results and comprehensive evaluation:

- TUM RGB-D dataset
- KITTI dataset

Table 10
Benchmark comparison of performance metrics using the KITTI datasets.

Reference	Accuracy (%)	ATE (m)	RPE_t (m)	RPE_r (deg)
Qin et al. (2024)	✖	2.3136	0.0072	0.0020
Lv et al. (2024a)	✖	3.5343	1.8074	✖
Wang, Li, Shen and Cai (2020)	✖	1.3267	0.4815	✖
Esparza and Flores (2022)	✖	1.4493	0.0233	✖
Chen, Liu et al. (2022)	80.82	4.606	✖	✖
Singh et al. (2022)	✖	✖	0.87	✖

Table 11
Evaluation of semantic SLAM systems on TUM datasets using ATE performance metric.

Sequence	SG-SLAM (NCNN) (Cheng et al., 2023)		Dyna-SLAM (Mask R-CNN) (Bescos et al., 2018)		DS-SLAM (SegNet) (Yu et al., 2018a)		YOLO-SLAM (darknet19-yolov3) (Wu, Guo et al., 2022)		RDS-SLAM (Mask R-CNN) (Liu & Miura, 2021b)		RDS-SLAM (Segnet) (Liu & Miura, 2021b)	
Results	Original	Ours	Original	Ours	Original	Ours	Original	Ours	Original	Ours	Original	Ours
fr3_walking_xyz	0.0152	0.019	0.015	0.016	0.024	0.023	0.014	0.013	0.021	0.021	0.057	0.056
fr3_walking_static	0.007	0.008	0.007	0.006	0.008	0.0078	0.007	0.006	0.081	0.078	0.02	0.02
fr3_walking_rpy	0.032	0.034	0.136	0.135	0.443	0.444	0.216	0.223	0.146	0.145	0.16	0.159
fr3_walking_half	0.026	0.023	0.029	0.029	0.03	0.03	0.028	0.028	0.025	0.030	0.08	0.08

Table 12
Evaluation of semantic SLAM systems on TUM dataset using RPE_t performance metric.

Sequence	SG-SLAM (NCNN) (Cheng et al., 2023)		Dyna-SLAM (Mask R-CNN) (Bescos et al., 2018)		DS-SLAM (SegNet) (Yu et al., 2018a)		YOLO-SLAM (darknet19-yolov3) (Wu, Guo et al., 2022)		RDS-SLAM (Mask R-CNN) (Liu & Miura, 2021b)		RDS-SLAM (Segnet) (Liu & Miura, 2021b)	
Results	Original	Ours	Original	Ours	Original	Ours	Original	Ours	Original	Ours	Original	Ours
fr3_walking_xyz	0.0194	0.022	0.021	0.022	0.033	0.033	0.019	0.019	0.028	0.028	0.042	0.043
fr3_walking_static	0.010	0.013	0.008	0.009	0.0102	0.011	0.009	0.0087	0.041	0.042	0.022	0.022
fr3_walking_rpy	0.045	0.074	0.044	0.045	0.150	0.15	0.093	0.092	0.111	0.111	0.132	0.132
fr3_walking_half	0.027	0.032	0.028	0.028	0.029	0.029	0.026	0.027	0.028	0.027	0.048	0.051

8.4.2. Results

We present evaluation results for various semantic SLAM systems tested on both indoor and outdoor datasets. First, the results for the TUM RGB-D indoor dataset are presented, followed by the KITTI outdoor dataset, which further highlights the advantages of semantic SLAM in handling challenging outdoor scenarios.

- TUM dataset: The following evaluation results were obtained using the TUM RGB-D indoor dataset, with a focus on the sequences `fr3_walking_xyz`, `fr3_walking_static`, `fr3_walking_rpy`, and `fr3_walking_half`, as presented in Tables 11, 12, and 13, respectively. Key performance metrics such as ATE, RMSE, RPE_t, and RPE_R, are used to highlight the effectiveness of integrating semantic information in enhancing SLAM accuracy in dynamic environments. The integration of semantic information in dynamic SLAM systems significantly enhances their performance compared to traditional methods like ORB-SLAM2. This improvement is evident in the superior results across various metrics and sequences from the TUM dataset. For instance, SG-SLAM achieves an ATE of 0.019 m in the `fr3_walking_xyz` sequence, vastly outperforming ORB-SLAM2's 0.693 m. Additionally, SG-SLAM's RPE_t for the same sequence is 0.022 m, which is significantly better than ORB-SLAM2's 0.475 m.
- KITTI dataset:
  The KITTI dataset, particularly in outdoor environments, further highlights the benefits of semantic SLAM. For instance, VDO-SLAM with Mask R-CNN records an ATE of 1.2 m in the KITTI 00 sequence, compared to ORB-SLAM2's 1.3 m. Furthermore, VDO-SLAM achieves a RPE_t of 0.06 m in the same sequence, compared to ORB-SLAM2's 0.04 m. These results emphasize the importance of semantic data integration, which enhances scene understanding by allowing the system to distinguish between different objects and dynamic elements.

Tables 14, 15, and 16 present a comprehensive evaluation of the performance of different SLAM systems integrated with semantic information, including Dyna-SLAM and VDO-SLAM, both of which use Mask R-CNN for semantic integration, compared to the traditional ORB-SLAM2 system, which does not employ semantic data. The evaluation is based on KITTI outdoor dataset sequences, focusing on key metrics such as ATE, RPE_t, and RPE_r.

8.4.3. Benchmarking against ORB-SLAM2

To establish a comprehensive benchmarking framework, all tested systems were evaluated under identical conditions using widely accepted datasets (TUM RGB-D and KITTI) and metrics, including ATE, Translational RPE (RPE_t), and Rotational RPE (RPE_r). ORB-SLAM2, which lacks semantic integration, was chosen as the baseline system to provide a reference point for evaluating the impact of semantic methods.

The benchmarking results on the TUM RGB-D dataset show significant performance differences between ORB-SLAM2 and semantic SLAM systems. For the sequence `fr3_walking_xyz`, ORB-SLAM2 achieved an ATE of 0.693 m, whereas SG-SLAM recorded an impressive 0.019 m, reflecting a 97.3% improvement. Similarly, for the `fr3_walking_static` sequence, SG-SLAM achieved an ATE of 0.008 m compared to ORB-SLAM2's 0.392 m, highlighting its superior capability in both static and dynamic scenarios. These results are shown in Table 17.

The RPE_t results further emphasize the impact of semantic methods. In the `fr3_walking_xyz` sequence, ORB-SLAM2's RPE was 0.475 m/frame compared to SG-SLAM's 0.022 m/frame, demonstrating a significant reduction in translational drift due to semantic integration. This is clearly shown in Table 18.

ORB-SLAM2 performed moderately well on static sequences in the KITTI dataset but showed considerable limitations in dynamic scenarios. For instance, in the `KITTI 01` sequence, which involves high-speed

Table 13
Evaluation of semantic SLAM systems on TUM dataset using RPE_r performance metric.

Sequence	SG-SLAM (NCNN) (Cheng et al., 2023)		Dyna-SLAM (Mask R-CNN) (Bescos et al., 2018)		DS-SLAM (SegNet) (Yu et al., 2018a)		YOLO-SLAM (darknet19-yolov3) (Wu, Guo et al., 2022)		RDS-SLAM (Mask R-CNN) (Liu & Miura, 2021b)		RDS-SLAM (Segnet) (Liu & Miura, 2021b)	
Results	Original	Ours	Original	Ours	Original	Ours	Original	Ours	Original	Ours	Original	Ours
fr3_walking_xyz	0.504	0.504	0.628	0.627	0.826	0.834	0.598	0.588	0.723	0.028	0.922	0.919
fr3_walking_static	0.267	0.270	0.261	0.271	0.269	0.269	0.262	0.342	1.168	0.042	0.494	0.540
fr3_walking_rpy	0.956	0.957	0.989	1.002	3.012	3.000	1.823	1.823	9.319	0.111	13.170	13.210
fr3_walking_half	0.811	0.812	0.784	0.776	0.814	0.812	0.753	0.752	0.821	0.027	1.876	1.874

Table 14
Evaluation of semantic SLAM systems on KITTI dataset using ATE performance metric.

Seq	Dyna-SLAM (Mask R-CNN) (Bescos et al., 2018)		VDO-SLAM (Mask R-CNN) (Zhang, Henein, Mahony, & Ila, 2020)	
Results	Original	Ours	Original	Ours
KITTI 00	1.4	1.2	1.2	1.2
KITTI 01	9.4	10.1	8.9	8.7
KITTI 02	6.7	7.1	5.4	5.7
KITTI 03	0.6	0.6	0.6	0.6
KITTI 04	0.2	0.3	0.2	0.2

Table 15
Evaluation of semantic SLAM systems on KITTI dataset using RPE_t performance metric.

Seq	Dyna-SLAM (Mask R-CNN) (Bescos et al., 2018)		VDO-SLAM (Mask R-CNN) (Zhang et al., 2020)	
Results	Original	Ours	Original	Ours
KITTI 00	0.04	0.03	0.067	0.072
KITTI 01	0.05	0.04	0.044	0.044
KITTI 02	0.04	0.05	0.021	0.020
KITTI 03	0.06	0.04	0.03	0.04
KITTI 04	0.07	0.06	0.05	0.05

Table 16
Evaluation of semantic SLAM systems on KITTI datasets using RPE_r performance metric.

Seq	Dyna-SLAM (Mask R-CNN) (Bescos et al., 2018)		VDO-SLAM (Mask R-CNN) (Zhang et al., 2020)	
Results	Original	Ours	Original	Ours
KITTI 00	0.06	0.05	0.07	0.072
KITTI 01	0.04	0.03	0.012	0.003
KITTI 02	0.03	0.03	0.04	0.04
KITTI 03	0.04	0.04	0.08	0.078
KITTI 04	0.06	0.05	0.11	0.10

Table 17
ATE comparison on TUM RGB-D dataset.

Sequence	ORB-SLAM2 (m)	SG-SLAM (m)
fr3_walking_xyz	0.693	0.019
fr3_walking_static	0.392	0.008
fr3_walking_rpy	1.022	0.034

Table 18
RPE_t comparison on TUM RGB-D dataset.

Sequence	ORB-SLAM2 (m/frame)	SG-SLAM (m/frame)
fr3_walking_xyz	0.475	0.022
fr3_walking_static	0.361	0.013
fr3_walking_rpy	0.451	0.074

motion on highways, DynaSLAM recorded an ATE of 10.1 m compared to ORB-SLAM2's 10.4 m, showing a modest improvement. For the `KITTI 00` sequence, DynaSLAM achieved a slightly improved ATE of 1.2 m over ORB-SLAM2's 1.3 m. These results illustrate that semantic integration provides more robust performance in handling dynamic elements, though its impact varies depending on the complexity of the scene. Results are displayed in Table 19.

A comparative analysis reveals the following key insights:

Table 19
ATE comparison on KITTI dataset.

Sequence	ORB-SLAM2 (m)	SG-SLAM (m)
KITTI 00	1.3	1.2
KITTI 01	10.4	10.1
KITTI 02	5.7	7.1

- **Accuracy Improvements:** Semantic SLAM systems consistently outperformed ORB-SLAM2 across both datasets. SG-SLAM demonstrated a 97.3% improvement in ATE for the `fr3_walking_xyz` sequence, while DynaSLAM showed marginal gains for challenging KITTI sequences, such as `KITTI 01`.
- **Dynamic Object Handling:** Semantic methods significantly reduced errors in sequences with high dynamic content. For example, SG-SLAM reduced RPE_t in the `fr3_walking_xyz` sequence by over 95% compared to ORB-SLAM2.
- **Efficiency Trade-Offs:** Semantic systems, while more accurate, often require higher computational resources. RDS-SLAM, however, provided a balance between accuracy and efficiency, achieving notable reductions in processing time compared to ORB-SLAM2.

Table 20
Average processing time per frame (ms).

Systems	Average processing time per frame (ms)
ORB-SLAM2	59.26
SG-SLAM (NCNN)	65.71
YOLO-SLAM (darknet19-yolov3)	696.09
DS-SLAM (SegNet)	59.4
DynaSLAM (Mask R-CNN)	192.00
RDS-SLAM (Mask-RCNN)	57.5
RDS-SLAM (Segnet)	57.5

These results highlight the transformative potential of semantic integration in SLAM, enabling systems to handle dynamic environments more effectively and improve trajectory accuracy. The benchmarking also highlights the importance of standardized evaluation frameworks to ensure consistent comparisons and drive advancements in the field.

8.4.4. Processing time

Table 20 shows the average processing time per frame for various dynamic SLAM systems, each using different semantic algorithms and tested on the previously specified hardware. The results highlight the computational efficiency and performance variations across the systems, ranging from ORB-SLAM2, which does not use semantic algorithms, to more complex setups like YOLO-SLAM and DynaSLAM. The latter systems demonstrate significantly higher processing times due to their advanced dynamic object detection capabilities.

