Received: 31 October 2022 | Revised: 27 February 2023 | Accepted: 13 March 2023

DOI: 10.1002/rob.22170

SURVEY ARTICLE

WILEY

# A systematic literature review on long-term localization and mapping for mobile robots

Ricardo B. Sousa[1,2] | Héber M. Sobreira[2] | António Paulo Moreira[1,2]

[1]Electrical Engineering Department, Faculty of Engineering of the University of Porto, Porto, Porto, Portugal

[2]CRIIS—Centre for Robotics in Industry and Intelligent Systems, INESC TEC—Institute for Systems and Computer Engineering Technology and Science, Porto, Porto, Portugal

**Correspondence**
Ricardo B. Sousa, Electrical Engineering Department, Faculty of Engineering of the University of Porto, Porto, R. Dr. Roberto Frias, n/a,4200-465 Porto, Portugal.
Email: up201503004@edu.fe.up.pt

**Funding information**
Fundação para a Ciência e a Tecnologia

## Abstract

Long-term operation of robots creates new challenges to Simultaneous Localization and Mapping (SLAM) algorithms. Long-term SLAM algorithms should adapt to recent changes while preserving older states, when dealing with appearance variations (lighting, daytime, weather, or seasonal) or environment reconfiguration. When also operating robots for long periods and trajectory lengths, the map should readjust to environment changes but not grow indefinitely. The map size should depend only on updating the map with new information of interest, not on the operation time or trajectory length. Although several studies in the literature review SLAM algorithms, none of the studies focus on the challenges associated to lifelong SLAM. Thus, this paper presents a systematic literature review on long-term localization and mapping following the Preferred Reporting Items for Systematic reviews and Meta-Analysis guidelines. The review analyzes 142 works covering appearance invariance, modeling the environment dynamics, map size management, multisession, and computational topics such as parallel computing and timing efficiency. The analysis also focus on the experimental data and evaluation metrics commonly used to assess long-term autonomy. Moreover, an overview over the bibliographic data of the 142 records provides analysis in terms of keywords and authorship co-occurrence to identify the terms more used in long-term SLAM and research networks between authors, respectively. Future studies can update this paper thanks to the systematic methodology presented in the review and the public GitHub repository with all the documentation and scripts used during the review process.

**KEYWORDS**
lifelong SLAM, long-term autonomy, mobile robots, Simultaneous Localization and Mapping (SLAM)

## 1 | INTRODUCTION

An autonomous mobile robot requires a representation of its surroundings to localize itself relative to the environment. Simultaneous Localization and Mapping (SLAM) addresses this problem by incorporating the robot state estimation (pose and possibly other state variables) concurrently with the mapping process. The latter builds a representation of the environment perceived by the robot, originating a map incrementally built when exploring unknown areas or refined on passages through known locations.

In a static scene, the robot would only need to map once because the map would remain consistent with the environment. However, autonomous systems applications such as in industrial locations, outdoor environments, or service-oriented applications (shopping centers, homes) do not have static environments. Robots deployed in these environments deal with moving elements in the scene (humans, objects), environment reconfiguration (logistics locations, warehouses), and appearance variations (lighting, weather, seasonal, or daytime changes). Dynamic elements in the scene and environment reconfiguration may degrade the localization accuracy if the map is not updated accordingly (Pomerleau et al., 2014). Similarly, appearance variations cause inconsistencies in the localization and mapping processes. These inconsistencies result from observations perceived at two separate time instants with different signatures that may represent the same information in the environment. The changing conditions of the environment are a challenge for a SLAM system. The system should decide how (consider the more recent state, or permanent changes instead of temporary ones) and when (when the variations occur, or after a time period) to update the map. The challenge of performing SLAM in varying conditions is also related to the stability-plasticity dilemma. This dilemma states that a lifelong learning system should adapt to new changes and preserve old states over time (Biber & Duckett, 2009).

Furthermore, the map would grow indefinitely when gathering all the information perceived from the environment. This ever-growing problem poses another challenge for SLAM systems due to the limited computational resources of the mapping vehicle. The map size should depend on unobserved information from the environment, not on the trajectory length of the mapping process or operation time. Indeed, the map size should not grow unless the robot observes parts of the environment not yet represented in the map (Kretzschmar & Stachniss, 2012).

Long-term SLAM algorithms address the problem of performing SLAM over possibly unlimited periods of time (Biber & Duckett, 2009). One of the challenges to achieve lifelong SLAM remains the localization and mapping processes being invariant to the changing appearance of the scene. Lighting, weather, and seasonal variations impact the landscape and affect the data from sensor observations outdoors. Even though appearance variance highly changes visual data, ranging data may also be affected due to changes in the foliage outdoors. As for indoors, lighting changes, direct light, and dust may impact the sensor observations (Yin, Xu, Wang, et al., 2021). Another challenge of performing lifelong SLAM is modeling the dynamics of the environment. Moving elements and scene reconfiguration change the scene's appearance. However, dynamics modeling is more related to how to represent moving elements and structures on the map, how to identify and distinguish them from permanent and static observations, and possibly, how to predict the dynamics and influence the planning of the robot path. The ever-growing problem of the map is even more predominant in long periods of operation. A static environment raises the issue of removing redundant information from the map to manage its size. In addition to redundant information, a varying environment originates another problem of removing outdated information to improve the computational and memory efficiency of long-term SLAM algorithms (Boniardi et al., 2019). Other computational concerns related to improving the overall efficiency of the usage of computational resources may also arise in the long-term SLAM problem. These concerns include memory management, parallel computing, and using external computing devices. Moreover, long operation periods of long-term SLAM systems also create the challenge of how to handle the data acquired in each operation run. This challenge is related to multisession methodologies. The localization and mapping processes should be able to localize the robot in the existing map and handle the session data without assuming a prior initial pose (Labbé & Michaud, 2019).

Although several studies overview SLAM literature, only a subset of those studies mentions the challenges of performing long-term SLAM. Cadena et al. (2016) have a brief survey on the robustness and scalability of autonomous systems. The survey focuses on loop closure validation, dynamic environments, pose graph sparsification, and parallel and distributed computing for metric and semantic SLAM. In contrast, Lowry et al. (2016) limit the study to vision-based topological SLAM and discuss the challenge of varying conditions. Bresson et al. (2017) overview autonomous driving trends in terms of scalability, map updatability, and dynamicity. Still, the survey limits the discussion to algorithms that have both odometry and mapping modules, excluding localization-only works. Also, Bresson et al. (2017) focus more on the modules of the SLAM (relocalization, localization against a map). Kunze et al. (2018) give a brief overview of artificial intelligence (AI)-related works for robustness to appearance changes and learning dynamics of the environment. The discussion focuses on the areas in which AI enables long-term operation of autonomous systems. As for Zaffar et al. (2018), the study evaluates the long-term autonomy of sensors such as monocular and stereo cameras, and LiDAR. Overall, to the best of our knowledge, none of the existing studies overviews the trends to deal with the challenges in performing long-term SLAM. Also, the studies that overview some of the challenges of lifelong SLAM do not clarify the paper selection process of the cited works, not allowing other researchers to repeat the same process when searching in bibliographic databases.

So, this paper presents a systematic literature review on long-term localization and mapping following the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) (Page et al., 2021) statement. The systematic method followed in this review allows the replication and future updates of the results by other researchers following the paper selection scheme. The selection scheme leads to the 142 works included in this literature review for discussion and analysis. This review does not focus only on a single challenge of performing long-term SLAM. Instead, the discussion and analysis of the 142 included works cover the topics of invariance to changing appearance, dynamics modeling, map sparsification to account the possible ever-growing size effect on the map, multisession properties, and other computational concerns (e.g., parallel and distributed computing, memory management). Furthermore, this review considers all domains (air, ground, and water) and environments (indoors and outdoors), analyzes methods intended for different sensor modalities (e.g., cameras, 2D and 3D LiDARs, and radar), and focuses on works with long-term experiments (varying conditions in terms of lighting, weather, and seasonal changes, dynamic changing

environments, and experiments with increasing operation area). Finally, this paper makes available a public GitHub repository[1] with all the documentation and scripts used during the elaboration of this systematic review.

The main contributions of this paper relative to existing overviews and surveys on the SLAM literature are the following ones:

1. discussion on methodologies and trends focused on appearance invariance, dynamic elements, map sparsification and multisession techniques, and other computational concerns;
2. comparative analysis on the public data sets and experimental data used in the included works in terms of environment conditions, sensorization, and distance and time properties;
3. presentation of common evaluation metrics used by the included works in the experimental results.

### 1.1 | Paper organization

The rest of this review is organized as follows. Section 2 discusses the limitations of existent SLAM reviews in the context of long-term SLAM and presents the purpose and motivations of this paper. Section 3 explains the methodology followed in this review to search and select the included records. The methodology includes a data extraction process used for synthesis and analysis of the 142 included works. In Appendix A, Table A1 presents the data extraction results of the included records. Next, Section 4 analyzes the bibliographic information of the 142 works included in this review. The analysis identifies research networks and keywords related to the topic of this review, discusses the coverage of the data sources considered in the methodology relative to the identification results, and presents the evolution of the publication year and most relevant publication venues found in the included records. Section 5 discusses the methodologies of the 142 included records related to long-term localization and mapping. The discussion is organized by the following topics: appearance variance, dynamics modeling, map sparsification, multisession, and computational. Additionally, Section 5 includes an analysis of the experimental data (public data sets, distance and time considerations in the experiments, and ground-truth data) and evaluation metrics used by the authors. Then, Section 6 outlines challenges and future directions for long-term SLAM. Section 7 discusses possible limitations of this study. Lastly, Section 8 presents the conclusions of this review.

## 2 | PURPOSE OF THE STUDY

### 2.1 | Limitations of current SLAM reviews

The main studies reviewing the SLAM literature are presented in Table 1. In terms of introductions to the problem formulation, the existent studies focus on explaining different frameworks for performing SLAM. Durrant-Whyte and Bailey (2006) and Bailey and Durrant-Whyte (2006) perform an in-depth discussion of Bayesian-based probabilistic formulations, namely, the Extended Kalman Filter (EKF) and the Rao-Blackwellized particle filter frameworks. These frameworks are categorized as filtering approaches to solving the SLAM problem. Filtering approaches model SLAM as a state estimation problem, where the state being the robot pose (and possibly other variables) and the map. In contrast, Grisetti et al. (2010) explain in detail the smoothing formulation graph-based SLAM. This formulation estimates the full trajectory of the robot from the set of sensor measurements, also known as full SLAM. Thrun (2008) introduce both probabilistic (EKF and particle filter) and smoothing (graph) formulations of SLAM. Focusing on vision sensorization, Scaramuzza and Fraundorfer (2011) and Fraundorfer and Scaramuzza (2012) present an extensive tutorial on visual

**TABLE 1** Existent literature reviews, surveys, and tutorials on Simultaneous Localization and Mapping (SLAM).

| Year | Topic | Reference |
|---|---|---|
| 2006 | Probabilistic formulations (EKF, particle filter) | Bailey and Durrant-Whyte (2006), Durrant-Whyte and Bailey (2006) |
| 2008 | Probabilistic and pose graph formulations | Thrun (2008) |
| 2010 | GraphSLAM | Grisetti et al. (2010) |
| 2011 | Observability, convergence, consistency (feature-based SLAM) | Dissanayake et al. (2011) |
| 2012 | Visual odometry | Fraundorfer and Scaramuzza (2012), Scaramuzza and Fraundorfer (2011) |
| 2014 | Underwater navigation and localization | Paull et al. (2014) |
| 2015 | Visual SLAM | Yousif et al. (2015) |
| 2016 | Observability, convergence, consistency (feature and graph-based SLAM) | Huang and Dissanayake (2016) |
| | Multirobot SLAM | Saeedi et al. (2016) |
| | Visual place recognition | Lowry et al. (2016) |
| | SLAM literature overview | Cadena et al. (2016) |
| 2017 | Autonomous vehicles | Bresson et al. (2017) |
| 2018 | | Kuutti et al. (2018) |
| 2018 | AI for long-term autonomy | Kunze et al. (2018) |
| | Long-term sensorization | Zaffar et al. (2018) |
| 2020 | Deep learning | Fayyad et al. (2020) |
| | Multirobot search and rescue | Queralta et al. (2020) |
| 2021 | Self-driving vehicles | Badue et al. (2021) |
| 2022 | Underground navigation | Ebadi et al. (2022) |

[1] https://github.com/sousarbarb/slr-ltlm-mr

odometry for estimating relative motion from visual data. The tutorial covers the topics of camera modeling and calibration, motion estimation, and feature matching. Yousif et al. (2015) extend the previous visual odometry tutorial to include methodologies for vision-based SLAM. Although the introductions mentioned here provide comprehensive explanations of the problem formulation itself, none of these introductions focus the discussion on possible challenges of performing long-term SLAM.

Moreover, other studies focus on theoretical aspects of the SLAM formulation. Dissanayake et al. (2011) discuss the fundamental properties of SLAM, namely, observability (if the problem is solvable), convergence (if the state uncertainty converges to a finite value), and consistency (if the estimated state is unbiased). Huang and Dissanayake (2016) extend the previous work to provide an in-depth explanation of the fundamental properties. The study defines also criteria for the performance evaluation of SLAM algorithms in terms of consistency, accuracy, and computational efficiency. In contrast to Dissanayake et al. (2011) that focus mainly on filtering-based SLAM, Huang and Dissanayake (2016) also discuss the properties in the context of smoothing approaches. Both studies focus only on theoretical aspects, not discussing the problem of long-term localization and mapping.

In terms of surveys in the literature on SLAM trends, Cadena et al. (2016) provide a broad overview of metric and semantic SLAM works. The study also discusses localization and mapping robustness in terms of loop closure validation and dealing with a dynamic environment, the SLAM scalability concerning pose graph sparsification, and parallel and distributed computing. On the contrary, Lowry et al. (2016) focus on topological SLAM. The survey provides a comprehensive review of visual place recognition. Although the study discusses the challenges of navigating a robot in varying conditions, the discussion is limited to vision sensors. Bresson et al. (2017) survey trends regarding single and multivehicle SLAM and large-scale experiments for autonomous vehicles. The study compares the methods over accuracy, scalability, availability, recovery, map updatability, and scene dynamicity. However, Bresson et al. (2017) only refer to approaches composed at least by odometry and a mapping module, not discussing localization-only algorithms. Also, the discussion is more focused on loop closure and relocalization modules and leveraging existing data, not on the methodologies for dealing with the challenges of performing long-term SLAM. Similarly to Bresson et al. (2017), Kuutti et al. (2018) and Badue et al. (2021) analyze trends for self-driving vehicles. While Kuutti et al. (2018) focus on sensorization and cooperative localization between vehicles, Badue et al. (2021) concentrate on the architecture of autonomous driving systems. However, none of two studies discuss challenges to achieve lifelong SLAM. Saeedi et al. (2016) present a review on multirobot SLAM discussing several solutions and techniques. The review identifies large-scale and dynamic environments, multisession, and agent scalability as challenges for multirobot SLAM. Still, the study does not overview existent methodologies to tackle those problems. Paull et al. (2014) and Ebadi et al. (2022) also overview the SLAM literature but are more specific in terms of the domain. The former focuses on autonomous underwater navigation, and the latter on SLAM in extreme underground environments. However, those two studies do not discuss long-term SLAM.

The surveys of Kunze et al. (2018) and Zaffar et al. (2018) focus on some challenges of long-term autonomy. Kunze et al. (2018) provide a brief overview on how AI can enable the long-term operation of autonomous systems. Although the survey discusses challenges such as environments with varying appearance and learning the dynamics of moving elements, the discussion is limited to AI-related works. Also, the survey only briefly analyzes the localization and mapping tasks due to also focusing on reasoning, human-robot interaction, and planning. As for Zaffar et al. (2018), the study only focuses on discussing and comparing different sensors in terms of sensor lifetime, field operability, ease-of-replacement, and suitability to different types of environment.

Nevertheless, none of the existent studies presented in Table 1 describe their methodology to select the works for discussion. This limitation does not allow other researchers to repeat the reviewing process for, for example, updating existent studies to maintain an up-to-date knowledge of the current state of the art in SLAM.

## 2.2 | Motivation and goals

This study reviews the literature on long-term localization and mapping for mobile robots. The review follows the PRISMA (Page et al., 2021) guidelines defining a systematic methodology to ensure the repeatability of the paper selection and data extraction processes. The repeatability of the systematic method followed in this review allows future updates on the discussion and the methodology presented in this study. Furthermore, the literature review presented in this paper does not focus on any specific trend related to long-term SLAM or time interval of publication. These considerations improve the coverage of the review over the long-term localization and mapping topic.

In summary, this review intends to understand the following questions:

1. main challenges inherent to lifelong SLAM;
2. main strategies for accomplishing long-term operations with mobile robots;
3. public data sets commonly used for evaluating long-term localization and mapping algorithms;
4. how the researchers evaluate the performance of autonomous systems in long-term operations.

The review may be framed in the Population–Intervention–Comparison–Outcome (PICO) framework. The population item limits the coverage of the review to a given group of people or machines. The intervention defines which procedure or technology is under study in the review. The comparison item states alternative interventions to the one considered in the review. Then, the outcome defines the expected experimental results the review is interested in

(Borrego et al., 2014; Wohlin et al., 2012). One possible application of the PICO framework in literature reviews is to obtain an initial search query. The query would have an AND operator between items of the framework and OR within the keywords of each PICO item. (Freitas, 2014). The PICO template specific to this review's topic is the following one:

1. Population: mobile robots;
2. Intervention: localization, mapping, SLAM;
3. Comparison: *not applicable to this study*;
4. Outcome: long-term operation, lifelong autonomy.

Lastly, this review does not intend to review the fundamentals of SLAM nor the main formulations. The reader should refer to Durrant-Whyte and Bailey (2006) and Bailey and Durrant-Whyte (2006) for the EKF and particle filter probabilistic formulations, and to Grisetti et al. (2010) for the graph formulation of SLAM.

## 3 | METHODOLOGY

A systematic literature review uses explicit, rigorous, and reproducible systematic methods to synthesize the findings of studies related to a particular research question, topic area, or phenomenon of interest. This type of review assures the quality and trustworthiness of the review's findings by presenting a complete, organized, and summarized analysis of all works considered while allowing others to replicate or update the reviews. The most common standard for performing a systematic review is the PRISMA (Page et al., 2021) statement. Although the PRISMA statement has been designed originally for evaluating the effects of health interventions, the checklist items of the methodology are general and applicable to other subject areas (Benos et al., 2021; Chukwu & Garg, 2020; Moayedi et al., 2020; Qureshi et al., 2020). Thus, the methodology used in this systematic review follows the PRISMA (Page et al., 2021) guidelines.

This section presents the detailed methodology used in this study. First, the eligibility criteria determine the studies to include in the review. Next, the search strategy details the information sources considered in the review, the base query string, and the search fields used to perform the inquiry in each source. The paper selection process focuses on describing its stages following the PRISMA statement. This process includes the quality evaluation criteria used to select works for the synthesis and analysis phase of the review. Lastly, the data extraction process details the data collected for synthesis and analysis. Parsifal (Freitas, 2014) is the online tool used to support the literature review in designing the methodology protocol, removing duplicates, screening and selecting works including their quality assessment. Additional documentation and scripts developed within the scope of this review related to removing duplicates, checking and processing the bibliographic references, and data extraction are available in the public GitHub repository (see footnote 1).

### 3.1 | Eligibility criteria

Table 2 presents the exclusion criteria used to determine the eligible studies for the selection process. These eligibility criteria focus on the type of paper and availability. The index criterion rejects all publications not indexed in a scientific publication venue. This rejection guarantees that the eligible works were peer-reviewed by the scientific community. Also, the exclusion criteria reject short papers and gray, secondary, and tertiary literature. Short papers do not usually present a detailed methodology of their scientific contribution. As for only considering primary literature in the review, this criterion increases the relevance of search results by favoring original articles and ensuring peer-review of the works. In terms of language, only considering studies with English full-texts increases the scope and visibility of the review. Also, the eligibility criteria reject studies not available in digital libraries for reproducibility and accessibility reasons.

Another exclusion criterion considered in the review is relative to the studies' categorization of their subject areas by bibliographic databases. The ones considered in the review are Computer Science, Engineering, Mathematics, or Multidisciplinary areas. In the list provided by the Clarivate's Journal Citation Reports,[2] these four subject areas include the AI, interdisciplinary applications, electrical and computers engineering, robotics, and applied mathematics categories, among others. These categories are intrinsically related to the localization and mapping problem for long-term operation of mobile robots.

The final three criteria presented in Table 2 focus on the scientific contribution of the studies. The data set criterion rejects all works that focus only on sharing a data collection. Data sets are important for the evolution of localization and mapping algorithms in providing a benchmark for comparison and reference purposes. However, their scientific contribution is not directly comparable to research articles in the context of a literature review. Odometry-only approaches are unusable over long distances invalidating their use for long-term operations with mobile robots. As for the scope criterion, this review rejects papers that pass the other exclusion criteria but are not related to long-term localization and mapping.

### 3.2 | Search strategy

The search phase consists of identifying the data sources that could be relevant to this literature review. The search strategy also defines the base string and which fields to consider in the search process to obtain the records included in this review. Web of Science and Scopus are traditionally the two most widely used bibliographic databases. However, previous studies demonstrate that different databases differ significantly in their scientific coverage (Mongeon & Paul-Hus, 2016; V. K. Singh et al., 2021). Thus, the data sources

[2] https://jcr.clarivate.com/jcr/browse-categories

**TABLE 2** Exclusion criteria for the selection process.

| E# | Criteria | Statement |
|---|---|---|
| E1 | Index | Papers not indexed in a scientific publication venue |
| E2 | Language | Full-text of the papers not published in English |
| E3 | Subject Area | Papers not classified in the databases as Computer Science, Engineering, Mathematics, or Multidisciplinary |
| E4 | Short Papers | Papers classified as short papers accordingly to the publication venue |
| E5 | Gray, Secondary, and Tertiary Literature | Books, preprints, reports, reviews, thesis, … |
| E6 | Availability | Full-text of the papers not available in digital libraries |
| E7 | Data set | Papers that focus only on data collection |
| E8 | Coverage | Papers using only odometry for localization |
| E9 | Scope | Papers that focus on different and not related subjects |

considered in this review are the following ones: ACM Digital Library, Dimensions, IEEE Xplore, INSPEC, Scopus, and Web of Science.

Moreover, May 17, 2022, is the date of the last full inquiry. Future reviews on the topic of this study should consider this final date as their initial one. As for the inquiry process in the data sources, the base string used is the following one:

```
(robot* OR vehicle*)AND
((locali* AND map*) OR slam")AND
("long term" OR life long" OR lifelong)
```

The first terms, `robot* OR vehicle*`, attempt to focus the search results on the population defined in the PICO template (see Section 2.2). The two terms have multiple synonyms within the scope of autonomous mobile robots: mobile robots, autonomous vehicles, robotics, agricultural robots, intelligent robots, service robots, unmanned aerial/ground/underwater vehicles, among other terms. Therefore, by adding the asterisk to the end of the terms robot and vehicle (`robot*` and `vehicle*`, respectively), and by only considering the terms with asterisk in the inquiry, all the synonyms are covered for the desired population. Given the incompatibility of the Dimensions database with wildcards (e.g., using the asterisk), the first part of the base string becomes as follows when searching in this database: `robot OR robots OR robotics OR vehicle OR vehicles`.

The next part of the query focus on the intervention side of the systematic review. Given the interest of this review on searching for localization and mapping algorithms, `locali*` and `map*` summarize all the synonyms for the localization and mapping terms, respectively. For example, `locali*` not only is agnostic to the US versus UK spelling differences (localization vs. localisation, respectively) but also resumes several synonyms: localization, localize, or localizing. The term `map*` also attempts to cover its respective synonyms such as map, maps, or mapping. Also, the acronym `"slam"` is another alternative to search for localization and mapping algorithms. Even though its definition is compatible with `locali* AND map*`, some authors only refer to SLAM. Similarly to the inquiry's first part, the second one becomes as follows when searching in Dimensions:

```
((localize OR localization OR localizing OR localise OR localisation OR localising) AND (map OR maps OR mapping)) OR slam".
```

As for `"long term" OR life long" OR lifelong`, this part of the base string is relative to the outcome of the PICO framework. The reason for having both `"life long"` and `lifelong` terms is the existing confusion in which term is grammatically the correct one. Also, the use of double quotation marks in the term `"long term"` searches for approximate phrases. Indeed, using `"long term"` in the query will retrieve results with either *long term* or *long-term* terms.[3]

The Title, Abstract, and Keywords are the fields selected for obtaining the search results. The third one includes the author keywords, the indexed terms by the databases, and the uncontrolled ones, if available. The selection of these search fields for this review improves the relevance of the results compared to using all fields and searching in the full text. The selected search fields focus the search on the summary items of the works. Indeed, the main contributions of scientific works should be summarized in at least the title, abstract, or author keywords. The indexed terms also help to obtain records only related to the base string used in this review. However, not all data sources have available the search fields considered in the review. Some of the data sources require an adaptation when performing the search.

Furthermore, the ACM Digital Library allows searching within multiple search fields, including the ones considered in this review. However, the advanced search query on this library sets by default an AND operator between the different fields. The default setting must be changed manually in the query syntax to the desired OR operator. Also, there are two options to search items in the ACM Digital Library: *The ACM Full-Text Collection* and *The ACM Guide to Computing Literature*. Given that the latter includes all the content from the former, the identification process in this source performs

[3] https://service.elsevier.com/app/answers/detail/a_id/34325/

the search using *The ACM Guide to Computing Literature* option. Other than searching in the publications' full data, Dimensions only has the title and abstract search fields compatible with this review. IEEE Xplore imposes a limit to seven wildcards in the search query. Consequently, the search results of this digital library using the base string for the inquiry are the grouping of different searches considering only one search field at a time. The results relative to each search field are imported to Parsifal to remove the duplicates. As for INSPEC, Scopus, and Web of Science, these databases have available all the search fields considered in the review.

In terms of the publication date, this review does not restrict it to avoid ignoring important works and improve the discussion. Indeed, to best of our knowledge, there is not available in the literature a systematic review on long-term localization and mapping for mobile robots covering a certain time window, and thus, to provide an initial date for rejecting older publications. Even though the number of publications per year could indicate an initial date on when the topic gained relevance, the date filtering could still reject important works.

## 3.3 | Selection process

The selection process of this review summarized in Figure 1 has three phases: identification, screening, and quality assessment. The first phase consists of inquiring each data source discussed previously with the base string and adapting it if needed. The second phase requires screening the papers. In this review, screening is equivalent to reading the publications' title and abstract and deciding whether the study is eligible or not based on the exclusion criteria. Then, a set of evaluation criteria assesses the quality of the eligible records. The records obtained after the three phases of the selection process are the ones accepted for the data extraction phase.

### 3.3.1 | Identification

In the identification phase of this review, the search strategy is applied to all data sources. ACM Digital Library (n.d.), Dimensions (n.d.), INSPEC (n.d.), Scopus (n.d.), and Web of Science (n.d.) data sources only require a single inquiry to obtain the search results. Given the limitation of the IEEE Xplore (n.d.) for using wildcards mentioned in Section 3.2, the number of records for this source presented in Figure 1 represents the results of seven inquiries (using the fields title, abstract, author keywords, IEEE terms, INSPEC controlled terms, and the INSPEC uncontrolled ones, respectively) after removing the duplicates with the support of Parsifal. Although the total number of search results found is 2160, Parsifal is used to remove duplicates from different data sources, excluding 1339 records. Following the duplicates removal, the exclusion criteria defined in Section 3.2 exclude 232 works from the review. This

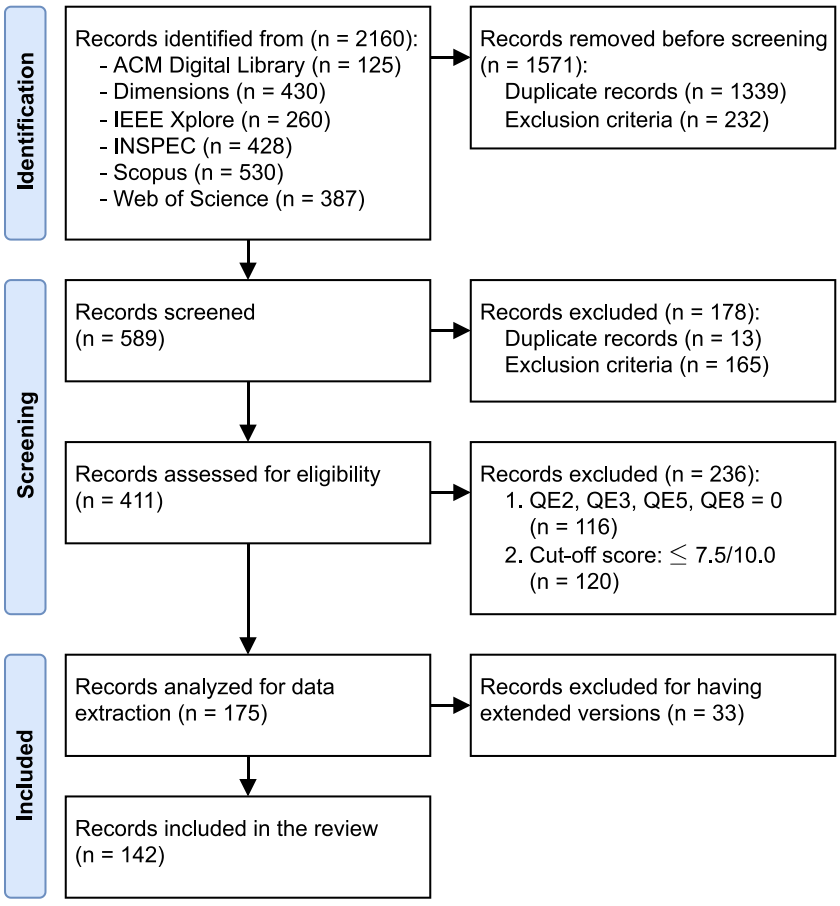


**FIGURE 1** PRISMA flow diagram for the selection process. [Color figure can be viewed at wileyonlinelibrary.com]

exclusion is possible due to INSPEC (n.d.), Scopus (n.d.), or Web of Science (n.d.) having filters related to the publication's type, subject area, and language.

The works excluded from the search results also include the ones that do not meet the exclusion criteria E4 and E7. For the first one, a Python script available in the GitHub repository[1] of this review searches studies with four or fewer pages. Even though short papers have a maximum number of three pages, the papers with four pages do not usually present a detailed methodology. As for the E7 exclusion criterion, some works are removed from the review by searching in their title for the term "data set." All articles excluded in this review have gone through a double-check procedure, including the records rejected by the Python script intended for the exclusion criterion E4. The double-check procedure is to verify if the exclusion criteria are correctly applied and rejected articles are not related to the topic of long-term SLAM. For example, articles published in the Remote Sensing journal from MDPI do not meet the E3 criterion. Indeed, the Journal Citations Reports from Clarivate classifies it by the following categories: Remote Sensing, Geosciences Multidisciplinary, Environmental Sciences, and Imaging Science & Photographic Technology. However, most search results from this journal found in the identification phase are directly related to the topic of this review and the respective subject areas. Thus, in these cases and other ones related to the remaining exclusion criteria, the decision is reverted to consider the initially rejected studies for the next phase of the review.

### 3.3.2 | Screening

Next, the screening phase in this review consists of reading the title and abstract of the publications and rejecting the ones that meet the exclusion criteria. However, the initially rejected papers have another assessment for validating the exclusion. This assessment includes the analysis of the results and conclusions to either confirm the exclusion decision or reverse it to eligible works for the quality assessment phase. As a result of the screening phase, 178 studies are rejected from the initially identified 589 works. The duplicate records found in screening and removed manually are due to titles with invalid characters originated by exporting the search results from the Dimensions (n.d.) database.

### 3.3.3 | Quality assessment

The quality evaluation of 411 eligible works for this review follows the 8 Quality Evaluation (QE) criteria presented in Table 3. All of them are subjective criteria derived from the analysis of the eligible works. The score column establishes the possible values for the QE criteria. The minimum, intermediate, and maximum values correspond to eligible works being none, partial, and fully compliant with each criterion, respectively. Furthermore, QE1, QE2, QE4, and QE8 focus on the details provided in the papers. More specifically, if the discussion of the related work, the proposed methodology, the experimental setup, and the results are detailed and thoroughly analyzed in the papers, respectively. The possible scores for QE3 are twice the value of QE1, QE2, QE4, and QE8. This difference is due to QE3 being related to the topic of this review. A work focusing on both localization and mapping problems will have a score of 2.0 (full compliance). If the study only focuses on one of these problems or none, the scores will be 1.0 or 0.0 (partial or no compliance, respectively). QE5 evaluates the long-term results of the eligible studies. This criterion is either 2.0 (full) or 0.0 (no compliance). QE5 has the same range as QE3 for similar reasons, given the focus of this review on long-term localization and mapping algorithms. The definition of long-term experiments for assigning full compliance in QE5 is the following one: dynamic changing environments (e.g., dynamic elements or semi-static ones), increasing environments or feature maps in terms of their size, redundant data removal, or varying conditions (e.g., different seasons of the year or lighting conditions). QE6 and QE7 can only be 1.0 or 0.0. QE6 intends to highlight works that compare their experimental results to the state of the art and/or ground-truth data. QE7 emphasizes the importance of having available either the implementation of the proposed

**TABLE 3** Quality evaluation criteria and score range.

| QE# | Criteria | Score |
|---|---|---|
| QE1 | Does the paper have an updated state of the art on long-term localization and mapping? | {0.0, 0.5, 1.0} |
| QE2 | Is the methodology appropriate and detailed? | {0.0, 0.5, 1.0} |
| QE3 | Does the methodology consider both localization and mapping problems? | {0.0, 1.0, 2.0} |
| QE4 | Is the hardware and/or software used in the experiments detailed? | {0.0, 0.5, 1.0} |
| QE5 | Does the paper presents any kind of long-term experimental results? | {0.0, 2.0} |
| QE6 | Does the paper presents comparative results with other methods and/or ground-truth data? | {0.0, 1.0} |
| QE7 | Does the work's implementation and/or the data used in the experiments are publicly available? | {0.0, 1.0} |
| QE8 | Is the discussion of the results and conclusions appropriate and detailed? | {0.0, 0.5, 1.0} |

methodology or the data used in the experiments. The data and/or implementation availability would allow other works to perform comparisons of the proposed methodologies. Lastly, considering the possible scores for the QE criteria in Table 3, each work can only have a maximum score of 10.0.

After evaluating the 411 eligible works accordingly to the previously discussed QE criteria (the scores of each record are available in the GitHub repository [see footnote 1]), our first conclusion is that works with a nondetailed or not appropriate methodology, results' discussion, or conclusions should not be included in the review. Another conclusion is relative to rejecting works that do not consider either localization or mapping problems, or do not present any long-term experimental results. This rejection is due to the focus of this review on the long-term localization and mapping problem for mobile robots. Furthermore, the quality assessment phase should consider a cut-off score to filter works with low quality scores. Consequently, the assessment phase considers the following two reasons to reject a record:

1. QE2, QE3, QE5, QE8: reject works with a 0.0 (no compliance) score;
2. cut-off score: reject works with a score lower or equal to 7.5/10.0.

The distribution of the evaluation scores and the QE criteria itself justify the selection of a 7.5/10.0 cut-off score. Figure 2 illustrates the scores distribution for all eligible works versus the scores of the ones that pass the first criterion defined previously for the QE phase (related to the compliance on the QE2, QE3, QE5, and QE8 criteria). The assessment of this criterion rejects 116 records (28%) of the 411 eligible works (see Figure 1). Even though the distribution of the evaluation scores changes significantly in the range of scores lower or equal to 7.5/10.0, as observed between Figure 2a,b, only one work with a score higher than 7.5 is rejected due to not having a detailed and appropriate discussion of the results. This result indicates that interesting works are associated with high scores, as intended when using a quality assessment methodology. The result also suggests that the range between 8.0 and 10.0 have the most interesting and quality works compatible with the focus of this review on long-term localization and mapping. Although only assessing the eligible works would seem to lead to the same results in terms of records included in the review, the rejection criterion on QE2/3/5/8 prevents outliers related to the quality assessment. From the remaining 295 eligible works, cut-off scores from 7.5 up to 8.5 have the following corresponding rejection rates:

1. 7.5/10.0 → reject 120 records (40.7%) → include 175 records
2. 8.0/10.0 → reject 160 records (54.2%) → include 135 records
3. 8.5/10.0 → reject 203 records (68.8%) → include 92 records

The 8.5 cut-off score would not be suitable because methods that focus only on localization or mapping, or not having either the implementation or the experimental data publicly available would be obligated to have maximum scores in the other criteria to be included in the review. In these cases, a work would have a maximum score of 9.0 due to partial compliance on QE3 or no compliance on the QE7 criteria. Likewise, a cut-off score of 8.0 would only leave a margin for having a single partial compliance on QE1, QE2, QE4 or QE8 criteria in similar cases, even though it would reject 160/295 (54%) records. Therefore, the 7.5/10.0 cut-off score is more appropriate for the quality assessment phase in this review by leaving margin for works to have partial compliance in more than one criterion. Indeed, this cut-off score allows an article with no public data and/or implementation (e.g., due to confidentiality agreements) to have up to four criteria with partial compliance, depending on the criterion's maximum score or if the work has available the experiments data and/or implementation. Another example is articles that only focus on localization or mapping. In these cases, the work could have no public implementation, even though requiring a maximum score on all other criteria. Also, if the work has public data or implementation available, two other criteria could have partial compliance.

Overall, as illustrated in Figure 1, the quality assessment of the 411 eligible works considering the two criteria mentioned previously leads to rejecting a total of 236 (57%) records. As a result, the remaining 175 records will be analyzed for data extraction.

## 3.4 | Data extraction

The data extraction process analyzes the records selected after the quality assessment phase and extracts information from these works. In the scope of this review, the Data Extraction (DE) items required for each record are the following ones:

- [DE1] Long-term considerations—long-term factors the works consider in their proposed approach and experiments. Considering the knowledge obtained in the previous phases of this review's methodology, we considered the following factors for categorizing the included works:
  - appearance: varying conditions, appearance changes;
  - dynamics: environment dynamics, dynamic elements;
  - sparsity: map pruning, redundant data removal;
  - multisession: map management;
  - computational: memory management, efficiency.
- [DE2] Localization—how the robot localizes itself and the type of localizer;
- [DE3] Mapping—type of the map;
- [DE4] Multirobot—if the proposed methodologies consider multi-robot systems;
- [DE5] Execution mode—offline, online, if requires both, or if no information on this item;
- [DE6] Environment and domain—type of environment (indoor, outdoor) and domains (air, ground, water) tested with the proposed methodologies;
- [DE7] Sensory setup—which sensors considered in the methodologies;

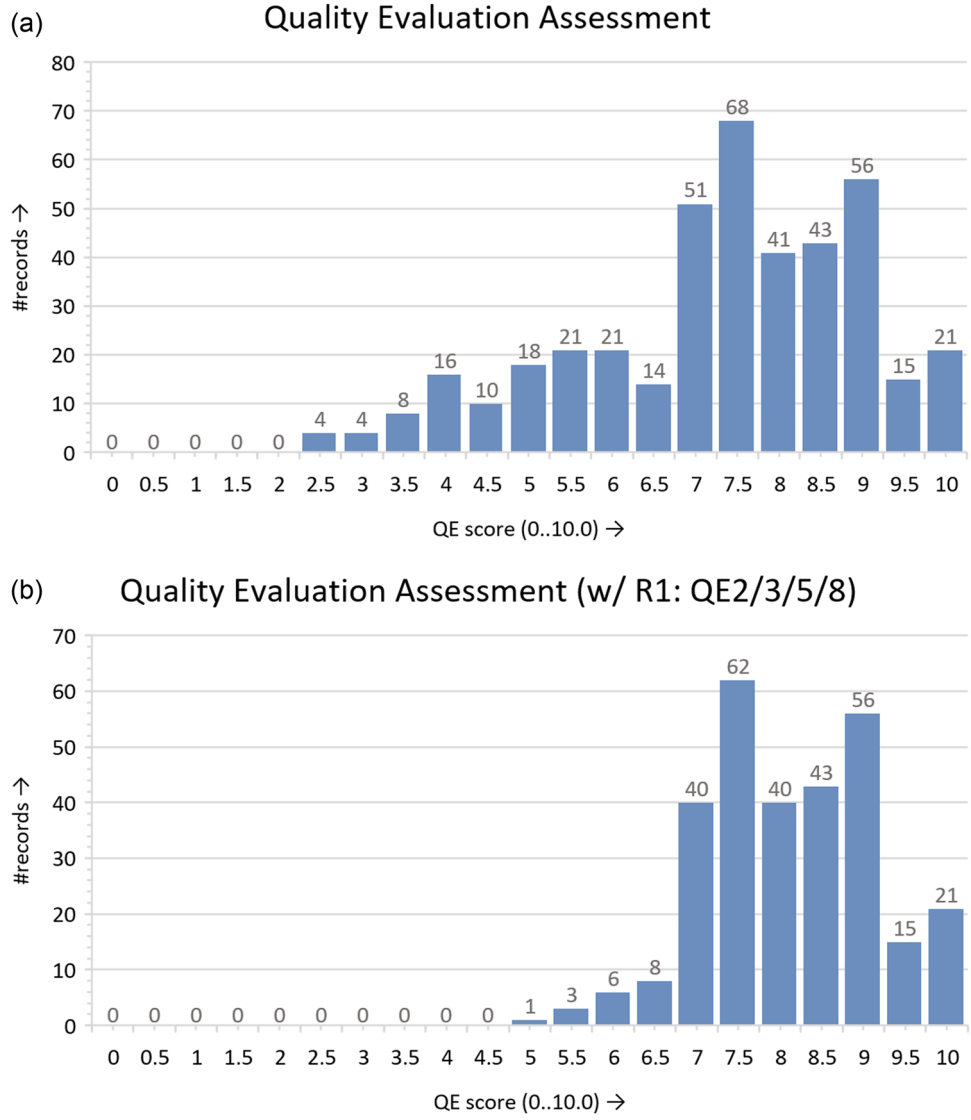


**FIGURE 2** Distribution of the quality evaluation scores obtained from assessing the eligible works considered in the review: (a) all eligible works; (b) works that pass the rejection criterion during the QE assessment related to QE2/3/5/8 = 0.0 (no compliance). [Color figure can be viewed at wileyonlinelibrary.com]

- [DE8] Nonpublic experiments—if the authors performed experiments or tests with nonpublic data;
- [DE9] Ground-truth—how ground-truth for nonpublic data is obtained or its type, if available;
- [DE10] Distance and time characteristics—relative to the nonpublic experiments if available, as follows:
  - total distance (km) of the nonpublic experiments;
  - path (km), in the case of repetitive paths;
  - total time (h) in terms of continuous operation;
  - time interval (day/week/month/year, or d/w/m/y) between the first and the last run.
- Data sets—if and which public data sets are used in the experiments;
- Evaluation metrics—which metrics are used for evaluation.

In Section 5.6, a comparison table of the public data sets identified by the DE11 item will contain the sensory setup, ground-truth data availability from the data sets, and the distance and time characteristics. The columns of the data sets comparison table are similar to the data extraction items for nonpublic data, among other aspects. As a result, the distinction between public and nonpublic data availability represented in DE8, DE9, and DE10 allows presenting the distance and time characteristics of nonpublic data independently from the public data sets.

Although the data extraction phase in a systematic literature review usually does not remove any records, 33 of the analyzed 175 works have extended versions of the proposed methodologies, more detailed ones, or equivalent methods applied in different conditions. Thus, these records are not included in the review to improve the

discussion section in terms of singularity and originality of proposed approaches for the long-term localization and mapping problem. The extracted information helped identifying the corresponding extended and more complete versions of these works. A document containing the association of the removed versions to the records included in the review is available in the public GitHub repository[1], including their bibliographic references. Consequently, 142 original works are included in this review for an overview of these records in Section 4, and their synthesis and discussion in Section 5. The information relative to the 12 data items for each of the included records is available in Appendix A and also in the repository[1]. The included works represent 34.55% of the 411 eligible records for this review. This result indicates that the methodology followed in this review led to a high percentage of quality results.

## 4 | RESULTS OVERVIEW

In this section, the main goal is to overview the results in terms of their bibliographic data, not relative to their scientific contribution. First, a statistical analysis assesses the coverage of the data sources considered in this review. Next, the tool VOSviewer (van Eck & Waltman, 2010, 2014) is used to obtain the co-occurrence results for the keywords and the authors. The former focuses on the keywords' recency and their occurrence in the 142 included works. The latter discusses the research networks between the authors and the ones with more publications in long-term localization and mapping. Lastly, two analyses are presented relative to the evolution of the records' publication year and most relevant publication venues.

### 4.1 | Data source

The results on the identification phase are exported to BibTeX files from each data source. This exportation considers all the information available in the data sources, such as citation (e.g., author, title, publication venue, and type of record) and bibliographic (e.g., affiliation and the publisher) information of each record, the abstract, and author and indexed keywords. Next, using the `bibtexparser`[4] Python library, the BibTeX files are processed to identify incomplete records. For example, the DOI must be specified and, if not available, the record's information must be manually completed with a corresponding URL. Then, considering the 142 included records in this review, a Python script searches each record in the BibTeX files corresponding to each data source. This search uses the DOI, URL, and title data to identify if a data source had in its identification results the searched record. Given that these three data fields can contain lower and upper letters, the respective strings must be compared only after converting them to lower cases. As a result, the number of identified records by each data source of the 142 included ones in the review are the following ones:

1. ACM Digital Library (n.d.): 25 records (17.6%);
2. Dimensions (n.d.): 84 records (59.2%);
3. IEEE Xplore (n.d.): 67 records (47.2%);
4. INSPEC (n.d.): 102 records (71.8%);
5. Scopus (n.d.): 120 records (84.5%);
6. Web of Science (n.d.): 105 records (73.9%).

The database Scopus (n.d.) is the source that identifies the greatest number of included records in this review. This result is expected given that Scopus (n.d.) is considered as one of the largest curated databases (V. K. Singh et al., 2021), indexing more than 25,000 active titles (e.g., conferences proceedings, journals) and 7000 publishers.[5] Two other sources with more than 70% of identified records are INSPEC (n.d.) and Web of Science (n.d.). Similarly to Scopus (n.d.), these two databases index also records from thousands of journals, conferences, and publishers.[6,7] Although Dimensions (n.d.) is also a bibliographic database covering millions of publications from thousands of sources, this database is the newest one (created in 2018) relative to the other three considered in this review (INSPEC (n.d.), Scopus (n.d.), and Web of Science (n.d.)). The fact of being the newest database could be a factor to why Dimensions (n.d.) obtained a lower percentage (59.2%) than the other three databases. Another possible reason is that Scopus (n.d.) and Web of Science (n.d.) have the majority of their coverage in Life Sciences, Physical Sciences, and Technology Area (including the Engineering subject area related to the topic of this review), while Dimensions (n.d.) has better coverage in Social Sciences and Arts & Humanities (V. K. Singh et al., 2021). Even though IEEE Xplore (n.d.) is a digital library and only indexes works published by IEEE and its partners, this data source returns 47.2% of the include records in the review. The main reason is that this library indexes publications related to electrical engineering and computer science.[8] These two subject areas are related to the long-term localization and mapping topic. Finally, the ACM Digital Library (n.d.) using *The ACM Guide to Computing Literature* collection only finds records published by ACM, and possibly links to other records in the literature focused exclusively on computing,[9] not directly related to the Computer Science or Engineering subject areas. This limitation of the ACM Digital Library (n.d.) explains why this source obtained a lower coverage percentage than the other sources in this review.

Furthermore, Table 4 presents a coverage analysis of the identified results from each data source for the 142 included records in this review. Table 4a presents the pairwise overlap between sources. The corresponding percentage is the ratio of records identified by both sources to the one between the two that has

[4] https://bibtexparser.readthedocs.io/en/master/

[5] https://www.elsevier.com/solutions/scopus/how-scopus-works
[6] https://www.elsevier.com/solutions/engineering-village/content/inspec
[7] https://clarivate.com/webofsciencegroup/solutions/web-of-science/
[8] https://ieeexplore.ieee.org/Xplorehelp/overview-of-ieee-xplore/about-ieee-xplore
[9] https://libraries.acm.org/digital-library/acm-guide-to-computing-literature

**TABLE 4** Pairwise coverage analysis of the data sources considered in the review over the 142 included records: (a) identification only on both pairwise sources (#{$A \cap B$}/min{#$A$, #$B$}); (b) on either ones (#{$A \cup B$}/#records).

| (a) | | | | | | |
|---|---|---|---|---|---|---|
| $A \cap B$ | acm | dim | ieee | insp | scop | wos |
| acm | – | 96.0% | 44.0% | 88.0% | 96.0% | 96.0% |
| dim | – | – | 68.7% | 77.4% | 97.6% | 96.4% |
| ieee | – | – | – | 89.6% | 91.0% | 74.6% |
| insp | – | – | – | – | 87.3% | 69.6% |
| scp | – | – | – | – | – | 89.5% |
| wos | – | – | – | – | – | – |
| **(b)** | | | | | | |
| $A \cup B$ | acm | dim | ieee | insp | scop | wos |
| acm | – | 59.9% | 57.0% | 73.9% | 85.2% | 74.6% |
| dim | – | – | 73.9% | 85.2% | 85.9% | 76.1% |
| ieee | – | – | – | 76.8% | 88.7% | 85.9% |
| insp | – | – | – | – | 93.7% | 95.8% |
| scp | – | – | – | – | – | 92.3% |
| wos | – | – | – | – | – | – |

Abbreviations: dim, Dimensions (n.d.); ieee, IEEE Xplore (n.d.); insp, INSPEC (n.d.); scop, Scopus (n.d.); wos, Web of Science (n.d.).

the smallest number of results: #{$A \cap B$}/min{#$A$, #$B$}, where #$A$ and #$B$ is the number of results for a data source $A$ and $B$, respectively, and #{$A \cap B$} is the intersection results between the two sources. For example, if the pairwise results is 100%, it means that the data source with more records found was capable of obtaining all the results, that is, had full coverage over the other source. Table 4b reports the percentage of records identified by at least one of two data sources over all 142 included records: #{$A \cup B$}/142, where $A \cup B$ is the union correspondence results of the sources $A$ and $B$. This percentage represents the joint coverage of two databases over the 142 included records.

Analyzing the coverage results in Table 4, the first observation is that the pairwise union results of two sources increase the independent coverage of each source. This observation validates the need identified in the methodology (see Section 3) to consider several data sources in the identification phase of a review. Moreover, the pairwise union coverage of INSPEC (n.d.), Scopus (n.d.), and Web of Science (n.d.) is greater than 90% of the included records. When evaluating the joint coverage of these three databases, they identify all 142 of the included records, that is, a 100% coverage. Although this result could indicate that those three sources guarantee full coverage of the long-term localization and mapping research topic, it is always advisable to consider as most as possible sources in the methodology. Another observation is relative to the overlap of Scopus (n.d.) with the other sources, which is greater than 85%. This overlap indicates that Scopus (n.d.) covers results not only on the topic of this review but also the results obtained by the other sources considered in the methodology. Lastly, INSPEC (n.d.) and Web of Science (n.d.) achieve a pairwise overlap percentage of 69.6% between themselves, while their union represents 95.8% of the included records. This discrepancy indicates that these two sources identify unique results between themselves. Indeed, INSPEC (n.d.) identifies 31/142 records not found by Web of Science (n.d.), and vice-versa for Web of Science (n.d.), with 34/142 unique records.

## 4.2 | Keywords co-occurrence

Next, VOSviewer (van Eck & Waltman, 2010, 2014) is used to analyze the co-occurrence of keywords in the included articles. This co-occurrence is the relatedness of items determined based on the number of documents in which the keywords occur together. For this analysis, first, a Python script processes the BibTeX file containing the citation and bibliographic information, the author and the indexed keywords, and the abstract of the records to join the author with the indexed keywords in the same `keywords` field. Then, an online tool[10] converts this processed BibTeX to a RIS file. Even though VOSviewer supports file types directly exported from Dimensions (n.d.), Scopus (n.d.), or Web of Science (n.d.) as input, none of these data sources obtained all the 142 included records of the review in the identification phase. Given that VOSviewer does not support BibTeX files, the conversion to a RIS file is required for usage as an input. The disadvantage of using this file format in VOSviewer is only allowing to perform co-occurrence of items (e.g., keywords or authors), while bibliographic data from Dimensions (n.d.), Scopus (n.d.), or Web of Science (n.d.) in CSV files would allow other analysis such as citation, co-citation, or bibliographic coupling. However, the creation of these CSV files follow different templates depending on the data source. So, RIS files allow the integration of all 142 included records for obtaining the two co-occurrence analysis presented in this review (namely, keywords and co-authorship).

In Figure 3a, the network presents the overlay visualization of the keywords co-occurrence in the included records weighted by the number of occurrences of each term, using full counting for the links' strength. The latter computes the strength of the links directly by the number of co-occurrences of the respective two terms. The overlay visualization colors the keywords differently according to the average publication year of the included records in which each of the keywords appears. This coloring allows analyzing which are the ones that are associated with the most recent publications. As for the keywords' weighting, the number of occurrences dictates the size of the circles. Furthermore, the minimum number of occurrences of a keywords set in VOSviewer for obtaining the graph is 5 originating the 34 keywords illustrated in Figure 3a. This parameter was selected for visualization purposes while also filtering uninteresting keywords. Similarly, setting the attraction and repulsion parameters to 2 and 0,

[10] https://www.bibtex.com/c/bibtex-to-ris-converter/

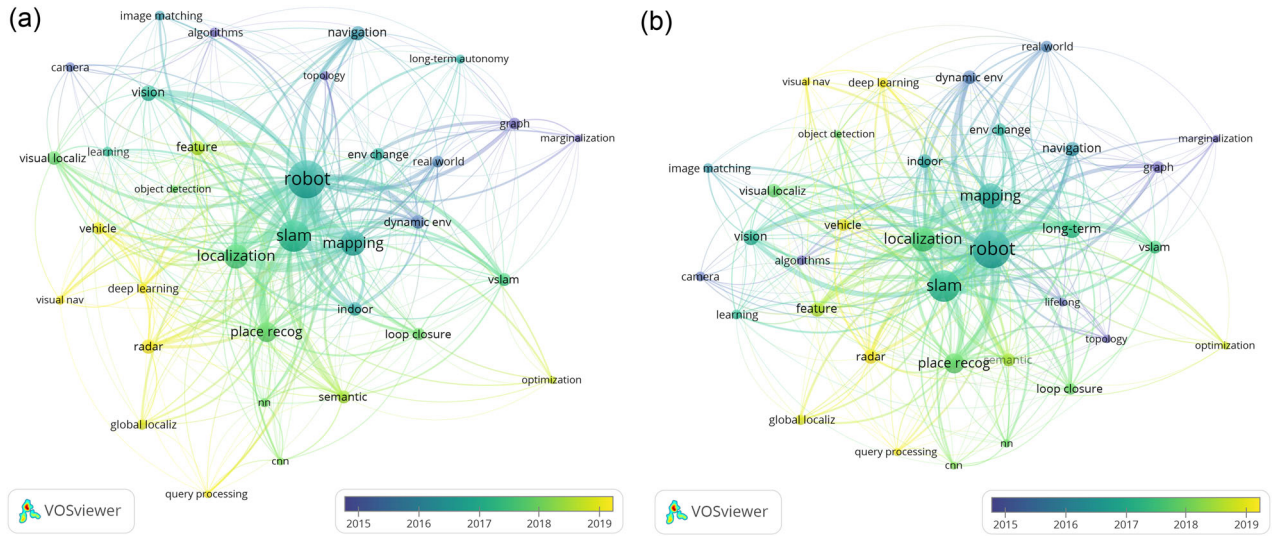


**FIGURE 3** Keywords co-occurrence analysis on the 142 included records generated by VOSviewer with overlay visualization by the average publication year: (a) original keywords; (b) all keywords containing long-term and lifelong summarized by the terms themselves. Parameters used for generating the co-occurrence network: minimum number of occurrences = 5, attraction = 2, repulsion = 0, scale = 1.49, circles size variation = 0.5, lines size validation = 1.0. `cnn`, Convolutional Neural Networks; `env`, environment; `localiz`, localization, `nav`, navigation; `nn`, Neural Networks, `recog`, recognition; `vslam`, visual SLAM. [Color figure can be viewed at wileyonlinelibrary.com]

respectively, distances the terms more from each other than using the values recommended in the VOSviewer manual[11] (2 and 1, respectively). These two parameters only interfere in the localization of the terms in the map, not in the graph connections. Lastly, a thesaurus of the keywords (available in the repository [1]) is used to join similar terms: spelling differences (e.g., localization – localisation), full terms versus abbreviations (SLAM), while also allowing the concatenation of long keywords for visualization reasons.

Overall, the keyword `robot` is the one that appears more times in the included records: 109 occurrences, links with 33 other terms, and has a total link strength of 390 (sum of co-occurrences of all of its links). This result is expected due to the relation of this review's topic to robotics. Similarly, three other keywords in the network related to long-term localization and mapping topic with high values of occurrence, number of links, and total link strength are `slam` (74, 33, and 280), `mapping` (47, 32, and 196), and `localization` (46, 31, and 188, respectively). The methodology for the search strategy discussed in Section 3.2 considers all of these four keywords. Thus, the significant influence of `robot`, `slam`, `mapping`, and `localization` in the keywords co-occurrence analysis indicates that, after all the phases executed in this review's methodology, the 142 included records have a high correlation with the keywords considered in the search query. Given that the keywords are usually selected or indexed to capture the essence of the document, this correlation indicates that the search query is appropriate to obtain the search results, even considering only the keywords as search fields.

As for keywords related to the outcome of the PICO framework, `longterm autonomy` occurs only 6 times in the included records, linking with 16 other keywords and having a total link strength of 27. This low occurrence could indicate that the term `longterm autonomy` is not usually used by the authors nor indexed by the databases. However, the specific term of `longterm autonomy` does not summarize all the possibilities for the outcome of the PICO framework (see Section 2). Indeed, for this reason, the search query for the identification phase uses only the following single terms: `"long term"` and `"life long"` (resumes the possibility of having a space or a hyphen), and `lifelong`. Figure 3b presents the keywords co-occurrence analysis using the same parameters for obtaining Figure 3a. The difference to the latter network is using a thesaurus that summarizes all the keywords that contain `longterm` and `lifelong` into the terms themselves, obtaining 35 keywords with a minimum of 5 occurrences in the 142 included records. In terms of occurrences, number of links, and total link strength, the impact of the thesaurus keyword `longterm` is 25, 27, and 103, and for lifelong 6, 17, and 31, respectively. These values are much higher than the ones respective only to `longterm autonomy` from Figure 3a. The reason is that `longterm` in Figure 3b compiles the occurrences of keywords such as `longterm autonomy`, `long term mapping`, and `longterm localization` (6, 2, and 2 occurrences, respectively), and `lifelong` sum up, for example, three different versions of `lifelong learning` (using `lifelong`, `lifelong` and `life long` with 2, 1, and 2 occurrences, respectively) and `lifelong slam` (1 occurrence). Hence, these results proves that the third `AND` part of the search query (`"long term" OR life long" OR lifelong`) covers well the PICO framework's outcome. Plus, they also show no consensus among the

[11] https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.8.pdf

authors and by the databases indexation on how to define a keyword for the topic of long-term localization and mapping.

In terms of the average year of publication, analyzing the diagrams in Figure 3 on its colorization, the first observation is the recency of terms related to visual localization. The keywords visual SLAM (vslam), visual navigation (visual nav), and visual localization (visual localiz) have all an average publication year higher than 2017. This recency indicates that recent approaches related to the topic of this review, long-term localization and mapping, are more inclined to use vision as a sensorization input. Another sensor that appeared with high relevance in the network is radar, with 15 occurrences and an average publication year of 2019.20. This sensor is agnostic to the environment changes such as illumination and season changes intrinsically associated with vision and could be the reason why the recent works related to long-term localization and mapping are using it. Moreover, place recognition (place recog) stands out not only by its recency but importance. The keyword itself (place recog) occurs 31 times and an average publication year of 2017.77, with terms related to place recognition such as loop closure and global localization (global localiz) with recent average publication years (2017.82 and 2018.75, respectively) and strong link to place recognition (five co-occurrences for each of the links between loop closure and global localiz with place recog). Lastly, machine learning also seems to be used in recent works included in this review. The keyword learning occurs seven times with an average publication year of 2017.00. Neural Networks (nn), Convolutional Neural Networks (cnn), and deep learning have a similar number of occurrences (6, 5, and 8) and publication years higher than 2017 (2017.83, 2018.00, and 2019.12, respectively). These results could mean a more recent trend of using machine learning to improve the long-term autonomy of mobile robots.

Although the recency of keywords related to dynamic environments is lower than 2017 (2015.50 and 2016.75 for dynamic env and env change), they have a high occurrence (14 and 12, respectively), located close to each other in the network, and have a strong link between them (4 co-occurrences). Three keywords also located near each other are graph and marginalization while having similar average publication years (2014.70 and 2014.60, respectively). Even though the number of occurrences of these terms is low (10 and 5 for graph and marginalization, respectively), their map proximity could indicate a focus in the past on the topic of graph sparsity, that is, maintaining the graph in the long-term to only depend on the environment size and not on the robot's operation time.

The keywords co-occurrence analysis also relates to the categories of DE1 (see Section 3.4). Works associated with place recognition, global localization, and loop closure terms require invariance to the appearance changes in the environment, equivalent to the appearance category. The dynamics category is associated with works focused on dynamic environments. As for the other group of keywords with a high occurrence and strong links between each other, the ones related to graph and marginalization, the respective works focus on removing uninformative data from the map (Kretzschmar & Stachniss, 2012), which is related to map sparsification, and so, to the sparsity category of DE1. These relations between the appearance, dynamics, and sparsity categories to the semantic analysis of the keywords co-occurrence supports the categorization of DE1 considered in this review, while also indicating that the discussion on the proposed methodologies should focus on each one of the categories. Even though the two remaining categories of DE1 (multisession and computational) are not represented in the keyword analysis, the execution of the data extraction phase identified the need for having these two categories, given the importance of multisession handling and computational efficiency for long-term localization and mapping. However, each category of DE1 will be discussed in Section 5 in further detail.

## 4.3 | Co-authorship analysis

The other analysis obtained using VOSviewer is the co-authorship network presented in Figure 4. Similar to the keywords network illustrated in Figure 3, the co-occurrence of the authors' names creates links among them in the graph. The strength of these links is dictated by the number of documents the two authors of a link are co-authors in the same record, and the number of co-authored works determines the size of the circles respective to each author in the graph. In contrast to Figure 3, the network in Figure 4 does not have any overlay specific to coloring depending on the average publication year. Instead, the main goal of the co-authorship analysis in this review is to present possible research networks detected in the 142 included records. Thus, the coloring in Figure 4 represents the clusters of authors detected by VOSviewer. This network only considers authors with a minimum of 3 works for relevance and visualization reasons, resulting in 27 authors. Also, authors identified only by the initial of the first name and by the surname can lead to incorrect correspondences in terms of co-authorship. VOSviewer detects 392 authors in the 142 included records using the original RIS file used in Section 4.2 compared with 413 after checking the authors names. Indeed, a manual check is performed on all authors of the included records to guarantee no false correspondences for the co-authorship analysis with VOSviewer. This manual check ensures each author has its full first and surname and any middle initials while also using the same name for an author in different records.