The evaluation of the tested dynamic SLAM systems highlights the significant influence of semantic methods on performance, particularly in handling dynamic environments. Among the systems evaluated, SG-SLAM (using NCNN) and DynaSLAM (leveraging Mask R-CNN) demonstrated superior performance. SG-SLAM excelled in indoor scenarios, such as those represented by the TUM dataset, where its robust semantic segmentation effectively filtered dynamic elements, achieving a notably low ATE of 0.019 m in the `fr3_walking_xyz` sequence compared to ORB-SLAM2's 0.693 m. DynaSLAM, on the other hand, proved highly effective in both indoor and outdoor environments, as demonstrated by its robust results on the KITTI dataset. Its use of Mask R-CNN facilitated accurate detection and exclusion of dynamic objects, yielding reliable mapping and localization even in complex scenes, such as in the `KITTI 00` sequence, where it achieved an ATE of 1.4 m.

Conversely, systems like YOLO-SLAM (darknet19-yolov3) and ORB-SLAM2 underperformed in key areas. YOLO-SLAM, despite its potential for real-time processing, was hindered by high computational demands and less refined semantic segmentation, resulting in higher processing times (696.09 ms per frame) and reduced accuracy in dynamic scenarios. ORB-SLAM2, lacking semantic integration, consistently struggled to distinguish between static and dynamic elements, leading to significant errors in trajectory estimation and mapping in challenging environments.

In specific use cases, certain systems emerged as better suited for particular tasks. For real-time or resource-constrained applications, RDS-SLAM (SegNet) offered a balanced approach, achieving competitive accuracy, such as an ATE of 0.02 m in the `fr3_walking_static` sequence, while maintaining low processing times (57.5 ms per frame). For highly dynamic scenes, DynaSLAM with Mask R-CNN stood out as the most robust system due to its superior semantic capabilities. These findings underscore the critical role of semantic integration in dynamic SLAM, emphasizing the need for future research to focus on optimizing these methods for real-time applications while maintaining accuracy and robustness.

8.4.5. Challenges in replicating results using custom datasets

The replication of results in semantic SLAM, particularly when applying existing systems to custom datasets, poses significant challenges. While open-source systems provide the foundational tools, their effective deployment often requires addressing several technical and methodological hurdles. These challenges span dependencies and versioning issues, as well as pre-processing and postprocessing requirements.

One of the primary challenges in implementing semantic SLAM systems is the complexity associated with managing software dependencies and version compatibility. These systems typically rely on an integration of robotics frameworks, deep learning libraries, and hardware-specific configurations.

Robotics frameworks, such as DynaSLAM and SG-SLAM, are highly dependent on the Robot Operating System (ROS), which presents compatibility challenges due to its version-specific requirements with different operating systems and hardware platforms. Similarly, deep learning models employed for semantic segmentation, like Mask R-CNN and SegNet, necessitate specific versions of libraries such as TensorFlow or PyTorch. Inconsistencies in library versions can lead to model incompatibility or suboptimal performance. Additionally, hardware dependencies, including GPU compatibility, CUDA version alignment, and VRAM capacity, pose further obstacles, particularly for computationally intensive applications like YOLO-SLAM.

To address these issues, containerization technologies such as Docker are recommended, as they encapsulate all necessary dependencies within isolated environments, ensuring consistent performance across different systems. Moreover, comprehensive documentation detailing software and library version requirements can significantly enhance system reproducibility and compatibility.

Adapting custom datasets for integration with existing semantic SLAM systems often requires extensive preprocessing to meet the specific format and quality standards of these systems. Many SLAM frameworks are optimized for standardized datasets such as TUM or KITTI, necessitating the reformatting of custom datasets to include synchronized RGB-D or stereo image pairs, accurate timestamps, and ground truth pose information to ensure system compatibility. Additionally, real-world datasets commonly contain sensor noise, missing data, and other artifacts that can negatively impact performance. To address these issues, preprocessing techniques such as noise reduction, depth image interpolation, and data smoothing are essential for enhancing data quality and maintaining system accuracy. Furthermore, semantic SLAM systems rely heavily on precise object annotations for effective dynamic object detection and segmentation. Generating high-quality semantic labels for custom datasets is often a time-consuming process that requires a combination of automated annotation tools and manual verification to ensure both accuracy and consistency across the dataset.

After processing a custom dataset with a semantic SLAM system, several challenges emerge in postprocessing and result evaluation. One of the primary difficulties is the calculation of performance metrics, as custom datasets often lack predefined ground truth data necessary for evaluating standard SLAM metrics such as Absolute Trajectory Error (ATE) and Relative Pose Error (RPE). In such cases, it becomes essential to generate high-quality ground truth data using alternative methods, such as motion capture systems or precise manual annotations, to ensure accurate performance assessment. Additionally, maintaining semantic alignment between the system's outputs and the dataset's object categories can be complex, particularly when the custom dataset employs a taxonomy that differs from that used by the pretrained models. This misalignment may require additional mapping or reclassification to ensure meaningful comparisons. Furthermore, effective visualization of the results is critical for comparing performance across different systems. Standardized visualization techniques are necessary to clearly illustrate differences in mapping accuracy, localization performance, and semantic segmentation outcomes, thereby facilitating a comprehensive evaluation of the system's capabilities. By analyzing studies that have made their implementations publicly available, we can assess not only the reproducibility of proposed techniques but also their potential for real-world adoption and further advancements in Semantic SLAM research. Building on these insights, the next section discusses key open challenges and promising directions for future research in the field.

9. Future work directions

Future research in semantic SLAM should prioritize the development of adaptive model architectures capable of dynamic weight adjustment in response to scene variations. As environments transition from static indoor spaces to dynamic outdoor scenarios, semantic segmentation models must intelligently recalibrate their parameters to maintain optimal performance. This adaptability is crucial for handling the diverse conditions encountered in real-world deployments, from lighting changes to varying object densities and movement patterns.

Real-time environmental adaptation represents a critical challenge requiring sophisticated methodologies. Online learning frameworks and domain adaptation techniques offer promising solutions for enabling dynamic recalibration of VSLAM systems. These approaches allow models to continuously update their understanding based on incoming sensory data, ensuring robust performance even in fluctuating and uncertain contexts. Such adaptability is particularly vital for autonomous systems operating in unstructured environments where pre-trained models may encounter previously unseen scenarios.

Temporal coherence of semantic information emerges as another fundamental requirement for robust semantic SLAM. Maintaining consistent semantic labels across consecutive frames not only alleviates positioning inaccuracies caused by discontinuities but also strengthens the reliability of loop closure detection. This semantic consistency over time proves essential for long-term robotic operations, where systems must recognize previously visited locations despite temporal changes. Applications requiring continuous mobility in repetitive tasks particularly benefit from this temporal stability, as it enables more accurate global localization and map consistency.

Computational efficiency remains a primary concern for real-time semantic SLAM deployment. Future research must focus on developing lightweight semantic segmentation models that balance accuracy with resource constraints. Advanced optimization techniques, including network pruning, knowledge distillation, and quantization, show promise for creating models suitable for edge computing platforms. These approaches enable sophisticated semantic understanding on computationally limited autonomous robots without compromising real-time performance requirements.

Dynamic object handling presents unique challenges that demand specialized solutions. Integrating robust object detectors to identify and exclude moving entities from the mapping process significantly improves pose estimation accuracy. Dynamic environments, particularly those with low texture or repetitive patterns, pose substantial challenges as moving objects reduce the availability of reliable static features. Future systems should employ probabilistic approaches to distinguish between static and dynamic elements, potentially incorporating motion prediction models to anticipate and compensate for environmental changes.

The convergence of online learning and domain adaptation algorithms offers a pathway toward truly adaptive semantic SLAM systems. These techniques enable gradual model refinement in response to contextual changes encountered during extended operation periods. Online learning facilitates immediate incorporation of new environmental knowledge, while domain adaptation enables effective knowledge transfer between different operational contexts. This combination ensures that semantic SLAM systems remain accurate and relevant throughout their deployment lifecycle, adapting to seasonal changes, structural modifications, and evolving environmental conditions.

The ultimate objective is achieving seamless integration between semantic understanding and geometric mapping through end-to-end joint optimization. This holistic approach moves beyond treating semantic segmentation as an isolated module, instead fostering deep interconnections between all VSLAM components. Joint optimization frameworks should simultaneously refine semantic predictions, geometric estimates, and data associations, creating systems where each component benefits from and contributes to the others' performance. Such integration promises more robust and reliable VSLAM systems capable of operating effectively in complex, real-world scenarios while maintaining both geometric accuracy and semantic understanding.

10. Conclusion

This comprehensive review has examined the evolution and current state of semantic SLAM, demonstrating how the integration of semantic understanding with traditional geometric mapping has transformed robotic perception and navigation. Through our analysis of various approaches across different sensor modalities, from monocular to multimodal systems, we have highlighted both the significant advances achieved and the challenges that remain.

The evaluation of existing semantic SLAM systems reveals consistent improvements in robustness and accuracy when semantic information is properly integrated. The reproducibility studies conducted on benchmark datasets confirm that semantic-enhanced systems outperform traditional geometric SLAM in dynamic environments, though at the cost of increased computational complexity. These findings underscore the importance of balancing semantic richness with real-time performance requirements.

Several key insights emerge from this survey. First, the choice of sensor modality significantly impacts both the quality of semantic understanding and computational efficiency. While RGB-D and LiDAR systems provide rich geometric information, monocular approaches demonstrate surprising effectiveness when combined with advanced deep learning techniques. Second, the handling of dynamic environments remains a critical differentiator among approaches, with recent methods showing promising results through probabilistic modeling and temporal consistency constraints. Third, the gap between laboratory demonstrations and real-world deployment persists, particularly in terms of long-term reliability and computational constraints.

Several fundamental challenges must be addressed to realize the full potential of semantic SLAM. Dynamic model adaptation stands as a primary research direction, requiring systems that can adjust their parameters in response to environmental changes without manual intervention. Temporal coherence of semantic information presents another crucial area, as maintaining consistent semantic understanding over extended periods is essential for reliable long-term operation. Additionally, the development of lightweight yet accurate semantic segmentation models remains vital for deploying these systems on resource-constrained platforms.

The integration of online learning and domain adaptation techniques offers promising avenues for creating truly adaptive semantic SLAM systems. These approaches would enable continuous improvement and adaptation to new environments, moving beyond the current paradigm of fixed, pre-trained models. Furthermore, end-to-end joint optimization of semantic and geometric components represents the ultimate goal, promising systems where perception and mapping are seamlessly integrated rather than loosely coupled.

In conclusion, semantic SLAM has moved from research laboratories to practical implementation in autonomous systems. While significant challenges remain in computational efficiency, dynamic scene handling, and long-term reliability, the rapid progress documented in this survey suggests that robust, real-world semantic SLAM systems are within reach. As the field continues to mature, we anticipate that the convergence of advanced machine learning, efficient computing architectures, and novel algorithmic approaches will enable the next generation of intelligent robotic systems capable of truly understanding and navigating complex, dynamic environments. The future of autonomous navigation lies not just in knowing where things are, but in understanding what they are and how they relate to the robot's objectives.

CRediT authorship contribution statement

Houssein Kanso: Writing – original draft, Visualization. **Abhilasha Singh:** Writing – original draft, Investigation, Formal analysis. **Etaf El Zarif:** Software, Validation, Writing – review & editing, Implementation. **Nooruldeen Almohammed:** Writing – original draft, Methodology, Investigation. **Jinane Mounsef:** Conceptualization, Writing –

review & editing, Supervision, Project administration. **Noel Maalouf:** Writing – reviewing & editing, Conceptualization. **Bilal Arain:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Abate, M., Chang, Y., Hughes, N., & Carlone, L. (2024). Kimera2: Robust and accurate metric-semantic SLAM in the real world. arXiv:2401.06323.

Abate, M., Schwartz, A., Wong, X. I., Luo, W., Littman, R., Klinger, M., et al. (2023). Multi-camera visual-inertial simultaneous localization and mapping for autonomous valet parking. In *International symposium on experimental robotics* (pp. 567–581). Springer.

Abdelnasser, H., Mohamed, R., Elgohary, A., Alzantot, M. F., Wang, H., Sen, S., et al. (2016). SemanticSLAM: Using environment landmarks for unsupervised indoor localization. *IEEE Transactions on Mobile Computing*, *15*(7), 1770–1782.

Ahmed, M. F., Masood, K., Fremont, V., & Fantoni, I. (2023). Active slam: A review on last decade. *Sensors*, *23*(19), 8097.

Ai, Y., Sun, Q., Xi, Z., Li, N., Dong, J., & Wang, X. (2023a). Stereo SLAM in dynamic environments using semantic segmentation. *Electronics*, *12*(14), http://dx.doi.org/10.3390/electronics12143112, URL https://www.mdpi.com/2079-9292/12/14/3112.

Ai, Y., Sun, Q., Xi, Z., Li, N., Dong, J., & Wang, X. (2023b). Stereo SLAM in dynamic environments using semantic segmentation. *Electronics*, *12*(14), 3112. http://dx.doi.org/10.3390/electronics12143112, URL https://www.mdpi.com/2079-9292/12/14/3112.

An, L., Pan, X., Li, T., & Wang, M. (2022). A visual dynamic-SLAM method based semantic segmentation and multi-view geometry. In Y. Wang, & S. Chen (Eds.), *vol. 12162*, *International conference on high performance computing and communication*. SPIE, International Society for Optics and Photonics, Article 1216214. http://dx.doi.org/10.1117/12.2628175.

Antonini, A., Guerra, W., Murali, V., Sayre-McCord, T., & Karaman, S. (2020). The blackbird uav dataset. *The International Journal of Robotics Research*, *39*(10–11), 1346–1364.

Arshad, S., & Kim, G.-W. (2024). SLGD-loop: A semantic local and global descriptor-based loop closure detection for long-term autonomy. *IEEE Transactions on Intelligent Transportation Systems*.

Arth, C., Pirchheim, C., Ventura, J., Schmalstieg, D., & Lepetit, V. (2015). Instant outdoor localization and slam initialization from 2.5 d maps. *IEEE Transactions on Visualization and Computer Graphics*, *21*(11), 1309–1318.

Asgharivaskasi, A., & Atanasov, N. (2023). Semantic octree mapping and shannon mutual information computation for robot exploration. *IEEE Transactions on Robotics*, *39*(3), 1910–1928.

Atanasov, N., Zhu, M., Daniilidis, K., & Pappas, G. J. (2016). Localization from semantic observations via the matrix permanent. *The International Journal of Robotics Research*, *35*(1–3), 73–99.

Azzam, R., Alkendi, Y., Taha, T., Huang, S., & Zweiri, Y. (2021). A stacked LSTM-based approach for reducing semantic pose estimation error. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–14.

Azzam, R., Taha, T., Huang, S., & Zweiri, Y. (2020). Feature-based visual simultaneous localization and mapping: A survey. *SN Applied Sciences*, *2*, 1–24.

Bajpai, A., Burroughes, G., Shaukat, A., & Gao, Y. (2016). Planetary monocular simultaneous localization and mapping. *Journal of Field Robotics*, *33*(2), 229–242.

Bavle, H., De La Puente, P., How, J. P., & Campoy, P. (2020). VPS-SLAM: Visual planar semantic SLAM for aerial robotic systems. *IEEE Access*, *8*, 60704–60718.

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., et al. (2019). SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *2019 IEEE/CVF international conference on computer vision*. ICCV, IEEE.

Bescos, B., Cadena, C., & Neira, J. (2021). Empty cities: A dynamic-object-invariant space for visual SLAM. *IEEE Transactions on Robotics*, *37*(2), 433–451.

Bescos, B., Campos, C., Tardós, J. D., & Neira, J. (2021). DynaSLAM II: Tightly-coupled multi-object tracking and SLAM. *IEEE Robotics and Automation Letters*, *6*(3), 5191–5198. http://dx.doi.org/10.1109/LRA.2021.3068640.

Bescos, B., Facil, J. M., Civera, J., & Neira, J. (2018). DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, *3*(4), 4076–4083.

Beul, M., Droeschel, D., Nieuwenhuisen, M., Quenzel, J., Houben, S., & Behnke, S. (2018). Fast autonomous flight in warehouses for inventory applications. *IEEE Robotics and Automation Letters*, *3*(4), 3121–3128.

Bloesch, M., Omari, S., Hutter, M., & Siegwart, R. (2015). Robust visual inertial odometry using a direct EKF-based approach. In *2015 IEEE/RSJ international conference on intelligent robots and systems*. IROS, IEEE.

Bowman, S. L., Atanasov, N., Daniilidis, K., & Pappas, G. J. (2017). Probabilistic data association for semantic SLAM. In *2017 IEEE international conference on robotics and automation*. ICRA, IEEE.

Bresson, G., Alsayed, Z., Yu, L., & Glaser, S. (2017). Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, *2*(3), 194–220.

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., et al. (2016). Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, *32*(6), 1309–1332.

Cai, Y., Ou, Y., & Qin, T. (2024). Improving SLAM techniques with integrated multi-sensor fusion for 3D reconstruction. *Sensors*, *24*(7), http://dx.doi.org/10.3390/s24072033, URL https://www.mdpi.com/1424-8220/24/7/2033.

Campos, C., Elvira, R., Rodr'iguez, J., Montiel, J. M. M., & Tardós, J. D. (2020). ORB-SLAM3: An accurate open-source library for visual, Visual2013Inertial, and multimap SLAM. *IEEE Transactions on Robotics*, *37*, 1874–1890. http://dx.doi.org/10.1109/TRO.2021.3075644.

Cao, A.-Q., & de Charette, R. (2021). MonoScene: Monocular 3D semantic scene completion. In *2022 IEEE/CVF conference on computer vision and pattern recognition* (pp. 3981–3991). http://dx.doi.org/10.1109/CVPR52688.2022.00396.

Cao, L., Liu, J., Lei, J., Zhang, W., Chen, Y., & Hyyppä, J. (2025). Real-time motion state estimation of feature points based on optical flow field for robust monocular visual-inertial odometry in dynamic scenes. *Expert Systems with Applications*, Article 126813.

Cao, H., Xu, J., Li, D., Shangguan, L., Liu, Y., & Yang, Z. (2022). Edge assisted mobile semantic visual SLAM. *IEEE Transactions on Mobile Computing*, *22*(12), 6985–6999.

Chai, W., Li, C., & Li, Q. (2023). Multi-sensor fusion-based indoor single-track semantic map construction and localization. *IEEE Sensors Journal*, *23*(3), 2470–2480.

Chen, S., Jian, Z., Huang, Y., Chen, Y., Zhou, Z., & Zheng, N. (2019). Autonomous driving: cognitive construction and situation understanding. *Science China. Information Sciences*, *62*, 1–27.

Chen, S., Li, C., Jiang, Q., Zhuang, X., Zhang, B., Zhou, B., et al. (2025). TextGeo-SLAM: A LiDAR SLAM with text semantics and geometric constraints-based loop closure. *IEEE Internet of Things Journal*.

Chen, L., Ling, Z., Gao, Y., Sun, R., & Jin, S. (2023). A real-time semantic visual SLAM for dynamic environment based on deep learning and dynamic probabilistic propagation. *Complex & Intelligent Systems*, *9*(5), 5653–5677.

Chen, K., Liu, J., Chen, Q., Wang, Z., & Zhang, J. (2022). Accurate object association and pose updating for semantic SLAM. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 25169–25179.

Chen, Y., Pan, D., Pan, Y., Liu, S., Gu, A., & Wang, M. (2015). Indoor scene understanding via monocular RGB-D images. *Information Sciences*, *320*, 361–371.

Chen, S., Shao, D., Zhang, L., & Zhang, C. (2022). Learning depth-aware features for indoor scene understanding. *Multimedia Tools and Applications*, *81*(29), 42573–42590.

Chen, N., Wei, D., Lin, D., & Lin, L. (2025). Semantic SLAM using laser-vision data fusion: Enhancing autonomous navigation in unstructured environments. *Alexandria Engineering Journal*, *127*, 606–618.

Chen, K., Xiao, J., Liu, J., Tong, Q., Zhang, H., Liu, R., et al. (2025). Semantic visual simultaneous localization and mapping: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *26*(6), 7426–7449. http://dx.doi.org/10.1109/TITS.2025.3556928.

Chen, X., Xue, J., Fang, J., Pan, Y., & Zheng, N. (2020). Using detection, tracking and prediction in visual SLAM to achieve real-time semantic mapping of dynamic scenarios. In *2020 IEEE intelligent vehicles symposium* (pp. 666–671). IEEE.

Chen, Y., Zhang, L., Zhao, S., & Zhou, Y. (2025). Online indoor visual odometry with semantic assistance under implicit epipolar constraints. *Pattern Recognition*, *159*, Article 111150.

Chen, B., Zhuang, Y., & Wang, S. (2024). A real-time and lightweight monocular 3-D object detector on CPU-based edge devices for UGV's indoor SLAM systems. *IEEE/ASME Transactions on Mechatronics*.

Cheng, S., Sun, C., Zhang, S., & Zhang, D. (2023). SG-SLAM: A real-time RGB-D visual SLAM toward dynamic scenes with semantic and geometric information. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–12. http://dx.doi.org/10.1109/TIM.2022.3228006, URL https://ieeexplore.ieee.org/document/9978699/.

Cheng, J., Wang, C., Mai, X., Min, Z., & Meng, M. Q.-H. (2021). Improving dense mapping for mobile robots in dynamic environments based on semantic information. *IEEE Sensors Journal*, *21*(10), 11740–11747.

Chghaf, M., Rodriguez, S., & Ouardi, A. E. (2022). Camera, LiDAR and multi-modal SLAM systems for autonomous ground vehicles: A survey. *Journal of Intelligent and Robotic Systems*, *105*(1).

Cho, S., Kim, C., Park, J., Sunwoo, M., & Jo, K. (2020). Semantic point cloud mapping of LiDAR based on probabilistic uncertainty modeling for autonomous driving. *Sensors*, *20*(20), 5900.

Choe, S., Seong, H., & Kim, E. (2022). Indoor place category recognition for a cleaning robot by fusing a probabilistic approach and deep learning. *IEEE Transactions on Cybernetics*, *52*(8), 7265–7276.

Choi, W., Chao, Y.-W., Pantofaru, C., & Savarese, S. (2015). Indoor scene understanding with geometric and semantic contexts. *International Journal of Computer Vision*, *112*, 204–220.

Choi, S., Zhou, Q.-Y., Miller, S., & Koltun, V. (2016). A large dataset of object scans. arXiv:1602.02481.

Choudhary, S., Carlone, L., Nieto, C., Rogers, J., Christensen, H. I., & Dellaert, F. (2017). Distributed mapping with privacy and communication constraints: Lightweight algorithms and object-based models. *The International Journal of Robotics Research*, *36*(12), 1286–1311.

Civera, J., Galvez-Lopez, D., Riazuelo, L., Tardos, J. D., & Montiel, J. M. M. (2011). Towards semantic SLAM using a monocular camera. In *2011 IEEE/RSJ international conference on intelligent robots and systems*. IEEE.

Cornejo-Lupa, M. A., Ticona-Herrera, R. P., Cardinale, Y., & Barrios-Aranibar, D. (2020). A survey of ontologies for simultaneous localization and mapping in mobile robots. *ACM Computing Surveys*, *53*(5), 1–26.

Dang, Y., Chen, P., Liang, R., Huang, C., Tang, Y., Yu, T., et al. (2019). Real-time semantic plane reconstruction on a monocular drone using sparse fusion. *IEEE Transactions on Vehicular Technology*, *68*(8), 7383–7391.

Davison, A. J. (2003). Real-time simultaneous localisation and mapping with a single camera. In *Proceedings of the ninth IEEE international conference on computer vision* (pp. 1403–1410). IEEE.

Deng, M., Hu, J., Wen, J., Zhang, X., & Jin, Q. (2025). Object detection based visual SLAM optimization method for dynamic scene. *IEEE Sensors Journal*.

Deng, W., Huang, K., Chen, X., Zhou, Z., Shi, C., Guo, R., et al. (2020). Semantic RGB-D SLAM for rescue robot navigation. *IEEE Access*, *8*, 221320–221329. http://dx.doi.org/10.1109/ACCESS.2020.3031867.

Deng, L., Yang, M., Hu, B., Li, T., Li, H., & Wang, C. (2019). Semantic segmentation-based lane-level localization using around view monitoring system. *IEEE Sensors Journal*, *19*(21), 10077–10086.

Digital Science (2024). Dimensions.ai – Research insights platform. URL https://www.dimensions.ai (Accessed 12 December 2024).

Dissanayake, G., Newman, P., Clark, S., Durrant-Whyte, H., & Csorba, M. (2001). A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, *17*(3), 229–241.

Dong, N., Qin, M., Chang, J., Wu, C.-H., Ip, W.-H., & Yung, K.-L. (2022). Weighted triplet loss based on deep neural networks for loop closure detection in VSLAM. *Computer Communications*, *186*, 153–165.

Duan, R., Feng, Y., & Wen, C.-Y. (2022). Deep pose graph-matching-based loop closure detection for semantic visual SLAM. *Sustainability*, *14*(19), 11864.

Dube, R., Cramariuc, A., Dugas, D., Sommer, H., Dymczyk, M., Nieto, J., et al. (2020). SegMap: Segment-based mapping and localization using data-driven descriptors. *The International Journal of Robotics Research*, *39*(2–3), 339–355.

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping: part I. *IEEE Robotics & Automation Magazine*, *13*(2), 99–110.

Engel, J., Koltun, V., & Cremers, D. (2017). Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(3), 611–625.

Engel, J., Schöps, T., & Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular SLAM. In *European conference on computer vision* (pp. 834–849). Springer.