Analyzing Figure 4, the co-authorship network presents eight clusters. These clusters are separated from each other, that is, no link exists between authors from different clusters. However, this separation does not mean that there is not any co-authorship between authors from different clusters only indicating that for a minimum of three co-authored documents there is not a connection between the authors of these eight clusters. Even so, the graph presented in Figure 4 allows the identification of the most relevant research networks in terms of number of co-authored documents and in the context of long-term localization and mapping, considering the 142 records included in this review. As a results, the following

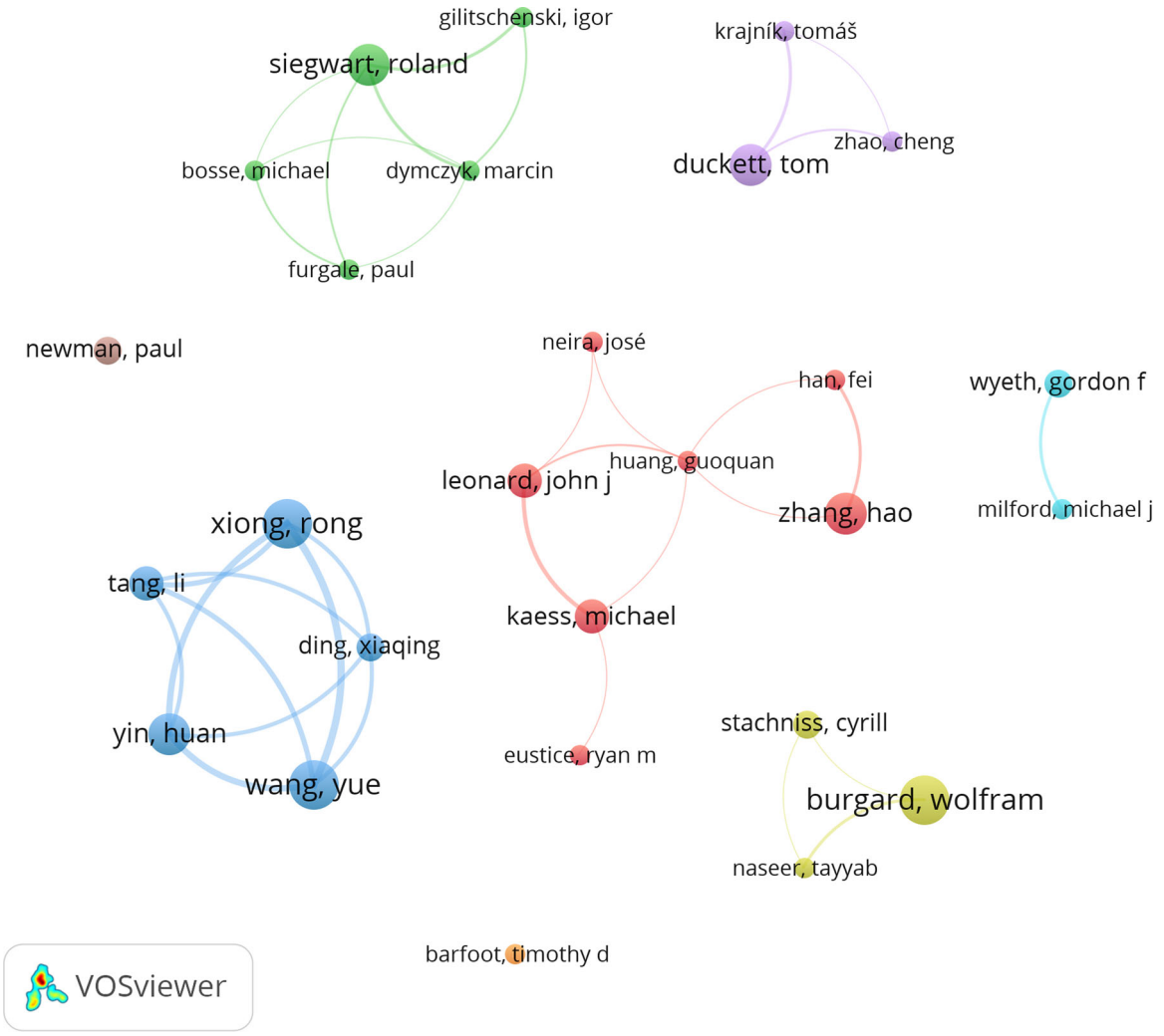


**FIGURE 4** Co-authorship analysis on the 142 included records generated by VOSviewer. Parameters used for generating the co-occurrence network: minimum number of occurrences = 3, attraction = 4, repulsion = −2, scale = 1.49, circles size variation = 1.0, lines size validation = 1.0. [Color figure can be viewed at wileyonlinelibrary.com]

enumeration presents the authors that belong to each cluster in the format of <author[ORCiD], [Google Scholar](#coauthored documents)>, where [] means options / if available:

1. Rong Xiong https://orcid.org/0000-0001-9318-9014, https://scholar.google.pt/citations?user=1hI9bqUAAAAJ (7), Yue Wang https://orcid.org/0000-0002-0981-935X, https://scholar.google.pt/citations?user=N543LSoAAAAJ (7), Huan Yin https://orcid.org/0000-0002-0872-8202, https://scholar.google.pt/citations?user=1fNc3vUAAAAJ (6), Li Tang https://orcid.org/0000-0003-2590-6872 (5), and Xiaqing Ding https://orcid.org/0000-0001-7802-0130, https://scholar.google.pt/citations?user=6u5OHUcAAAAJ (4);
2. Hao Zhang https://scholar.google.pt/citations?user=Ug2VxyUAAAAJ (6), John J. Leonard https://orcid.org/0000-0002-8863-6550, https://scholar.google.pt/citations?user=WPe7vWwAAAAJ (5), Michael Kaess https://scholar.google.pt/citations?user=27eupmsAAAAJ (5), Fei Han https://orcid.org/0000-0002-8619-3987 (3), Guoquan Huang https://scholar.google.pt/citations?user=trMUyZIAAAAJ (3), Jos? Neira https://orcid.org/0000-0003-0668-977X, https://scholar.google.pt/citations?user=scoMbR8AAAAJ (3), and Ryan M. Eustice https://orcid.org/0000-0002-9989-4942, https://scholar.google.pt/citations?user=WroYmiAAAAAJ (3);
3. Wolfram Burgard https://orcid.org/0000-0002-5680-6500, https://scholar.google.pt/citations?user=zj6FavAAAAAJ (7), Cyrill Stachniss https://orcid.org/0000-0003-1173-6972, https://scholar.google.pt/citations?user=8vib2IAAAAAJ (4), and Tayyab Naseer https://orcid.org/0000-0002-3350-3005, https://scholar.google.pt/citations?user=1FePZqEAAAAJ (3);
4. Roland Siegwart https://orcid.org/0000-0002-2760-7983, https://scholar.google.pt/citations?user=MDIyLnwAAAAJ (6), Igor Gilitschenski https://orcid.org/0000-0001-6426-365X, https://scholar.google.pt/citations?user=Nuw1Y4oAAAAJ (3), Marcin Dymczyk https://orcid.org/0000-0003-3667-8764, https://scholar.google.pt/citations?user=XYHy7U8AAAAJ (3),

Michael Bosse https://scholar.google.pt/citations?user=eopb1VgAAAAJ (3) and Paul Furgale https://orcid.org/0000-0002-7367-1046, https://scholar.google.pt/citations?user=RNDtSG8AAAAJ (3);
5. Tom Duckett https://orcid.org/0000-0003-2971-7905, https://scholar.google.pt/citations?user=et1GU2EAAAAJ (6), Cheng Zhao https://orcid.org/0000-0001-8502-3233, https://scholar.google.pt/citations?user=EAC-8m0AAAAJ (3), and Tomáš Krajník https://orcid.org/0000-0002-4408-7916, https://scholar.google.pt/citations?user=Qv3nqgsAAAAJ (3);
6. Gordon F. Wyeth https://orcid.org/0000-0002-4996-3612, https://scholar.google.pt/citations?user=yfXZfXEAAAAJ (4) and Michael J. Milford https://orcid.org/0000-0002-5162-1793, https://scholar.google.pt/citations?user=TDSmCKgAAAAJ (3);
7. Paul Newman https://scholar.google.pt/citations?user=BtO5fTUAAAAJ (4);
8. Timothy D. Barfoot https://orcid.org/0000-0003-3899-631X, https://scholar.google.pt/citations?user=N_vPIhoAAAAJ (3).

When analyzing the affiliations of the authors mentioned previously at the time of publication, all authors of the first cluster belonged to the State Key Laboratory of Industrial Control and Technology (SKLICT) and the Institute of Cyber-Systems and Control at Zhejiang University in China. Even though Huan Yin, Yue Wang, Xiaqing Ding, Li Tang, and Rong Xiong mention their affiliation to the Joint Centre for Robotics Research between Zhejiang University, China, and the University of Technology Sydney, Sydney, in the work Yin et al. (2020), this specific affiliation only appeared in this article. The total link strength (sum of all links weights) of each of the authors in that cluster is higher than 16, meaning a high co-authorship between them. Indeed, all five authors have links between all of them. Similar to the first cluster, the third, fourth, fifth, and sixth clusters have common affiliations within each one: the Autonomous Intelligent Systems at the University of Freiburg in Germany, the Autonomous Systems Lab (ASL) at ETH Zürich in Switzerland, the Lincoln Centre for Autonomous Systems (LCAS) at the University of Lincoln in UK, and the School of Electrical Engineering and Computer Science at Queensland University of Technology (QUT) in Australia, respectively. However, the interlinking between the authors is not as strong as in the first cluster, as shown in Figure 4 by the authors of these clusters not being connected between all the ones within each cluster. Even so, the common affiliation shows there is considerable interest by these research units in the long-term localization and mapping topic.

The affiliation analysis in the second cluster is more complex given that there was no affiliation common to all authors at the time of the records' publication. Instead, the following affiliations were found: Fei Han and Hao Zhang with the Department of Computer Science at Colorado School of Mines in the USA, Guoquan Huang with the Department of Mechanical Engineering at the University of Delaware in the USA, John J. Leonard and Michael Kaess with the Computer Science and Artificial Intelligence Laboratory (CSAIL) at the Massachusetts Institute of Technology (MIT) in the USA, Ryan M. Eustice with the Perceptual Robotics Laboratory (PeRL) at the University of Michigan in the USA, and José Neira with the Instituto de Investigación en Ingeniería de Aragón (I3A) at the Universidad de Zaragoza in Spain. Although there are five different affiliations to which the seven authors stated in the respective records, four of the research institutions noted for the second cluster are in the USA, indicating a possible reason for facilitating the linkage between these authors from different research units.

In terms of the clusters composed of single authors, the affiliations of Paul Newman and Timothy D. Barfoot are the Oxford Robotics Institute at the University of Oxford in UK and the Autonomous Space Robotics Laboratory (ASRL) at the University of Toronto Institute for Aerospace Studies (UTIAS) in Canada, respectively. Even though these two authors are not linked with any others in the network, the co-authorship analysis indicates that they have an interest in long-term localization and mapping. This interest is shown by their number of co-authored records: 4 and 3 by Paul Newman and Timothy D. Barfoot, respectively.

As for the number of co-authored publications, considering the 142 included records, the authors that appeared to have more research on the review's topic are Rong Xiong, Yue Wang, and Wolfram Burgard, given the 7 co-authored publications of each one. However, Rong Xiong and Yue Wang have co-authored the 7 documents attributed to each of them. This relation and similar ones can bias the analysis of which authors are having more impact in the review's topic. The clustering shown in Figure 4 allows a more unbiased analysis relative to the co-authorship links between authors. Thus, based on the clustering and which author from each cluster has the most co-authored publications, the most influential authors in long-term localization and mapping are the following ones: Rong Xiong (or Yue Wang), Hao Zhang, Wolfram Burgard, Roland Siegwart, Tom Duckett, Gordon F. Wyeth, Paul Newman, and Timothy D. Barfoot.

### 4.4 | Year of publication

The relevance of the long-term localization and mapping topic can be evaluated by the evolution of the number of publications. Figure 5 presents this evolution from the earliest year of publication of the included records to the year at the time of writing this article. The latter has its respective data dashed to indicate that the last year is not completed at the time of writing. Analyzing Figure 5, this review's topic seems to have gain relevance in 2009 with six works, compared with only one publication in 2007 and another in 2002 in the previous years to 2009. From that year onwards, the graph has an almost linear tendency reaching a maximum of 23 records in 2021, while already having eight publications in 2022 until May 17, 2022. This tendency shows that long-term localization and mapping is gaining interest throughout the years and, consequently, supports the importance and relevance of this review for the scientific community.

### 4.5 | Publication venue

Finally, the last overview of the 142 included records in the review is relative to the publication venue. Table 5 presents the venues with

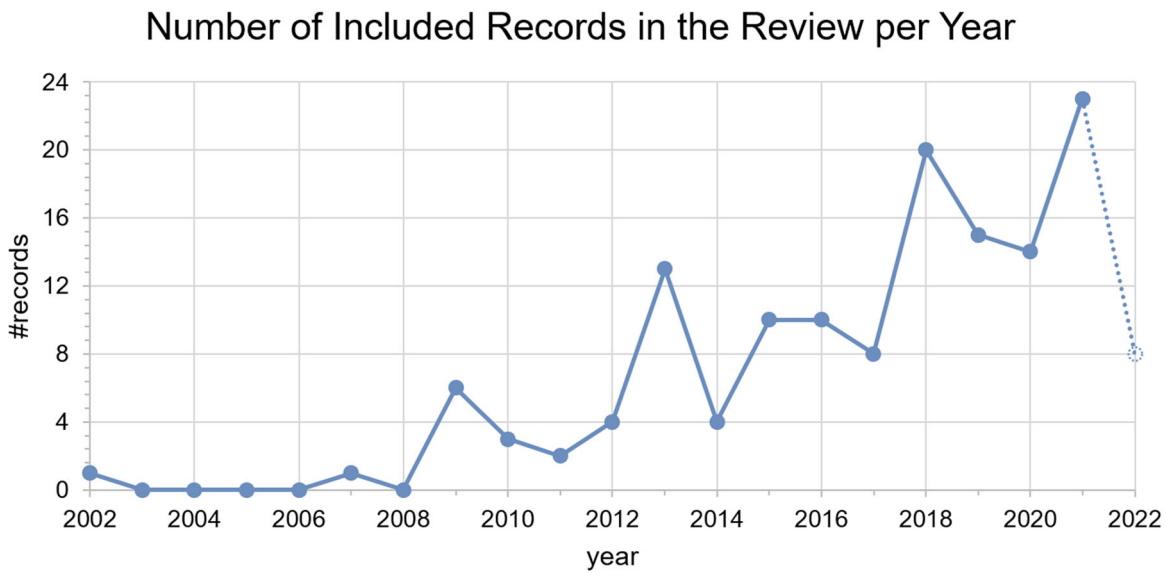


**FIGURE 5** Evolution of published records per year considering the 142 included records in this review. The time interval is between the smallest publication year found in the included records (2002) and the year of last full inquiry's date (2022). The latter is with a dotted line due to the fact that the last full inquiry does not consider the whole year. [Color figure can be viewed at wileyonlinelibrary.com]

more than 1 publication, separating the journals and conferences in two different tables (Tables 5a,b, respectively). The columns $\mu$ present the average year of publication of the records associated to a certain venue, while max columns display the publishing recency by the year of the most recent publication in the venue. For comparing to the average value ($\mu$), the third column ($\sigma$) of each table presents the standard deviation based on the publication year data. The last column state the number of records published in the venue from the 142 records included in the review for discussion.

In terms of journals, the *Robotics and Autonomous Systems*, *IEEE Robotics and Automation Letters*, and the International Journal of Robotics stand out with more than 10 publications. Also, these journals have a high standard deviation (greater than 1.5), indicating that the publications spread out throughout the years. In the case of the *IEEE Robotics* and Automation Letters, these results gain more relevance indicating a recent trend on publishing on this journal, considering that its creation was only on 2015.[12] With more than five publications, the *Journal of Field Robotics and the Autonomous Robots* have recent average of publication (2017) with a high standard deviation (greater than 2.0), similarly indicating that authors have been publishing in these two journals along the years. In contrast, the *IEEE Transactions on Intelligent Transportation Systems and Sensors* journals have a standard deviation lower than 1 year, with an average publication year of at least 2019. The recency of publication on these two journals with a very low deviation suggests a recent interest of the authors to publish in these two journals works related to long-term localization and mapping.

As for conferences, the data in Table 5b shows a high discrepancy in the number of publications related to this review's topic in ICRA and IROS compared to the other venues. Indeed, all the other conferences have only a maximum of three records published in them, compared with 22 and 17 papers in ICRA and IROS, respectively. When considering that 60 of the 142 included records are published in conferences, ICRA and IROS with a total of 39 published works related to this review's topic represent 65% of works published in conferences and 27.5% of all included records. This result expresses the high relevance of ICRA and IROS in the topic of long-term localization and mapping.

## 5 | DISCUSSION

The main goal of this review is to synthesize methodologies focused on long-term localization and mapping. Therefore, the discussion first analyzes the techniques proposed in the 142 included works for the five categories of DE1 (see Section 3.4). Section 5.1 discusses methodologies related to dealing with the varying appearance of environments for localization and place recognition. Section 5.2 analyzes works focused on modeling the environment dynamics or identifying dynamic objects within the environment. Section 5.3 focuses on approaches to remove redundant data from the map or identify novelty data to bound the map size. Section 5.4 discusses how methods handle multisession in terms of mapping. Section 5.5 reviews works related to computation concerns over long-term localization and mapping, in addition to the ones relative to map sparsification discussed in Section 5.3. The organization of topics in each of the five subsections mentioned previously follows our observations retrieved for each included work during the review and

[12] https://www.ieee-ras.org/publications/ra-l

**TABLE 5** Publication venues of the included records in this review with more than one record published in the venue: (a) journals; (b) conferences.

| (a) | | | | |
|---|---|---|---|---|
| | Year | | | |
| Journal | μ | σ | Max | # |
| *Robotics and Autonomous Systems* | 2016 | 3.9 | 2021 | 13 |
| *IEEE Robotics and Automation Letters* | 2019 | 1.7 | 2022 | 12 |
| *International Journal of Robotics Research* | 2014 | 3.2 | 2022 | 11 |
| *Journal of Field Robotics* | 2017 | 3.5 | 2022 | 8 |
| *Autonomous Robots* | 2017 | 2.2 | 2020 | 7 |
| *IEEE Transactions on Intelligent Transportation Systems* | 2021 | 0.8 | 2022 | 4 |
| *Sensors* | 2019 | 0.8 | 2020 | 4 |
| *IEEE Transactions on Robotics* | 2017 | 3.1 | 2022 | 4 |
| *IEEE Sensors Journal* | 2020 | 1.5 | 2021 | 2 |
| *International Journal of Advanced Robotic Systems* | 2020 | 1.5 | 2021 | 2 |

| (b) | | | | |
|---|---|---|---|---|
| | Year | | | |
| Conference | μ | σ | Max | # |
| IEEE International Conference on Robotics and Automation (ICRA) | 2016 | 3.9 | 2021 | 22 |
| IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 2017 | 3.6 | 2021 | 17 |
| IEEE International Conference on Robotics and Biomimetics (ROBIO) | 2019 | 2.1 | 2021 | 3 |
| IEEE International Intelligent Transportation Systems Conference (ITSC) | 2018 | 2.4 | 2021 | 3 |
| European Conference on Mobile Robots (ECMR) | 2014 | 0.9 | 2015 | 3 |
| IEEE Intelligent Vehicles Symposium (IV) | 2019 | 0.5 | 2019 | 2 |
| International Conference on 3D Vision (3DV) | 2018 | 1.5 | 2019 | 2 |
| International Conference on Advanced Robotics (ICAR) | 2011 | 2.0 | 2013 | 2 |

Abbreviations: μ, average year of publication; σ, standard deviation of the publication year; max, maximum year of publication; #, number of records published at a certain venue.

analysis processes. The organizational topics presented at the beginning of the subsections represent trends commonly found among the papers discussed to tackle the several challenges of performing long-term SLAM. Those trends may be technologies, research hypotheses, and/or motivations considered in the included works. All the data and annotations retrieved during the review process used to elaborate the discussion of the 142 included records are available in the GitHub repository[1]. This section also discusses how the included works evaluate their results in long-term operations. Thus, Section 5.6 analyzes the experimental data and data sets used in the experiments, and Section 5.7 presents the evaluation metrics used to evaluate the proposed methodologies.

## 5.1 | Appearance variance

Next, the discussion focuses on the included works categorized in DE1 as appearance (see Table A1). The different methodologies found in these works deal with variable lighting changes, perspective or viewpoint variance, moving elements in the scene, different weather conditions, or changes caused by the year's seasons. To improve the discussion, the analysis of the proposed techniques related with appearance invariance is organized into the following topics: (1) experience maps for treating different appearances as multiples experiences, (2) illumination invariant transformations, (3) handcrafted features, (4) features extracted using Convolutional Neural Networks (CNN), (5) assessment of feature stability, (6) multimodal features, (7) exploiting temporal coherence using image sequence matching, and (8) a discussion of the different sensors modalities used in the included works for appearance invariance.

### 5.1.1 | Experience maps

One way to deal with the appearance variance of environments is by treating different conditions as multiple experiences. The biologically

inspired RatSLAM (Ball et al., 2013) introduces the experience map as a semi-metric topological map. Each experience in the map is a view of the environment at a certain position. Then, wheel odometry provides the relative pose of the links between experiences. New experiences are created when none of the previous ones saved in the map are sufficiently similar in appearance to the current scene. Glover et al. (2010) combine the mapping of RatSLAM with the place recognition of FAB-MAP (Cummins & Newman, 2008). The latter improves the loop closure detection of the original RatSLAM, given that FAB-MAP has light invariant characteristics compared to RatSLAM. This invariance of FAB-MAP is achieved by learning a generative model for the Bag of Words (BoW) model (Sivic & Zisserman, 2003). Both RatSLAM (Ball et al., 2013) and the hybrid RatSLAM+FAB-MAP (Glover et al., 2010) systems use visual data to retrieve information from the environment. Although Martini et al. (2020) also use experience-based mapping, the main sensor is a radar instead of a camera. The point clouds from the sensor and the point descriptors retrieved from the radar data represent an experience at a given pose. Radar is known for being less affected by environment changes such as different illumination or weather conditions compared to vision sensors (Hong et al., 2022).

Similar to experience-based mapping, another concept used by Konolige and Bowman (2009) and Tang et al. (2019) is adding the environment changes to the map upon degradation in the localization estimation. Konolige and Bowman (2009) implement a keyframe SLAM with a Visual Odometry (VO) module, where each keyframe represents a view of the environment. The place recognition module tries to match the current frame to similar views already represented in the map for loop closure. Tang et al. (2019) applies a similar idea to experience maps based on the 2D manifold assumption for locally smooth navigation. Even though the proposed topological local-metric framework encodes geometric information in the edges, the nodes do not require global pose. Consequently, the manifold framework does not force to exist a global consistency in the map. New nodes are triggered from localization failure. The goal is to restrict the erroneous alignment computed from odometry locally.

Instead of considering an experience as a location or a view of the current scene, Churchill and Newman (2013) define it as a whole sequence of saved poses and related features directly obtained from the VO module. A connection in the topological map implies in the proposed approach that two experiments observe the same space. However, links between experiences do not have any geometric meaning. The method does not implement a specific place recognition module for loop closure. The authors assumed that the robot will subsequently return to a place that can have successful localization. Gadd and Newman (2016) build on the work of Churchill and Newman (2013) for multirobot systems. The proposed method adds FAB-MAP (Cummins & Newman, 2008) for place recognition in the existing map. The latter is maintained by a centralized versioning framework. The selection of the most relevant experiences by the centralized framework for multiagent localization assumes that appearance change is only driven by the passage of day time.

Another example of experience maps is Visual Teach & Repeat systems using spatial-temporal pose graphs, as implemented in MacTavish et al. (2018) and Zhang, Warren, et al. (2018). Similar to Churchill and Newman (2013), an experience is the output of the VO module defining the appearance of a scene throughout a path. In the teaching phase, the robot is teleoperated by humans creating privileged experiences in the graph. Autonomous experiences are the ones relative to the repetition phase. These experiences are linked either temporally or spatially if they are sequential in time or related metrically by multiexperience matching, respectively. Unlike Churchill and Newman (2013), new experiences have a known metric pose relative to the others in the pose graph.

In general, experience-based mapping methods try to generate new experiences if the environment changes. These methods expect that at a certain point in time the robot will be able to localize itself relative to previous experiences, not requiring new ones to be added to the map. One characteristic of experience-based methods is their appearance invariant properties due to the accumulated knowledge of different changes in the same location. However, these approaches are not scalable in the long-term time frame nor able to deal with dynamic elements, even using central servers as in Gadd and Newman (2016) with more computational resources than the robots. For example, views with moving elements in the scene would be treated as different experiences. Consequently, the map size and the computation time possibly would increase in those varying conditions. Pruning algorithms would be required to remove redundant or outdated information, as in Konolige and Bowman (2009) or Tang et al. (2019). Also, other methods should be employed to deal not only with long-term appearance changes (weather conditions or seasonal changes) but also with dynamic elements in the scene.

### 5.1.2 | Illumination transformations

As a preprocessing step, illumination invariant transformations can be applied on color images to increase the robustness of visual localization to changing lighting conditions and shadows. One example is the illumination invariant space that combines the log-responses of the three color channels (red, green, and blue) into one-dimensional space with a weighting parameter. The latter parameter is conditioned by the peak spectral responses of each color channel and is usually available in the camera specifications. This one-dimensional space is only dependent on the sensor and elements in the scene, while being independent of the intensities and colors. Both works of Arroyo et al. (2018) and Yang et al. (2021) use this transformation for preprocessing the color images into grayscale ones. The works showed that the preprocessing step improved the lighting invariance properties of the proposed methodologies compared with using the color images.

An alternative to using predefined illumination invariant transformations is to learn them. Clement et al. (2020) learn a nonlinear transformation mapping function from the RGB color space to

grayscale also combining the three-channel log-responses. However, the method relaxes the constraints of the one-dimensional space from to the original weighting parameter used in Arroyo et al. (2018) and Yang et al. (2021). Instead of using the same parameters independently of the image content, Clement et al. (2020) train an encoder to predict the optimal transformation weighting parameters of the three-channel log-responses. The objective function chosen for maximization is based on the number of inlier feature matches from a vision localization pipeline. The learned nonlinear RGB to grayscale transformation helped achieving a full-day cycle using a single mapping experience and applying the optimized transformation to the color images.

Even though the Gamma correction (Gonzalez & Woods, 2002) does not transform an image to an invariant color space, this transformation can be used to strengthen low-illumination changes. Sun et al. (2021) use the Gamma transform to synthesize low-illumination night-time images from daytime ones. Applying the transformation in the HSV (Hue, Saturation, Value) space, the gamma parameter adjusts the value channel without distorting the colors. Then, the synthesized images are used for training the DarkPoint descriptor proposed by Sun et al. (2021) to improve day-to-night matching.

### 5.1.3 | Handcrafted features

Many localization and mapping algorithms rely on the detection and extraction of features. The designation of handcrafted features refers to properties derived from the sensor data as a two-step process. First, a keypoint detector algorithm finds the location of features in the sensor data. Next, a descriptor is computed for each of them. This descriptor should be able to uniquely distinguish each feature from one another (Nanni et al., 2017). Algorithms for long-term localization and mapping using handcrafted feature should be robust to changing conditions such as illumination, appearance, weather and seasonal changes.

*Visual features*

A way to improve long-term feature-based visual localization is to enhance the descriptiveness of visual feature descriptors and their long-term stability. Kawewong et al. (2013) define the Position Invariant Robust Features (PIRF). In a sliding window framework, PIRF tracks the motion of local features such as Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) or Speeded Up Robust Features (SURF) (Bay et al., 2006) selecting the stable ones. Using an incremental tree-like PIRF (with inverted index as in BoW) dictionary, the method was robust to viewpoint variance and unstable features. Also, PIRF-based localization improved the recall on place recognition over FAB-MAP (Cummins & Newman, 2008).

Moreover, Histogram of Oriented Gradients (HOG) features have been used in the literature to improve the invariance of feature-based localization and mapping in changing appearance conditions. This improvement is achieved due to the HOG descriptors capturing local gradient information robust to seasonal changes (Naseer et al., 2015). Li et al. (2015) compute local HOG descriptors from visually-salient image patch features in an underwater environment. Using a trained Support-Vector Machine (SVM) (Boser et al., 1992) to classify the matching between corresponding patches, the method achieved approximately 80% accuracy with dramatic appearance changes. Although Naseer et al. (2015) compute HOG descriptors from each cell of a partitioned image, a global descriptor for the whole image joins all the cell features. The global descriptor proved to be robust to foliage color changes, occlusions, and seasonal changes. Vysotska et al. (2015) use the same global HOG descriptor as in Naseer et al. (2015). However, Vysotska et al. (2015) apply the descriptor in the context of image sequence matching. The method requires a rough global pose estimation for the images (e.g., GPS) for efficient matching in terms of computation time.

Local Difference Binary (LDB) (Yang & Cheng, 2014) features also include gradient comparisons. These features are used in the Able for Binary-appearance Loop-closure Evaluation (ABLE) (Arroyo et al., 2018) approach to achieve higher descriptiveness power for appearance invariance. ABLE outperformed FAB-MAP (Cummins & Newman, 2008) in terms of precision-recall evaluation metrics. An advantage of using binary features such as LDB is the possibility of using the Hamming distance to compute descriptor similarity, improving the computational efficiency of this process over cosine similarity or Euclidean distance.

Another work from the included records focused on improving the long-term performance of handcrafted visual features is from Karaoğuz and Bozma (2016). Their approach uses bubble descriptors for preserving the relative $S^2$ geometry of visual features, being rotationally invariant. The experimental results demonstrated improvements on viewpoint and illumination invariance of bubble features-based localization.

Instead of preserving the long-term appearance-invariance of visual descriptors, Neubert et al. (2015) introduce the SuperPixel-based Appearance Change Prediction (SP-ACP) to predict extreme appearance changes across seasons. SP-ACP extracts descriptors (combination of color histogram in Lab color space with upright SURF descriptor) from the image superpixels. Then, the method clusters the descriptors into seasonal-specific vocabularies using hierarchical $k$-means. With training images with pixel-accurate alignment between images, the known pixel association creates a translation dictionary between seasons to synthesize a predicted image for cross-season place recognition. SP-ACP was able to improve cross-season place recognition compared to not comparing with the predicted image. However, the method has the limitation of requiring pixel-wise alignment in training.

The work of Griffith and Pradalier (2017) considers GPS and compass data in addition to visual data. Griffith and Pradalier (2017) build on SIFT Flow (Liu et al., 2011) to find dense correspondences among images for survey registration in long-term lakeshore monitoring. SIFT Flow combines the precision of point-based feature matching with the robustness of whole-image matching. The proposed method also considers GPS data, the feature tracks from

a visual SLAM, and the compass measurements to bias image registration. Griffith and Pradalier (2017) were able to match images from different surveys separated by several months with dramatic changes relative to lighting, occlusions, seasonal changes, and even the sun glare.

Even though Cao et al. (2018) and Cao et al. (2021) require a 2D or a 3D laser for place recognition and not a visual sensor, these methods use 2D image representations of a point cloud to extract visual handcrafted features. Cao et al. (2018) transform the 3D point clouds of a 3D laser into 2D images using the bearing angle 2D representation (image according to the relative position among adjacent laser points, without projecting the point cloud onto a certain surface). Using a BoW approach with the dictionary learned using Oriented FAST and Rotated BRIEF (ORB) (Rublee et al., 2011) features, the query image is matched to the database ones, while performing geometric verification by reprojecting the ORB features into the 3D coordinate frame. One main advantage of using 3D LiDAR in the experiments was its less sensitivity to lighting conditions relative to visual sensors while not being incapacitated in the dark environments. The proposed method outperformed Multiview 2D Projection (M2DP) (L. He et al., 2016)—global descriptor for point clouds—, given that M2DP could not deal with situations where the point clouds distributions were centralized and similar to each other. As for Cao et al. (2021), the proposed method accepts also 2D laser data by accumulating a sequence of scans. The 2D representation used differs from Cao et al. (2018) by projecting the point cloud into cylindrical coordinates and using the centroid of the point cloud to ensure viewpoint invariance. Using Gabor filters (Gabor, 1946) to detect and describe the contours of the images, Cao et al. (2021) generate Binary Robust Independent Elementary Features (BRIEF) (Calonder et al., 2010) descriptors for matching images using a nearest neighbors search. In addition to showing the seasonal appearance variance in laser data (e.g., different foliage in the scene), the proposed methodology outperforms SeqSLAM (Milford & Wyeth, 2012) (sequential place recognition) and PointNetVLAD (Uy & Lee, 2018) (CNN-based place recognition for 3D point clouds) on precision-recall.

In terms of visual features from radar data, Hong et al. (2022) extract visual features used for tracking using a blob detector based on a Hessian matrix. These features are extracted from a 2D cartesian image obtained by transforming the radar data to the equivalent polar image representation. The method also compensates the distortion from the vehicle's motion in the radar data. As for loop closure detection, the peaks in intensity from the polar radar image are evaluated to remove noise of areas without a real object due to speckle noise. Then, the processed polar image is transformed into a point cloud and the M2DP descriptor adapted to 2D point clouds is used to detect loop closure. The proposed methodology improved the radar odometry tracking and outperformed ORB-SLAM2 (Mur-Artal & Tardós, 2017).

*Environment structure features*

The structure of the environment defined by its geometry is more robust to appearance variance than the appearance itself. Common structure features extracted from sensors data are line and edge features. Biswas and Veloso (2013) extract 2D line segments corresponding to the walls from depth and 2D laser sensors. The line segment-based localization had a low failure rate on an over-a-year long-term indoor deployment even in areas with movable objects, due to the long-term stability of the line segment features. Nuske et al. (2009) extract 3D edge features of the scenes using a monocular camera to get the edges of the buildings in the environment, while employing an exposure control to maximize the strength of edges corresponding to the mapped ones. The proposed method was able to successfully track the edges of the buildings along an all-day outdoor experiment. Instead of using the walls of the buildings, An et al. (2016) formulate a visual node descriptor based on ceiling salient edge points. Even though the method achieved good results in lighting changing conditions, the method's performance decreases using low and inclined ceilings, due to the image perspective effect that may lead to matching failure in the implemented Iterative Closest Point (ICP) (Besl & McKay, 1992).

Furthermore, Meng et al. (2021) extract edge and planar features by evaluating the large and small values of the local surface smoothness over the points of a 3D laser, respectively. ICP estimates the laser odometry. Then, the histogram cross-correlation of the Normal Distribution Transform (NDT) (Biber & Strasser, 2003) that computes local probability density functions of the surface smoothness identifies the loop closures. The proposed method outperformed an ICP-based SLAM approach on Absolute Trajectory Error (ATE) in the experiments. As for Bosse and Zlot (2009), 2D point clouds segmented into connected components are clustered at regions of high curvature to get high curvature keypoints from multiple scans. The proposed descriptor based on the moment grid improves outdoor place recognition relative to SIFT (Lowe, 2004) or Hough transform peaks (Tomono, 2004) due to the moment grid descriptor including higher order of moments relative to other descriptors.

Poles are also scene structures used for long-term localization. Schaefer et al. (2021) retrieve the 2D coordinates of poles registered with a 3D laser. Results demonstrated the ability of reliable long-term localization over more than one year. In addition to poles, Berrio et al. (2019) extract also corner features from the 3D laser point cloud, being able to localize over a 6 month experiment at different times of the day.

Another possible application of environment structure features found in the included works is in crop fields for agriculture. Chebrolu et al. (2018) formulate an aerial image registration algorithm based on the positions of the crops assuming that the gaps between them remain the same over time. The method computes a vegetation mask by exploiting the Excess Green Index (ExG) of RGB images. Using the Hough transform (Rosenfeld, 1969) to find lines between vegetation, the center of the crops are the peaks on vegetation histograms perpendicular to the rows. The experimental results in Chebrolu et al. (2018) demonstrated invariance of the registration algorithm to changing conditions caused by weather and crop growth over 1 month.

### 5.1.4 | Convolutional neural networks (CNN)

A more recent direction noted in the included works is the use of CNN. The evolution of deep learning in computer vision led to researching how

CNN could be used for generating feature representations robust to appearance variance, as an alternative to handcrafted features. CNN-based features are known to offer more discriminative power compared to handcrafted features while being able to be more robust in challenging environments (Taisho & Kanji, 2016). The feature representations can be retrieved from the layers of CNN, with earlier ones usually extract low-level features such as edges or corners, while deeper layers extract high-level ones such as semantic structures (Chen et al., 2018). In addition to using the CNN feature maps, the included works also used CNN for semantic segmentation to extract semantic information from sensor data and appearance-content disentanglement for generating appearance-invariant descriptors.

*CNN feature maps*

One application of CNN features is for image place recognition as a classification task. Instead of comparing pairs or triplets of images, place recognition is formulated as a classification problem (Chen et al., 2018). In Taisho and Kanji (2016), the layer *fc6* (fully connected) of AlexNet (Krizhevsky et al., 2012) extracts 4096-dimensional CNN features from box regions in the query image, then reduced to 128-dimensional features with Principal Component Analysis (PCA). Comparing these features to the ones extracted from the reference images in a cross-domain library (collected in different routes and seasons), Taisho and Kanji (2016) define the query image as a set of nearest neighbor library features (similar to BoW) and employs the image-to-class distance with the Naive Bayes Nearest Neighbor (NBNN) method. The proposed PCA-NBNN descriptor outperformed BoW (Sivic & Zisserman, 2003) and FAB-MAP (Cummins & Newman, 2008) on a cross-season experiment in precision-recall metrics. Chen et al. (2018) also formulate a classification task for place recognition, using a VGG16 (Simonyan & Zisserman, 2015) network for generating local features. The method adds a convolutional, a fully-connected, and a softmax layer to learn the correct label output for classification. The proposed architecture outperformed FABMAP and SeqSLAM (Milford & Wyeth, 2012) on seasonal changing conditions.

Place recognition can also be formulated as a coarse to fine image matching problem. An initial set of reference image candidates is obtained based on nearest neighbor distances of image-wise global descriptors (Camara et al., 2020; Liu et al., 2021; Xin et al., 2017). Then, local features are used to obtain a more accurate estimation based on spatial matching (Camara et al., 2020; Xin et al., 2017) or geometrical verification (Liu et al., 2021). Xin et al. (2017) extract both global and local features using a convolutional layer (*conv3*) of the AlexNet network. The local features are extracted from regions of the image with candidate regions sorted by the objectness score (improves viewpoint invariance). Instead of using AlexNet, Camara et al. (2020) use layers from VGG16 for feature extraction, specifically, *conv5-2* and *conv4-2* layers for global and local features, respectively. As for Liu et al. (2021), the MobileNetV2 (Sandler et al., 2018) network is selected for global feature extraction due to its computational efficiency. However, the proposed work uses grid-based motion statistics and ORB (Rublee et al., 2011) descriptors to define local features, instead of using CNN features.

Deep features can be combined with handcrafted features and preprocessing techniques to facilitate learning and further enhance their discriminative properties. Zhang et al. (2022) use the Key.Net (Laguna et al., 2019) network for keypoint generation. This network combines handcrafted and learned filters to detect keypoints at different scale levels, reducing the number of learnable parameters. Combined with HardNet (Mishchuk et al., 2017) for descriptor extraction, the method outperformed a BoW approach in viewpoint and illumination changing conditions. Yin et al. (2020) propose a handcrafted rotational invariant feature to be the input of a LocNet (Gidaris & Komodakis, 2016) network for 3D laser-based place recognition. The proposed handcrafted feature reduced the complexity of the network and improved the efficiency on similarity evaluation. As for preprocessing techniques to help in training, Sun et al. (2021) use a the Gamma transform (Gonzalez & Woods, 2002) and other transformations (translation, scale, in-plane rotation, and symmetric perspective distortion) to generate day-nigh image pairs from daytime ones. These images are used for training the proposed visual descriptor DarkPoint on the keypoints generated by the SuperPoint (DeTone et al., 2018) keypoint detector. DarkPoint achieved approximately 1.7× more inliers during navigation than the original SuperPoint in day-night experiments.

Given that feature maps can extract different types of features depending on the deepness of the respective layers, Zhu et al. (2018) extract features from three layers (*conv3-3*, *conv4-4*, *conv5-3*) of a VGG16 network and concatenates these to form a global descriptor for an image. A cross-season experiment showed an increasing performance in precision-recall when the single layer gets deeper. These results are conformal to ones obtained in Yang et al. (2021). The layer *conv5-3* achieved higher accuracy and improved place recognition compared to using the feature maps from *conv4-4* and *conv3-3*. This result indicates that the spatial information increases in deeper layers. Zhu et al. (2018) also showed that fusing the three layers used in their work by concatenating them into a global descriptor improves even further the place recognition performance. Moreover, Yu et al. (2019) chose DenseNet (Huang et al., 2017) for feature extraction due to this network reusing feature maps, that is, connecting all layers with the same map sizes directly with each other. Then, Yu et al. (2019) use the Weighted Vector of Locally Aggregated Descriptor (WVLAD) encoding for obtaining a global descriptor of the image. The proposed descriptor improved precision-recall over other architectures (VGG16, ResNet50 [K. He et al., 2016]) and to a BoW place recognition method.

The included works also focus on 3D LiDAR and radar place recognition with CNN features. However, the raw point cloud data is not directly suitable for the CNN inputs. The most common solution is to project the point clouds onto the surface plane, the so-called Bird's-Eye View (BEV). Yin et al. (2018) encode directly the BEV of a 3D LiDAR into a low dimensional global feature using a bidirectional Generative Adversarial Network (GAN) (Donahue et al., 2017). Using the extracted features within the SeqSLAM (Milford & Wyeth, 2012) framework, the proposed method improved the precision-recall metrics over the original SeqSLAM in changing conditions. Similarly,

Martini et al. (2020) extract a global descriptor from the BEV using NetVLAD but with the point cloud from a radar sensor. Kim et al. (2019) formulate the point cloud descriptor Scan Context Image (SCI), also known as ScanContext. The 3D point cloud is converted to a polar representation of BEV named Scan Context (SC) matrix, where each cell of the 2D matrix contains the maximum height of points around a scene. Using the jet colormap to transform the SC into the SCI as a three-channel image suitable for the CNN inputs, Kim et al. (2019) use a LeNet (Lecun et al., 1998) network for feature extraction and place classification. The proposed architecture outperforms PointNetVLAD (Uy & Lee, 2018) and the handcrafted point cloud feature M2DP (L. He et al., 2016) in precision-recall. Based on SCI (Kim et al., 2019), Xu et al. (2021) propose the Differentiable Scan Context with Orientation (DiSCO) descriptor. This method distinguishes from SCI by applying the Fast Fourier Transformation (FFT) to convert the polar BEV representation to the frequency domain. Given that frequency spectrum is translation-invariant, DiSCO becomes rotation invariant. The results showed a superior performance to SCI and PointNetVLAD in changing conditions. Similar to DiSCO (Xu et al., 2021), Yin, Xu, Wang, et al. (2021) also use SCI and FFT for feature extraction of point clouds. The difference is the use of a shared U-Net (Ronneberger et al., 2015) architecture to extract features of 3D LiDAR and radar data, training simultaneously the radar-to-radar, LiDAR-to-radar, and LiDAR-to-LiDAR place recognition tasks. The proposed method had similar or improved performance in these three recognition tasks relative to SCI and DiSCO. In addition to BEV, Yin, Xu, Zhang, et al. (2021) also use the spherical view. Using two separated 2D CNN following the convolutional layers in VGG16 to encode local features, a VLAD layer extracts place features from each view (BEV and spherical). A tightly-coupled fusion network fuses the features of each view. The proposed FusionVLAD descriptor outperformed PointNetVLAD (Uy & Lee, 2018) and M2DP (L. He et al., 2016) on the recall metric in appearance variant conditions.

Lastly, a trend found in the included works to improve the discriminative power of CNN features in environment varying conditions is the use of triplets (Liu et al., 2021; Martini et al., 2020; Piasco et al., 2021; Sun et al., 2021; Yin, Xu, Wang, et al., 2021; Yin, Xu, Zhang, et al., 2021) to train the networks. A triplet consists of an anchor image, a positive match correspondence, and an unrelated negative example. Triplet loss tries to minimize the matching distance between positive pairs (anchor, positive) and maximize that between negative ones (anchor, negative), in contrast to only considering positive correspondences in training (Sun et al., 2021). The goal is to learn the similarity between two corresponding images and the dissimilarity between unrelated and unmatched image pairs to improve the appearance invariant properties of the networks. Additionally to visual data, Piasco et al. (2021) use depth information during training. Depth maps and geometric information are expected to remain more stable across time than visual data. A CNN encoder aggregates local features to produce a global descriptor, while a decoder reconstructs the scene geometry from the features obtained by the encoder. Then, triplet loss training uses the fusion of image and depth map descriptors. In the experiments, the depth map training supervision provided building shapes understanding while improving the performance compared to not using side information.

*Semantic segmentation*

Instead of using the feature maps of CNN, the networks can also segment raw data to extract semantic information. Naseer et al. (2017) use the Fast-Net (Oliveira et al., 2016) network to extract salient maps for stable structures. These structures considered in training are man-made ones such as buildings or signs that are presumable to be stable in long-term. Then, the salient maps boost the importance of features retrieved from a convolutional layer (*conv3*) for place recognition. The proposed method improved the precision-recall metrics compared to HOG (Dalal & Triggs, 2005) and place recognition without boosting stable structures on a cross-season experiment.

The included works also use semantic features from pixel-wise labeling of image data. T. Qin et al. (2020) modify an U-Net (Ronneberger et al., 2015) for semantic feature detection specifically trained for parking lots. This network generates pixel-wise segmentation of lanes, parking lines, guide signs, speed bumps, free space, obstacles, and wall, used in both localization and feature mapping. In the experiments, the semantic features were robust to light changes, texture-less-regions, motion blur, and appearance change. Berrio et al. (2021) also segment an image with pixel-wise labels, discriminating 12 classes: pole, building, road, vegetation, undrivable road, pedestrian, rider, sky, fence, vehicle, and unknown (unlabeled and void). Using the extrinsic parameters of the 3D laser–camera, the pixel-wise semantic information from the labeled images is transferred to the 3D point cloud. Then, pole and corner features are retrieved from the projected point cloud onto the horizontal plane based on the Inertial Measurement Unit (IMU) data for localization and mapping. The long-term evaluation of the map corrections showed a decrease over time of the corrections demonstrating the stability of these features in outdoor environments. In addition to pixel-wise segmentation, G. Singh et al. (2021) connect the regions of each instance of the semantic classes to characterize them in terms of their centroid in 3D camera coordinates (using also depth information from a stereo camera) and connections to other regions. The proposed global semantic-geometric descriptor defines a location in terms of how the pairs of semantic entities are distributed in the scene. The proposed method obtained higher accuracy when compared to SeqSLAM (Milford & Wyeth, 2012), FAB-MAP (Cummins & Newman, 2008), and a BoW (Sivic & Zisserman, 2003)-based place recognition methods in a highly dynamic outdoor experiment.

Similar to G. Singh et al. (2021), graph embedding of semantic features also tries to integrate the geometric relationships between features for improving the robustness of place recognition. Han, Beleidy, et al. (2018) propose the Holism-And-Landmark Graph Embedding (HALGE) descriptor. In the training phase, an image is represented by its global HOG descriptor and semantic features (static or stable elements such as houses, traffic signs, trees). A graph

relates the training images from different domains and locations. The nodes are images or the semantic classes, whereas the edges represent the presence of a semantic class in an image or if two images represent the same location. Then, HALGE learns a projection matrix of each template database image from the graph to generate an appearance invariant feature from the original global HOG descriptor. The proposed method improved the performance over HOG, SURF (Bay et al., 2006), and color and AlexNet (Krizhevsky et al., 2012)-based descriptors in changing conditions. As for Gao and Zhang (2020), the proposed method formulated the place recognition task as a graph matching problem. The graph represents each semantic feature (same classes as in G. Singh et al. (2021)) by its central position in the image coordinate frame, while the edges represent the relations between features. These relations are spatial distances, angular relations, and appearance similarities (Euclidean distance of local HOG descriptors). Then, a graph optimization optimizes a correspondence matrix between the features in the query to the ones in the template images for obtaining the final matching scores, assuming a long-term worst-case scenario (maximizes the distance and angular similarities of features that have the least similar appearance). The proposed method outperformed Han, Wang, et al. (2018) and an HOG-based place recognition method on recall at higher precision in outdoor experiments with seasonal and weather changing conditions.

Semantic information can also be retrieved from sensors beyond cameras, such as a 3D LiDAR. Wang et al. (2021) use the RangeNet++ (Milioto et al., 2019) network to infer semantic labels of 3D point clouds from 3D LiDAR data. Even though the network can label 10 categories, the method only used the categories representative of pole-like objects (poles, tree trunks). The method achieved a higher localization accuracy than SCI (Kim et al., 2019) in an outdoor experiment with moving elements and dense vegetation.

*Appearance-content disentanglement*

When considering a place perceived by a vision sensor, the image retrieved from the sensor may be characterized by the place's appearance and content information. The location appearance may vary due to weather or seasonal changing conditions in the long-term perspective, among other factors. However, the content information is expected to remain more constant and invariant to appearance changes in the environment. The included works studied the possibility of learning the appearance-content disentanglement in feature representation. This disentanglement assumes that the images' latent space may be decomposed into the appearance and content spaces (C. Qin et al., 2020).

Oh and Eoh (2021) adopt the Variational AutoEncoders (VAE) (Kingma & Welling, 2013) architecture. This architecture uses an encoder to generate the appearance and content feature vectors, while a decoder reconstructs the original image from these vectors. Instead of using a single encoder, C. Qin et al. (2020) propose the Feature Disentanglement Network (FDNet) consisting of independent content and an appearance encoders, a decoder, and also an appearance discriminator to ensure the vectors are unrelated. Even though the content feature vector demonstrated to be invariant to seasonal changes, the method significantly reduced its performance on high viewpoint variance, where the content vector changed greatly while the appearance one did not changed at all. This results indicates that viewpoint change is considered to be content in the proposed algorithm. With a similar architecture to C. Qin et al. (2020), Tang et al. (2021) also consider a place domain discriminator to ensure that the content discriminator only contains the place information and does not include the location appearance. The method uses data augmentation in training to increase the robustness of the content discriminator against viewpoint changes. In the experiments, all images generated from a zero-appearance feature vector looked similar, while their place information remains conserved between the zero-appearance images. This result indicates that the proposed method can disentangle the input image across appearance changes.

Even though Hu et al. (2022) do not extract appearance and content independent features from the images, the proposed architecture builds on the same assumption of appearance-content disentanglement that a content representation of a location is shared across multiple domains. Hu et al. (2022) adopt a multidomain image-to-image architecture that expands the CycleGAN (Zhu et al., 2017) to multiple domains, with domain-specific encoder-decoder pairs and discriminators. For obtaining a shared-latent feature across different domains, the descriptor is learned using the feature consistency loss for domain-invariance. In the experiments, even though night-time images were not included in the training, the model was able to learn the content space of the places and outperformed FAB-MAP (Cummins & Newman, 2008) for place recognition.

### 5.1.5 | Feature stability

Although long-term handcrafted or CNN-based features intend to remain invariant to changing conditions of the environment, their long-term stability is not guaranteed to be the same for all detected features. In this context, Dymczyk, Stumm, et al. (2016) propose a CNN architecture based on AlexNet (Krizhevsky et al., 2012) for evaluating the feature stability for long-term visual localization. The network is trained using a set of labeled data pairs (image patch around the feature keypoint, label) or triplets (adds depth information). As for data labeling, the classification of the feature stability is a binary one: stable or unstable. The labels used in training are computed by assessing the number of feature observations over multiple sessions. In the experiments of over 15 months and changing conditions, the proposed method outperformed the random selection of features for localization in terms of f-score. Also, adding depth information improved the method's performance for place recognition.

Other approaches in the included works define predictor functions for evaluating the feature stability. Berrio et al. (2019) define the following predictors to evaluate the pole and corner features extracted from a 3D laser: the number of observations,

maximum detected and possible spanning angle, maximum length driven while observing the feature, maximum detection area, and concentration ratio. A regression algorithm adjusts the weights of each predictor based on the number of observations across sessions to define the scoring function. A threshold based on the histogram of the feature scores determines which features to include in the long-term map. Although Berrio et al. (2021) also use the concentration ratio and maximum driven length as predictors, their approach simplifies the selection strategy. The proposed method only includes observed features in the map if they have been observed for more than 1m and conserving the ones in sparse density areas to avoid localization failures. Berrio et al. (2021) also define a visibility measure related to the maximum range from where the feature is detected at a particular angle and the respective probability of detection. The method only computes the feature metrics when it has a match observation or if both not detected and not occluded. The latter is based on the visibility measure mentioned previously.

Furthermore, Egger et al. (2018) and Derner et al. (2021) propose methodologies for updating the map upon detecting changing conditions of the environment. Egger et al. (2018) define a minimum time interval between evaluations and the number of reconfirmations before updating the map with new stable and persistent features. The change in the conditions is determined by an overlap measure between the current view and the existing map. This overlap measure represents the relative amount of matched surfels extracted from a 3D laser. The proposed methodology led to a successful deployment of a robot over 18 months in changing conditions. Even though Derner et al. (2021) do not add features after creating the visual database used as a map, the proposed method updates the feature weights saved that represent their stability and reliability for localization. After computing the transformation between the current view and the best database match, the descriptors of the latter are compared with their transformed counterparts. This comparison is performed by reprojecting the keypoints of the database on the query image using the transformation between current view and the best match and recomputing the respective descriptors. The descriptors similarity, a spatial and temporal constraints, and the number of successful matches determine if the environment changed. In the latter case, the feature weights are updated based on their previous value and on the descriptors similarity. The method outperformed the localization without the weights update.

Instead of assuming observability independence, the observation of the features may be correlated between them. Nobre et al. (2018) model the feature persistence using a Bayesian filter in a time-varying feature-based environmental model. The model considers the correlation between features without assuming no specific-sensor feature descriptor. The approach follows a survivability formulation where each map feature has a latent survival-time (represents the time when the feature ceases to exist) and a persistence variable. The marginal persistence is estimated probabilistically given the detection sequence of all features. This probabilistic estimation assumes that if a set of features is co-observed and geometrically close, the likelihood that they belong to the same semantic object is high. The marginal feature persistence weights the data associations. The method was able to maintain track of the localization and updating the map accordingly in a semi-static changing environment. Luthardt et al. (2018) propose the Long-term Landmarks (LLamas) as persistent features, where the candidate points are the inlier feature tracks from visual odometry (short-term stable points). Considering that the map holds quality and viewpoint information, the correlated quality between neighboring viewpoints is modeled by Markov Random Field (Thrun et al., 2005). The experiments showed that the identified LLamas over a 2 month experiment consisted on persistent structures in the environment such as curbstone, sign, or a street lamp, discarding varying structures like vegetation, parked carts or shadows. As for Bürki et al. (2019), the proposed appearance equivalence class measure models the probability of observing the feature given the past map sessions. This model expects to observe again the same features together with those already co-observed in the past. Although the proposed selection measure outperformed the random selection of features in changing environments, the method suffered from the lock-in effect due to abrupt changes in the environment not being reflected in the observation sessions.

### 5.1.6 | Multimodal features

Another type of approach to feature-based localization and mapping is the use of multimodal features, given that these features can be more discriminative than only considering a single feature space (Latif et al., 2017). Filliat (2007) propose a two-stage voting scheme for localization integrating 3 different feature spaces: SIFT (Lowe, 2004), and local color and normalized gray level histograms. First, each feature space votes for the estimated location based on an incremental dictionary, without considering features seen in all known locations. Then, the votes of the different modalities are joined into a score that determines which location is the correct one. On the contrary, Latif et al. (2017) tested the use of multimodal features—*gist* (Oliva & Torralba, 2001) descriptors and feature maps from a CNN—by concatenating their descriptors into a single vector. In both Filliat (2007) and Latif et al. (2017), using multiple feature spaces improved the localization performance over considering a single feature space.

The included works also cover a more specific approach to multimodal features by formulating the place recognition task as a regularized sparse optimization problem. The optimization uses training data for learning the weight of each feature modality when computing the matching score between the query and database images (Han, Wang, et al., 2018; Han et al., 2017; Siva et al., 2020; Siva & Zhang, 2018). Han et al. (2017) formulate the Shared Representative Appearance Learning (SRAL) for fusing multimodal visual features from six different spaces applied on downsampled images as scene descriptors: color histograms, *gist*, HOG (Dalal & Triggs, 2005), Local Binary Patterns (LBP) (Ojala et al., 1996), SURF (Bay et al., 2006), and AlexNet (Krizhevsky et al., 2012) (*conv3*). SRAL outperformed the individual feature spaces and also the

concatenation of the 6 spaces into a single descriptor. Han, Wang, et al. (2018b) propose the RObust Multimodal Sequence-based (ROMS) loop closure detection. The proposed method adapts the regularized optimization framework to image sequence matching. The modalities considered are LDB (Yang & Cheng, 2014), *gist*, Faster R-CNN (Ren et al., 2015), and ORB (Rublee et al., 2011). ROMS outperformed both FAB-MAP (Cummins & Newman, 2008) and SeqSLAM (Milford & Wyeth, 2012) in appearance changing conditions, while improving the performance over considering a single feature space. In addition to learn discriminative modalities, Siva and Zhang (2018) formulate the Fusion of Omnidirectional Multisensory Perception (FOMP) that learns the weights representative of discriminative views (omnidirectional vision) and considers both image and depth modalities of features. The feature spaces considered are *gist*, HOG, LBP, and AlexNet (*conv3*). In a cross-season experiment, the depth-related modalities had more importance than the image ones, indicating that the latter are more susceptible to appearance change. Also, FOMP outperformed feature concatenation and only using the front field of view. As for Siva et al. (2020), the proposed Voxel-Based Representation Learning (VBRL) method identifies representative feature modalities and voxels from 3D point cloud. The feature spaces considered are the HOG in the XY, XZ, and YZ planes, the subvoxel occupancy scene descriptors, and the covariance points containe within each voxel. VBRL outperforms only considering discriminative voxels or features, and also outperformed descriptor concatenation in changing conditions.

### 5.1.7 | Image sequence matching

The temporal coherence of a sequence of visual data improves the performance of long-term place recognition in appearance variant conditions due to higher discriminative properties when exploring the temporal sequential relationships of the images (Nguyen et al., 2013; Ouerghi et al., 2018). Ouerghi et al. (2018) build on SeqSLAM (Milford & Wyeth, 2012) by proposing the Sequence Matching Across Route Traversals (SMART) system. The original SeqSLAM defines a location as a sequence of images by searching first for the best sequence match and then performing a local search for place recognition. Given the SeqSLAM's drawback on lack of viewpoint invariance due to global matching, SMART introduces a variable offset in the image match to compare each frame with the database within a range of image offsets, while also fusing the place recognition with visual odometry using an Extended Kalman Filter (EKF). The fusion of topological with local metric localization improved the mean error distance error over visual odometry in changing conditions, while SeqSLAM only provides a location-wise estimation. Han, Wang, et al. (2018) compared frame-to-frame matching to the proposed ROMS algorithm. The latter models frame correlation and formulates the image sequence matching problem into a regularized sparse optimization (in addition to learning the features modalities). ROMS improved the place recognition over frame-to-frame matching and outperformed SeqSLAM (Milford & Wyeth, 2012) and FAB-MAP (Cummins & Newman, 2008).

Moreover, Vysotska et al. (2015) define image sequence matching between a query and a database as a data association graph, encoding in the graph the cost proportional to the similarity between two images given by a HOG (Dalal & Triggs, 2005) descriptor. Instead of formulating the sequence matching as a network flow optimization problem, Vysotska et al. (2015) estimate the shortest path in the graph. This approach requires a rough global pose estimation for the images (e.g., GPS) to search efficiently through the graph for possible matches. Naseer et al. (2015) leverage the temporal sequence of images by requiring ordered sequential images in the database. The state transition model of the Bayes filter allows transitions between all places but modeled with different probabilities, while a sequence filtering searches for sequences of local peaks of matching images. Sequential information is accounted by imposing a minimum sequence length and maximum gap in frames between two matches to avoid false-positives. Both Vysotska et al. (2015) and Naseer et al. (2015) outperformed SeqSLAM (Milford & Wyeth, 2012) and network flow in the experiments.

Although an image sequence is a set of images, the sequence itself can be described by a descriptor. In both Arroyo et al. (2018) and Zhu et al. (2018), the sequence descriptor is the concatenation of the single images, and the sequence matching is the computation of Hamming distance between the descriptors. Arroyo et al. (2018) use the LDB (Yang & Cheng, 2014) binary descriptors for single images, and the experiments showed a lower accuracy for single image matching in long-term compared to the sequence descriptor. Also, the proposed method outperformed FAB-MAP (Cummins & Newman, 2008) and SeqSLAM (Milford & Wyeth, 2012) in terms of precision-recall metrics. As for Zhu et al. (2018), the feature maps from VGG16 (Simonyan & Zisserman, 2015) are normalized into a binary descriptor. The method outperformed FAB-MAP, SeqSLAM, and ABLE (Arroyo et al., 2018) in a cross-season experiment.

Lastly, Nguyen et al. (2013) propose an approach to identify topological places based on an image stream. The method uses a clustering scheme K-iteration Fast Learning Neural Network (FLANN) (Phuan & Prakash, 2002) to organize the visual input images into scene tokens. These tokes are the input to a Spatio-Temporal Long-Term Memory (LTM) architecture equivalent to an NN-based memory structure. This architecture represents topological locations as image sequences stored in the memory structure (LTM cells). Then, the proposed architecture models the topological structure of an environment by linking the scene clusters into a temporally ordered sequence using a one-shot learning mechanism and only requiring a single representation of the sequence. A pooling system determines the current topological location of the robot. The method was able to localize different topological sequences in appearance changing conditions.

### 5.1.8 | Sensor modalities

The appearance variance in the environments affects visual data as well as ranging data obtained from sensors such as 2D/3D lasers or radar. Visual data is affected by the illumination changes of day-night

situations, the weather changing conditions, and the changes on visual data caused by the different seasons of the year. Laser-based localization does not suffer from illumination variance. However, the laser is affected by low reflections or occlusions in unfavorable conditions such as fog, direct light, or moving elements in the scene. As for radar, the sensor is invariant to lighting and weather changes. Still, noisy measurements affect the performance of radar-based localization and mapping in long-term scenarios (Yin, Xu, Wang, et al., 2021).

Consequently, long-term localization and mapping algorithms should also consider fusing different sensor modalities to use the advantages of each one and improve the overall robustness of those algorithms to appearance changes. In addition to the works already discussed previously, Pérez et al. (2015), Coulin et al. (2022), and Nguyen et al. (2022) also focus on appearance invariance upon changing environments while using more than one modality. Pérez et al. (2015) introduce an appearance-based particle injection in the Monte Carlo Localization (MCL) framework to account the visual place recognition of FAB-MAP (Cummins & Newman, 2008). The BoW (Sivic & Zisserman, 2003) model of FAB-MAP is created using visual data recorded at different hours and changing conditions. Then, using the BoW model and a 2D occupancy grid as prior, the MCL fuses the odometry (wheel encoders and IMU data), the 2D laser, and the loop closure detection from FAB-MAP. The method did not need any manual recovery even in the case of global localization in a crowded environment with significant lighting changing conditions. Coulin et al. (2022) propose the use of a magnetic map with a Multi-State Constraint Kalman Filter (MSCKF). The magnetic map is built offline using visual-inertial SLAM in conjunction with global optimization to provide ground-truth positions for the map readings. As for localization, the tightly-coupled visual-inertial MSCKF reuses the magnetic map, while simultaneously estimating the magnetometer bias to avoid calibrating it every session. The experiments compared the proposed method to a visual-inertial SLAM algorithm with a visual map on a run one year after the creation of the map. The proposed method outperformed the other one given that visual data was variant to appearance changes in the environment, while reducing the ATE from 2.4 to 0.033 m compared with using vision-only in the MSCKF. As for Nguyen et al. (2022), the proposed Visual-Inertial-Ranging-Lidar (VIRAL) sensor fusion algorithm includes an IMU, 3D LiDAR, a camera, and Ultra-Wide-Band (UWB) data to localize an aerial vehicle in indoor environments. The vehicle odometry is estimated using the first three sensor modalities. UWB data is used to obtain an absolute position of the aerial vehicle in the world frame. VIRAL formulates cost functions of the sensors evaluated at every time step for inclusion in the optimization. The method improved over ORB-SLAM3 (Campos et al., 2021) in changing lighting conditions.

## 5.2 | Dynamics modeling

This section analyzes included works focused on modeling and identifying dynamic elements in the environment, categorized in DE1 as dynamics (see Table A1). Even though Section 5.1 already discusses appearance changes in the environment that can include moving elements in the scene, this section focuses on how the methods identify these elements and handle them for long-term localization and mapping. The discussion on dynamics modeling is organized into the following topics: (1) specific map representations used to model or deal with dynamic elements in the scene, (2) identification of dynamic elements by matching the current observation to the current map, (3) future prediction of dynamic properties of scene elements, and (4) semantic identification of dynamic objects.

### 5.2.1 | Map representation

Inspired by the human memory, Dayoub et al. (2011) and Bacca et al. (2013) adapt the multistore model of Atkinson and Shiffrin (1968) for robot mapping. This model divides the memory into three stores: Sensory Memory (SM) to save the perceived information, Short-Term Memory (STM), and Long-Term Memory (LTM). Three mechanisms move information between memories: selective attention for SM to STM, rehearsal to commit information from STM to LTM or which one is forgotten, and the retrieval mechanism to move unused information from LTM back to STM. Dayoub et al. (2011) implement two types of state machines for rehearsal and retrieval mechanisms of the STM and LTM. In rehearsal, a STM feature moves closer to LTM or moves back to the initial state (or forgotten if already in that state) when observed consecutively or if not, respectively. Similar for retrieval, where a feature in LTM moves to the initial state or closer to forget if observed in the current view or not. Consequently, LTM and STM save the most static and dynamic features based on their observability in the current view, respectively. In a changing environment, the method decreased the localization failure rate compared to a static view. Instead of using a state machine, Bacca et al. (2013) implement a Feature Stability Histogram (FSH) depending on the feature observability to distinguish between STM and LTM features using $k$-means clustering. This modification allows that an input feature in SM can bypass STM and become part of LTM depending on the feature strength. The method was able to filter out pedestrians from 2D laser and camera data and achieved a more accurate representation of the environment compared to a static approach.

Although Biber and Duckett (2009) do not adopt specific memory mechanisms, they implement STM and LTM maps. The method implements a dynamic map as a set of local maps, each maintaining submaps representing different timescales. The timescale parameter of each submap determines probabilistically when to add samples from 2D laser scans. The dynamics of the environment are represented by using five different timescales. The smaller one (~3.1 s) represents an STM map updated at every instant. The other 4 timescales (~0.43 runs, 0.43 days, 3.1 days, and 13.5 day) are LTM submaps updated after each season or daily. Instead of only localizing on the LTM maps as Dayoub et al. (2011) and Bacca et al. (2013), Biber and Duckett (2009) select the best representation of both STM and LTM maps that best explain the sensor data. In a 5 weeks

experiment, the localization with a dynamic map improved while a static representation degraded over time. The use of timescales led to static parts as walls emerging in LTM and dynamic elements disappearing from the STM maps.

Similar to Dayoub et al. (2011) and Bacca et al. (2013), two maps can represent a more stable and a more dynamic representations of the environment. In Walcott-Bryant et al. (2012), an active map represents the most current state of the environment. The active map includes parts that did not change from previous passes and objects added to the environment. A dynamic map only saves the points of a 2D laser scan that changed over time. Wang et al. (2019) use a tracking map with short-term static points and a long-term map. The latter only contains long-term static points identified by a semantic segmentation module with ORB-SLAM2 (Mur-Artal & Tardós, 2017). Also, Zhu et al. (2021) create offline a semi-dynamic map and a static one. The former has semi-dynamic objects (parked cars in a parking lot environment) and the latter map has both static and semi-dynamic objects. The main goal of representing two different dynamics is usually to favor the most stable one in the long-term. Walcott-Bryant et al. (2012) and Wang et al. (2019) only use static parts in the most current representation of the environment (active and the long-term maps, respectively) for localization. In the experiments, Walcott-Bryant et al. (2012) showed that their method was able to identify static parts. However, the method was affected by false positives and negatives, and by the blur effect in the 2D grid map. Wang et al. (2019) improved the ATE in a dynamic environment over ORB-SLAM2 (Mur-Artal & Tardós, 2017) and DynaSLAM (Bescos et al., 2018). As for Zhu et al. (2021), its MCL framework reduces the weight corresponding to observations of moved semi-dynamic objects. The method improved the localization in a parking lot compared to a standard MCL.