Eslamian, A., & Ahmadzadeh, M. R. (2022). Det-SLAM: A semantic visual SLAM for highly dynamic scenes using Detectron2. In *2022 8th Iranian conference on signal processing and intelligent systems* ICSPIS, (pp. 1–5). Behshahr, Iran, Islamic Republic of: IEEE, http://dx.doi.org/10.1109/ICSPIS56952.2022.10043931, URL https://ieeexplore.ieee.org/document/10043931/.

Esparza, D., & Flores, G. (2021). The STDyn-SLAM: A stereo vision and semantic segmentation approach for SLAM in dynamic outdoor environments. arXiv:2010.09857, URL https://arxiv.org/abs/2010.09857.

Esparza, D., & Flores, G. (2022). The STDyn-SLAM: A stereo vision and semantic segmentation approach for VSLAM in dynamic outdoor environments. *IEEE Access*, *10*, 18201–18209. http://dx.doi.org/10.1109/ACCESS.2022.3149885, URL https://ieeexplore.ieee.org/document/9706470/.

Fan, Y., Zhang, Q., Liu, S., Tang, Y., Jing, X., Yao, J., et al. (2020). Semantic SLAM with more accurate point cloud map in dynamic environments. *IEEE Access*, *8*, 112237–112252. http://dx.doi.org/10.1109/ACCESS.2020.3003160, URL https://ieeexplore.ieee.org/document/9119407/.

Fan, Y., Zhang, Q., Tang, Y., Liu, S., & Han, H. (2022). Blitz-SLAM: A semantic SLAM in dynamic environments. *Pattern Recognition*, *121*, Article 108225.

Fang, B., Mei, G., Yuan, X., Wang, L., Wang, Z., & Wang, J. (2021). Visual SLAM for robot navigation in healthcare facility. *Pattern Recognition*, *113*, Article 107822.

Fernandez-Cortizas, M., Bavle, H., Perez-Saura, D., Sanchez-Lopez, J. L., Campoy, P., & Voos, H. (2024). Multi S-graphs: An efficient distributed semantic-relational collaborative SLAM. *IEEE Robotics and Automation Letters*, *9*(6), 6004–6011. http://dx.doi.org/10.1109/LRA.2024.3399997.

Fraifer, M. A., Coleman, J., Maguire, J., Trslić, P., Dooly, G., & Toal, D. (2025). Autonomous forklifts: State of the art—Exploring perception, scanning technologies and functional systems—A comprehensive review. *Electronics*, *14*(1), http://dx.doi.org/10.3390/electronics14010153, URL https://www.mdpi.com/2079-9292/14/1/153.

Fu, M., Lu, X., Jin, Y., Zhang, W.-A., Prakapovich, R., & Sychou, U. (2023). Semantic map-based visual localization with consistency guarantee. *IEEE Sensors Journal*, *24*(1), 1065–1078.

Galagain, C., Poreba, M., & Goulette, F. (2025). Is semantic SLAM ready for embedded systems ? A comparative survey. arXiv:2505.12384, URL https://arxiv.org/abs/2505.12384.

Gao, Y., Hu, M., Chen, B., Yang, W., Wang, J., & Wang, J. (2024). Multi-mask fusion-based RGB-D SLAM in dynamic environments. *IEEE Sensors Journal*.

Ge, F., Zhang, Y., Wang, L., Coleman, S., & Kerr, D. (2023). Double-domain adaptation semantics for retrieval-based long-term visual localization. *IEEE Transactions on Multimedia*, *26*, 6050–6064.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Conference on computer vision and pattern recognition*.

Gong, Z., Li, J., Luo, Z., Wen, C., Wang, C., & Zelek, J. (2021). Mapping and semantic modeling of underground parking lots using a backpack LiDAR system. *IEEE Transactions on Intelligent Transportation Systems*, *22*(2), 734–746.

Gonzalez, M., Marchand, E., Kacete, A., & Royan, J. (2022). TwistSLAM: Constrained SLAM in dynamic environment. *IEEE Robotics and Automation Letters*, *7*(3), 6846–6853. http://dx.doi.org/10.1109/LRA.2022.3178150.

Gou, R., Chen, G., Yan, C., Pu, X., Wu, Y., & Tang, Y. (2022). Three-dimensional dynamic uncertainty semantic SLAM method for a production workshop. *Engineering Applications of Artificial Intelligence*, *116*, Article 105325.

Grinvald, M., Furrer, F., Novkovic, T., Chung, J. J., Cadena, C., Siegwart, R., et al. (2019). Volumetric instance-aware semantic mapping and 3D object discovery. *IEEE Robotics and Automation Letters*, *4*(3), 3037–3044.

Guan, P., Cao, Z., Chen, E., Liang, S., Tan, M., & Yu, J. (2020). A real-time semantic visual SLAM approach with points and objects. *International Journal of Advanced Robotic Systems*, *17*(1), Article 1729881420905443.

Guerrero-Font, E., Bonin-Font, F., Martin-Abadal, M., Gonzalez-Cid, Y., & Oliver-Codina, G. (2021). Sparse Gaussian process for online seagrass semantic mapping. *Expert Systems with Applications*, *170*, Article 114478.

Guo, L., & Fan, G. (2022). Holistic indoor scene understanding by context-supported instance segmentation. *Multimedia Tools and Applications*, *81*(25), 35751–35773.

Gupta, S., Arbeláez, P., Girshick, R., & Malik, J. (2015). Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, *112*, 133–149.

Habibpour, M., Nemati, A., Meghdari, A., Taheri, A., & Nazari, S. (2024). RSV-SLAM: Toward real-time semantic visual SLAM in indoor dynamic environments. In *Lecture notes in networks and systems*, *Lecture notes in networks and systems* (pp. 832–844). Cham: Springer Nature Switzerland.

Han, X., Li, S., Wang, X., & Zhou, W. (2021a). Semantic mapping for mobile robots in indoor scenes: A survey. *Information*, *12*(2), 92. http://dx.doi.org/10.3390/info12020092, URL https://www.mdpi.com/2078-2489/12/2/92.

Han, X., Li, S., Wang, X., & Zhou, W. (2021b). Semantic mapping for mobile robots in indoor scenes: A survey. *Information*, *12*(2), http://dx.doi.org/10.3390/info12020092, URL https://www.mdpi.com/2078-2489/12/2/92.

Han, B., Wei, J., Zhang, J., Meng, Y., Dong, Z., & Liu, H. (2023). GardenMap: Static point cloud mapping for Garden environment. *Computers and Electronics in Agriculture*, *204*, Article 107548.

Han, S., & Xi, Z. (2020). Dynamic scene semantics SLAM based on semantic segmentation. *IEEE Access*, *8*, 43563–43570.

Han, X., & Yang, L. (2023). Sq-slam: Monocular semantic slam based on superquadric object representation. *Journal of Intelligent and Robotic Systems*, *109*(2), 29.

Hardegger, M., Roggen, D., Calatroni, A., & Tröster, G. (2016). S-SMART: A unified bayesian framework for simultaneous semantic mapping, activity recognition, and tracking. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *7*(3), 1–28.

He, X., Ding, L., & Lan, Y. (2024). DSK-SLAM: A dynamic SLAM system combining semantic information and a novel geometric method based on K-means clustering. *IEEE Sensors Journal*.

He, J., Li, M., Wang, Y., & Wang, H. (2023). OVD-SLAM: An online visual SLAM for dynamic environments. *IEEE Sensors Journal*, *23*(12), 13210–13219.

He, G., Zhang, Q., & Zhuang, Y. (2022). Online semantic-assisted topological map building with LiDAR in large-scale outdoor environments: Toward robust place recognition. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–12.

Hempel, T., & Al-Hamadi, A. (2022). An online semantic mapping system for extending and enhancing visual SLAM. *Engineering Applications of Artificial Intelligence*, *111*, Article 104830.

Herb, M., Weiherer, T., Navab, N., & Tombari, F. (2021). Lightweight semantic mesh mapping for autonomous vehicles. In *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6732–6738). IEEE, http://dx.doi.org/10.1109/ICRA48506.2021.9560996.

Hu, D., Gan, V. J., & Yin, C. (2023). Robot-assisted mobile scanning for automated 3D reconstruction and point cloud semantic segmentation of building interiors. *Automation in Construction*, *152*, Article 104949.

Hu, Z., Qi, W., Ding, K., Qi, H., Zhao, Y., Zhang, X., et al. (2025). Optimized feature points and keyframe methods for VSLAM in high-dynamic indoor environments. *IEEE Transactions on Intelligent Transportation Systems*.

Hu, X., Wu, Y., Zhao, M., Yang, L., Zhang, X., & Ji, X. (2025). PAS-SLAM: A visual SLAM system for planar-ambiguous scenes. *IEEE Transactions on Circuits and Systems for Video Technology*.

Hu, L., Zhang, Y., Wang, Y., Jiang, Q., Ge, G., & Wang, W. (2022). A simple information fusion method provides the obstacle with saliency labeling as a landmark in robotic mapping. *Alexandria Engineering Journal*, *61*(12), 12061–12074.

Huang, X., Chen, X., Zhang, N., He, H., & Feng, S. (2024). ADM-SLAM: Accurate and fast dynamic visual SLAM with adaptive feature point extraction, deeplabv3pro, and multi-view geometry. *Sensors*, *24*(11), http://dx.doi.org/10.3390/s24113578, URL https://www.mdpi.com/1424-8220/24/11/3578.

Huang, L., Wang, Z., Yun, J., Jiang, D., & Gong, C. (2025). Dynamic feature rejection based on geometric constraint for visual SLAM in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*.

Huang, L., Zhu, Z., Yun, J., Xu, M., Liu, Y., Sun, Y., et al. (2023). Semantic loopback detection method based on instance segmentation and visual SLAM in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, *25*(3), 3118–3127.

Hughes, N., Chang, Y., Hu, S., Talak, R., Abdulhai, R., Strader, J., et al. (2024). Foundations of spatial perception for robotics: Hierarchical representations and real-time systems.

Isele, S. T., Haas-Fickinger, F., & Zöllner, J. M. (2021). SERALOC: SLAM on semantically annotated radar point-clouds. In *2021 IEEE international intelligent transportation systems conference* (pp. 2917–2924). IEEE.

Islam, Q. U., Ibrahim, H., Chin, P. K., Lim, K., & Abdullah, M. Z. (2024). MVS-SLAM: Enhanced multiview geometry for improved semantic RGBD SLAM in dynamic environment. *Journal of Field Robotics*, *41*(1), 109–130. http://dx.doi.org/10.1002/rob.22248, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.22248, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.22248.

Ji, X., Liu, P., Niu, H., Chen, X., Ying, R., & Wen, F. (2023). Object SLAM based on spatial layout and semantic consistency. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–12.

Ji, T., Wang, C., & Xie, L. (2021). Towards real-time semantic RGB-D SLAM in dynamic environments. In *2021 IEEE international conference on robotics and automation*. IEEE.

Jiang, Z., Guo, Y., Jiang, K., Hu, M., & Zhu, Z. (2021). Optimization of intelligent plant cultivation robot system in object detection. *IEEE Sensors Journal*, *21*(17), 19279–19288.

Jiang, H., Xu, Y., Li, K., Feng, J., & Zhang, L. (2024). RoDyn-SLAM: Robust dynamic dense RGB-D SLAM with neural radiance fields. *IEEE Robotics and Automation Letters*.

Jiao, J., Wang, C., Li, N., Deng, Z., & Xu, W. (2022). An adaptive visual dynamic-SLAM method based on fusing the semantic information. *IEEE Sensors Journal*, *22*(18), 17414–17420.

Jin, S., Chen, L., Sun, R., & McLoone, S. (2020). A novel vSLAM framework with unsupervised semantic segmentation based on adversarial transfer learning. *Applied Soft Computing*, *90*, Article 106153.

Judd, K. M., & Gammell, J. D. (2024). Multimotion visual odometry. *The International Journal of Robotics Research*, *43*(8), 1250–1278. http://dx.doi.org/10.1177/02783649241229095.

Jung, J., Kim, S., Seo, B., Jang, W., Lee, S., Shin, J., et al. (2025). An energy-efficient processor for real-time semantic LiDAR SLAM in mobile robots. *IEEE Journal of Solid-State Circuits*.

Karpyshev, P., Kruzhkov, E., Yudin, E., Savinykh, A., Potapov, A., Kurenkov, M., et al. (2022). Mucaslam: Cnn-based frame quality assessment for mobile robot with omnidirectional visual slam. In *2022 IEEE 18th international conference on automation science and engineering* (pp. 368–373). IEEE.

Kim, U.-H., Kim, S.-H., & Kim, J.-H. (2022). Simvodis: Simultaneous visual odometry, object detection, and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(1), 428–441.

Klein, G., & Murray, D. (2007). Parallel tracking and mapping for small AR workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality* (pp. 225–234). IEEE.

Kong, X., Liu, S., Taher, M., & Davison, A. J. (2023). Vmap: Vectorised object mapping for neural field slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 952–961).

Kostavelis, I., & Gasteratos, A. (2015). Semantic mapping for mobile robotics tasks: A survey. *Robotics and Autonomous Systems*, *66*, 86–103.

Kostavelis, I., & Gasteratos, A. (2017). Semantic maps from multiple visual cues. *Expert Systems with Applications*, *68*, 45–57.

Kuang, B., Yuan, J., & Liu, Q. (2022). A robust RGB-D SLAM based on multiple geometric features and semantic segmentation in dynamic environments. *Measurement Science and Technology*, *34*(1), Article 015402. http://dx.doi.org/10.1088/1361-6501/ac92a0.

Lai, K., Bo, L., & Fox, D. (2014). Unsupervised feature learning for 3D scene labeling. In *2014 IEEE international conference on robotics and automation* (pp. 3050–3057). http://dx.doi.org/10.1109/ICRA.2014.6907298.

Laina, S. B., Boche, S., Papatheodorou, S., Schaefer, S., Jung, J., & Leutenegger, S. (2025). FindAnything: Open-vocabulary and object-centric mapping for robot exploration in any environment. arXiv preprint arXiv:2504.08603.

Lee, J., Back, M., Hwang, S. S., & Chun, I. Y. (2023a). Improved real-time monocular SLAM using semantic segmentation on selective frames. *IEEE Transactions on Intelligent Transportation Systems*, *24*(3), 2800–2813. http://dx.doi.org/10.1109/tits.2022.3228525.

Lee, J., Back, M., Hwang, S. S., & Chun, I. Y. (2023b). Improved real-time monocular SLAM using semantic segmentation on selective frames. *IEEE Transactions on Intelligent Transportation Systems*, *24*(3), 2800–2813.

Li, G., Fan, H., Jiang, G., Jiang, D., Liu, Y., Tao, B., et al. (2023). RGBD-SLAM based on object detection with two-stream YOLOv4-MobileNetv3 in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, *25*(3), 2847–2857.

Li, F., Fu, C., Sun, D., Li, J., & Wang, J. (2024). SD-SLAM: A semantic SLAM approach for dynamic scenes based on LiDAR point clouds. *Big Data Research*, *36*, Article 100463.

Li, F., Fu, C., Wang, J., & Sun, D. (2025). Dynamic semantic SLAM based on panoramic camera and LiDAR fusion for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*.

Li, W., Gu, J., Dong, Y., Dong, Y., & Han, J. (2020). Indoor scene understanding via RGB-D image segmentation employing depth-based CNN and CRFs. *Multimedia Tools and Applications*, *79*, 35475–35489.

Li, R., Gu, D., Liu, Q., Long, Z., & Hu, H. (2018). Semantic scene mapping with spatio-temporal deep neural network for robotic applications. *Cognitive Computation*, *10*, 260–271.

Li, M., Guo, Z., Deng, T., Zhou, Y., Ren, Y., & Wang, H. (2025). DDN-SLAM: Real time dense dynamic neural implicit SLAM. *IEEE Robotics and Automation Letters*.

Li, B., Hao, V. C., Stuckey, P. J., Reid, I., & Rezatofighi, H. (2025). Hier-slam++: Neuro-symbolic semantic slam with a hierarchically categorical gaussian splatting. arXiv preprint arXiv:2502.14931.

Li, J., Hu, S., Li, Q., Chen, J., Leung, V. C., & Song, H. (2021). Global visual and semantic observations for outdoor robot localization. *IEEE Transactions on Network Science and Engineering*, *8*(4), 2909–2921.

Li, Y., Jiang, L., Lei, B., Tang, B., & Zhu, J. (2025). Robust real-time localization system via semantic dimensional chains for degraded scenarios. *IEEE Internet of Things Journal*.

Li, L., Kong, X., Zhao, X., Huang, T., & Liu, Y. (2022). Semantic scan context: a novel semantic-based loop-closure method for LiDAR SLAM. *Autonomous Robots*, *46*(4), 535–551.

Li, L., Kong, X., Zhao, X., Li, W., Wen, F., Zhang, H., et al. (2021). SA-LOAM: Semantic-aided LiDAR SLAM with loop closure. In *2021 IEEE international conference on robotics and automation* (pp. 7627–7634). IEEE.

Li, M., Liu, S., Zhou, H., Zhu, G., Cheng, N., Deng, T., et al. (2024). Sgs-slam: Semantic gaussian splatting for neural dense slam. In *European conference on computer vision* (pp. 163–179). Springer.

Li, Y., Song, G., Hao, S., Mao, J., & Song, A. (2023). Semantic stereo visual SLAM toward outdoor dynamic environments based on ORB-SLAM2. *Industrial Robot: The International Journal of Robotics Research and Application*, *50*(3), 542–554. http://dx.doi.org/10.1108/IR-09-2022-0236, URL https://www.emerald.com/insight/content/doi/10.1108/IR-09-2022-0236/full/html.

Li, R., Wang, S., & Gu, D. (2018). Ongoing evolution of visual SLAM from geometry to deep learning: Challenges and opportunities. *Cognitive Computation*, *10*(6), 875–889.

Li, A., Wang, J., Xu, M., & Chen, Z. (2021). DP-SLAM: A visual SLAM with moving probability towards dynamic environments. *Information Sciences*, *556*, 128–142.

Li, T., Ye, L., Zhu, X., Chuai, S., Wu, J., Zhang, W., et al. (2025). Research on VSLAM algorithm based on improved YOLO algorithm and multi-view geometry. *Journal of Field Robotics*.

Li, J., Zhang, X., Li, J., Liu, Y., & Wang, J. (2020). Building and optimization of 3D semantic map based on lidar and camera fusion. *Neurocomputing*, *409*, 394–407.

Li, C., Zhou, B., & Li, Q. (2024). Semanticcslam: Using environment landmarks for cooperative simultaneous localization and mapping. *IEEE Internet of Things Journal*.

Li, B., Zou, D., Huang, Y., Niu, X., Pei, L., & Yu, W. (2023). Textslam: Visual slam with semantic planar text features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(1), 593–610.

Liao, Y., Xie, J., & Geiger, A. (2023). Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3292–3310.

Lin, X., Su, Z., Zhu, Z., Yuan, P., Zhu, H., & Zhou, X. (2025). Joint semantic-geometric mapping of unstructured environment for autonomous mobile robotic sprayers. *Journal of Field Robotics*.

Lin, S., Wang, J., Xu, M., Zhao, H., & Chen, Z. (2023). Contour-SLAM: A robust object-level SLAM based on contour alignment. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–12.

Lin, Z., Zhang, Q., Tian, Z., Yu, P., & Lan, J. (2024). DPL-SLAM: Enhancing dynamic point-line SLAM through dense semantic methods. *IEEE Sensors Journal*, *24*(9), 14596–14607. http://dx.doi.org/10.1109/JSEN.2024.3373892, URL https://ieeexplore.ieee.org/document/10477312/.

Lin, Z., Zhang, Q., Tian, Z., Yu, P., Ye, Z., Zhuang, H., et al. (2025). Slam2: Simultaneous localization and multimode mapping for indoor dynamic environments. *Pattern Recognition*, *158*, Article 111054.

Liu, X., Lei, J., Prabhu, A., Tao, Y., Spasojevic, I., Chaudhari, P., et al. (2024). SlideSLAM: Sparse, lightweight, decentralized metric-semantic SLAM for multi-robot navigation. arXiv preprint arXiv:2406.17249.

Liu, R., Mi, L., & Chen, Z. (2021). AFNet: Adaptive fusion network for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(9), 7871–7886.

Liu, Y., & Miura, J. (2021a). RDMO-SLAM: Real-time visual SLAM for dynamic environments using semantic label prediction with optical flow. *Ieee Access*, *9*, 106981–106997.

Liu, Y., & Miura, J. (2021b). RDS-SLAM: Real-time dynamic SLAM using semantic segmentation methods. *IEEE Access*, *9*, 23772–23785. http://dx.doi.org/10.1109/ACCESS.2021.3050617, URL https://ieeexplore.ieee.org/document/9318990/.

Liu, W., Sun, W., & Liu, Y. (2021). Dloam: Real-time and robust lidar slam system based on cnn in dynamic urban environments. *IEEE Open Journal of Intelligent Transportation Systems*.

Liu, X., Wen, S., Jiang, Z., Tian, W., Qiu, T. Z., & Othman, K. M. (2023). A multisensor fusion with automatic vision–LiDAR calibration based on factor graph joint optimization for SLAM. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–9.

Liu, D., Wu, J., Du, Y., Zhang, R., & Cong, M. (2024). SBC-SLAM: Semantic bio-inspired collaborative SLAM for large-scale environment perception of heterogeneous systems. *IEEE Transactions on Instrumentation and Measurement*.

Liu, Q., Yuan, J., & Kuang, B. (2024). Sia-slam: a robust visual slam associated with semantic information in dynamic environments. *Multimedia Tools and Applications*, *83*(18), 53531–53547.

Lou, L., Li, Y., Zhang, Q., & Wei, H. (2023). SLAM and 3D semantic reconstruction based on the fusion of lidar and monocular vision. *Sensors*, *23*(3), http://dx.doi.org/10.3390/s23031502, URL https://www.mdpi.com/1424-8220/23/3/1502.

Luo, Y., Rao, Z., & Wu, R. (2023). FD-SLAM: A semantic SLAM based on enhanced fast-SCNN dynamic region detection and DeepFillv2-driven background inpainting. *IEEE Access*, *11*, 110615–110626. http://dx.doi.org/10.1109/ACCESS.2023.3322453, URL https://ieeexplore.ieee.org/document/10273401/.

Lv, J., Yao, B., Guo, H., Gao, C., Wu, W., Li, J., et al. (2024a). MOLO-SLAM: A semantic SLAM for accurate removal of dynamic objects in agricultural environments. *Agriculture*, *14*(6), 819. http://dx.doi.org/10.3390/agriculture14060819, URL https://www.mdpi.com/2077-0472/14/6/819.

Lv, J., Yao, B., Guo, H., Gao, C., Wu, W., Li, J., et al. (2024b). MOLO-SLAM: A semantic SLAM for accurate removal of dynamic objects in agricultural environments. *Agriculture*, *14*(6), http://dx.doi.org/10.3390/agriculture14060819, URL https://www.mdpi.com/2077-0472/14/6/819.

Mascaro, R., Teixeira, L., & Chli, M. (2022). Volumetric instance-level semantic mapping via multi-view 2D-to-3D label diffusion. *IEEE Robotics and Automation Letters*, *7*(2), 3531–3538. http://dx.doi.org/10.1109/LRA.2022.3146502.

McCormac, J., Handa, A., Davison, A., & Leutenegger, S. (2016). SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In *2017 IEEE international conference on robotics and automation* (pp. 4628–4635). http://dx.doi.org/10.1109/ICRA.2017.7989538.

McCormac, J., Handa, A., Davison, A., & Leutenegger, S. (2017). SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In *2017 IEEE international conference on robotics and automation*. IEEE.

Meilland, M., & Comport, A. I. (2013). On unifying key-frame and voxel-based dense visual SLAM at large scales. In *2013 IEEE/RSJ international conference on intelligent robots and systems*. IEEE.

Memon, A. R., Iqbal, M., & Almakhles, D. (2024). DisView: A semantic visual IoT mixed data feature extractor for enhanced loop closure detection for UGVs during rescue operations. *IEEE Internet of Things Journal*.

Montemerlo, M., Thrun, S., Koller, D., & Wegbreit, B. (2002). FastSLAM: A factored solution to the simultaneous localization and mapping problem. *Aaai/Iaai*, *593598*, 593–598.

Motlagh, H. D. K., Lotfi, F., Taghirad, H. D., & Germi, S. B. (2019). Position estimation for drones based on visual SLAM and IMU in GPS-denied environment. In *2019 7th international conference on robotics and mechatronics* (pp. 120–124). IEEE.

Mukherjee, A., Das, S. D., Ghosh, J., Chowdhury, A. S., & Saha, S. K. (2021). Semantic segmentation of surface from lidar point cloud. *Multimedia Tools and Applications*, *80*, 35171–35191.

Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, *31*(5), 1147–1163.

Mur-Artal, R., & Tardos, J. D. (2017). ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, *33*(5), 1255–1262.

Muthu, S., Tennakoon, R., Rathnayake, T., Hoseinnezhad, R., Suter, D., & Bab-Hadiashar, A. (2020). Motion segmentation of rgb-d sequences: Combining semantic and motion information using statistical inference. *IEEE Transactions on Image Processing*, *29*, 5557–5570.

Narita, G., Seno, T., Ishikawa, T., & Kaji, Y. (2019). PanopticFusion: Online volumetric semantic mapping at the level of stuff and things. In *2019 IEEE/RSJ international conference on intelligent robots and systems* (pp. 4205–4212). http://dx.doi.org/10.1109/IROS40897.2019.8967890.

Nie, Y., Guo, S., Chang, J., Han, X., Huang, J., Hu, S.-M., et al. (2020). Shallow2Deep: Indoor scene modeling by single image understanding. *Pattern Recognition*, *103*, Article 107271.