### 5.2.2 | Map matching

Environment dynamics can be identified by comparing the current observation to the map. Assuming a prior vector map as a permanent map, Biswas and Veloso (2017) determine the probability of observed features being long-term ones by the 2D laser scan-to-map matching distance. Short-term features are determined by the scan-to-scan matching distance, while the remaining ones are considered dynamic features and not considered for localization. Compared to MCL with a static map and to Tipaldi et al. (2013), Biswas and Veloso (2017) had lower localization error in a parking-lot environment. However, the method would not handle semi-static changes. Instead of using a permanent map, Zhang, Chen, et al. (2019) maintain a Signed Distance Field (SDF) (Curless & Levoy, 1996) representation based on a prior occupancy map. The method rejects dynamic points identified by range flow and updates the SDF-based map with semi-static changes observed in the scan-to-map difference. Compared to MCL in a semi-static environment, the proposed method had lower pose errors and an improved representation of the environment. Boniardi et al. (2019) detect semi-static changes leveraging the ICP (Besl & McKay, 1992) scan-to-map consistency and a CAD prior of the environment, and updates the map accordingly. The method was capable of maintaining a consistent map when dealing with substantial reconfiguration of the environment. Du et al. (2022) minimize the Gibbs energy defined on the proposed Long-term Consistent Conditional Random Field (LC-CRF) for detecting dynamic points. The method considers that dynamic points have often a large reprojection error in frame-to-map matching. Also, the proposed LC-CRF assumes that points tend to have the same dynamic properties as the neighbor ones. In a dynamic scene, LC-CRF achieved lower ATE than ORB-SLAM (Mur-Artal et al., 2015).

Furthermore, Pan et al. (2019) and Ding et al. (2020) leverage clustering properties of the observations evaluating the observations count. Pan et al. (2019) segment the points of a 3D LiDAR point cloud into different clusters. The proposed segmentation assumes that dynamic points do not appear frequently in the same place. The map only considers clusters that appear in same location more than 10 times. As for Ding et al. (2020), the method build on the assumption that dynamic and static parts of the environment have a clustering property relative to its neighbors (similar to Du et al. (2022)). The number of observations in different sessions combined with its consistency relative to its neighbors determine if a map point is static throughout the sessions. Both representations of the environment in Pan et al. (2019) and Ding et al. (2020) were stable to structural changes in the environment.

The concept of ray tracing is also used by the included works to handle dynamic changes. Lázaro et al. (2018) use ray tracing to exploit the free space information. When comparing two 2D point clouds from a viewpoint, the ray tracing evaluation identifies new objects added to the scene (observed point closer to viewpoint than the old one) and outdated information (observed point further way). The ray tracing evaluation allows the identification of dynamic changes and having an up-to-date representation for localization. Given that ray tracing in 3D is expensive in terms of memory and requires dense map representations, Pomerleau et al. (2014) use directly the sparse point cloud from a 3D laser. The map points are associated with each single reading in small conical apertures in spherical coordinates, updating the observed points closer than the mapped ones and the further ones are left untouched. The approximation of ray tracing results are used to update the probability of points in the map to be dynamic, based on a Bayesian approach. The probability of being dynamic can be used in ICP to not trust dynamic points. Indeed, Pomerleau et al. (2014) had a more precise and cleaner map of the environment than using a standard ICP matching in the experiments. Instead of weighting the map points, An et al. (2016) propose the Dynamic Edge Link (DEL) to model the dynamics in the edges of a pose graph instead on the data itself. The observation of moving obstacles between two poses change the weight of the respective edge, decreasing gradually the weight until not detecting the obstacle. Integrating DEL in an exploration scheme, nodes with a edge weight average lower than a threshold are not considered for exploration due to the robot may be unable to move to that position. The lower weight average means

that there are frequent moving obstacles or changed structures near the node.

Although the standard NDT (Biber & Strasser, 2003) representation does not model free space, Einhorn and Gross (2013), Saarinen et al. (2013), and Einhorn and Gross (2015) use NDT with occupancy maps to model explicitly the free space. The two works also adopt exponential weighted moving average and covariance for new measurements having an higher influence than old ones. Einhorn and Gross (2015) propose a generic 2D/3D mapping using NDT and occupancy maps. The hit cells considering the current observation are updated incrementally with exponential weighting. The other cells along the sensor beam potentially empty are updated using the standard update rule of occupancy maps based on the log-odds of the occupancy value and on the inverse range sensor model. Instead of using the standard occupancy map update, the sensor model in Saarinen et al. (2013) depend on the inconsistency between observation and map. Also, the occupancy value describes the confidence of the NDT based on past observations. As for Einhorn and Gross (2015), the method defines two probabilities for the occupancy map: occupancy and statically occupied. The first probability is updated based on the sensor model (2D/3D generic beam sensor), and the second one is adapted slowly to high probability for static objects in the environment. The statically occupancy probability follows the proposed ad-hoc model that is parameterized to control how fast the static occupancy probabilities are adapted, depending also on the occupancy probability itself. Einhorn and Gross (2013) and Einhorn and Gross (2015) were able to handle semi-static and dynamic changes having a consistent and up-to-date representation of the environment, while Saarinen et al. (2013) favored long-term static structures in dynamic environments.

### 5.2.3 | Prediction modeling

In the included works, Markov processes are used to predict the dynamics of the environment. Tipaldi et al. (2013) use a dynamic occupancy grid and exploit the stationary distribution and the state holding time associated with Hidden Markov Models (HMM) (Rabiner, 1989) on a 2D grid. The method uses past observations for each run to learn the state transition probabilities iteratively to estimate the HMM parameters. Then, the localization can infer how often is expected to see a dynamic object in the environment and for how long. Comparing the proposed HMM-based localization to MCL using a standard grid, the former had a lower localization failure rate than MCL, capable of dealing with high dynamics (moving cars) and lower ones (parked cars). Rapp et al. (2015) implement a semi-Markov process extended by a Levy process to model a time dependency on the state holding time of Markov processes The proposed method also predicts the expected retention time for each cell being in a state as Tipaldi et al. (2013). In the experiments, the proposed model integrated in MCL improved the classic MCL in a dynamic environment.

The environment dynamics can have periodic patterns associated with them. Assuming periodic changing patterns, Krajník et al. (2017) propose the FreMEn (Frequency Map Enhancement) to model the probability of occupancy or feature visibility in a grid as a combination of harmonic functions related to periodic processes. FreMEn uses spectral analysis (Fourier transform) to compute the harmonic functions and predict future state with a given confidence. In a changing environment, FreMEn outperformed a static map and experience maps (Churchill & Newman, 2013) in terms of localization error by selecting the most likely visible features at each location for localization. Santos et al. (2016) adopt the FreMEn within an exploration scheme. The proposed planner predicts which areas are more likely to change at a certain time and generate the subsequent locations to explore. The experimental results showed that considering the environment dynamics increases the amount of information gathered compared to static models. Unlike FreMEn, Wang et al. (2020) model both aperiodic and periodic changes by an Auto-regressive Moving Average Model (ARMA) (Hassler, 2016). This model describes time series as stationary stochastic processes in terms of polynomials. While FreMEn is able to update recursively its model online, ARMA only is updated once a day based on past observations. However, the model achieved a higher prediction accuracy than FreMEn and a lower localization failure rate than both FreMEn and Tipaldi et al. (2013).

Instead of modeling the dynamics in the map, Thomas et al. (2021) use a KPConv (Thomas et al., 2019) network to predict online dynamic motion labels of points with single 3D laser scans as input. The method is a self-supervised learning approach with two main modules: PointMap and PointRay. PointMap is an ICP-based SLAM algorithm to provide a point cloud map for the annotation process. PointRay uses a similar approach to Pomerleau et al. (2014) to approximate ray tracing using spherical coordinates and detect dynamic observations. Then, the PointRay module annotates the data for training with the following dynamic labels: permanent (static points over all sessions), ground (to avoid ray tracing ground samples), and long-term (still objects in single sessions but relocated between sessions) and short-term (dynamic objects) movables, with the localization not considering the latter two. In the simulation experiments, PointMap with the proposed prediction module led to lower localization pose errors than an MCL algorithm.

### 5.2.4 | Dynamic objects detection

In terms of detecting dynamic objects, Yue et al. (2020) propose a collaborative dynamic mapping to detect humans using visual and thermal images and a 3D LiDAR. The YOLOv3 (Redmon & Farhadi, 2018) algorithm extracts the bounding boxes from the images relative to humans. The 3D point cloud projection onto the images allows the creation of a static point cloud for localization and mapping of each robot by filtering out the points corresponding to humans. In the experiments, the dynamic objects removal generated a more accurate relative transformation of the collaborative maps compared to not removing those objects. Zhu et al. (2021) also use YOLOv3 to extract bounding boxes of dynamic classes (parked cars)

from a RGB image. The projection of 3D LiDAR points allows the creation of the static and semi-dynamic maps required for localization. Even though Zhu et al. (2021) do not discard the dynamic objects, the localization module reduce the importance of weight of the corresponding observations. Instead of using visual data, Sun et al. (2018) adapt the PointNet (Charles et al., 2017) for object recognition (pedestrian, cyclist, car, or background) to classify the scan points of a 3D LiDAR. The proposed Recurrent-OctoMap maintains the occupancy and semantic information in the map cells, whereas the latter specifies the cell semantic state and the probability of the prediction. The transition between states is learned by a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) (Hochreiter & Schmidhuber, 1997). In a long-term experiment, the method was able to improve its 3D semantic map compared to a standard Bayes update.

Moreover, pixel-wise semantic segmentation is another way to identify dynamic objects. Additionally to the proposed semantic-descriptor in G. Singh et al. (2021), the method sets lower weights to features detected on sky and dynamic classes (person, car, etc.) from the semantic segmentation of EdgeNet (Plastiras et al., 2019). Instead of identifying object classes, Song et al. (2019) propose the MD-Net CNN to segment a grayscale image into unstable, static, and moving pixel points, only using static points for localization. The localization error was reduced compared to not estimated the pixel dynamic attribute. Ganti and Waslander (2019) propose the Semantically Informed Visual Odometry (SIVO) to improve the performance of ORB-SLAM2 (Mur-Artal & Tardós, 2017) by using the Bayesian neural network SegNet (Alex Kendall & Cipolla, 2017) for segmentation and computation of the network uncertainty. SegNet is trained to distinguish different object classes for identifying dynamic objects (sky, car, truck/bus, person/rider, motorcycle/bicycle, and void) from static ones (road, traffic sign, building, wall/fence, pole, vegetation, sidewalk, traffic light, and terrain). Only static keypoints that reduce the most of the state's uncertainty (considering the network uncertainty) are considered as input for ORB-SLAM2. SIVO was able to remove uninformative and dynamic keypoints from the current frame. However, the rejection of potential dynamic objects without verifying if they are moving reduced the localization performance of the module in certain scenarios.

Dynamic objects identification can be improved by verifying geometric constraints. Bescos et al. (2018) propose DynaSLAM as a front end for ORB-SLAM2 (Mur-Artal & Tardós, 2017) to segment potential dynamic classes using a Mask R-CNN (He et al., 2020). The semantic labeling is improved using a multiview geometry verification. DynaSLAM outperformed ORB-SLAM2 in highly dynamic scenarios while having similar accuracy in static ones. However, its performance reduced in slower dynamics. Similar to DynaSLAM, Wang et al. (2019) implement a front end for ORB-SLAM2 to identify movable objects with a ResNet (K. He et al., 2016)-based network for segmentation. The segmentation of the previous frame and a geometric verification step based on the reprojection error improves the labeling of dynamic objects. The method improved the ATE over DynaSLAM and ORB-SLAM2 in a scenario with movable objects. Instead of using semantic segmentation, the Semantic and Geometric Contraints Visual SLAM (SGC-VSLAM) (Yang et al., 2020) uses YOLOv3 to extract bounding boxes of dynamic objects for also improving ORB-SLAM2. A constraint based on epipolar geometry improves the labeling. SGC-VSLAM decrease the Root Mean Square Error (RMSE) of the ATE by 96% compared to ORB-SLAM2 in highly dynamic environments. However, similar to DynaSLAM, its performance decreased in lower dynamics. Finally, Xing et al. (2022) propose the DE-SLAM to deal with Short-Term Dynamics (STD) and Long-Term Dynamics (LTD) at the same time. A MobileNetv2 (Sandler et al., 2018) identifies bounding boxes of movable objects (cars, persons, etc.) classified as STD. A motion check of STD elements recognizes all moving objects in the current keyframe. As for LTD, DE-SLAM uses HOG (Dalal & Triggs, 2005) features extracted from ORB (Rublee et al., 2011) keypoints to improve its invariance to illumination changes. In the experiments, DE-SLAM improved the localization over ORB-SLAM2 in a changing environment. All of these methods using geometric constraints to improve the identification of dynamic objects only use static features for localization and mapping.

## 5.3 | Map sparsification

The next subject in this discussion is the analysis of the included works categorized as sparsity in DE1 (see Table A1). The methodologies proposed in those works manage the map size of the environment representation perceived by the mapping agent, where the size should be dependent on the operation area and not on the trajectory length of the robots. Hence, the discussion on map sparsification is organized into the following topics: (1) sparsification of graph SLAM to remove redundant nodes or outdated environment observations, (2) management of the keyframe graph and its features relative to the keyframe formulation of the SLAM problem, and (3) generic sparsification methods proposed in the included works for feature maps.

### 5.3.1 | Pose graph SLAM

In graph-based SLAM, the constant update of the map leads to the ever-growing problem of the pose graph, where most basic approaches grow with the length of the trajectory or operation time. However, this growth should be bounded only by the size of the mapped environment. Information-theoretic methods focus on removing redundant nodes based on the concept of mutual information for limiting graph growth (Kretzschmar & Stachniss, 2012). Outdated nodes should also be removed to limit the graph size and update the current map representation upon changes in the environment. Additionally, the spatial distribution, time recency, and fusion of information onto existing nodes can also enable the reduction of nodes over time.

*Mutual information*

The concept of mutual information from information theory can be used to determine which nodes to remove from the pose graph. Kretzschmar

et al. (2010) and Kretzschmar and Stachniss (2012) estimate the expected information gain of a node based on its entropy contributing to the robot's pose belief in the current pose's neighborhood. The nodes with the lowest information gain are removed until the value is greater than a threshold. Kretzschmar and Stachniss (2012) also set a limit on the total number of nodes of the graph. However, these two methods differ on how to marginalize the edges. Kretzschmar et al. (2010) remove all $N$ edges of a removed node and adds $N - 1$ edges between the removed one and a neighbor node. The latter is selected as the node that minimizes the edges length of the affected nodes by the removal. Effectively, the method only decreases by 1 the total number of edges per node removal. As for Kretzschmar and Stachniss (2012), this method summarizes the information of the original edges into the nodes that are kept using an approximate marginalization that preserves sparsity. This approximation is based on using Chow-Liu trees (Chow & Liu, 1968) to approximate a local probability distribution of the graph minimizing the relative entropy, or also known as the Kullback-Leibler Divergence (KLD), to reduce the number of edges locally. In the experiments, both Kretzschmar et al. (2010) and Kretzschmar and Stachniss (2012) stabilize the number of nodes and edges over time, limiting the computational requirements of online execution, while full marginalization (densely connected graph after removal) leads to an increasing number of edges.

The works focused on edge marginalization assume the prior selection of the node for removal. Carlevaris-Bianco et al. (2014) propose the Generic Linear Constraints (GLC) to produce a set of constraints over the subset of nodes affected by the node removal. These constraints can produce either the full marginalization (dense GLC) or a sparse approximation using a Chow-Liu tree (sparse GLC). The repeated application of a sparse GLC node removal only had a low difference in both mean pose error and KLD compared to the full graph. Ozog et al. (2016) apply the same graph marginalization as Carlevaris-Bianco et al. (2014) on a pose graph map obtained with an underwater vehicle. The system also had similar KLD compared to full graph. Huang et al. (2013) formulate an $l_1$-regularized optimization problem to minimize the KLD of the approximation estimating GLC from the discarded measurements. The proposed method did not impact the localization error while also reducing by 77% the nonzero elements of the information matrix, improving the sparsity of the graph. As for Mazuran et al. (2016), this work proposes the Nonlinear Factor Recovery (NFR) edge marginalization. The proposed marginalization approach searches for the set of nonlinear factors that best represent the marginal distribution of the subset of nodes affected in the removal in terms of KLD. The method considers both global and local linearization points. Mazuran et al. (2016) demonstrated that NFR is equivalent to GLC when using only relative measurements. In the experiments, NFR tended to achieve similar or improved performance relative to GLC between the sparsified and full graph, having similar results or improved KLD depending on good or poor linearization points, respectively.

Other methods in the included works do not require edge marginalization. Maddern et al. (2012) do not impose global geometric corrections on loop closure to ensure similar odometric sequences on different passages in the same locations, only requiring the update on existent odometric edges. The method sets a maximum limit on the number of nodes eliminating the ones with lowest relative information content computed by the negative log of odometric and appearance-based matching likelihoods. The algorithm stabilized at a constant execution time and memory occupation due to the limit of nodes. Ila et al. (2017) propose an incremental solution to decide whether a node should be added or not. Only nodes part of informative links or establishing informative links are added to the graph, avoiding to add unnecessary edges and thus not requiring edge marginalization. Although the method was able to slow the growth rate of nodes and edges in the experiments, the method was not able to bound it. Egger et al. (2018) filter all poses with an overlap with their closest neighbors of the respective submaps higher than a threshold. The removal also updates the scores of affected neighbors relative to the number of observations required for evaluating the stability of the 3D LiDAR surfel features. Considering an overlap threshold of 0.6, the resulting map in the experiments was 9.4MB compared with the 5GB of the initial point cloud map.

### *Outdated information*

Instead of selecting nodes based on mutual information, the removal of outdated nodes based on the current information can limit the size of the pose graph. Walcott-Bryant et al. (2012) remove inactive nodes that do not represent the current state of the environment, considering the creation of nodes for each run to be able to create the active and dynamic maps required for dealing with dynamic environments. The removed points labeled in the dynamic map (points no longer present in the active map) over time allow the identification of inactive nodes. In the experiments, the method was only able to remove approximately 50% of the nodes and edges compared to the full graph. Tang et al. (2019) filter submaps in the proposed manifold navigation considering the number of successful localization in each location versus the attempted ones to indicate the outdated submaps. Even though the number of nodes stabilizes over time, the stabilization only happens on the third day, possibly due to the proposed manifold navigation treating new conditions of the environment as new nodes. Boniardi et al. (2019) also evaluate the current localization of the robot to select nodes for removal. The method prunes outdated nodes when the pose's belief drops below a tolerance level, possibly related to changes in the environment. Additionally, upon loop detection, the method builds a local map from the subset of nodes candidate for loop closure. This local map allows for discarding candidates that are not topologically consistent with the local environment surrounding the robot using ray tracing, avoiding the addition of unnecessary edges. The method was able to limit the graph size in the same environment over multiple runs, being dependent on the size of the operational area and not on the operation time.

### *Spatial density and time recency*

The spatial distribution of the nodes is another approach to evaluate the graph sparsity. Johannsson et al. (2013) propose an incremental approach for managing the addition of new nodes, similar to Ila et al.

(2017). However, Johannsson et al. (2013) only add a new node if there is no existing node in the spatial proximity of the current position, instead of being based on mutual information. If a loop is detected, though the first one would not generate any new node due to the spatial constraint, in the case of a second loop, the method compounds the chain of constraints between the first and second loops to add a new constraint. In contrast to Johannsson et al. (2013), Zeng and Si (2021) add nodes upon revisiting locations for optimizing the pose graph. Then, the method identifies redundant nodes clustering loop closure edges to identify similar trajectories. Both works of Johannsson et al. (2013) and Zeng and Si (2021) showed reduced growth compared to the full graph in the number of nodes. The growth also seemed to stabilize in the experiments.

In addition to the spatial density, the time recency of the node can be another factor for selecting nodes for removal. Kurz et al. (2021) try to keep the spatial density of nodes below a certain threshold across the entire map. The proposed scale-invariant density measure determines to remove the nodes with the highest densities, until the density is lower than a threshold. The edge marginalization in the graph follows a similar approach to Kretzschmar et al. (2010). The removal process keeps the most recent nodes from being removed, even though the density measure is computed considering all nodes. The method reduced the growth in the number of nodes compared with the full graph, having stabilized over time. Similarly, Ali et al. (2021) also favor older nodes for removal, moving those nodes after a certain traveled distance from the online graph to an offline database to not lose the information. Also, upon loop detection, the older submap is substituted with a new one. Compared to ORB-SLAM2 (Mur-Artal & Tardós, 2017), the proposed method had lower computational requirements in terms of CPU and memory usage.

*Information fusion*

The information fusion between nodes and/or with the current observation also allows the reduction of the graph growth. Both Einhorn and Gross (2013) and Einhorn and Gross (2015) fuse NDT (Biber & Strasser, 2003) map fragments that cover a similar region of the environment, only fusing nodes whose relative pose is known with low uncertainty. The marginalization of the affected edges is performed using the same approach as Kretzschmar et al. (2010). In the experiments, the number of vertices in both works did not increase significantly on revisiting locations. Assuming that a loop closure occurs due to spatial closeness having overlap between the affected nodes, Lázaro et al. (2018) merge loop closure-related nodes (including intra- and inter-sessions) following a similar methodology to the ray tracing one used for detecting dynamic changes in the environment. The oldest point cloud is used as a reference for refinement with the newest one based on their timestamps. Edge marginalization is based on condensed measurements where the remaining node is connected to the neighbors using a star-like topology. The method achieved a 50% node reduction in an experiment while retaining the localization accuracy. Karaoğuz and Bozma (2020) use a Topological Spatial Cognition (TSC) model to organize the visual place memory as a collection of appearances and respective descriptors for each robot, with a hierarchical organization to cluster places with a similar appearance. Based on the similarity of the descriptors, Karaoğuz and Bozma (2020) merge place memories of TSC models on a multirobot system, incorporating all places known by other robots but not known to itself. The merge is performed based on the nature of overlap of the descriptor hyperspheres in appearance space. Although the method was able to merge the TSC models between two robots leading to 18 final locations, the merged locations are more than the 15 predicted ones due to limited field of view and appearance changes.

### 5.3.2 | Keyframe SLAM

A specific formulation of the pose graph is the keyframe SLAM. This formulation selects keyframes from the frames usually captured by a camera. Then, the 3D map points are triangulated considering features extracted from the camera images, and the edges determine the keyframes that observe the map points. Odometry and shared observations of map points induce additional edges between keyframes (Schmuck & Chli, 2019). The keyframe SLAM also suffers the same ever-growing graph problem as the standard pose graph formulation. Thus, long-term localization and mapping methods using keyframe SLAM should employ policies to manage the growth of the number of keyframes and map points, or restricting the selection of the frames from the views captures from a sensor to ensure the graph sparsity.

*Keyframe graph management*

Clustering techniques are employed in the included works to identify keyframes for removal. Konolige and Bowman (2009) propose the Least-Recently Used (LRU) algorithm to limit the keyframes in a local neighborhood, while preserving their diversity and removing preferably older unmatched views. LRU clusters keyframes based on its feature matching closeness, assuming that keyframes capturing similar environment appearances will be in the same cluster. If the number of local keyframes exceeds a threshold, older ones are removed from each cluster. Then, having clusters only with a single view, LRU removes the oldest exemplars. LRU reduced significantly the number of views and edges relative to having no management rule for keyframes while preserving the mean number of clusters in local neighborhoods, that is, preserving the diversity. Instead of evaluating single keyframes, Bouaziz et al. (2022) exploit the similarity between traversals in the keyframe map that represent runs possibly in different environment conditions. Hierarchical clustering on a proposed similarity matrix between traversals identifies which one to remove when their number exceeds a predefined threshold, trying to maintain the map diversity as possible as in Konolige and Bowman (2009). The limitation on the number of traversals bound the computational requirements of the method.

Moreover, Gadd and Newman (2016) implement a merge process in its centralized versioning framework to measure the

relevance of discovered segments by every single agent using stereo matching from visual odometry. The merging strategy with multiple agents built the map in an experiment in 3.6 h, while a single agent would require 12.4 h. The size of the merged database obtained from multiagent mapping was smaller than a single agent due to the redundancy check. Similar to methods for sparsification of the standard pose graph, Ding et al. (2019) use the KLD measure to determine which keyframe to remove based on their contribution. When adding a new one, the method updates the KLD of each keyframe that has common features to the new frame. The keyframes with KLD lower than a threshold are removed using GLC (Carlevaris-Bianco et al., 2014) for edge marginalization if needed. The proposed sparsification approach reduced the map size for transmitting it between an external agent and the robot.

Although the keyframe management removes map points indirectly, for example, when the points are not well constrained with less than two observations (Schmuck & Chli, 2019), management techniques on the keyframe graph can also remove points directly from the feature map. LLamaSLAM (Luthardt et al., 2018) considers only high-quality Long-term LandMarks (LLama) points (persistent features selected from the tracked ones with VO) for adding to the map The method ensures the spatial coverage of map points in a 2D grid selecting only the best 10 points within each cell. The keyframe selection is based on the overall quality threshold of the observed LLama points in the frame. Furthermore, ORB-SLAM (Mur-Artal et al., 2015) implements keyframe and feature addition and removal rules. ORB-SLAM only adds a keyframe if the current view tracks at least 50 points in the sliding window and less than 90% of the points of the current reference keyframe. ORB-SLAM also discards all keyframes whose 90% points are seen in at least more than three other keyframes. As for map points, the points are only retained if the tracking finds them in more than 25% of the frames which are predicted to be observed in and must be observed from at least three keyframes. Zhang, Chen, et al. (2018) implement the same map management method as ORB-SLAM in a multirobot system to reduce redundant data, where each robot executes independently monocular SLAM and communications their map with other robots. Instead of using thresholds, Schmuck and Chli (2019) modify ORB-SLAM with a redundancy score to classify the map points. The method defines a maximum score of 1 for features seen in more than five keyframes (also removing these points from the map) and a score of 0 for the minimum of two observations. Then, the proposed approach scores the keyframes using the normalized sum of their features' scores. Considering a maximum limit of keyframes, the algorithm was able to compress up to 50% relative to no management while also outperforming the original ORB-SLAM in RMSE.

The selection process of features for removal in the keyframe graph can use more than a single scoring function. Dymczyk et al. (2015) score the map points on the number of observations (considering a lower bound for removal, while restraining the deletion in rarely visited areas and not well-constrained points), the descriptors variance (rejecting the ones with high variances), and the number of observations between different sessions. The method also sets a minimum number of keypoints to retain a keyframe and a total limit on the number of map points. These policies led to an approximately constant number of keyframes and map points in posterior sessions in the same environment.

*Representative keyframes*

Another approach to reduce the growth of keyframe graph found in the included works is to restrict the selection of the keyframes from the sensor live data. Pirker et al. (2011) only add keyframes to the graph if at least 55% of the image area is covered by keypoints used for tracking, in contrast with the standard rule of adding frames accordingly to a certain motion relative to the previous keyframe. Also, the method removes points based on their visibility accordingly to the Histogram of Oriented Cameras (HOC) (Pirker, 2010) descriptor that represents a map point. In the experiments, the method was able to keep the map size proportional to explored space stabilizing the number of features.

Topic-probabilistic models used in both Paul and Newman (2013) and Murphy and Sibley (2014) also try to identify the most representative keyframes from camera images. Paul and Newman (2013) apply Latent Dirichlet Allocation (LDA) (Blei et al., 2003) as a probabilistic topic model to identify perplexing observations and retrieve images similar in thematic content. The proposed approach provided a compact representation of the sampling set to improve the timing efficiency of FAB-MAP (Cummins & Newman, 2008). Topic models provide a low-dimensional representation of BoW (Sivic & Zisserman, 2003), capturing their thematic content via word-occurrences. Although the method reduced the number of generated keyframes, Paul and Newman (2013) did not implemented a deletion or forgetting rule for the map frames. As for Murphy and Sibley (2014), their approach implements probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) as the topic modeling engine in an incremental manner to expand the vocabulary and perform topic updating online. The topic clustering using DBSCAN (Ester et al., 1996) identifies temporally smoothed unique places. In the experiments, the proposed method retained up to 1.06% of the image stream while having similar precision and recall metrics relative to considering all frames, respectively.

### 5.3.3 | Features management

In addition to feature management policies applied in keyframe SLAM and discussed previously, the included works also propose techniques for sparsification of feature maps to improve the mapping scalability in the long-term time frame. Hochdorfer and Schlegel (2009) and Hochdorfer et al. (2009) focus on ensuring the spatial distribution of the features in the environment, removing the ones that cover nearly the same region. The first clusters SURF (Bay et al., 2006)-based features based on the $l_1$ distance of their 2D position in the map using $k$-means with the number of clusters as 25% of known features in the map. Considering feature information content as dependent on the covariance matrix, the proposed approach selects

the cluster with the maximum difference in information content and removes the feature with the lowest localization benefit. The method also limits the number of map features to 50. This limit constrains the computation requirements of the method. The second work uses DBSCAN (Ester et al., 1996) instead of $k$-means because the former is a density-based algorithm while the latter is a partitioning one. Even though both Hochdorfer and Schlegel (2009) and Hochdorfer et al. (2009) limit the number of map features and improve their spatial coverage, DBSCAN clustering led to a better spatial distribution of features compared to $k$-means.

The evaluation on matching observations with the features in the map versus the number of attempts could indicate unstable or dynamic features that could be removed from the map and reducing its size. Davison and Murray (2002) delete features after a predetermined number of matching fails when they should be visible and sets a maximum of two visible features (minimum for the robot to localize itself) in the map due to the computational limitations of the experimental setup. Even though the limit was to improve the computational efficiency with the resources available at the time, the map should have consider more features to improve localization robustness to feature occlusion. Similar to Davison and Murray (2002), the STM/LTM memory scheme implemented by Dayoub et al. (2011) and Bacca et al. (2013) imposes a consecutive observation of the features for retaining them in the map. Bacca et al. (2013) showed in the experiments that removing useless and old features avoided the ever-increasing number of features, leading to an approximately constant map size over different runs.

Furthermore, the use of multiple predictors for evaluating feature stability on changing conditions also helps the sparsification of the feature map. The removal of features with low scores on the predictor proposed by Berrio et al. (2019) had similar localization covariance in the experiments while removing up to 70% of the least valuable features in the map. Berrio et al. (2021) also evaluate the concentration ratio and maximum driven length predictors used in their work. Features in high concentration areas and low visibility in terms of the maximum driven length while observing them are discarded from the map. The removal method contributed to the map size being approximately constant in later runs throughout a 24 weeks experiment. Similar to Berrio et al. (2019), Dymczyk, Schneider, et al. (2016) formulate a regression to optimize the weights given to the predictors and combine all predictors into a single score. The method considers as predictors the number of frames the feature is re-observed, traveled distance while observing the landmark and the one between the two most distant keyframes while tracking, maximum angle between observation rays, the mean reprojection error, a gravity constant to favor anchored objects presumable more useful for localization, the vertical coordinate, and the descriptor appearance classification. The results showed a 80% reduction on data transfers with similar localization recall when selecting a subset of the map features compared to retaining all features. Also, Mühlfellner et al. (2016) create a Summary Map from a map gathered over multiple traversals in the environment by selecting a limited number of features, first, based on the observations in distinct traversals, and then, on the total number of observations. The authors compared the Summary Map with 1200 features to only selecting features seen during the most recent traversals, features seen in two or more traversals, and to the works of Konolige and Bowman (2009) and Dayoub et al. (2011). The Summary Map, Konolige and Bowman (2009), and retaining features seen in two or more traversals achieved the higher localization accuracy, while the Summary Map had higher accuracy than the other methods at the same map size.

In terms of managing a BoW (Sivic & Zisserman, 2003) dictionary, Tsintotas et al. (2021) present an incremental BoW model to remove multiple codewords of repetitive patterns representing the same environmental elements at different time instants. A spatial check identifies the redundant words upon loop closure. These words are ignored if not associated with the chosen loop image and are merged with the ones in the database accordingly to the median of the descriptors. The incremental approach reduced the model size compared to other BoW-based approaches while also improving the timing efficiency due to having fewer visual words for the search of loop closures. Instead of managing the addition and removal of words from the BoW model, Opdenbosch et al. (2018) propose a culling map point algorithm. This algorithm minimizes the points coding cost to keep map points that exhibit good compression properties. Also, the algorithm favors the map points with many visually similar observations when assigning the features descriptors to its closest visual word from a pretrained BoW model. The method allows the definition of a prior size for the desired BoW model to constrain the computation requirements. The integration of the proposed method in ORB-SLAM2 (Mur-Artal & Tardós, 2017) reduced by three times the map size (3 MB to 1 MB) having a similar localization success rate, while also reducing the number of points substantially (17,426 to 2370 points).

Lastly, the works of Schaefer et al. (2021) and Wang et al. (2021) that retrieve pole features from 3D LiDAR data for appearance invariance in changing conditions also employ feature management policies to avoid redundant points in the maps. Schaefer et al. (2021) merge ambiguous poles by projecting them onto the ground plane and evaluate their overlap. The merge process computes a weighted average over their center coordinates and widths over the mean pole score, determined by averaging over the scores of all voxels that touch the pole. Wang et al. (2021) segment the point clouds into clusters based on the semantic labels obtained with the RangeNet++ (Milioto et al., 2019) network. For each cluster, the label is voted by the statistical number of the point labels in the cluster. Considering the clusters of the global map versus the ones found in the current laser scan, each cluster of the current scan is searched by the closest neighbor and is only added into the map if the cluster is not found in the global map.