Osman, H., Darwish, N., & Bayoumi, A. (2023). PlaceNet: A multi-scale semantic-aware model for visual loop closure detection. *Engineering Applications of Artificial Intelligence*, *119*, Article 105797.

Pak, J., & Son, H. I. (2025a). 3D LiDAR-based semantic SLAM for intelligent irrigation using UAV. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Pak, J., & Son, H. I. (2025b). 3D LiDAR-based semantic SLAM for intelligent irrigation using UAV. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *18*, 7495–7508. http://dx.doi.org/10.1109/JSTARS.2025.3547717.

Palazzolo, E., Behley, J., Lottes, P., Giguère, P., & Stachniss, C. (2019a). ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. URL https://www.ipb.uni-bonn.de/pdfs/palazzolo2019iros.pdf.

Palazzolo, E., Behley, J., Lottes, P., Giguère, P., & Stachniss, C. (2019b). Refusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *2019 IEEE/RSJ international conference on intelligent robots and systems* (pp. 7855–7862). http://dx.doi.org/10.1109/IROS40897.2019.8967590.

Pan, Y., Hu, K., Cao, H., Kang, H., & Wang, X. (2024). A novel perception and semantic mapping method for robot autonomy in orchards. *Computers and Electronics in Agriculture*, *219*, Article 108769.

Peng, S., Ran, T., Yuan, L., Zhang, J., & Xiao, W. (2024). Robust perception-based visual simultaneous localization and tracking in dynamic environments. *IEEE Transactions on Cognitive and Developmental Systems*.

Peng, X., Tong, P., Yang, X., Wang, C., & Zou, A.-M. (2025). IDMF-VINS: Improving visual-inertial SLAM for complex dynamic environments with motion consistency and feature filtering. *IEEE Sensors Journal*.

Peng, C., Xu, C., Wang, Y., Ding, M., Yang, H., Tomizuka, M., et al. (2024). Q-slam: Quadric representations for monocular slam. arXiv preprint arXiv:2403.08125.

Peng, H., Zhao, Z., & Wang, L. (2024). A review of dynamic object filtering in SLAM based on 3D LiDAR. *Sensors*, *24*(2), http://dx.doi.org/10.3390/s24020645, URL https://www.mdpi.com/1424-8220/24/2/645.

Popovic, M., Thomas, F., Papatheodorou, S., Funk, N., Vidal-Calleja, T., & Leutenegger, S. (2021). Volumetric occupancy mapping with probabilistic depth completion for robotic navigation. *IEEE Robotics and Automation Letters*, *6*(3), 5072–5079. http://dx.doi.org/10.1109/LRA.2021.3070308.

Pu, H., Luo, J., Wang, G., Huang, T., & Liu, H. (2023). Visual SLAM integration with semantic segmentation and deep learning: A review. *IEEE Sensors Journal*, *23*(19), 22119–22138.

Pu, H., Luo, J., Wang, G., Huang, T., Wu, L., Xiao, D., et al. (2025). Visual inertial SLAM based on spatiotemporal consistency optimization in diverse environments. *Journal of Field Robotics*, *42*(3), 679–696.

Pugh, B., Chernak, D., & Jiddi, S. (2023). GeoSynth: A photorealistic synthetic indoor dataset for scene understanding. *IEEE Transactions on Visualization and Computer Graphics*, *29*(5), 2586–2595.

Qadri, M., & Kantor, G. (2021). Semantic feature matching for robust mapping in agriculture. arXiv:2107.04178, URL https://arxiv.org/abs/2107.04178.

Qi, H., Chen, X., Yu, Z., Li, C., Shi, Y., Zhao, Q., et al. (2025). Semantic-independent dynamic SLAM based on geometric re-clustering and optical flow residuals. *IEEE Transactions on Circuits and Systems for Video Technology*.

Qi, H., Hu, Z., Xiang, Y., Cai, D., & Zhao, Y. (2023). ATY-SLAM: A visual semantic SLAM for dynamic indoor environments. In *Advanced intelligent computing technology and applications* (pp. 3–14). Springer Nature Singapore, http://dx.doi.org/10.1007/978-981-99-4761-4_1.

Qian, Z., Fu, J., & Xiao, J. (2022). Towards accurate loop closure detection in semantic SLAM with 3D semantic covisibility graphs. *IEEE Robotics and Automation Letters*, *7*(2), 2455–2462.

Qian, Z., Patath, K., Fu, J., & Xiao, J. (2021). Semantic slam with autonomous object-level data association. In *2021 IEEE international conference on robotics and automation* (pp. 11203–11209). IEEE.

Qin, Y., Mei, T., Gao, Z., Lin, Z., Song, W., & Zhao, X. (2022). RGB-D SLAM in dynamic environments with multilevel semantic mapping. *Journal of Intelligent and Robotic Systems*, *105*(4), http://dx.doi.org/10.1007/s10846-022-01697-y.

Qin, L., Wu, C., Chen, Z., Kong, X., Lv, Z., & Zhao, Z. (2024). RSO-SLAM: A robust semantic visual SLAM with optical flow in complex dynamic environments. *IEEE Transactions on Intelligent Transportation Systems*, 1–16. http://dx.doi.org/10.1109/TITS.2024.3402241, URL https://ieeexplore.ieee.org/document/10542573/.

Qiu, Z., Zhuang, Y., Yan, F., Hu, H., & Wang, W. (2019). RGB-DI images and full convolution neural network-based outdoor scene understanding for mobile robots. *IEEE Transactions on Instrumentation and Measurement*, *68*(1), 27–37.

Ran, T., Yuan, L., Zhang, J., He, L., Huang, R., & Mei, J. (2021). Not only look but infer: Multiple hypothesis clustering of data association inference for semantic SLAM. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–9.

Ran, T., Yuan, L., Zhang, J., Tang, D., & He, L. (2021). RS-SLAM: A robust semantic SLAM in dynamic environments based on RGB-D sensor. *IEEE Sensors Journal*, *21*(18), 20657–20664.

Ran, T., Yuan, L., Zhang, J., Wu, Z., & He, L. (2022). Object-oriented semantic SLAM based on geometric constraints of points and lines. *IEEE Transactions on Cognitive and Developmental Systems*, *15*(2), 751–760.

Rosen, D. M., Doherty, K. J., Terán Espinoza, A., & Leonard, J. J. (2021). Advances in inference and representation for simultaneous localization and mapping. *Annual Review of Control, Robotics, and Autonomous Systems*, *4*(1), 215–242.

Rosinol, A., Abate, M., Chang, Y., & Carlone, L. (2019). Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE international conference on robotics and automation* (pp. 1689–1696). http://dx.doi.org/10.1109/ICRA40945.2020.9196885.

Rosinol, A., Abate, M., Chang, Y., & Carlone, L. (2020). Kimera: An open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE international conference on robotics and automation*. IEEE.

Rosinol, A., Leonard, J. J., & Carlone, L. (2023). Probabilistic volumetric fusion for dense monocular SLAM. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3097–3105).

Rosinol, A., Sattler, T., Pollefeys, M., & Carlone, L. (2019). Incremental visual-inertial 3D mesh generation with structural regularities. In *2019 international conference on robotics and automation*. IEEE.

Rosinol, A., Violette, A., Abate, M., Hughes, N., Chang, Y., Shi, J., et al. (2021). Kimera: From SLAM to spatial perception with 3D dynamic scene graphs. *The International Journal of Robotics Research*, *40*(12–14), 1510–1546.

Rosu, R. A., Quenzel, J., & Behnke, S. (2020). Semi-supervised semantic mapping through label propagation with semantic texture meshes. *International Journal of Computer Vision*, *128*(5), 1220–1238.

Ruan, C., Zang, Q., Zhang, K., & Huang, K. (2023). Dn-slam: A visual slam with orb features and nerf mapping in dynamic environments. *IEEE Sensors Journal*, *24*(4), 5279–5287.

Ruiz-Sarmiento, J. R., Galindo, C., & González-Jiménez, J. (2017). Robot@ home, a robotic dataset for semantic mapping of home environments. *The International Journal of Robotics Research*, *36*(2), 131–141.

Rünz, M., & Agapito, L. (2017). Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *2017 IEEE International Conference on Robotics and Automation* (pp. 4471–4478). http://dx.doi.org/10.1109/ICRA.2017.7989518.

Rünz, M., & Agapito, L. (2018). MaskFusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *2018 IEEE international symposium on mixed and augmented reality* (pp. 10–20). http://dx.doi.org/10.1109/ISMAR.2018.00024.

Sahili, A. R., Hassan, S., Sakhrieh, S., Mounsef, J., Maalouf, N., Arain, B., et al. (2023). A survey of visual SLAM methods. *IEEE Access*.

Samadzadeh, A., & Nickabadi, A. (2023). SRVIO: Super robust visual inertial odometry for dynamic environments and challenging loop-closure conditions. *IEEE Transactions on Robotics*, *39*(4), 2878–2891.

Sandstrom, E., Li, Y., Van Gool, L., & Oswald, M. R. (2023). Point-SLAM: Dense neural point cloud-based SLAM. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 18433–18444).

Sedaghat, N., & Brox, T. (2015). Unsupervised generation of a viewpoint annotated car dataset from videos. In *IEEE international conference on computer vision*. URL http://lmb.informatik.uni-freiburg.de//Publications/2015/SB15.

Shao, X., Liu, C., Lu, P., Li, Y., & Akbar, A. (2025). Landslide robotics: a prototype for interactive and sustainable geohazard investigation. *Landslides*, *22*(4), 1291–1308.

Shao, X., Shen, Y., Zhang, L., Zhao, S., Zhu, D., & Zhou, Y. (2023). Slam for indoor parking: A comprehensive benchmark dataset and a tightly coupled semantic framework. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(1), 1–23.

Shao, X., Zhang, L., Zhang, T., Shen, Y., & Zhou, Y. (2022). MOFIS SLAM: A multi-object semantic SLAM system with front-view, inertial, and surround-view sensors for indoor parking. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(7), 4788–4803.

Shi, X., Li, D., Zhao, P., Tian, Q., Tian, Y., Long, Q., et al. (2020). Are we ready for service robots? The OpenLORIS-Scene datasets for lifelong SLAM. In *2020 international conference on robotics and automation* (pp. 3139–3145).

Shi, J., Zha, F., Guo, W., Wang, P., & Li, M. (2020). Dynamic visual SLAM based on semantic information and multi-view geometry. In *2020 5th international conference on automation, control and robotics engineering* (pp. 671–679). http://dx.doi.org/10.1109/CACRE50138.2020.9230242.

Shoukat, M. U., Yan, L., Deng, D., Imtiaz, M., Safdar, M., & Nawaz, S. A. (2024). Cognitive robotics: Deep learning approaches for trajectory and motion control in complex environment. *Advanced Engineering Informatics*, *60*, Article 102370.

Singh, G., Wu, M., Do, M. V., & Lam, S.-K. (2022). Fast semantic-aware motion state detection for visual slam in dynamic environment. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 23014–23030.

Smith, R. C., & Cheeseman, P. (1986). On the representation and estimation of spatial uncertainty. *International Journal of Robotics Research*, *5*(4), 56–68.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., & Cremers, D. (2012). A benchmark for the evaluation of RGB-D SLAM systems. In *Proc. of the international conference on intelligent robot systems*.

Sumikura, S., Shibuya, M., & Sakurada, K. (2019). OpenVSLAM: A versatile visual SLAM framework. In *Proceedings of the 27th ACM international conference on multimedia*. http://dx.doi.org/10.1145/3343031.3350539.

Sun, Y., Ma, Z., Zhou, M., & Cao, Z. (2023). A topological semantic mapping method based on text-based unsupervised image segmentation for assistive indoor navigation. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–13.

Sun, H., Wang, P., Ni, C., & Li, J. M. (2024). Loop closure detection based on image semantic feature and bag-of-words. *Multimedia Tools and Applications*, *83*(12), 36377–36398.

Sun, Q., Yuan, J., & Zhang, X. (2021). IT-HYFAO-VO: Interpretation tree-based VO with hybrid feature association and optimization. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–18.

Takleh, T. T. O., Bakar, N. A., Rahman, S. A., Hamzah, R., & Aziz, Z. (2018). A brief survey on SLAM methods in autonomous vehicle. *International Journal of Engineering and Technology*, *7*(4), 38–43.

Tang, S., Huang, H., Zhang, Y., Yao, M., Li, X., Xie, L., et al. (2023). Skeleton-guided generation of synthetic noisy point clouds from as-built BIM to improve indoor scene understanding. *Automation in Construction*, *156*, Article 105076.

Tang, Y., Zhao, C., Wang, J., Zhang, C., Sun, Q., Zheng, W. X., et al. (2023). Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(12), 9604–9624.

Tardioli, D., Sicignano, D., Riazuelo, L., Romeo, A., Villarroel, J. L., & Montano, L. (2016). Robot teams for intervention in confined and structured environments. *Journal of Field Robotics*, *33*(6), 765–801.