## 5.4 | Multisession

This section analyzes works categorized as multisession in DE1 (see Table A1) focusing on methodologies for dealing with the start of the

robot in each operation session. A multisession system must handle the data acquired in each session by a robot without having a prior initial pose relative to the current map. This system should avoid the restart of the mapping procedure in all runs while being capable of localizing the robot in the existing map (Labbé & Michaud, 2019). In the context of long-term localization and mapping, a multisession solution is desirable to integrate new information acquired over different operation runs without requiring a known initial pose for accomplishing a continuous autonomous operation in a changing and dynamic environment. Even though global localization is required for multisession to localize the robot in a known map without any prior knowledge, Sections 5.1 and 5.2 already discuss methodologies for topological and metrical localization robust to changes and moving elements in the scene.

One methodology found in the included works is the implementation of global multisession where both localization and mapping processes consider a global common frame between all sessions (Ozog et al., 2016). Bürki et al. (2019) implement an offline process to localize a new mapping session against an existing map. This process generates an initial pose estimation for the vehicle in the new data set and the feature association between the new data set and the features in the map. New map points are created from unmatched features in the multisession association step. Then, the system optimizes the resulting global multisession map again with bundle adjustment.

Instead of assuming that the localization is always possible to perform in the current map similarly to global multisession, another possibility is considering independent sessions at the beginning of each run. Latif et al. (2012) propose the Realizing Reversing Recovering (RRR) algorithm that defines two types of loop closures: intra- and inter-session. If no inter-session loop, RRR considers all sessions unconnected between them. Assuming that the front-end of graph SLAM deals with odometry outliers, the graph error introduced by each constraint would be caused only by loop closing links. Then, RRR clusters loop hypotheses accordingly to its impact on the graph error, where small and similar errors versus greater and contradictory would be caused by correct and false loop closures, respectively. The method was able to recover from 600 wrong loop closures in a four session experiment. Oberländer et al. (2013) start a new mapping session in each run. Eventually, the independent graph will be connected to previous sessions by matching the submaps of the graph nodes, similar to the inter-loop considered by Latif et al. (2012). Similar implementations are considered by Mühlfellner et al. (2016), Lázaro et al. (2018), and Labbé and Michaud (2019), allowing inter-loop links connect different sessions when the data association module of localization and mapping finds a candidate loop between sessions. Also, the experimental results in Mühlfellner et al. (2016) showed an improvement in terms of increasing localization recoveries from failures when considering a multisession map compared to single-session.

The representation of independent mapping sessions can be represented in the same graph using anchor nodes as Ozog et al. (2016). Each robot session has an associated anchor node containing the transformation from the global to the session's reference frames. This representation allows individual sessions to optimize their pose graphs before any links are formed between sessions while allowing faster convergence than global multisession in the pose graph formulation of the SLAM problem. Additionally, Ozog et al. (2016) use GLC constraints (Carlevaris-Bianco et al., 2014) for graph sparsification, as discussed previously. The method was capable of merging a multisession experiment with 12 sessions 3 years apart.

## 5.5 | Computational

Next, this section discusses the works categorized as computational in DE1 (see Table A1). These works focus on computational concerns of long-term localization and mapping apart from map sparsification, as the latter subject is already discussed in Section 5.3 and belongs to another category in DE1 (sparsity). The discussion is organized by the following topics: (1) mechanisms to manage map storage, (2) techniques for reducing the descriptor dimensions, (3) parallel computing, and (4) timing efficiency improvements for place recognition in long-term localization and mapping.

### 5.5.1 | Map management

Although map sparsification techniques try to maintain constrained memory and processing requirements by removing redundant or outdated information from the map, these techniques could not be enough to guarantee computational stability in online execution. Thus, a methodology followed by Oberländer et al. (2013) and Labbé and Michaud (2019) only maintains a sampled version of the map in RAM while the remaining part or the whole map is saved in the disk memory. Oberländer et al. (2013) make the proposed multisession submap graph compatible with serialization, allowing its nodes to be transparently swapped out to disk. The online execution maintains only a smaller version of the entire graph in RAM. This version allows faster localization estimates, where the full-resolution map can be brought back into memory on demand when detailed comparisons are needed for localization. The oldest scans are moved to disk upon adding new ones to limit the memory to a constant size. As for Labbé and Michaud (2019), the proposed Real-Time Appearance-Based Mapping (RTAB-Map) implements a memory system resembling the one adopted by Dayoub et al. (2011): STM, Working Memory (WM), and LTM. However, these memories only define which nodes of the graph are considered in the online execution. Indeed, STM assembles the sensor data into a node for adding to the graph, WM is the nodes considered for operation, and LTM are nodes transferred from WM to satisfy the online requirements of RTAB-Map, where LTM is an SQLite offline database. This transfer is dictated by a weighting mechanism to favor frequently observed locations to be preserved in WM. Both works satisfied the online requirements of the respective localization and mapping algorithms in the experiments due to capping the memory and computation requirements of the online execution.

In the context of Earth-scale mapping, Kim et al. (2021) present a Geodetic Normal Distribution (GND) map structure. A geodetic quad-tree tiling organizes the Earth's surface into spatial tiles with the same angular size in latitude and longitude directions. The unique identification number inferred by location allows real-time searching. This tiling organization allows large-scale localization and a way to manage submaps of different locations on the Earth. Also, the 3D LiDAR point cloud conversion into normal distributions compresses the map size of each location. The method was able to compress the map size by 85% relative to only considering a point cloud while satisfying the localization requirements of the experimental setup. Also, the method was able to map and localize three different continents (Europe, Asia, and America) with the GPS information inferring the quad-key tile of vehicle's localization.

### 5.5.2 | Descriptor dimension reduction

The reduction of features descriptor dimensions can reduce the overall map size and increase the efficiency in feature searching and matching. An example is the work of Bosse and Zlot (2009) that performs a nonlinear normalization on each descriptor while reducing the proposed moment grid descriptor's dimensions to remove elements with low signal-to-noise ratio. Both normalization functions and dimension reduction require training their respective parameters in map data from a similar environment to the expected one. The method showed that reducing the dimension of feature descriptors reduces the computation required for nearest searching neighbors and the map size.

One of the most used techniques to reduce the descriptors' dimensions of CNN-based features is Principal Component Analysis (PCA). This compression algorithm requires learning its model using extracted features from a database of example images as training data. Both Taisho and Kanji (2016) and Camara et al. (2020) used PCA for reducing from a 4096-dim AlexNet (Krizhevsky et al., 2012) and a 25088-dim VGG16 (Simonyan & Zisserman, 2015) descriptors to 128 and 100 dimensions, respectively. The results presented in both works showed that PCA does not affect significantly the accuracy of the place recognition methods while reducing the computation time. Also, PCA can be combined with whitening as in Piasco et al. (2021) to make the features less correlated with each other and have all the same variance while reducing the descriptor dimensions.

Furthermore, the random selection of descriptor components to reduce its dimensions is employed in Naseer et al. (2017) and Xin et al. (2017) due to not requiring a learning phase, unlike PCA. Naseer et al. (2017) use Sparse Random Projection to embed its FastNet (Oliveira et al., 2016)-based descriptor with approximately 130,000 dimensions into a reduced 4096-dim descriptor. The compression procedure achieved similar f-score, and precision-recall metrics compared with the uncompressed descriptor. Xin et al. (2017) evaluated the compressing of the 64,896-dim AlexNet-based feature used in their work using a random selection of the descriptor's components. The results showed similar precision while speeding up 17 times the matching speed and achieving a compression ratio of 93.7%, considering descriptors with 4096 dimensions. Also, max-pooling the descriptors of CNN-based features by channel is another technique for dimension reduction. Yu et al. (2019) use 4-max pooling to reduce 1024-dim descriptors into 256 dimensions. Even though the accuracy of place recognition was slightly lower than PCA in the experiments, the pooling reduction scheme had lower computational complexity.

Semantic hashing implemented in Ikeda and Kanji (2010) focuses on learning a compact binary code for image retrieval. The method uses *gist* (Oliva & Torralba, 2001) scene descriptors as input to the network architecture that progressively maps the high-dimensional input vector to lower dimensions. The network's output is binarized by a threshold learned in the training phase to obtain the final lower-dim binary descriptor. The binary code allows searching directly in a hash table. In the experiments, the method achieved a compact feature representation scalable for large environments, with an 8 KB visual dictionary, and a 5.3 MB of visual words on approximately 20 km in mapping and localization trajectories. However, semantic hashing had the same problems as pretrained dictionaries as BoW (Sivic & Zisserman, 2003), where there is a semantic gap between the data set used for learning the dictionary and the one for localization.

### 5.5.3 | Parallel computing

In terms of improving the overall efficiency of the computational resources, parallel computing allows the simultaneous execution of algorithms. Williams et al. (2014) propose concurrent filtering and smoothing for pose graph formulations of SLAM to achieve faster updates of the current solution while optimizing simultaneously the full graph even in presence of loop closures. The method factorizes the graph into three groups: a small number of most recent states, a large group of states for global smoothing, and separator states to make the filter and smoother ones conditionally independent. If the computation of the filtering stage exceeds a real-time threshold, the algorithm moves the older states in filtering to the smoother thread for further optimization. Periodic synchronization exchanges updated information between the filter and smoother threads after concurrent updates while also accounting for delays. The method achieved a constant filtering time update while concurrently performing full optimization of the graph. Even though the smoother update time increased over time, the smoother optimization required to improve the consistency of the graph over time did not interfere with the filtering stage. Yang et al. (2021) also use multithreading computing. The multithread approach splits an image into multiple grids for high parallelism and feature extraction for loop closure. The results showed a reduction in computational complexity by checking four candidates in parallel for each frame.

The use of external computation resources allows parallel computation of different algorithms, even though delays may occur. Ding et al. (2019) divide the localization task between the robot and the cloud. The latter is responsible to maintain a map and refine the localization estimation of the robot considering visual-inertial

odometry constraints sent to the cloud. This refinement performs an alignment of the data received from the robot to the laser and visual points of the cloud map. As for the robot, only maintains a local sliding window and implements a delayed state EKF to account the refined localization estimators computed by the cloud. Even with a network latency of 5 s, the EKF was able to converge and achieve a robust localization in terms of ATE. Similarly, Ali et al. (2020) store the global map in an external computing system, in their case, an edge device. The method adapts ORB-SLAM2 (Mur-Artal & Tardós, 2017) for mobile-edge parallel execution, where only the most recent data is kept in the mobile device and the edge device performs the heavier computation tasks such as global bundle adjustment and optimization of the graph relations. The method achieved constant memory usage and execution time on the mobile device in the experiments due to outsourcing computation tasks to the edge device.

#### 5.5.4 | Timing efficiency

Finally, three other included works focus on improving the timing efficiency related to the place recognition process. Mohan et al. (2015) consider two discretization levels for the words of a BoW (Sivic & Zisserman, 2003) model: a finer level representing the images and a coarse one representing environments. The method implements two nested levels of inverted indexes for fast computation. One index encodes the co-occurrence of words in a high-level environment index, and the other one stores the BoW image words of the environments. As for place recognition, the query image is transformed into two BoW vectors representing the fine and coarse levels to perform a coarse-to-fine search in the respective indexes. Compared with a single inverted-index as in a standard BoW model, the hierarchical inverted index achieves similar accuracy in loop recognition while decreasing the execution time, allowing large dictionaries for the same execution time. The second work focused on timing efficiency is Latif et al. (2017). This method formulates a sparse $l_1$ minimization problem that is convex for place recognition, instead of using nearest neighbor search to find a vector whose elements best explain the image also represented by a vector. This formulation is independent of the representations (e.g., BoW, scene features) leveraging fast convergent optimizers to ensure real-time generation of loop closure hypotheses. As for the vector solution of the optimization, it is expected to be very sparse (only one nonzero, if current image matches perfectly one of the dictionary). However, the method only declares a valid loop only when it is globally unique, to avoid false positives. The experiments showed similar precision to a BoW-based approach and lower recall due to being more conservative on loop closure detection, while being compatible with a real-time implementation (116.45 ms mean time for 4800-dim image descriptors in a database of 8358 images). As for the third work, Wu and Wu (2019) use a deep supervised hashing network to learn hash codes for direct access of similar features in a hash table, similar to Ikeda and Kanji (2010). Even when using brute force for matching hash codes, the method achieved a 6.67 ms matching time with a 8192-dim descriptor in a 10,000 image database. The proposed method also had similar precision-recall and f-score compared to SeqSLAM (Milford & Wyeth, 2012) or AlexNet-based place recognition.

### 5.6 | Long-term experimental data

While the previous subsections of the discussion in this review focus on the methodologies for performing long-term localization and mapping, this section focuses on the analysis of the experimental data used by the authors in their experiments. The analysis is organized as follows: (1) identification and discussion of the public data sets used by the included works in the review, (2) discussion of the distance and time characteristics of the experimental data, and (3) identification of the types of ground-truth data more used to evaluate the proposed methodologies.

#### 5.6.1 | Public data sets

The analysis of the data extraction item DE11 (see Section 3.4 and Table A1) shows that 89 of the 142 (62.7%) included works in this review used public data sets to evaluate the proposed methodologies. This percentage indicates the importance of using data sets in the experiments, possibly due to facilitating the accessibility to the experimental data and allowing comparisons between works that use the same data sets and evaluation metrics. Even considering the 77 works that also performed private experiments, 24 (31%) of those also used data sets for evaluating the proposed methods. Also, Figure 6 presents the usage of public data sets over time versus the records published per year. This graph shows similar linear tendencies between the two series and an usage over than 65% in the past 5 years. The latter percentage strengths the indication of the importance of using public data sets in the experiments given by the included works in this review.

Although DE11 indicates which data sets are used in the experiments, this item does not characterize the data sets. Thus, Table 6 presents a comparison table with the following items: long-term characteristics of the data set in terms of the environment conditions (lighting, day and night sequences, weather and seasonal changing conditions, dynamic elements, and sparsity), type of environment (indoor, outdoor, or both), the domain of the agent used for acquiring data (ground, air, or water, and the commercial unit used if indicated), sensorization, if the data set provides intrinsic and extrinsic calibration of the sensor setup used, type of ground-truth data, format, and long-term characteristics in terms of distance, time, and the number of runs. Next, the discussion focuses on comparing the data sets based on the column items presented in Table 6 and correlating their usage in terms of DE1.

*Environment*

The outdoor environment is the most seen one in the 43 data sets, with 27 being acquired outdoors compared with 19 indoors, and 3

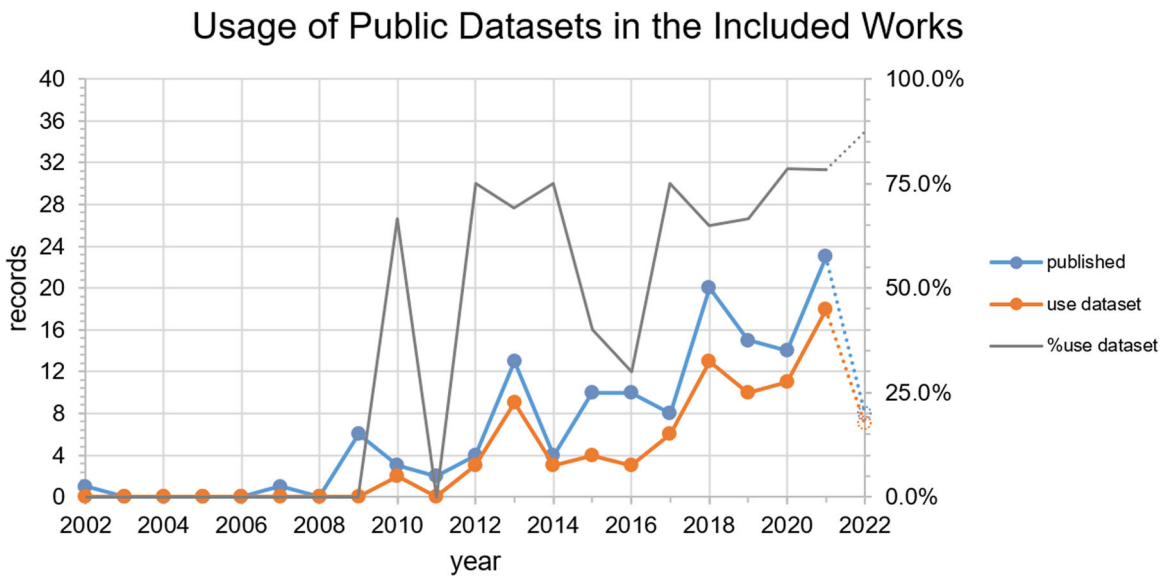


**FIGURE 6** Evolution of the usage of public data sets per year considering the 142 included records in this review. he time interval is between the smallest publication year found in the included records (2002) and the year of last full inquiry's date (2022). The latter is with a dotted line due to the fact that the last full inquiry does not consider the whole year.[Color figure can be viewed at wileyonlinelibrary.com]

data sets (*Gardens Point Campus of QUT* (Glover, 2014), *NCLT* [Carlevaris-Bianco et al., 2016], and *NTU VIRAL* [Nguyen et al., 2021]) having indoor and outdoor sequences. The environment changing conditions more present in indoor data sets are lighting changes and dynamic elements, for example, in office environments where the exterior and artificial light influence the visual perception and moving people increase environment dynamics (not only the people, but moving objects taken by persons). Although night periods, weather and seasonal changes also influence indoor conditions, this influence is mostly in the lighting conditions and only appear in four indoor-only data sets (*COLD* (Pronobis & Caputo, 2009), *CoBots long-term* (Biswas & Veloso, 2013), *MIT Stata Center* (Fallon et al., 2013), and *Witham Wharf RGBD* (Krajník et al., 2014)), accordingly to the respective data set descriptions. Similarly, the outdoor data sets are more affected by changing lighting and moving objects. However, these data sets consider more frequently and are more influenced by other changes. This influence is not only in lighting conditions but also in visual perception (color of the leaves in different seasons) and moving elements in the scene (water of the rain or moving tree branches due to strong wind). In terms of recency, in the past 5 years, only 2 of the 14 data sets released during that period are in indoor locations. This recent tendency and the fact of 27/43 data sets having outdoor sequences indicate more interest in this type of environment by the included works in this review.

As for the diversity of the acquisition conditions, the most diverse data sets are *COLD* (Pronobis & Caputo, 2009), *Witham Wharf RGBD* (Krajník et al., 2014), *NCLT* (Carlevaris-Bianco et al., 2016), *USyd Campus* (Zhou et al., 2020), *RADIATE* (Sheeny et al., 2021), *IPLT* (Bouaziz et al., 2021), and *Oxford RobotCar* (Maddern et al., 2017). The latter two have all changing conditions in the environment, that is, lighting, day/night sequences, dynamic elements, and weather and seasonal changes. Even though the remaining diverse data sets do not consider one of these conditions, the data sets are still interesting in the long-term localization and mapping context with a high diversity of environment conditions.

The data sets categorized as sparsity are intended for testing map maintenance algorithms to constrain the graph size in the graph SLAM formulation to the operation area and not to the trajectory length due to usually being available the full graph map of the data set. Although these data sets are useful for evaluating map maintenance, they normally lack several other changing conditions that influence long-term localization and mapping while also all of those data sets being indoors. Only *CoBots long-term* (Biswas & Veloso, 2013) and *MIT Stata Center* (Fallon et al., 2013) data sets seem to be more diverse in terms of environment conditions by capturing sequences with different lighting conditions and dynamic elements in the scene.

### *Sensorization*

In terms of the type of sensors used for acquiring data, the ones utilized in the data sets are odometry (wheeled, visual, inertial, laser, or a combination of different odometric sources), cameras (monocular, stereo, omnidirectional, RGBD, or thermal), lasers (2D/3D), radar, sonar, IMU, and GPS, similar to the sensors found in the data extraction phase of this review. The more common type of sensor is camera used in 37/43 (86%) data sets. This predominance is conformal with the high usage in 104/142 (73.2%) included works and occurrence of related keywords in the analysis presented in Figure 3 (both vision and camera keywords appear in the graph with 18 and 7 occurrences, respectively), indicating an interest of using camera sensors in data acquisition and long-term localization and mapping. Also, the omnidirectional vision used in five data sets can be

accomplished by using an hyperbolic mirror (*Bicocca (indoor)* (Fontana et al., 2009), *COLD* (Pronobis & Caputo, 2009)), joining the image of several cameras and using their extrinsic calibrated parameters, or using an omnidirectional camera (*Ford Campus* (Pandey et al., 2011), *NCLT* (Carlevaris-Bianco et al., 2016), *New College* (Smith et al., 2009)) such as the Point Grey LadyBug 2 5-view. Although the thermal camera is only present in *KAIST* (Choi et al., 2015), this sensor can be interesting for building inspection (Yue et al., 2020).

Moreover, data sets used in the included works recently released also use 3D laser (or 3D LiDAR) and radar sensors. Although the older data set with 3D laser is from 2011 (*Ford Campus* (Pandey et al., 2011)), 5/10 (50%) data sets using this sensor were released over the past 5 years from a total of 14 data sets released in this same period. This trend is noted also when analyzing the DE7 item of the included works, where 24/27 (89%) methods using a 3D laser were proposed since 2018, indicating a recent importance of this sensor for long-term localization and mapping. As for radar data, all three data sets using the sensor (*MulRan* (Kim et al., 2020), *Oxford Radar RobotCar* (Barnes et al., 2020), and *RADIATE* (Sheeny et al., 2021)) were released since 2020. A corresponding recency is noted in included works with 3/4 (75%) methods (Martini et al. (2020), Yin, Xu, Wang, et al. (2021), and Hong et al. (2022)) using the sensor are also from 2020 onwards. This recent usage indicates a recent interest of using radar data within the scope of this review's topic, probably due to being less affected by changing lighting or weather conditions compared to visual sensors (Hong et al., 2022).

As for the other sensors used in the data set, odometry data, IMU, 2D laser, and GPS are also extensively used in the data sets. The first two provide relative motion information of the vehicle and are used in 16 and 17 data sets and 33 and 19 included works, respectively. Although the 2D laser is used in 17 data sets, 15 of those are from 2016 and previous years. However, the sensor is still used in the included works over the years, especially in indoor environments, with 21/25 works for indoors using 2D lasers. As for GPS data, this sensor is usually used as ground-truth data, as will be discussed later.

Sensor calibration is important for achieving long-term localization and mapping, not only for avoiding the propagation of inconsistency pose errors between sensors through time, but also to process the perceived data from the environment in the same coordinate referential frame. The intrinsic calibration is usually relative to camera sensors, where 25/37 (68%) data sets with this sensor provide the intrinsic parameters to the user. Some of the data sets with cameras do not provide those parameters due to being intended only for image-based place recognition (e.g., *City Center (FAB-MAP)* (Cummins & Newman, 2008), *CBD* (Zhang, Yan, et al., 2019), or *Freiburg Across Seasons* (Naseer et al., 2018)). In terms of extrinsic calibration, 24/43 (56%) data sets provide these parameters, being useful for evaluating methods where the parameters are required to be processed in the same reference frame.

The data sets more diverse in terms of their sensor setup are *Bicocca (indoor)* (Fontana et al., 2009), *Oxford RobotCar* (Maddern et al., 2017), and *Oxford Radar RobotCar* (Barnes et al., 2020), with 7, 6, and 6 different types of sensors, respectively.

### *File format*

Most of the data sets used by the included works define a specific format for organizing the respective data. These formats use standard file types such as plain text, CSV, or binary files having the advantage of not being tied to any particular software. Even so, there are common characteristics between those specific formats. Images are usually saved in JPG or PNG files, whereas PNG files are also used in the data sets for saving depth information of RGBD sensors. The laser data is usually saved in binary files due to its easiness for parsing by different programming languages and size considerations (Geiger et al., 2013; Maddern et al., 2017). Another common aspect of specific formats is the use of plain text or CSV files to save IMU, GPS, and/or odometry data. As for radar data, the respective polar representations are saved in PNG files.

However, the data sets also make available standard log formats compatible with different types of sensor data. The most used one is ROSbag from the Robot Operating System (ROS) framework in 9 datasets. This log format is compatible with common messages defined in ROS for different sensors.[13] The other standard format used in more than one data set is CARMEN log files defined in the CARMEN robot navigation toolkiT.[14] Although this log format supports different sensor data such as odometry or lasers, the CARMEN navigation toolkit is not updated since 2008 (version 0.7.4-beta), considered to be deprecated.

### *Usage relation with the included records*

As for relating the data sets usage with this review's included works, the data sets can be related with the DE1 categorization of the records. From the appearance category in DE1, 50/75 (67%) works use 32 different public data sets from Table 6 to evaluate the proposed methodologies. The data sets most used are *KITTI* (Geiger et al., 2013), *Nordlandsbanen* (Skrede, 2013), *NCLT* (Carlevaris-Bianco et al., 2016), *St Lucia Brisbane* (Milford & Wyeth, 2008), and *Oxford RobotCar* (Maddern et al., 2017) (13, 11, 9, 8, and 8 usages). *KITTI* (Geiger et al., 2013) is also the most used overall, given the 26 works utilizing it for evaluation. However, this data set and *St Lucia Brisbane* (Milford & Wyeth, 2008) do not have seasonal nor weather changing conditions that greatly influence the appearance invariance of the methods, as discussed in Section 5.1, even though those data sets have high usage by the appearance-related works. Indeed, more recent data sets such as *NCLT* (Carlevaris-Bianco et al., 2016) or *Oxford RobotCar* (Maddern et al., 2017) already widely used for evaluation, or also *USyd Campus* (Zhou et al., 2020) and *RADIATE* (Sheeny et al., 2021) would be suitable for evaluating the appearance invariance of the localization and mapping algorithms due to the data sets' diversity in terms of varying conditions.

[13] http://wiki.ros.org/common_msgs
[14] https://carmen.sourceforge.net/

**TABLE 6** Data sets used in the 142 included records for long-term localization and/or mapping experiments.

| | Long-term | | | | | | | | Sensor | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Camera | | | | | | | Laser | | | | | |
| Dataset | Lighting | Day/night | Weather | Seasonal | Dynamics | Sparsity | Environ. | Domain | Odo | Gray | Color | Monocular | Stereo | Omni | RGBD | Thermal | 2D | 3D | Radar | Sonar | IMU | GPS |
| FHW (Burgard et al., 2003) | | | | | | x | Indoor (museum) | Ground (TOURBOT) | x | | | | | | | | x | | | | | |
| FR079 (Stachniss, 2003a) | | | | | | x | Indoor (office) | Ground (robot) | x | | | | | | | | x | | | | | |
| FR101 (Stachniss, 2003b) | | | | | | x | Indoor (office) | ground (robot) | x | | | | | | | | x | | | | | |
| Intel Research Lab 2003 (Hähnel, 2003) | | | | | | x | Indoor (office) | Ground (robot) | x | | | | | | | | x | | | | | |
| MIT Killian Court (Bosse et al., 2004) | | | | | | x | Indoor (office) | Ground (robot) | x | | | | | | | | x | | | x | | |
| City Center (FAB-MAP) (Cummins & Newman, 2008) | x | | | | x | | Outdoor (urban) | Ground (robot) | | | x | x | | | | | | | | | | x |
| Lip6Indoor (Angeli et al., 2008) | | | | | | | Indoor (office) | Ground (handheld) | | | x | x | | | | | | | | | | |
| Lip6Outdoor (Angeli et al., 2008) | x | | | | x | | Outdoor (campus) | Ground (handheld) | | | x | x | | | | | | | | | | |
| New College (FAB-MAP) (Cummins & Newman, 2008) | x | | | | x | | Outdoor (campus) | Ground (robot) | | | x | x | | | | | | | | | | x |
| St Lucia Brisbane 2007 (Milford & Wyeth, 2008) | x | | | | x | | Outdoor (urban) | Ground (car) | | x | | x | | | | | | | | | | |
| Bicocca (indoor) (Fontana et al., 2009) | x | | | | x | x | Indoor (office) | Ground (Robocom) | x | x | x | x | x | x | | | x | | | x | x | |
| COLD (Pronobis & Caputo, 2009) | x | x | x | | x | | Indoor (office) | Ground (Pioneer 3, ATRV Mini, PeopleBot) | x | | x | x | | x | | | x | | | | | |
| Malaga 2009 (Blanco et al., 2009) | x | | | | x | | Outdoor (parking, campus) | Ground (car) | | | x | x | | | | | x | | | | x | x |
| New College (Smith et al., 2009) | x | | | | x | | outdoor (campus) | Ground (Segway) | x | x | x | | x | x | | | x | | | | x | x |
| albert-b-laser-vision (Stachniss, 2010) | x | | | | | x | Indoor (office) | Ground (iRobot B21r) | x | | x | x | | | | | x | | | | | |
| CMU-VL (Badino et al., 2011) | x | | | x | x | | Outdoor (urban) | Ground (car) | | | x | x | | | | | | | | | | x |
| Ford Campus (Pandey et al., 2011) | x | | | | x | | Outdoor (campus, urban) | Ground (car) | | | x | | | x | | | x | x | | | x | x |
| UTIAS Multi-Robot (Leung et al., 2011) | x | | | | | | Indoor (empty space) | Ground (iRobot Create) | x | | x | x | | | | | | | | | | |
| Alderley Brisbane (Milford & Wyeth, 2012) | x | | x | | x | | Outdoor (urban) | Ground (car) | | | x | x | | | | | | | | | | |
| TUM RGBD Dynamic (Sturm et al., 2012) | x | | | | x | | Indoor (office, industrial hall) | Ground (handheld, Pioneer 3) | | | x | | | | x | | | | | | x | |

| Calib. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Intrinsic | Extrinsic | GT data | Format | Dist. (km) | Path (km) | Time (h) | Int. (d/w/m/y) | Seq. |
| | | – | CARMEN | – | – | 1.98 | – | 1 |
| | | – | CARMEN | – | – | 0.29 | – | 1 |
| | | – | CARMEN | – | – | 0.29 | – | 1 |
| | | – | CARMEN | 0.506 | – | 0.75 | – | 1 |
| | | – | CARMEN | 2.2 | – | 2.5 | – | 1 |
| x | | GPS, manual | Plain text (non-img), jpg (img) | 2 | – | – | – | 1 |
| x | | Manual | ppm (imgs) | – | – | 0.11 | – | 1 |
| x | | Manual | ppm (imgs) | – | – | 0.3 | – | 1 |
| x | | GPS, manual | Plain text (non-img), jpg (img) | 1.9 | – | – | – | 1 |
| | | – | – | 66 | – | 1.67 | – | 1 |
| x | x | Map model, laser-based | Plain text (non-img), png (img) | – | – | 2.5 | 3d | 5 |
| | | Laser-based, manual | Plain text (non-img), jpg (img) | 0.92 | – | 0.99 | – | 76 |
| x | x | RTK-GPS | Rawlog MRPT | 6.358 | – | – | – | 6 |
| x | x | GPS | Plain text (non-img), png, jpg (img) | 2.2 | – | 0.73 | – | 1 |
| | | – | CARMEN (non-img), jpg (img) | – | – | 0.18 | – | 1 |
| x | x | GPS | – | – | 8.5 | – | 1y | 16 |
| x | x | RTK-GPS | LCM log | – | – | – | 2m | – |
| x | | External tracking system | jpg (img), dat (non-img) | – | – | 4.78 | – | 9 |
| | | Manual | – | 16 | 8 | – | – | 2 |
| x | x | External tracking system | Plain text (non-img), png (img + depth), ROS bag | 0.285 | – | 0.35 | – | 15 |

(Continues)

**TABLE 6** (Continued)

| Dataset | Long-term | | | | | | | | Sensor | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Camera | | | | | | | Laser | | | | | |
| | Lighting | Day/night | Weather | Seasonal | Dynamics | Sparsity | Environ. | Domain | Odo | Gray | Color | Monocular | Stereo | Omni | RGBD | Thermal | 2D | 3D | Radar | Sonar | IMU | GPS |
| CoBots long-term (Biswas & Veloso, 2013) | x | | | x | x | x | Indoor (office) | Ground (robot) | x | | x | | | | x | | x | | | | | |
| KITTI (Geiger et al., 2013) | x | | | | x | | Outdoor (urban) | Ground (car) | | x | x | x | x | | | | | x | | | x | x |
| MIT Stata Center (Fallon et al., 2013) | x | | | x | x | x | Indoor (office) | Ground (PR2) | x | | x | | x | | x | | x | | | | x | |
| Nordland (Skrede, 2013) | x | | x | x | | | Outdoor (railway) | Ground (train) | | | x | x | | | | | | | | | | x |
| Gardens Point Campus of QUT (Glover, 2014) | x | x | | | x | | Indoor, outdoor (campus) | Ground (handheld) | | | x | x | | | | | | | | | | |
| Witham Wharf RGBD (LCAS STRANDS) (Krajník et al., 2014) | x | x | | x | x | | Indoor (office) | Ground (SCITOS-G5) | | | x | | | | x | | x | | | | | |
| KAIST (Choi et al., 2015) | x | x | | | x | | Outdoor (urban) | Ground (car) | | | x | | x | | | x | | | | | x | x |
| EuRoC (Burri et al., 2016) | x | | | | | | Indoor (industrial hall, office) | Air (AscTec Firefly) | | x | | x | x | | | | | | | | x | |
| NCLT (Carlevaris-Bianco et al., 2016) | x | | x | x | x | | Indoor, outdoor (campus) | Ground (Segway) | | | x | | | x | | | x | x | | | x | x |
| Berlin Kudamm (Chen et al., 2017) | x | | | | x | | Outdoor (urban) | Ground (car) | | | x | x | | | | | | | | | | |
| Oxford RobotCar (Maddern et al., 2017) | x | x | x | x | x | | Outdoor (urban) | Ground (car) | x | | x | x | x | | | | x | x | | | x | x |
| YQ21 (Tang, 2017) | x | | | | x | | Outdoor (campus) | Ground (car) | | | x | x | x | | | | | x | | | x | x |
| CMU-Seasons (Sattler et al., 2018) | x | | | x | x | | Outdoor (urban) | Ground (car) | | | x | x | | | | | | | | | | |
| Freiburg Across Seasons (Naseer et al., 2018) | x | | | x | x | | Outdoor (urban) | Ground (car) | | | x | | x | | | | | | | | | x |
| RobotCar Seasons (Sattler et al., 2018) | x | | x | | x | | Outdoor (urban) | Ground (car) | | | x | x | x | | | | | | | | | |
| Bonn RGB-D Dynamic (Palazzolo et al., 2019) | | | | | x | | Indoor (office) | Ground | | | x | | | | x | | | | | | | |
| CBD (Zhang, Yan, et al., 2019) | x | | | | x | | Outdoor (urban) | Ground | | | x | x | x | | | | | | | | | |
| MulRan (Kim et al., 2020) | | | | | x | | Outdoor (urban) | Ground (car) | | | | | | | | | | x | x | | x | x |
| Oxford Radar RobotCar (Barnes et al., 2020) | x | | x | | x | | Outdoor (urban) | Ground (car) | x | | x | x | x | | | | | x | x | | x | x |

| Calib. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Intrinsic | Extrinsic | GT data | Format | Dist. (km) | Path (km) | Time (h) | Int. (d/w/m/y) | Seq. |
| | x | – | ROS bag | 131 | – | 260 | 2y3m | 1082 |
| x | x | RTK-GPS | png (img), binary (laser), plain text (imu, gps) | – | – | 1.18 | 8d | 61 |
| | x | Map model | ROS bag | 42 | – | 38 | 1y9m | 84 |
| | | GPS | mp4 (video stream), plain text (gps) | 2916 | 729 | 39.74 | – | 4 |
| | | Ground-plane position | png (imgs), plain text (ground plane) | – | – | – | – | 3 |
| | x | – | ROS bag | – | – | – | 1y1m | 368 |
| x | x | RTK-GPS | png (imgs), plain text (imu, gps) | 84 | – | – | 18d | 36 |
| x | x | External tracking system | ROS bag | 0.8936 | – | 0.37 | – | 11 |
| x | x | RTK-GPS, SLAM-based | Binary (laser), tiff (img), plain text (non-laser or img) | 147.4 | – | 34.9 | 1y4m | 27 |
| | | Manual | jpg (img) | – | – | – | – | 2 |
| x | x | RTK-GPS | png (img), binary (laser), plain text (imu, gps, odo) | 1010.46 | 10 | – | 1y8m | 133 |
| x | x | RTK-GPS | Binary (laser), jpg (imgs), plain text (gps) | 23 | – | 6.5 | 1w | 21 |
| x | x | Manual | jpg (img) | – | 8.5 | – | 330d | 17 |
| | | GPS, manual | jpg (img) | 110 | – | – | 3y | 3 |
| x | x | Manual | jpg (img) | – | 10 | – | 178d | 10 |
| x | x | External tracking system | png (imgs, depth), plain text (imu, gps) | – | – | – | – | 26 |
| | | Manual | png (imgs) | – | – | – | – | 1 |
| | x | SLAM-based | Binary (laser), CSV (global pose, radar ray), png (radar polar img) | 41.2 | – | – | 2m15d | 12 |
| x | x | RTK-GPS, SLAM-based | png (img, raw laser, radar), binary (laser), plain text (imu, gps, odo) | 280 | 10 | – | 1m | 32 |

(Continues)

**TABLE 6** (Continued)

| Dataset | Long-term | | | | | | | | Sensor | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Camera | | | | | | | Laser | | | | | |
| | Lighting | Day/night | Weather | Seasonal | Dynamics | Sparsity | Environ. | Domain | Odo | Gray | Color | Monocular | Stereo | Omni | RGBD | Thermal | 2D | 3D | Radar | Sonar | IMU | GPS |
| USyd Campus (Zhou et al., 2020) | x | | x | x | x | | Outdoor (campus) | Ground (car) | x | | x | x | | | | | | x | | | x | x |
| IPLT (Bouaziz et al., 2021) | x | x | x | x | x | | Outdoor (parking) | Ground (car) | x | x | | x | | | | | x | | | | | x |
| RADIATE (Sheeny et al., 2021) | x | x | x | | x | | Outdoor (parking, urban) | Ground (car) | | | x | x | x | | | | | x | x | | x | x |
| NTU VIRAL (Nguyen et al., 2021) | x | | | | | | Indoor, outdoor (campus) | Air (DJI M600) | | x | | x | x | | | | | x | | | x | |

Abbreviations: dist., total distance length of the data set; int., time interval between the start and end acquisition dates/time instants (d/w/m/y equivalentto day/week/month/year, 0 if only 1 run); odo, odometry (wheeled,laser, visual, inertial, or a combination of odometry sources); path, total path distance if repeated several times; seq., number of sequences of the data set; time, total operation time.

Although the works categorized as dynamics and sparsity also use public data sets for evaluation, the usage is slightly lower than for appearance-related methods (44% and 51%, respectively, compared with 67%). *KITTI* (Geiger et al., 2013), *TUM RGBD Dynamic* (Sturm et al., 2012), and *Witham Wharf RGBD* (Krajník et al., 2014) are the only data sets used in more than one work categorized as dynamic (7, 6, and 2 usages, respectively), considering a total of 10 different data sets used by these works. However, *KITTI* (Geiger et al., 2013) and *TUM RGBD Dynamic* (Sturm et al., 2012) could be not the most suitable data sets for evaluating the performance over different levels of dynamics in the environment due to the smaller total operation time of 1.18 and 0.35 h, respectively, compared to other data sets classified as having dynamic elements in Table 6. For example, *Witham Wharf RGBD* (Krajník et al., 2014) has a time frame of 1 year and a month in an indoor office environment capturing different motion frequencies or habits of the persons working at the scene with an average of 1 daily acquisition run. *Oxford RobotCar* (Maddern et al., 2017) and *USyd Campus* (Zhou et al., 2020) are also interesting due to the long time frames of the data acquisition (1 year and 8 months, and 1 year, with 133 and 52 runs, respectively). Also, *IPLT* (Bouaziz et al., 2021) is captured in a parking lot environment capturing semi-static and dynamic moving cars in the scene. As for sparsity-related works, 22 different data sets are used in the experiments, whereas *KITTI* (Geiger et al., 2013), *Intel Research Lab* (Hähnel, 2003), *MIT Killian Court* (Bosse et al., 2004), *MIT Stata Center* (Fallon et al., 2013), and *EuRoC* (Burri et al., 2016) being the most utilized ones (6, 5, 3, 3, and 3 usages, respectively). *KITTI* (Geiger et al., 2013) and *EuRoC* (Burri et al., 2016) are used for evaluating feature management techniques, though these data sets have small operation time frames and, possibly, trajectory lengths. Although *MIT Stata Center* (Fallon et al., 2013) would seem like a good data set for evaluating the sparsity due to the long trajectory length and time frame (42 km and 38 h, respectively), the data set's description indicates that hardware and calibration problems in the data acquisition setup may have created inconsistencies in the data. As for *Intel Research Lab* (Hähnel, 2003), *MIT Killian Court* (Bosse et al., 2004), and other data sets classified as sparsity in Table 6, these are widely used for graph sparsification due to repeated passages in same locations with a total of 15 usages, even though those data sets have usually only 1 data sequence. Other recent data sets could also be interesting for evaluation sparsification techniques of the map such as *Oxford RobotCar* (Maddern et al., 2017) and *MulRan* (Kim et al., 2020), given the repeated passages over the 10km path and the long trajectory length of 41.2 km, respectively.

Furthermore, 5/7 multisession works used data sets, whereas the *MIT Stata Center* (Fallon et al., 2013) and *Intel Research Lab* (Hähnel, 2003) being the most used ones with two usages each. While each data sequence of *MIT Stata Center* (Fallon et al., 2013) may represent a single session (Lázaro et al., 2018), the unique data sequence of *Intel Research Lab* (Hähnel, 2003) can be split into different sessions (Latif et al., 2012). This approach is valid for applying to other data sets in Table 6. As for the computational categorization on DE1, this category does not relate to the data sets used in the experiments because the computational efficiency is more dependent on the proposed localization and/or mapping algorithm than on the data.

In terms of multirobot works (Gadd & Newman, 2016; Karaoğuz & Bozma, 2020; Yue et al., 2020; Zhang, Chen, et al., 2018) identified by the DE4, even though the data set *UTIAS Multi-Robot* (Leung et al., 2011) collects data from FIVE robots and being the only multirobot data set in Table 6, it is only used in Nobre et al. (2018) to test the reconfiguration of landmarks in the scene in different runs for single-robot localization and mapping. The only data sets used in multirobot works are *KITTI* (Geiger et al., 2013) in Zhang, Chen, et al. (2018) and *COLD* (Pronobis & Caputo, 2009) in Karaoğuz and Bozma (2020).

| Calib. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Intrinsic | Extrinsic | GT data | Format | Dist. (km) | Path (km) | Time (h) | Int. (d/w/m/y) | Seq. |
| x | x | GPS | ROS bag | – | – | – | 1y | 52 |
| x | x | GPS | ROS bag | – | 0.2 | – | 2y | 127 |
| x | x | RTK-GPS | ROS bag | – | – | 4.98 | – | – |
| x | x | External tracking system | ROS bag | 1.845 | – | 0.9 | – | 9 |

These works assume that different data sequences are acquired by different agents or segment the sequence in data subsets, similar to the evaluation with data sets of multisession works. However, the fact that only four multirobot works are included in this review and only one public data set used in the included records is acquired with multiple vehicles could indicate that the use of multirobot systems is not yet widely studied in the long-term localization and mapping topic.

### 5.6.2 | Distance and time considerations

Next, analyzing the total trajectory length of the private experiments on the included records (DE10 on Table A1) and public data sets (see Table 6), five works and six data sets have a length greater than 100 km. Higher values on the trajectory length indicate possibly more interesting data for evaluating sparsity management techniques discussed in Section 5.3, given that the desired behavior of a mapping algorithm is its scalability being only dependent on the environment size and not on the trajectory length. Although the total trajectory length does not necessarily relates directly with the environment area, the latter is rarely seen in the experiments description, and even for the trajectory length, only 36/77 works that perform private experiments and 22/43 data sets indicate the length.

The other distance measure considered in this review to characterize the experimental data is the one relative to repeating the same path, with 7/8 data sets and 3/8 included works that specify this metric having a repetitive path distance greater than 8 km and more than 1 run. These low numbers do not necessarily indicate incomplete information in the experimental description due to a data acquisition can be performed on nonrepetitive routes. Even so, repeating the same exact path under different environment conditions (i.e., appearance variance) could be a case study for evaluating the appearance invariance of localization and mapping algorithms discussed in Section 5.1.

In terms of time-related long-term characteristics of experimental data, longer total operation times indicate a robust evaluation of the proposed localization and mapping algorithms over long continuous periods, and greater time interval suggests data acquired under severe changing conditions (not only in the environment appearance but also semi-static modifications in the scene). However, only 2/10 works performing private experiments and indicating the total operation time test their methods over a total of 8h (equivalent to a work day), while also only 4/23 data sets that define the total log time in their description have more than 8 h of data. On the contrary, 41/77 works and 18/43 data sets characterize the interval between the first and the last data sequence, which of those 29 works performing private experiments and 17 data sets have at least a 1 week interval. These results indicate that even though the included works in this review use experimental data with greater time intervals, often several days or weeks, not so much importance is given toward the total operation time.

### 5.6.3 | Ground-truth data

As for the types of ground-truth data found in the included works (see DE9 in Table A1) and the public data sets used for evaluation (see Table 6), the manual annotation is one of the most used types of ground-truth including image to image association (e.g., useful for evaluating image-based place recognition), manual alignment of maps (Biswas & Veloso, 2013), or manually segmenting images (Geiger et al., 2013). Although GPS-based data is also widely used in the experiments, whereas the RTK-GPS variant improves the pose

precision compared with the basic positioning system, GPS is meant for use in outdoor environments. The alternative for indoor environments used in the experiments is external tracking systems using, for example, reflective markers put on the robot to track them through systems such as OptiTrack[15] or Vicon[16] to provide precise measurements of the robot's pose. Simulation data used in 10 included works can also provide precise ground-truth data for the robot's pose or other types of information, even though not in a real environment.

Moreover, SLAM or laser-based ground-truth data are also found in the experimental evaluation of the included works and public data sets. The experimental methodology uses localization and mapping algorithms other than the one being evaluated to provide ground-truth data usually using a different sensor setup, or using the same algorithm but including all data sessions or global optimization over the entire pose graph. Specifically, laser-based localization is widely used in the included works as ground-truth data to evaluate vision-based methods (Nuske et al., 2009). Similarly to SLAM-based ground-truth data, the experimental results of localization and mapping methods without pruning are also used as a reference for evaluating sparsification algorithms (Carlevaris-Bianco et al., 2014).

As for model-based ground-truth data, the work of Boniardi et al. (2019) and the data sets *Bicocca (indoor)* (Fontana et al., 2009) and *MIT Stata Center* (Fallon et al., 2013) propose the use of floor plans as a model of the environment to align the current estimation with the model and obtain a ground-truth for the trajectory of the robot on the map. Ozog et al. (2016) is the other method using map models in the experimental evaluation but in the context of ship hull inspection, where the current map estimation is aligned with the ship hull CAD model for obtaining the trajectory ground-truth data. However, the use of map-based ground-truth data to evaluate the mapping process is not seen in the included works, other than performing a visual quality evaluation over the mapping results and the sensor data.

## 5.7 | Evaluation metrics

The final topic of discussion over the included works in this review is the analysis on evaluation metrics used for assessing the performance of the proposed methodologies (see DE12 in Table 6). This analysis is organized by (1) metrics intended for place recognition, (2) evaluation of the robot pose, (3) assessment of map sparsity, and (4) computational performance.

### 5.7.1 | Place recognition

The evaluation of place recognition in the context of long-term localization and mapping is intrinsically related to the invariance of the evaluation metrics to changing conditions. The most used metrics for place recognition in the included works are precision and recall, where 50/142 works use these metrics in the experimental evaluation, and from those works 39 are categorized as appearance. These metrics characterize the performance of recognizing successfully different places related to the number of true and false positives and true and false negatives, and also include the precision-recall curve where a greater area under the curve indicates a better classifier for place recognition.

Other metrics less used in the included works but also important are the confusion matrix, the localization success rate, and the f-score and f-beta measures. The confusion matrix associates the predicted place to the true value in the case of each data entry representing a unique place, and a unique diagonal in the matrix would be the ideal result. This matrix is also used for comparing the data stream versus a reference database, where the appearance of multiple diagonals in the matrix indicates the capability of the place recognition algorithm to perform loop closure relative to the database. The localization rate is the ratio between the number of successful localization versus and the attempts. As for the f-score and f-beta metrics, these measures combines the precision and recall in an unique value, where the f-beta allows the weighting of precision versus recall depending on which is more important for the method's use case.

### 5.7.2 | Robot pose

Robot pose-related metrics are widely used in works categorized as appearance, dynamic, and sparsity methods (21/75, 18/32, and 16/45, respectively). The pose error indicates if the localization algorithm is affected by changing conditions over time. For dynamic-related methods, the pose error is also useful to show the influence of moving elements in the scene on the localization performance. As for sparsification techniques, the pose error can characterize the influence of the map pruning algorithm over the localization estimator.

In terms of evaluating the robot pose, the pose error is also one of the most used metrics in the included works (50/142). These works evaluate the pose error metric in terms of its instant measurement over time or relative to a data sequence in terms of the mean, standard deviation, and/or RMSE values of the pose error. Also, a specific measure of pose error used in 15/142 works is the Absolute Trajectory Error (ATE) usually computed over an entire trajectory. This metric requires the time alignment between the localization estimation and the ground-truth data and computes the mean and standard deviation of the estimation differences between samples with the same time instant.[17]

The covariance of the pose estimation is also considered in the included works for evaluating the robot pose error, where the covariance matrix represents the uncertainty of the robot's pose over an experiment. Also, in Hochdorfer and Schlegel (2009), the covariance matrix's eigenvalues are used to evaluate the uncertainty

[15] https://optitrack.com/
[16] https://www.vicon.com/

[17] See definition of ATE in the *TUM RGBD Dynamic* (Sturm et al., 2012) data set (https://vision.in.tum.de/data/datasets/rgbd-dataset/tools#absolute_trajectory_error_ate) and in the RAWSEEDS benchmarking toolkit (http://www.rawseeds.org/rs/methods/view//9).

of the estimator, where greater values represent greater uncertainties. Instead of computing the eigen values from the covariance matrix, Dayoub and Duckett (2008) compute these values from the inverse covariance matrix, and so, the logic also inverts relative to Hochdorfer and Schlegel (2009), where smaller eigen values would mean smaller uncertainties in that case.

#### 5.7.3 | Map sparsity

As for evaluating the map sparsity, this evaluation is inherently related to the sparsity category of DE1. In terms of metrics, the analysis of the evolution of the number of nodes is widely used in 19/45 sparsity-related works. This metric is useful to study the evolution of the graph size in the graph formulation of SLAM over the operation time and/or trajectory length. The number of edges over time or the edge reduction ratio compared with no pruning data also indicate growth over time of the edges, while the gamma index of a graph (ratio between the number of existing edges and the possible ones) indicates the current sparsity over the graph connectivity. Another important metric for evaluating the graph sparsification is the Kullback-Leibler Divergence (KLD) measure that defines the difference between two probabilistic distributions. The included works use the KLD to compare the information loss between the sparse graph and the one without pruning, in which a 0 value of KLD would mean that the two distributions have identical information, and so the graph pruning algorithm was able to remove only redundant data. More generally, the evaluation of the number of map points over time is also presented in the results of sparsity-related works.

#### 5.7.4 | Computational performance

Finally, the evaluation of the computational performance is widely analyzed in the included works. This evaluation is not necessarily related to only the computational category of DE1 because an algorithm's computational performance impacts its online execution. Considering the 110 methods with online execution modes identified in Table 1 by DE5, 76 (69.1%) works evaluate the computational resources required for online execution of the proposed methodology, indicating the importance of this type of experimental analysis in the included works. In terms of computational performance metrics, the execution time measurement is most used one being evaluated in 69 works. However, other metrics such as the runtime memory or the computational complexity are considered in the included works for evaluating the computational resources required to execute the proposed methodologies.

## 6 | CHALLENGES AND FUTURE DIRECTIONS

The growing interest in mobile robots and their usage in different applications and complex environments stress the importance of improving the robustness of autonomous systems to varying conditions and long operation periods. Although the localization and mapping algorithms included in this study help achieve long-term operations, there are still open challenges in lifelong SLAM. So, in addition to the challenges discussed in Section 5, other potential challenges related to lifelong SLAM and research directions are listed below.

### 6.1 | Vision-based global place recognition

Given the limited field of view characteristic of vision sensors (apart from omnidirectional cameras), the analysis of the included records shows that it is still challenging to recognize places using vision-based global descriptors. The limited field of view influences the visual content of the image, as shown in the experimental results of C. Qin et al. (2020), where the method significantly reduced its performance due to viewpoint variance.

One possible direction could be the usage of data augmentation as in Tang et al. (2021) for learning global visual descriptors, even though the latter work does not clarify to what extent augmented data helped in viewpoint variance. Another possible solution would be the use of omnidirectional vision. However, the networks traditionally used for learning CNN-based features (considered more discriminative compared to handcrafted features, as previously discussed in Section 5.1.4) may not be directly applicable due to the different aspect ratio of omnidirectional images retrieve from sensors such as the Point Grey LadyBug 2 5-view.

### 6.2 | Dynamics modeling

Most of the included works modeling the environment dynamics determine the observations as static (permanent features), semi-static (short-term static object or static at the current observation instant), or dynamic (moving object in the scene) by either representing them in maps associated with different discrete meanings of dynamics or reasoning the relation between their semantic class and the expected dynamics. However, the determination of a dynamics value for the observations could be interesting to observe its evaluation over time for predicting the environment dynamics or accounting them in the localization and mapping processes.

In the included works, Tipaldi et al. (2013) and Rapp et al. (2015) use Markov-based processes for predicting environment dynamics. However, these works assume the independence of observations, which could not be valid because static and dynamic objects may influence the dynamics of their surroundings. While FreMEn (Krajník et al., 2017) estimates the dynamicity through spectral analysis, this method assumes only periodic changes in the environment. Even though ARMA (Wang et al., 2020) models both aperiodic and periodic changes, its offline operation does not allow an online estimation of the observations' dynamicity value. So, it remains a challenge estimating online the dinamicity of environment observations unless the localization and mapping algorithms assume discrete levels for dynamics.

### 6.3 | Online graph sparsification

In the graph formulation for the SLAM problem, the methods GLC (Carlevaris-Bianco et al., 2014) and NFR (Mazuran et al., 2016) stand out in terms of their graph sparsification results. These methods obtain a graph growth approximately dependent only on the environment area and not on the operation time or trajectory length. However, the two methods are mostly intended for offline execution (e.g., between operation sessions) due to their additional computational cost when operating online. The work of Ila et al. (2017) seems to be an interesting alternative by proposing an incremental solution focused on the computational cost of graph sparsification. However, experimental results only showed that the method slows the graph's growth rate instead of bounding it when operating in the same environment area. Even though Boniardi et al. (2019) achieve a bounded computation time by pruning nodes based on topological consistency, it remains to be seen the results of graph sparsification without the CAD prior and in more highly dynamic environments. Thus, online graph sparsification is still an open research question and important to achieve continuous operation for extended time periods.

### 6.4 | Decentralized computation

Given the computational complexity inherent to SLAM, an alternative to running locally in the robot is decentralizing the algorithm's execution, offloading some parts to external agents with more computational power. In the included works, while Ali et al. (2020) implement a mobile-edge parallel execution bounding the computation time in the local device, the execution time and memory of the edge device grow over time. Furthermore, the state of the communication link influences the quality of localization and mapping, as shown by Ding et al. (2019) when evaluating the proposed cloud-based visual localization system with different network delays and packet losses. Although the solution proposed by Ding et al. (2019) can deal with delays up to 5 s, the method requires a permanent link with the cloud due to the robot only performing localization.

Overall, the topic of decentralized computing either by using edge devices or cloud-based solutions is still not well studied in the context of long-term localization and mapping. For example, the external devices could be able to perform global optimizations and searches, improving the initial estimations given by the robot. Another use case for decentralized long-term SLAM would be the external agent keeping observations of the same location at different time instants to evaluate the appearance and dynamic changes in the scene, while the robot would access the most updated, invariant, and stable map for localization.

### 6.5 | Multirobot long-term SLAM

Most of the current research discussed in this review focuses on single-robot long-term SLAM. Extending the current research for multirobot systems would be interesting for optimizing the autonomous systems operation. However, the consideration of multirobots also creates new challenges. One of would could be the decentralized and distribution SLAM execution within the multirobot system (e.g., which information to exchange between robots) and the possibility of having external agents (e.g., edge or cloud devices) to the multirobot system for offloading computation tasks. Another challenge would be considering the heterogeneous characteristics of the robots (domain, sensors, motion constraints) in merging information.

In the context of lifelong SLAM, one possible direction and an open challenge for multirobot systems could be exploiting different sensor modalities in heterogeneous robots to achieve appearance invariance in changing conditions. The usage of different modalities has the benefit of utilizing the advantages of each sensor and improving the overall robustness to appearance changes. For example, instead of equipping a robot with a LiDAR, a camera, and radar, the autonomous system could be three robots and each with one of the three sensors mentioned previously.

### 6.6 | Active exploration

Information-driven exploration is an interesting research topic consisting of actively planning the locations and times for the robot to visit. In the context of lifelong localization and mapping, active navigation could plan the robot trajectory, for example, for avoiding locations that the robot predicts to be highly dynamic or for generating specific mapping tasks in locations known to be very susceptible to appearance changes to maintain an up to date representation of the environment. One example found in the included works is Santos et al. (2016) that use the dynamic prediction provided by the FreMEn (Krajník et al., 2017) module for the planner to predict which areas are more likely to change and define the locations to explore. However, achieving active navigation would require tightly coupling the planning process with the robot's localization and mapping estimations. Also, the reasoning and modeling of the environments changes would play an important role for planning the tasks.

### 6.7 | Human-to-machine interfaces

The persons interacting daily with autonomous systems may also play a role in achieving a successful long-term operation. An operator of the system can provide prior information useful for the localization and mapping tasks. For example, if the relocalization estimation was not successful based on the current environment representation, the operator could define the robot's initial pose. Another example could be considering the operator input to specify the dynamics expected in certain areas to help the autonomous system modeling the environment dynamics.

However, the interaction with the user should not be through raw sensor data due to possibly increasing the need for special training for the operator, especially in industrial applications. Boniardi et al. (2019) present an interesting work in terms of using CAD prior for localization and mapping while facilitating the interaction with the user. Another potential direction could be using higher levels of information for interaction such as high-level geometric and semantic features perceived by the robot.

## 7 | LIMITATIONS OF THE STUDY

Although this study follows a systematic methodology for selecting the included works in the review, the methodology followed by this study has limitations. One limitation is related to the goal of overview the long-term SLAM topic instead of providing an in-depth analysis, discussion, and comparison focused on a single challenge intrinsic to lifelong autonomy. This limitation leads to an extensive and complex discussion of the included works that may possibly not cover all the details of the proposed methods in the included records. However, none of the existent reviews in SLAM focus on the long-term localization and mapping problem nor clarify the selection methodology of the works included in their studies. Also, not focusing on a single challenge related to long-term SLAM provides a broader and interesting discussion of the topic, given that some challenges may be related between each other. For example, removing elements from the current map estimation considered to be outdated due to appearance changes is related to both environment changing conditions and map management. Even so, the broader discussion itself is a limitation of the study, and future ones may prefer to focus only on methodologies related to a single challenge of lifelong SLAM.

The search strategy of this review may also be a possible limitation of this study. For example, if a paper does not mention any of the keywords that compose the search query (e.g., long-term, lifelong, or SLAM) in the abstract, title, or authors' keywords, our methodology (see Section 3) may not identify the paper. The search fields for the inquiry in all data sources considered in this review include indexed keywords. Still, the risk exists of excluding records if the databases do not index the papers with keywords in the search query and those obey the query logic expression. However, the search query does not focus on only a single challenge of performing long-term SLAM. This approach of having a more generic query string intends also to reduce the risk of not identifying records related to the topic of this review. Also, the fact that the methodology followed in this review identifies 142 works related to lifelong SLAM (and 411 eligible ones after the screening phase) indicates the extended reach of the search strategy.

Furthermore, the quality assessment in the selection phase of this study only considers two criteria associated with the topic of this study, namely, QE3 and QE5. While the remaining quality criteria evaluate the eligible records in terms of their scientific contribution, only two out of nine being related to the topic of the review may be considered a limitation of this study. This limitation is related to the previous one. Indeed, the QE not considering specific challenges and characteristics of long-term autonomy tries to avoid the a priori knowledge of the authors to the review on lifelong SLAM biasing the methods' selection methodology. However, even though the followed methodology obtained two distinct peaks in the QE scores (see Figure 2) that could be interpreted as belonging to two different clusters—records to exclude versus the ones to include in the review—, the inclusion of more QE criteria would perhaps improve the distinction between the clusters.

The other limitation of this study is the discussion of the public data sets. Instead of considering a different query to find and select data sets, the discussion focused only on the ones used in the experiments by the included works. While this selection approach allowed the identification of 43 different data sets, it does not mean that these data sets are the best to use to evaluate methodologies related to long-term SLAM due to the identification of the data sets discussed in this study may be biased by the included works themselves. Although the main focus of this systematic literature review is on ways to achieve lifelong autonomy and not only on the experimental data, future studies should consider to review separately the data sets from the methodologies. Still, the inclusion of an analysis of common used data sets by the included works improves the interest of this review for researchers interested in long-term localization and mapping.

## 8 | CONCLUSIONS

This paper presents a systematic literature review on long-term localization and mapping for mobile robots. The review selects 142 works from the literature covering the main strategies to achieve lifelong SLAM and discussing the experimental data (including private experiments and public data sets) and evaluation metrics commonly used to assess the performance of autonomous systems in long-term operations. The discussion analyzes the included works in terms of appearance invariance to changing conditions in the environment, dealing with dynamic elements in the scene, multisession strategies, map sparsification techniques to bound the computational resources, and other computational-related topics. Also, an overview over the bibliographic data of the included works identifies the most used terms in long-term SLAM and research networks between the authors of the included records using the VOSviewer (van Eck & Waltman, 2010, 2014) tool. The overview over the bibliographic data also discusses the evolution of the number of publications over time and the publication venues with more records related to long-term SLAM.

Overall, the methodologies discussed in this study are a step forward to achieve lifelong SLAM. In terms of dealing with appearance changes in the environment, CNN-based features are more discriminative compared to handcrafted features. The consideration of both appearance and geometric cues in the feature descriptor also improves its invariance to changing conditions. Semantic features and considering different sensorization sources (e.g., cameras, LiDAR, and/or radar) can also

increase the robustness to appearance changes in the environment. For dealing with dynamic elements, the most common approach is to distinguish between static, dynamic, and semi-static changes in the perceived environment. Usually, only static changes and stable features are used for localization to improve the reliability of the pose estimator. As for constraining the map size to the explored environment area instead of the trajectory length, the evaluation of the mutual information based on information theory or heuristics is the most used technique to bound the map size.

However, there are still challenges and future directions for the research on long-term SLAM. Vision-based global recognition is affected by viewpoint variance. The use of omnidirectional sensors to increase the field-of-view relative to perspective cameras may decrease that variance. Decentralized computation architectures may be an interesting research direction to offload heavier computation tasks from the mapping agent and improve the scalability of long-term SLAM algorithms. For example, cloud and edge computing devices allow the parallelization of tasks and increase the availability of computational resources, even though communication constraints should be accounted. Also, the extension of the methodologies discussed in this study to multirobots systems may optimize the operation of autonomous systems while also increasing their robustness in the long-term.

Lastly, this review can be updated in future studies thanks to following a systematic review methodology based on the PRISMA (Page et al., 2021) statement. The methodology defines explicitly the time interval in the selection process (the data of the last full inquiry is May 17, 2022) and the search query used to identify records from the literature. Also, all documentation and scripts using during the review process are available in the public GitHub repository (see footnote 1) referred in Section 1 to facilitate replicating of the results and future updates to this review. This paper may be extended to focus on single challenges of long-term SLAM for providing an in-depth comparison with experimental results between different methodologies.

## ACKNOWLEDGMENTS

This study was supported by FCT—Fundação para a Ciência e a Tecnologia and INESC TEC – Institute for Systems and Computer Engineering, Technology and Science. This work is financed by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. This work is also financed by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia, within scholarship 2021.04591.BD.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.

## ORCID

*Ricardo B. Sousa* http://orcid.org/0000-0003-4537-5095
*Héber M. Sobreira* http://orcid.org/0000-0002-8055-1093
*António Paulo Moreira* http://orcid.org/0000-0001-8573-3147

## REFERENCES

*ACM Digital Library*. (n.d.) Available at: https://dl.ACM.org/ [Accessed 6th May 2022].

Alex Kendall, V.B. & Cipolla, R. (2017) Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: Kim, T. K., Zafeiriou, S., Brostow, G. & Mikolajczyk, K. (Eds.) *British Machine Vision Conference (BMVC 2017)*, pp. 57.1–57.12. https://doi.org/10.5244/C.31.57

Ali, A.J.B., Hashemifar, Z.S. & Dantu, K. (2020) Edge-SLAM: edge-assisted visual simultaneous localization and mapping. In: *MobiSys'20: Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, pp. 325–337. https://doi.org/10.1145/3386901.3389033

Ali, W., Liu, P., Ying, R. & Gong, Z. (2021) A life-long SLAM approach using adaptable local maps based on rasterized LIDAR images. *IEEE Sensors Journal*, 21(19), 21740–21749. Available from: https://doi.org/10.1109/JSEN.2021.3100882

An, S.-Y., Lee, L.-K. & Oh, S.-Y. (2016) Ceiling vision-based active SLAM framework for dynamic and wide-open environments. *Autonomous Robots*, 40(2), 291–324. Available from: https://doi.org/10.1007/s10514-015-9453-0.

Angeli, A., Filliat, D., Doncieux, S. & Meyer, J.-A. (2008) Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 24(5), 1027–1037. Available from: https://doi.org/10.1109/TRO.2008.2004514

Arroyo, R., Alcantarilla, P.F., Bergasa, L.M. & Romera, E. (2018) Are you ABLE to perform a life-long visual topological localization? *Autonomous Robots*, 42(3), 665–85. Available from: https://doi.org/10.1007/s10514-017-9664-7.