Tchuiev, V., & Indelman, V. (2023). Epistemic uncertainty aware semantic localization and mapping for inference and belief space planning. *Artificial Intelligence*, *319*, Article 103903.

Teed, Z., & Deng, J. (2021). DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. (pp. 16558–16569).

Tian, Y., Chang, Y., Arias, F. H., Nieto-Granda, C., How, J. P., & Carlone, L. (2022). Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems. *IEEE Transactions on Robotics*, *38*(4).

Tian, Y., Chang, Y., Quang, L., Schang, A., Nieto-Granda, C., How, J., et al. (2023). Resilient and distributed multi-robot visual SLAM: Datasets, experiments, and lessons learned. In *IEEE/RSJ intl. conf. on intelligent robots and systems*.

Tian, G., Liu, L., Ri, J., Liu, Y., & Sun, Y. (2019). ObjectFusion: An object detection and segmentation framework with RGB-D SLAM and convolutional neural networks. *Neurocomputing*, *345*, 3–14.

Tian, Y., Wan, X., Zhang, S., Zuo, J., Shao, Y., Liu, B., et al. (2025). Lo-SLAM: Lunar target-oriented SLAM using object identification, relative navigation and multi-level mapping. *IEEE Transactions on Geoscience and Remote Sensing*.

Tian, L., Yan, Y., & Li, H. (2023). SVD-SLAM: Stereo visual SLAM algorithm based on dynamic feature filtering for autonomous driving. *Electronics*, *12*(8), http://dx.doi.org/10.3390/electronics12081883, URL https://www.mdpi.com/2079-9292/12/8/1883.

Tian, R., Zhang, Y., Yang, L., Zhang, J., Coleman, S., & Kerr, D. (2024). Dynaquadric: Dynamic quadric slam for quadric initialization, mapping, and tracking. *IEEE Transactions on Intelligent Transportation Systems*.

Tiozzo Fasiolo, D., Scalera, L., Maset, E., & Gasparetto, A. (2023). Towards autonomous mapping in agriculture: A review of supportive technologies for ground robotics. *Robotics and Autonomous Systems*, *169*, Article 104514. http://dx.doi.org/10.1016/j.robot.2023.104514, URL https://www.sciencedirect.com/science/article/pii/S0921889023001537.

Trejos, K., Rincón, L., Bolaños, M., Fallas, J., & Marín, L. (2022). 2D slam algorithms characterization, calibration, and comparison considering pose error, map accuracy as well as cpu and memory usage. *Sensors*, *22*(18), 6903.

Tschopp, F., Nieto, J., Siegwart, R., & Cadena, C. (2021). Superquadric object representation for optimization-based semantic SLAM. arXiv preprint arXiv:2109.09627.

Vasilopoulos, V., Pavlakos, G., Schmeckpeper, K., Daniilidis, K., & Koditschek, D. E. (2022). Reactive navigation in partially familiar planar environments using semantic perceptual feedback. *The International Journal of Robotics Research*, *41*(1), 85–126.

Venator, M., Bruns, E., & Maier, A. (2020). Robust camera pose estimation for unordered road scene images in varying viewing conditions. *IEEE Transactions on Intelligent Vehicles*, *5*(1), 165–174.

Vishnyakov, B., Sgibnev, I., Sheverdin, V., Sorokin, A., Masalov, P., Kazakhmedov, K., et al. (2021). Real-time semantic slam with dcnn-based feature point detection, matching and dense point cloud aggregation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 399–404.

Wan, K., & Luo, J. (2025). SPVL-vSLAM: Visual SLAM for autonomous driving vehicles based on semantic patch-NetVLAD loop closure detection in semi-static scenes. *IEEE Transactions on Intelligent Transportation Systems*.

Wang, K., Guo, J., Chen, K., & Lu, J. (2025). An in-depth examination of SLAM methods: Challenges, advancements, and applications in complex scenes for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, *26*(7), 11066–11087. http://dx.doi.org/10.1109/TITS.2025.3545479.

Wang, Z., Li, W., Shen, Y., & Cai, B. (2020). 4-D SLAM: An efficient dynamic Bayes network-based approach for dynamic scene understanding. *IEEE Access*, *8*, 219996–220014. http://dx.doi.org/10.1109/ACCESS.2020.3042339, URL https://ieeexplore.ieee.org/document/9279261/.

Wang, N., Lu, H., Zheng, Z., Wang, H., Liu, Y.-H., & Chen, X. (2025). Leveraging semantic graphs for efficient and robust LiDAR SLAM. arXiv preprint arXiv:2503.11145.

Wang, P., Luo, W., Liu, J., Zhou, Y., Li, X., Zhao, S., et al. (2025). Real-time semantic SLAM-based 3D reconstruction robot for greenhouse vegetables. *Computers and Electronics in Agriculture*, *237*, Article 110582.

Wang, K., Ma, S., Ren, F., & Lu, J. (2021). SBAS: Salient bundle adjustment for visual SLAM. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–9.

Wang, J., Runz, M., & Agapito, L. (2021). DSP-SLAM: Object oriented SLAM with deep shape priors. In *2021 international conference on 3D vision* (pp. 1362–1371). London, United Kingdom: IEEE, http://dx.doi.org/10.1109/3DV53792.2021.00143, URL https://ieeexplore.ieee.org/document/9665919/.

Wang, Y., Tian, Y., Chen, J., Chen, C., Xu, K., & Ding, X. (2025). MSSD-SLAM: Multi-feature semantic RGB-D inertial SLAM with structural regularity for dynamic environments. *IEEE Transactions on Instrumentation and Measurement*.

Wang, Y., Tian, Y., Chen, J., Xu, K., & Ding, X. (2024). A survey of visual SLAM in dynamic environment: The evolution from geometric to semantic approaches. *IEEE Transactions on Instrumentation and Measurement*.

Wang, Z., Tian, G., & Liu, T. (2025). Object-level semantic metric mapping for robot object search in home environment. *IEEE Transactions on Industrial Electronics*.

Wang, Y., Wu, Y., Li, D., & Yu, W. (2024). Millimeter wave radar and vision fusion based semantic simultaneous localization and mapping. *IEEE Antennas and Wireless Propagation Letters*.

Wang, C., Zhang, Y., & Li, X. (2020). Pmds-slam: Probability mesh enhanced semantic slam in dynamic environments. In *2020 5th international conference on control, robotics and cybernetics* (pp. 40–44). IEEE.

Wang, S., Zheng, D., & Li, Y. (2023). LiDAR-SLAM loop closure detection based on multi-scale point cloud feature transformer. *Measurement Science and Technology*, *35*(3), Article 036305.

Wei, S., Chen, G., Chi, W., Wang, Z., & Sun, L. (2023). Object clustering with Dirichlet process mixture model for data association in monocular SLAM. *IEEE Transactions on Industrial Electronics*, *70*(1), 594–603.

Wei, W., Huang, K., Liu, X., & Zhou, Y. (2023). GSL-VO: A geometric-semantic information enhanced lightweight visual odometry in dynamic environments. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–13.

Wei, K., Ni, P., Li, X., Hu, Y., & Hu, W. (2024). LiDAR-based semantic place recognition in dynamic urban environments. *IEEE Sensors Journal*.

Wei, H., & Wang, L. (2018). Understanding of indoor scenes based on projection of spatial rectangles. *Pattern Recognition*, *81*, 497–514.

Wei, Y., Zhou, B., Duan, Y., Liu, J., & An, D. (2023). DO-SLAM: research and application of semantic SLAM system towards dynamic environments based on object detection. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(24), 30009–30026. http://dx.doi.org/10.1007/s10489-023-05070-w, URL https://link.springer.com/10.1007/s10489-023-05070-w.

Wen, S., Li, X., Liu, X., Li, J., Tao, S., Long, Y., et al. (2023). Dynamic SLAM: A visual SLAM in outdoor dynamic scenes. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–11. http://dx.doi.org/10.1109/TIM.2023.3317378, URL https://ieeexplore.ieee.org/document/10256128/.

Wen, C., Tan, J., Li, F., Wu, C., Lin, Y., Wang, Z., et al. (2021). Cooperative indoor 3D mapping and modeling using LiDAR data. *Information Sciences*, *574*, 192–209.

Wen, S., Zhao, Y., Liu, X., Sun, F., Lu, H., & Wang, Z. (2020). Hybrid semi-dense 3D semantic-topological mapping from stereo visual-inertial odometry SLAM with loop closure detection. *IEEE Transactions on Vehicular Technology*, *69*(12), 16057–16066.

Whelan, T., Salas-Moreno, R. F., Glocker, B., Davison, A. J., & Leutenegger, S. (2016). ElasticFusion: Real-time dense SLAM and light source estimation. *The International Journal of Robotics Research*, *35*(14), 1697–1716.

Wijaya, B., Jiang, K., Yang, M., Wen, T., Tang, X., & Yang, D. (2022). Crowd-sourced road semantics mapping based on pixel-wise confidence level. *Automotive Innovation*, *5*(1), 43–56.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution* (pp. 196–202). Springer.

Wu, W., Guo, L., Gao, H., You, Z., Liu, Y., & Chen, Z. (2022). YOLO-SLAM: A semantic SLAM system towards dynamic environment with geometric constraint. *Neural Computing and Applications*, *34*(8), 6011–6026. http://dx.doi.org/10.1007/s00521-021-06764-3, URL https://link.springer.com/10.1007/s00521-021-06764-3.

Wu, C., Li, X., Kong, D., Hu, Y., Ni, P., & Liu, X. (2024). GDO-SLAM: visual-based ground-aware decoupling optimized SLAM for UGV in outdoor environments. *IEEE Sensors Journal*.

Wu, S., Zhang, X., Zhang, S., Song, Z., Wang, R., & Yuan, J. (2025). MPOC-SLAM: An RGB-D SLAM system with motion probability and object category in high dynamic environments. *IEEE/ASME Transactions on Mechatronics*.

Wu, Y., Zhang, Y., Zhu, D., Deng, Z., Sun, W., Chen, X., et al. (2023). An object slam framework for association, mapping, and high-level tasks. *IEEE Transactions on Robotics*, *39*(4), 2912–2932.

Wu, Y., Zhang, Y., Zhu, D., Feng, Y., Coleman, S., & Kerr, D. (2020). EAO-SLAM: Monocular semi-dense object SLAM based on ensemble data association. In *2020 IEEE/RSJ international conference on intelligent robots and systems* (pp. 4966–4973). Las Vegas, NV, USA: IEEE, http://dx.doi.org/10.1109/IROS45743.2020.9341757, URL https://ieeexplore.ieee.org/document/9341757/.

Wu, H., Zhao, J., Xu, K., Zhang, Y., Xu, R., Wang, A., et al. (2022). Semantic SLAM based on deep learning in endocavity environment. *Symmetry*, *14*(3), 614.

Xia, L., Li, S., Yi, L., Ruan, H., & Zhang, D. (2024). Measure for semantics and semantically constrained pose optimization: A review. *IEEE Transactions on Instrumentation and Measurement*.

Xiao, H., Hu, Z., Lv, C., Meng, J., Zhang, J., & You, J. (2025). Progressive multi-modal semantic segmentation guided SLAM using tightly-coupled LiDAR-visual-inertial odometry. *IEEE Transactions on Intelligent Transportation Systems*.

Xie, W., Liu, P. X., & Zheng, M. (2021). Moving object segmentation and detection for robust RGBD-SLAM in dynamic environments. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–8.

Xie, X., Qin, Y., Zhang, Z., Yan, Z., Jin, H., Xu, M., et al. (2024). GY-SLAM: A dense semantic SLAM system for plant factory transport robots. *Sensors (Basel, Switzerland)*, *24*, http://dx.doi.org/10.3390/s24051374.

Xie, Y., Zhang, Y., Chen, L., Cheng, H., Tu, W., Cao, D., et al. (2022). RDC-SLAM: A real-time distributed cooperative SLAM system based on 3D lidar. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 14721–14730. http://dx.doi.org/10.1109/TITS.2021.3132375.

Xing, Z., Zhu, X., & Dong, D. (2022). DE-SLAM: SLAM for highly dynamic environment. *Journal of Field Robotics*, *39*(5), 528–542.

Xiong, J., Liang, J., Zhuang, Y., Hong, D., Zheng, Z., Liao, S., et al. (2023). Real-time localization and 3D semantic map reconstruction for unstructured citrus orchards. *Computers and Electronics in Agriculture*, *213*, Article 108217.

Xu, L., Feng, C., Kamat, V. R., & Menassa, C. C. (2020). A scene-adaptive descriptor for visual SLAM-based locating applications in built environments. *Automation in Construction*, *112*, Article 103067. http://dx.doi.org/10.1016/j.autcon.2019.103067, URL https://www.sciencedirect.com/science/article/pii/S0926580519309306.

Xu, B., Li, W., Tzoumanikas, D., Bloesch, M., Davison, A., & Leutenegger, S. (2019). MID-fusion: Octree-based object-level multi-instance dynamic SLAM. In *2019 international conference on robotics and automation* (pp. 5231–5237). http://dx.doi.org/10.1109/ICRA.2019.8794371.

Xu, B., Zheng, Z., Pan, Z., & Yu, L. (2025). HMC-SLAM: A robust VSLAM based on RGB-D camera in dynamic environment combined hierarchical multidimensional clustering algorithm. *IEEE Transactions on Instrumentation and Measurement*.

Yan, L., Hu, X., Zhao, L., Chen, Y., Wei, P., & Xie, H. (2022). DGS-SLAM: A fast and robust RGBD SLAM in dynamic environments combined by geometric and semantic information. *Remote Sensing*, *14*(3), http://dx.doi.org/10.3390/rs14030795, URL https://www.mdpi.com/2072-4292/14/3/795.

Yan, F., Wang, J., He, G., Chang, H., & Zhuang, Y. (2020). Sparse semantic map building and relocalization for UGV using 3D point clouds in outdoor environments. *Neurocomputing*, *400*, 333–342.

Yang, L., & Cai, H. (2024). Enhanced visual SLAM for construction robots by efficient integration of dynamic object segmentation and scene semantics. *Advanced Engineering Informatics*, *59*, Article 102313.

Yang, H., Chen, Y., Liu, J., Zhang, Z., & Zhang, X. (2023). A 3D lidar SLAM system based on semantic segmentation for rubber-tapping robot. *Forests*, *14*(9), http://dx.doi.org/10.3390/f14091856, URL https://www.mdpi.com/1999-4907/14/9/1856.

Yang, D., Gao, Y., Wang, X., Yue, Y., Yang, Y., & Fu, M. (2025). OpenGS-SLAM: Open-set dense semantic SLAM with 3D Gaussian splatting for object-level scene understanding. arXiv preprint arXiv:2503.01646.

Yang, C., He, L., Zhuang, H., Wang, C., & Yang, M. (2023). Pseudo-anchors: Robust semantic features for lidar mapping in highly dynamic scenarios. *IEEE Transactions on Intelligent Transportation Systems*, *24*(2), 1619–1630.

Yang, B., Ran, W., Wang, L., Lu, H., & Chen, Y.-P. P. (2022). Multi-classes and motion properties for concurrent visual slam in dynamic environments. *IEEE Transactions on Multimedia*, *24*, 3947–3960.

Yang, X., Wang, T., Wang, Y., Lang, C., Jin, Y., & Li, Y. (2025). FND-SLAM: A SLAM system using feature points and NeRF in dynamic environments based on RGB-D sensors. *IEEE Sensors Journal*.

Yang, J., Xu, J., Li, K., Lai, Y.-K., Yue, H., Lu, J., et al. (2020). Learning to reconstruct and understand indoor scenes from sparse views. *IEEE Transactions on Image Processing*, *29*, 5753–5766.

Yang, L., Ye, J., Zhang, Y., Wang, L., & Qiu, C. (2024). A semantic SLAM-based method for navigation and landing of UAVs in indoor environments. *Knowledge-Based Systems*, *293*, Article 111693.

Yang, L., Zhang, Y., Tian, R., Liang, S., Shen, Y., Coleman, S., et al. (2023). Fast, robust, accurate, multi-body motion aware SLAM. *IEEE Transactions on Intelligent Transportation Systems*, *25*(5), 4381–4397.

Ying, Z., & Li, H. (2023). IMM-SLAMMOT: Tightly-coupled SLAM and IMM-based multi-object tracking. *IEEE Transactions on Intelligent Vehicles*, *9*(2), 3964–3974.

Ying, Z., Yuan, X., Song, B., Song, Y., Zhou, F., & Sheng, W. (2023). Accurate and efficient 3D panoptic mapping using diverse information modalities and multidimensional data association. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(6), 4489–4502.

You, M., Luo, C., Zhou, H., & Zhu, S. (2023). Dynamic dense CRF inference for video segmentation and semantic SLAM. *Pattern Recognition*, *133*, Article 109023.

You, Y., Wei, P., Cai, J., Huang, W., Kang, R., & Liu, H. (2022). MISD-SLAM: Multimodal semantic SLAM for dynamic environments. *Wireless Communications and Mobile Computing*, *2022*(1), Article 7600669. http://dx.doi.org/10.1155/2022/7600669, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/7600669, URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/7600669.

Yousif, K., Bab-Hadiashar, A., & Hoseinnezhad, R. (2015). An overview to visual odometry and visual SLAM: Applications to mobile robotics. *Intelligent Industrial Systems*, *1*(4), 289–311.

Yu, C., Liu, Z., Liu, X.-J., Xie, F., Yang, Y., Wei, Q., et al. (2018a). DS-SLAM: A semantic visual SLAM towards dynamic environments. In *2018 IEEE/RSJ international conference on intelligent robots and systems* (pp. 1168–1174). Madrid: IEEE, http://dx.doi.org/10.1109/IROS.2018.8593691, URL https://ieeexplore.ieee.org/document/8593691/.

Yu, C., Liu, Z., Liu, X., Xie, F., Yang, Y., Wei, Q., et al. (2018b). DS-SLAM: A semantic visual SLAM towards dynamic environments. In *2018 IEEE/RSJ international conference on intelligent robots and systems* (pp. 1168–1174). http://dx.doi.org/10.1109/IROS.2018.8593691.

Yu, J., Xiang, Z., & Su, J. (2022). Hierarchical multi-level information fusion for robust and consistent visual SLAM. *IEEE Transactions on Vehicular Technology*, *71*(1), 250–259.

Zhai, H., Huang, G., Hu, Q., Li, G., Bao, H., & Zhang, G. (2024). Nis-slam: Neural implicit semantic rgb-d slam for 3d consistent scene understanding. *IEEE Transactions on Visualization and Computer Graphics*.

Zhang, J., Gui, M., Wang, Q., Liu, R., Xu, J., & Chen, S. (2019). Hierarchical topic model based object association for semantic SLAM. *IEEE Transactions on Visualization and Computer Graphics*, *25*(11), 3052–3062.

Zhang, W., Guo, Y., Niu, L., Li, P., Wan, Z., Shao, F., et al. (2024). Lp-slam: language-perceptive RGB-D SLAM framework exploiting large language model. *Complex & Intelligent Systems*, *10*(4), 5391–5409.

Zhang, J., Henein, M., Mahony, R., & Ila, V. (2020). VDO-SLAM: A visual dynamic object-aware SLAM system. arXiv preprint arXiv:2005.11052.

Zhang, H., Huo, J., Huang, Y., & Liu, Q. (2025). Real-time dynamic visual-inertial SLAM and object tracking based on lightweight deep feature extraction matching. *IEEE Transactions on Instrumentation and Measurement*.

Zhang, Q., & Li, C. (2023). Semantic SLAM for mobile robots in dynamic environments based on visual camera sensors. *Measurement Science and Technology*, *34*(8), Article 085202. http://dx.doi.org/10.1088/1361-6501/acd1a4, URL https://iopscience.iop.org/article/10.1088/1361-6501/acd1a4.

Zhang, J., & Singh, S. (2014). LOAM: Lidar odometry and mapping in real-time. In *Robotics: Science and systems x*. Robotics: Science and Systems Foundation.

Zhang, Z., Song, Y., Pang, B., Yuan, X., Xu, Q., & Xu, X. (2025). SSF-SLAM: Real-time RGB-D visual SLAM for complex dynamic environments based on semantic and scene flow geometric information. *IEEE Transactions on Instrumentation and Measurement*.

Zhang, X., Wang, L., & Su, Y. (2021). Visual place recognition: A survey from deep learning perspective. *Pattern Recognition*, *113*, Article 107760.

Zhang, X., Wang, X., & Zhang, R. (2022). Dynamic semantics SLAM based on improved mask R-CNN. *IEEE Access*, *10*, 126525–126535.

Zhang, J., Yuan, L., Ran, T., Tao, Q., & Wu, Z. (2023). Outlier elimination for monocular object SLAM based on spatiotemporal consistency constraints. *IEEE Sensors Journal*, *23*(8), 8887–8898.

Zhang, C., Zhang, R., Jin, S., & Yi, X. (2022). PFD-SLAM: A new RGB-D SLAM for dynamic indoor environments based on non-prior semantic segmentation. *Remote Sensing*, *14*(10), http://dx.doi.org/10.3390/rs14102445, URL https://www.mdpi.com/2072-4292/14/10/2445.

Zhang, H., Zhang, Y., Liu, Y., Naixue Xiong, N., & Li, Y. (2024). Slam loop closure detection algorithm based on MSA-SG. *Cluster Computing*, *27*(7), 9283–9301.

Zhao, Y., Xiong, Z., Zhou, S., Peng, Z., Campoy, P., & Zhang, L. (2022). KSF-SLAM: A key segmentation frame based semantic SLAM in dynamic environments. *Journal of Intelligent and Robotic Systems*, *105*(1), 3. http://dx.doi.org/10.1007/s10846-022-01613-4, URL https://link.springer.com/10.1007/s10846-022-01613-4.

Zhao, Z., Zhang, W., Gu, J., Yang, J., & Huang, K. (2019). Lidar mapping optimization based on lightweight semantic segmentation. *IEEE Transactions on Intelligent Vehicles*, *4*(3), 353–362.

Zhao, X., Zuo, T., & Hu, X. (2021). OFM-SLAM: A visual semantic SLAM for dynamic indoor environments. *Mathematical Problems in Engineering*, *2021*(1), Article 5538840. http://dx.doi.org/10.1155/2021/5538840, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/5538840, URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/5538840.

Zheng, C., Du, Y., Xiao, J., Sun, T., Wang, Z., Eynard, B., et al. (2025). Semantic map construction approach for human-robot collaborative manufacturing. *Robotics and Computer-Integrated Manufacturing*, *91*, Article 102845.

Zheng, Z., Lin, S., & Yang, C. (2024). RLD-SLAM: A robust lightweight VI-SLAM for dynamic environments leveraging semantics and motion information. *IEEE Transactions on Industrial Electronics*.

Zheng, S., Wang, J., Rizos, C., Ding, W., & El-Mowafy, A. (2023). Simultaneous localization and mapping (SLAM) for autonomous driving: Concept and analysis. *Remote Sensing*, *15*(4), http://dx.doi.org/10.3390/rs15041156, URL https://www.mdpi.com/2072-4292/15/4/1156.

Zhi, M., Deng, C., Li, B., Zhang, H., & Hong, C. (2024). A dynamic visual-inertial-wheel odometry with semantic constraints and denoised IMU-odometer prior for autonomous driving. *IEEE Sensors Journal*.

Zhong, Y., Hu, S., Huang, G., Bai, L., & Li, Q. (2022). WF-SLAM: A robust VSLAM for dynamic scenarios via weighted features. *IEEE Sensors Journal*, *22*(11), 10818–10827.

Zhou, Y., Mei, C., Liu, T., & Bai, L. (2023). Robust multi-sensor fusion via factor graph and variational Bayesian inference. In *Lecture notes in electrical engineering, Proceedings of 2022 international conference on autonomous unmanned systems* (pp. 11–22). Singapore: Springer Nature Singapore.

Zhou, Y., Tao, F., Fu, Z., Zhu, L., & Ma, H. (2023). RVD-SLAM: a real-time visual SLAM toward dynamic environments based on sparsely semantic segmentation and outlier prior. *IEEE Sensors Journal*, *23*(24), 30773–30785.

Zhou, Y., & Wang, X. (2025). Classified-SLAM: A real-time stereo SLAM based on scenario classification and outlier probability propagation for outdoor autonomous vehicle. *IEEE Transactions on Instrumentation and Measurement*.

Zhou, W., Yue, Y., Fang, M., Qian, X., Yang, R., & Yu, L. (2023). BCINet: Bilateral cross-modal interaction network for indoor scene understanding in RGB-D images. *Information Fusion*, *94*, 32–42.

Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., et al. (2022). NICE-SLAM: Neural implicit scalable encoding for SLAM. In *2022 IEEE/CVF conference on computer vision and pattern recognition* (pp. 12776–12786). New Orleans, LA, USA: IEEE, http://dx.doi.org/10.1109/CVPR52688.2022.01245, URL https://ieeexplore.ieee.org/document/9878912/.

Zhu, D., Wang, Z., Lu, T., & Jiang, X. (2024). PMF-SLAM: Pose-guided and multiscale feature interaction-based semantic SLAM for autonomous wheel loader. *IEEE Sensors Journal*, *24*(7), 11625–11638.

Zhu, Q., Xiao, J., & Fan, L. (2025). IndoorMS: A multispectral dataset for semantic segmentation in indoor scene understanding. *IEEE Sensors Journal*.

Zhu, W., Yuan, J., Zhang, X., & Chen, F. (2025). Bridging the gap between semantics and geometry in SLAM: A semantic-geometric tight-coupling monocular visual object SLAM system. *IEEE Transactions on Robotics*.

Zobeidi, E., Koppel, A., & Atanasov, N. (2022). Dense incremental metric-semantic mapping for multiagent systems via sparse Gaussian process regression. *IEEE Transactions on Robotics*, *38*(5), 3133–3153.