Atkinson, R. & Shiffrin, R. (1968) Human memory: a proposed system and its control processes. In: K. W., Spence & J. T., Spence (Eds.). *Psychology of learning and motivation*. Academic Press, pp. 89–195. Available from: https://doi.org/10.1016/S0079-7421(08)60422-3

Bacca, B., Salví, J. & Cufí, X. (2013) Long-term mapping and localization using feature stability histograms. *Robotics and Autonomous Systems*, 61(12), 1539–1558. Available from: https://doi.org/10.1016/j.robot.2013.07.003

Badino, H., Huber, D. & Kanade, T. (2011) Visual topometric localization. In: *2011 IEEE Intelligent Vehicles Symposium (IV)*, pp. 794–799. Available from: https://doi.org/10.1109/IVS.2011.5940504

Badue, C., Guidolini, R., Carneiro, R.V., Azevedo, P., Cardoso, V.B., & Forechi, A. et al. (2021) Self-driving cars: a survey. *Expert Systems with Applications*, 165, 113816. Available from: https://doi.org/10.1016/j.eswa.2020.113816

Bailey, T. & Durrant-Whyte, H. (2006) Simultaneous localization and mapping (SLAM): part II. *IEEE Robotics & Automation Magazine*, 13(3), 108–117. Available from: https://doi.org/10.1109/MRA.2006.1678144

Ball, D., Heath, S., Wiles, J., Wyeth, G.F., Corke, P. & Milford, M.J. (2013) OpenRatSLAM: an open source brain-based SLAM system. *Autonomous Robots*, 34(3), 149–176. Available from: https://doi.org/10.1007/s10514-012-9317-9

Barnes, D., Gadd, M., Murcutt, P., Newman, P. & Posner, I. (2020) The Oxford Radar RobotCar dataset: a radar extension to the Oxford RobotCar dataset. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6433–6438. Available from: https://doi.org/10.1109/ICRA40945.2020.9196884

Bay, H., Tuytelaars, T. & Van Gool, L. (2006) SURF: speeded up robust features. In: *Computer Vision—ECCV 2006*, pp. 404–417. Available from: https://doi.org/10.1007/11744023_32

Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D. & Bochtis, D. (2021) Machine learning in agriculture: a comprehensive updated review. *Sensors*, 21(11), 3758. Available from: https://doi.org/10.3390/s21113758

Berrio, J.S., Ward, J., Worrall, S. & Nebot, E. (2019) Identifying robust landmarks in feature-based maps. In: *2019 IEEE Intelligent Vehicles*

*Symposium (IV)*, pp. 1166–1172. Available from: https://doi.org/10.1109/IVS.2019.8814289

Berrio, J.S., Worrall, S., Shan, M. & Nebot, E. (2021) Long-term map maintenance pipeline for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(8), 10427–10440. Available from: https://doi.org/10.1109/TITS.2021.3094485

Bescos, B., Fácil, J.M., Civera, J. & Neira, J. (2018) DynaSLAM: tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, 3(4), 4076–4083. Available from: https://doi.org/10.1109/LRA.2018.2860039

Besl, P. & McKay, N.D. (1992) A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2), 239–256. Available from: https://doi.org/10.1109/34.121791

Biber, P. & Strasser, W. (2003) The normal distributions transform: a new approach to laser scan matching. In: *2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. vol. 3. pp. 2743–2748. Available from: https://doi.org/10.1109/IROS.2003.1249285

Biber, P. & Duckett, T. (2009) Experimental analysis of sample-based maps for long-term SLAM. *International Journal of Robotics Research*, 28(1), 20–33. Available from: https://doi.org/10.1177/0278364908096286

Biswas, J. & Veloso, M.M. (2013) Localization and navigation of the CoBots over long-term deployments. *International Journal of Robotics Research*, 32(14), 1679–1694. Available from: https://doi.org/10.1177/0278364913503892

Biswas, J. & Veloso, M.M. (2017) Episodic non-Markov localization. *Robotics and Autonomous Systems*, 87, 162–176. Available from: https://doi.org/10.1016/j.robot.2016.09.005

Blanco, J.-L., Moreno, F.-A. & González, J. (2009) A collection of outdoor robotic datasets with centimeter-accuracy ground truth. *Autonomous Robots*, 27(4), 327. Available from: https://doi.org/10.1007/s10514-009-9138-7

Blei, D.M., Ng, A. Y. & Jordan, M.I. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022. Available from: https://www.jmlr.org/papers/v3/blei03a.html

Boniardi, F., Caselitz, T., Kümmerle, R. & Burgard, W. (2019) A pose graph-based localization system for long-term navigation in CAD floor plans. *Robotics and Autonomous Systems*, 112, 84–97. Available from: https://doi.org/10.1016/j.robot.2018.11.003

Borrego, M., Foster, M.J. & Froyd, J.E. (2014) Journal of engineering education. *Systematic Literature Reviews in Engineering Education and Other Developing Interdisciplinary Fields*, 103(1), 45–76. Available from: https://doi.org/10.1002/jee.20038

Boser, B.E., Guyon, I. M. & Vapnik, V.N. (1992) A training algorithm for optimal margin classifiers. In: *COLT92: 5th Annual Workshop on Computational Learning Theory*, pp. 144–152. Available from: https://doi.org/10.1145/130385.130401

Bosse, M., Newman, P., Leonard, J.J. & Teller, S. (2004) Simultaneous localization and map building in large-scale cyclic environments using the Atlas framework. *International Journal of Robotics Research*, 23(12), 1113–1139. Available from: https://doi.org/10.1177/0278364904049393].

Bosse, M. & Zlot, R. (2009) Keypoint design and evaluation for place recognition in 2D lidar maps. *Robotics and Autonomous Systems*, 57(12), 1211–1224. Available from: https://doi.org/10.1016/j.robot.2009.07.009

Bouaziz, Y., Royer, E., Bresson, G. & Dhome, M. (2021) Over two years of challenging environmental conditions for localization: the IPLT dataset. In: *International Conference on Informatics in Control, Automation and Robotics (ICINCO 2021)*, pp. 383–387. Availabe from: https://doi.org/10.5220/0010518303830387

Bouaziz, Y., Royer, E., Bresson, G. & Dhome, M. (2022) Map management for robust long-term visual localization of an autonomous shuttle in changing conditions. *Multimedia Tools and Applications*, 81(16), 22449–22480. Available from: https://doi.org/10.1007/s11042-021-11870-4

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. Available from: https://doi.org/10.1109/TIV.2017.2749181

Burgard, W., Trahanias, P., Hähnel, D., Moors, M., Schulz, D., Baltzakis, H. et al. (2003) Tele-presence in populated exhibitions through web-operated mobile robots. *Autonomous Robots*, 15(3), 299–316. Available from: https://doi.org/10.1023/A:1026272605502

Bürki, M., Cadena, C., Gilitschenski, I., Siegwart, R. & Nieto, J. (2019) Appearance-based landmark selection for visual localization. *Journal of Field Robotics*, 36(6), 1041–1073. Available from: https://doi.org/10.1002/rob.21870

Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., et al. (2016) The EuRoC micro aerial vehicle datasets. *International Journal of Robotics Research*, 35(10), 1157–1163. Available from: https://doi.org/10.1177/0278364915620033

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. Available from: https://doi.org/10.1109/TRO.2016.2624754

Calonder, M., Lepetit, V., Strecha, C. & Fua, P. (2010) BRIEF: binary robust independent elementary features. In: *Computer Vision—ECCV 2010*. pp. 778–792. Available from: https://doi.org/10.1007/978-3-642-15561-1_56

Camara, L.G., Gäbert, C. & Přeučil, L. (2020) Highly robust visual place recognition through spatial matching of CNN features. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3748–3755. Available from: https://doi.org/10.1109/ICRA40945.2020.9196967

Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M.M. & Tardós, J.D. (2021) ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 37(6), 1874–1890. Available from: https://doi.org/10.1109/TRO.2021.3075644

Cao, F., Yan, F., Wang, S., Zhuang, Y. & Wang, W. (2021) Season-invariant and viewpoint-tolerant LiDAR place recognition in GPS-denied environments. *IEEE Transactions on Industrial Electronics*, 68(1), 563–574. Available from: https://doi.org/10.1109/TIE.2019.2962416

Cao, F., Zhuang, Y., Zhang, H. & Wang, W. (2018) Robust place recognition and loop closing in laser-based SLAM for UGVs in urban environments. *IEEE Sensors Journal*, 18(10), 4242–4252. Available from: https://doi.org/10.1109/JSEN.2018.2815956

Carlevaris-Bianco, N., Kaess, M. & Eustice, R.M. (2014) Generic node removal for factor-graph SLAM. *IEEE Transactions on Robotics*, 30(6), 1371–1385. Available from: https://doi.org/10.1109/TRO.2014.2347571

Carlevaris-Bianco, N., Ushani, A. & Eustice, R.M. (2016) The University of Michigan North Campus Long-Term vision and LIDAR dataset. *International Journal of Robotics Research*, 35(9), 1023–1035. Available from: https://doi.org/10.1177/0278364915614638

Charles, R.Q., Su, H., Kaichun, M. & Guibas, L.J. (2017) PointNet: deep learning on point sets for 3D classification and segmentation. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85. Available from: https://doi.org/10.1109/CVPR.2017.16

Chebrolu, N., Läbe, T. & Stachniss, C. (2018) Robust long-term registration of UAV images of crop fields for precision agriculture. *IEEE Robotics and Automation Letters*, 3(4), 3097–3104. Available from: https://doi.org/10.1109/LRA.2018.2849603

Chen, Z., Liu, L., Sa, I., Ge, Z. & Chli, M. (2018) Learning context flexible attention model for long-term visual place recognition. *IEEE Robotics*

*and Automation Letters*, 3(4), 4015–4022. Available from: https://doi.org/10.1109/LRA.2018.2859916

Chen, Z., Maffra, F., Sa, I. & Chli, M. (2017). Only look once, mining distinctive landmarks from convnet for visual place recognition. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 9–16. Available from: https://doi.org/10.1109/IROS.2017.8202131

Choi, Y., Kim, N., Park, K., Hwang, S., Yoon, J.S. & Kweon, I.S. (2015) *KAIST all-day visual place recognition benchmark and baseline*. Korea Advanced Institute of Science and Technology.

Chow, C. & Liu, C. (1968) Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3), 462–467. Available from: https://doi.org/10.1109/TIT.1968.1054142

Chukwu, E. & Garg, L. (2020) A systematic review of blockchain in healthcare: frameworks, prototypes, and implementations. *IEEE Access*, 8, 21196–21214. Available from: https://doi.org/10.1109/ACCESS.2020.2969881

Churchill, W. & Newman, P. (2013) Experience-based navigation for long-term localisation. *International Journal of Robotics Research*, 32(14), 1645–1661. Available from: https://doi.org/10.1177/0278364913499193

Clement, L., Gridseth, M., Tomasi, J. & Kelly, J. (2020) Learning matchable image transformations for long-term metric visual localization. *IEEE Robotics and Automation Letters*, 5(2), 1492–1499. Available from: https://doi.org/10.1109/LRA.2020.2967659

Coulin, J., Guillemard, R., Gay-Bellile, V., Joly, C. & de La Fortelle, A. (2022) Tightly-coupled magneto-visual-inertial fusion for long term localization in indoor environment. *IEEE Robotics and Automation Letters*, 7(2), 952–959. Available from: https://doi.org/10.1109/LRA.2021.3136241

Cummins, M. & Newman, P. (2008) FAB-MAP: probabilistic localization and mapping in the space of appearance. *International Journal of Robotics Research*, 27(6), 647–665. Available from: https://doi.org/10.1177/0278364908090961

Curless, B. & Levoy, M. (1996) A volumetric method for building complex models from range images. In: *SIGGRAPH'96: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 303–312. Available from: https://doi.org/10.1145/237170.237269

Dalal, N. & Triggs, B. (2005) Histograms of oriented gradients for human detection. In: *2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 886–893. Available from: https://doi.org/10.1109/CVPR.2005.177

Davison, A.J. & Murray, D.W. (2002) Simultaneous localization and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 865–880. Available from: https://doi.org/10.1109/TPAMI.2002.1017615

Dayoub, F., Cielniak, G. & Duckett, T. (2011) Long-term experiments with an adaptive spherical view representation for navigation in changing environments. *Robotics and Autonomous Systems*, 59(5), 285–295. Available from: https://doi.org/10.1016/j.robot.2011.02.013

Dayoub, F. & Duckett, T. (2008) An adaptive appearance-based map for long-term topological localization of mobile robots. In: *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3364–3369. Available from: https://doi.org/10.1109/IROS.2008.4650701

Derner, E., Gomez, C., Hernandez, A.C., Barber, R. & Babuška, R. (2021) Change detection using weighted features for image-based localization. *Robotics and Autonomous Systems*, 135, 103676. Available from: https://doi.org/10.1016/j.robot.2020.103676

DeTone, D., Malisiewicz, T. & Rabinovich, A. (2018) SuperPoint: self-supervised interest point detection and description. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 337–349. Available from: https://doi.org/10.1109/CVPRW.2018.00060

*Dimensions*. (n.d.) Available at: https://app.dimensions.ai/discover/publication [Accessed 6th May 2022].

Ding, X., Wang, Y., Tang, L., Yin, H. & Xiong, R. (2019) Communication constrained cloud-based long-term visual localization in real time. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2159–2166. Available from: https://doi.org/10.1109/IROS40897.2019.8968550

Ding, X., Wang, Y., Xiong, R., Li, D., Tang, L., Yin, H., et al. (2020) Persistent stereo visual localization on cross-modal invariant map. *IEEE Transactions on Intelligent Transportation Systems*, 21(11), 4646–4658. Available from: https://doi.org/10.1109/TITS.2019.2942760

Dissanayake, G., Huang, S., Wang, Z. & Ranasinghe, R. (2011) A review of recent developments in simultaneous localization and mapping. In: *2011 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pp. 477–482. Available from: https://doi.org/10.1109/ICIINFS.2011.6038117

Donahue, J., Krähenbühl, P. & Darrell, T. (2017) Adversarial feature learning. In: *International Conference on Learning Representations (ICLR 2017)*, pp. 1–18. Available from: https://doi.org/10.48550/arXiv.1605.09782

Du, Z.-J., Huang, S.-S., Mu, T.-J., Zhao, Q., Martin, R.R. & Xu, K. (2022) Accurate dynamic SLAM using CRF-based long-term consistency. *IEEE Transactions on Visualization and Computer Graphics*, 28(4), 1745–1757. Available from: https://doi.org/10.1109/TVCG.2020.3028218

Durrant-Whyte, H. & Bailey, T. (2006) Simultaneous localization and mapping (SLAM): part I. *IEEE Robotics & Automation Magazine*, 13(2), 99–110. https://doi.org/10.1109/MRA.2006.1638022

Dymczyk, M., Lynen, S., Cieslewski, T., Bosse, M., Siegwart, R. & Furgale, P. (2015) The gist of maps - summarizing experience for lifelong localization. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)* (June), pp. 2767–2773. Available from: https://doi.org/10.1109/ICRA.2015.7139575

Dymczyk, M., Schneider, T., Gilitschenski, I., Siegwart, R. & Stumm, E. (2016) Erasing bad memories: Agent-side summarization for long-term mapping. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4572–4579. Available from: https://doi.org/10.1109/IROS.2016.7759673

Dymczyk, M., Stumm, E., Nieto, J., Siegwart, R. & Gilitschenski, I. (2016) Will it last? Learning stable features for long-term visual localization. In: *2016 International Conference on 3D Vision (3DV)*, pp. 572–581. Available from: https://doi.org/10.1109/3DV.2016.66

Ebadi, K., Bernreiter, L., Biggie, H., Catt, G., Chang, Y., Chatterjee, A. et al. (2022) Present and future of SLAM in extreme underground environments [Preprint]. Available from: https://doi.org/10.48550/arXiv.2208.01787

Egger, P., Borges, P.V.K., Catt, G., Pfrunder, A., Siegwart, R. & Dubé, R. (2018) PoseMap: lifelong, multi-environment 3D LiDAR localization. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 3430–3437. Available from: https://doi.org/10.1109/IROS.2018.8593854

Einhorn, E. & Gross, H.-M. (2013) Generic 2D/3D SLAM with NDT maps for lifelong application. In: *2013 European Conference on Mobile Robots (ECMR)*. pp. 240–247. Available from: https://doi.org/10.1109/ECMR.2013.6698849

Einhorn, E. & Gross, H.-M. (2015) Generic NDT mapping in dynamic environments and its application for lifelong SLAM. *Robotics and Autonomous Systems*, 69(1), 28–39. Available from: https://doi.org/10.1016/j.robot.2014.08.008

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. pp. 226–231. Available from: https://www.aaai.org/Library/KDD/1996/kdd96-037.php

Fallon, M., Johannsson, H., Kaess, M. & Leonard, J.J. (2013) The MIT Stata Center data set. *International Journal of Robotics Research*, 32(14), 1695–1699. Available from: https://doi.org/10.1177/0278364913509035

Fayyad, J., Jaradat, M.A., Gruyer, D. & Najjaran, H. (2020) Deep learning sensor fusion for autonomous vehicle perception and localization: a review. *Sensors*, 20(15), 4220. Available from: https://doi.org/10.3390/s20154220

Filliat, D. (2007) A visual bag of words method for interactive qualitative localization and mapping. In: *2007 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3921–3926. Available from: https://doi.org/10.1109/ROBOT.2007.364080

Fontana, G., Matteucci, M., Sorrenti, D.G., Marzorati, D., Giusti, A., Taddei, P. et al. (2009) *Bicocca (indoor)*. Università degli Studi di Milano-Bicocca.

Fraundorfer, F. & Scaramuzza, D. (2012) Visual odometry. part II: matching, robustness, optimization, and applications. *IEEE Robotics & Automation Magazine*, 19(2), 78–90. Available from: https://doi.org/10.1109/MRA.2012.2182810

Freitas, V. (2014) *Parsifal*. Available at: https://parsif.al/ [Accessed 12th May 2022].

Gabor, D. (1946) Theory of communication. *Journal of the Institution of Electrical Engineers*, 93(26), 429–441. Available from: https://doi.org/10.1049/ji-3-2.1946.0074

Gadd, M. & Newman, P. (2016) Checkout my map: version control for fleetwide visual localisation. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5729–5736. Available from: https://doi.org/10.1109/IROS.2016.7759843

Ganti, P. & Waslander, S.L. (2019) Network uncertainty informed semantic feature selection for visual SLAM. In: *2019 Conference on Computer and Robot Vision (CRV)*, pp. 121–128. Available from: https://doi.org/10.1109/CRV.2019.00024

Gao, P. & Zhang, H. (2020) Long-term place recognition through worst-case graph matching to integrate landmark appearances and spatial relationships. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1070–1076. Available from: https://doi.org/10.1109/ICRA40945.2020.9196906

Geiger, A., Lenz, P., Stiller, C. & Urtasun, R. (2013) Vision meets robotics: the KITTI dataset. *International Journal of Robotics Research*, 32(11), 1231–1237. Available from: https://doi.org/10.1177/0278364913491297

Gidaris, S. & Komodakis, N. (2016) LocNet: improving localization accuracy for object detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 789–798. Available from: https://doi.org/10.1109/CVPR.2016.92

Glover, A.J. (2014) *Day and night with lateral pose change datasets (Gardens Point Campus of QUT)*. Queensland University of Technology. Available from: https://doi.org/10.5281/zenodo.4561862

Glover, A.J., Maddern, W.P., Milford, M.J. & Wyeth, G.F. (2010) FAB-MAP + RatSLAM: appearance-based SLAM for multiple times of day. In: *2010 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 3507–3512. Available from: https://doi.org/10.1109/ROBOT.2010.5509547

Gonzalez, R.C. & Woods, R. E. (2002) *Digital image processing*. Prentice Hall.

Griffith, S. & Pradalier, C. (2017) Survey registration for long-term natural environment monitoring. *Journal of Field Robotics*, 34(1), 188–208. Available from: https://doi.org/10.1002/rob.21664

Grisetti, G., Kümmerle, R., Stachniss, C. & Burgard, W. (2010) A tutorial on graph-based SLAM. *IEEE Intelligent Transportation Systems Magazine*, 2(4), 31–43. Available from: https://doi.org/10.1109/MITS.2010.939925

Hähnel, D. (2003) *Intel Research Lab (Seattle)*. Intel Labs.

Han, F., Beleidy, S.E., Wang, H., Ye, C. & Zhang, H. (2018) Learning of holism-landmark graph embedding for place recognition in long-term autonomy. *IEEE Robotics and Automation Letters*, 3(4), 3669–3676. Available from: https://doi.org/10.1109/LRA.2018.2856274

Han, F., Wang, H., Huang, G. & Zhang, H. (2018) Sequence-based sparse optimization methods for long-term loop closure detection in visual SLAM. *Autonomous Robots*, 42(7), 1323–1335. Available from: https://doi.org/10.1007/s10514-018-9736-3

Han, F., Yang, X., Deng, Y., Rentschler, M., Yang, D. & Zhang, H. (2017) SRAL: shared representative appearance learning for long-term visual place recognition. *IEEE Robotics and Automation Letters*, 2(2), 1172–1179. Available from: https://doi.org/10.1109/LRA.2017.2662061

Hassler, U. (2016) Autoregressive moving average processes (ARMA). In: *Stochastic Processes and Calculus: an Elementary Introduction with Applications*, pp. 45–75. Available from: https://doi.org/10.1007/978-3-319-23428-1_3

He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2020) Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 386–397. Available from: https://doi.org/10.1109/TPAMI.2018.2844175

He, K., Zhang, X., Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. Available from: https://doi.org/10.1109/CVPR.2016.90

He, L., Wang, X. & Zhang, H. (2016) M2DP: a novel 3D point cloud descriptor and its application in loop closure detection. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 231–237. Available from: https://doi.org/10.1109/IROS.2016.7759060

Hochdorfer, S., Lutz, M. & Schlegel, C. (2009) Lifelong localization of a mobile service-robot in everyday indoor environments using omnidirectional vision. In: *2009 IEEE International Conference on Technologies for Practical Robot Applications (TePRA)*, pp. 161–166. Available from: https://doi.org/10.1109/TEPRA.2009.5339626

Hochdorfer, S. & Schlegel, C. (2009) Towards a robust visual SLAM approach: addressing the challenge of life-long operation. In: *2009 International Conference on Advanced Robotics (ICAR)*, pp. 1–6. Available at: https://ieeexplore.ieee.org/document/5174794

Hochreiter, S. & Schmidhuber, J. (1997) Long short-term memory. *Neural Computation*, 9(8), 1735–1780. Available from: https://doi.org/10.1162/neco.1997.9.8.1735

Hofmann, T. (1999) Probabilistic latent semantic indexing. In: *SIGIR'99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57. Available from: https://doi.org/10.1145/312624.312649

Hong, Z., Petillot, Y., Wallace, A. & Wang, S. (2022) RadarSLAM: A robust simultaneous localization and mapping system for all weather conditions. *International Journal of Robotics Research*, 41(5), 519–542. Available from: https://doi.org/10.1177/0278364921080483

Hu, H., Wang, H., Liu, Z. & Chen, W. (2022) Domain-invariant similarity activation map contrastive learning for retrieval-based long-term visual localization. *IEEE/CAA Journal of Automatica Sinica*, 9(2), 313–328. Available from: https://doi.org/10.1109/JAS.2021.1003907

Huang, G., Kaess, M. & Leonard, J.J. (2013) Consistent sparsification for graph optimization. In: *2013 European Conference on Mobile Robots (ECMR)*, pp. 150–157. Available from: https://doi.org/10.1109/ECMR.2013.6698835

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017) Densely connected convolutional networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269. Available from: https://doi.org/10.1109/CVPR.2017.243

Huang, S. & Dissanayake, G. (2016) A critique of current developments in simultaneous localization and mapping. *International Journal of Advanced Robotic Systems*, 13(5), 1–13. Available from: https://doi.org/10.1177/1729881416669482

*IEEE Xplore*. (n.d.) Available at: https://ieeexplore.ieee.org/Xplore/home.jsp [Accessed 6th May 2022].

Ikeda, K. & Kanji, T. (2010) Visual robot localization using compact binary landmarks. In: *2010 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4397–4403. Available from: https://doi.org/10.1109/ROBOT.2010.5509579

Ila, V., Polok, L., Solony, M. & Svoboda, P. (2017) SLAM++—a highly efficient and temporally scalable incremental SLAM framework. *International Journal of Robotics Research*, 36(2), 210–230. https://doi.org/10.1177/0278364917691110

*INSPEC*. (n.d.) Available at: https://www.engineeringvillage.com/search/quick.url [Accessed 6th May 2022].

Johannsson, H., Kaess, M., Fallon, M. & Leonard, J.J. (2013) Temporally scalable visual SLAM using a reduced pose graph. In: *2013 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 54–61. Available from: https://doi.org/10.1109/ICRA.2013.6630556

Karaoğuz, H. & Bozma, H.I. (2016) An integrated model of autonomous topological spatial cognition. *Autonomous Robots*, 40(8), 1379–1402. Available from: https://doi.org/10.1007/s10514-015-9514-4

Karaoğuz, H. & Bozma, H.I. (2020) Merging Available from: of appearance-based place knowledge among multiple robots. *Autonomous Robots*, 44(6), 1009–1027. Available from: https://doi.org/10.1007/s10514-020-09911-2

Kawewong, A., Tongprasit, N. & Hasegawa, O. (2013) A speeded-up online incremental vision-based loop-closure detection for long-term SLAM. *Advanced Robotics*, 27(17), 1325–1336. Available from: https://doi.org/10.1080/01691864.2013.826410

Kim, C., Cho, S., Sunwoo, M., Resende, P., Bradaï, B. & Jo, K. (2021) A geodetic normal distribution map for long-term LiDAR localization on earth. *IEEE Access*, 9, 470–484. Available from: https://doi.org/10.1109/ACCESS.2020.3047421

Kim, G., Park, B. & Kim, A. (2019) 1-day learning, 1-year localization: long-term LiDAR localization using scan context image. *IEEE Robotics and Automation Letters*, 4(2), 1948–1955. Available from: https://doi.org/10.1109/LRA.2019.2897340

Kim, G., Park, Y.S., Cho, Y., Jeong, J. & Kim, A. (2020) MulRan: multimodal range dataset for urban place recognition. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6246–6253 Available from: https://doi.org/10.1109/ICRA40945.2020.9197298

Kingma, D.P. & Welling, M. (2013) Auto-encoding variational Bayes. In: *International Conference on Learning Representations (ICLR 2013)*, pp. 1–14. Available from: https://doi.org/10.48550/arXiv.1312.6114

Konolige, K. & Bowman, J. (2009) Towards lifelong visual maps. In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1156–1163. Available from: https://doi.org/10.1109/IROS.2009.5354121

Krajník, T., Fentanes, J.P., Mozos, O.M., Duckett, T., Ekekrantz, J. & Hanheide, M. (2014) Long-term topological localisation for service robots in dynamic environments using spectral maps. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4537–4542. Available from: https://doi.org/10.1109/IROS.2014.6943205

Krajník, T., Fentanes, J.P., Santos, J.M. & Duckett, T. (2017) FreMEn: frequency map enhancement for long-term mobile robot autonomy in changing environments. *IEEE Transactions on Robotics*, 33(4), 964–977. Available from: https://doi.org/10.1109/TRO.2017.2665664

Kretzschmar, H., Grisetti, G. & Stachniss, C. (2010) Lifelong map learning for graph-based SLAM in static environments. *KI—Kunstliche Intelligenz*, 24(3), 199–206. Available from: https://doi.org/10.1007/s13218-010-0034-2

Kretzschmar, H. & Stachniss, C. (2012) Information-theoretic compression of pose graphs for laser-based SLAM. *International Journal of Robotics Research*, 31(11), 1219–1230. Available from: https://doi.org/10.1177/0278364912455072

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems. Neural Information Processing Systems (NIPS 2012)*, 25, 1–9.

Kunze, L., Hawes, N., Duckett, T., Hanheide, M. & Krajník, T. (2018) Artificial intelligence for long-term robot autonomy: a survey. *IEEE Robotics and Automation Letters*, 3(4), 4023–4030. Available from: https://doi.org/10.1109/LRA.2018.2860628

Kurz, G., Holoch, M. & Biber, P. (2021) Geometry-based graph pruning for lifelong SLAM. In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3313–3320. Available from: https://doi.org/10.1109/IROS51168.2021.9636530

Kuutti, S., Fallah, S., Katsaros, K., Dianati, M., Mccullough, F. & Mouzakitis, A. (2018) A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications. *IEEE Internet of Things Journal*, 5(2), 829–846. Available from: https://doi.org/10.1109/JIOT.2018.2812300

Labbé, M. & Michaud, F. (2019) RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of Field Robotics*, 36(2), 416–446. Available from: https://doi.org/10.1002/rob.21831

Laguna, A.B., Riba, E., Ponsa, D. & Mikolajczyk, K. (2019) Key.Net: Keypoint detection by handcrafted and learned CNN filters. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5835–5843. Available from: https://doi.org/10.1109/ICCV.2019.00593

Latif, Y., Cadena, C. & Neira, J. (2012) Realizing, reversing, recovering: Incremental robust loop closing over time using the iRRR algorithm. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4211–4217. Available from: https://doi.org/10.1109/IROS.2012.6385879

Latif, Y., Huang, G., Leonard, J.J. & Neira, J. (2017) Sparse optimization for robust and efficient loop closing. *Robotics and Autonomous Systems*, 93, 13–26. Available from: https://doi.org/10.1016/j.robot.2017.03.016

Lázaro, M.T., Capobianco, R. & Grisetti, G. (2018) Efficient long-term mapping in dynamic environments. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 153–160. Available from: https://doi.org/10.1109/IROS.2018.8594310

Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. Available from: https://doi.org/10.1109/5.726791

Leung, K.Y.K., Halpern, Y., Barfoot, T.D. & Liu, H.H.T. (2011) The UTIAS multi-robot cooperative localization and mapping dataset. *International Journal of Robotics Research*, 30(8), 969–974. Available from: https://doi.org/10.1177/0278364911398404

Li, J., Eustice, R.M. & Johnson-Roberson, M. (2015) High-level visual features for underwater place recognition. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3652–3659. Available from: https://doi.org/10.1109/ICRA.2015.7139706

Liu, B., Tang, F., Fu, Y., Yang, Y. & Wu, Y. (2021) A flexible and efficient loop closure detection based on motion knowledge. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11241–11247. Available from: https://doi.org/10.1109/ICRA48506.2021.9561126

Liu, C., Yuen, J. & Torralba, A. (2011) SIFT flow: dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5), 978–994. Available from: https://doi.org/10.1109/TPAMI.2010.147

Lowe, D.G. (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Available from: https://doi.org/10.1023/B:VISI.0000029664.99615.94

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P. & Milford, M. J. (2016) Visual place recognition: a survey. *IEEE Transactions on Robotics*, 32(1), 1–19. Available from: https://doi.org/10.1109/TRO.2015.2496823

Luthardt, S., Willert, V. & Adamy, J. (2018) LLama-SLAM: learning high-quality visual landmarks for long-term mapping and localization. In: *2018 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 2645–2652. Available from: https://doi.org/10.1109/ITSC.2018.8569323

MacTavish, K., Paton, M. & Barfoot, T.D. (2018) Selective memory: recalling relevant experience for long-term visual localization. *Journal of Field Robotics*, 35(8), 1265–1292. Available from: https://doi.org/10.1002/rob.21838

Maddern, W., Milford, M. J. & Wyeth, G.F. (2012) Capping computation time and storage requirements for appearance-based localization with CAT-SLAM. In: *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 822–827. Available from: https://doi.org/10.1109/ICRA.2012.6224622

Maddern, W., Pascoe, G., Linegar, C. & Newman, P. (2017) 1 year, 1000 km: the Oxford RobotCar dataset. *International Journal of Robotics Research*, 36(1), 3–15. https://doi.org/10.1177/0278364916679498

Martini, D.D., Gadd, M. & Newman, P. (2020) KRadar++: coarse-to-fine FMCW scanning radar localisation. *Sensors*, 20(21), 1–23. https://doi.org/10.3390/s20216002

Mazuran, M., Burgard, W. & Tipaldi, G.D. (2016) Nonlinear factor recovery for long-term SLAM. *International Journal of Robotics Research*, 35(1-3), 50–72. https://doi.org/10.1177/0278364915581629

Meng, Q., Guo, H., Zhao, X., Cao, D. & Chen, H. (2021) Loop-closure detection with a multiresolution point cloud histogram mode in Lidar odometry and mapping for intelligent vehicles. *IEEE/ASME Transactions on Mechatronics*, 26(3), 1307–1317. Available from: https://doi.org/10.1109/TMECH.2021.3062647

Milford, M.J. & Wyeth, G.F. (2008) Mapping a suburb with a single camera using a biologically inspired SLAM system. *IEEE Transactions on Robotics*, 24(5), 1038–1053. Available from: https://doi.org/10.1109/TRO.2008.2004520

Milford, M.J. & Wyeth, G. F. (2012) SeqSLAM: visual route-based navigation for sunny summer days and stormy winter nights. In: *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1643–1649. Available from: https://doi.org/10.1109/ICRA.2012.6224623

Milioto, A., Vizzo, I., Behley, J. & Stachniss, C. (2019) RangeNet++: fast and accurate LiDAR semantic segmentation. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4213–4220. Available from: https://doi.org/10.1109/IROS40897.2019.8967762

Mishchuk, A., Mishkin, D., Radenovic, F. & Matas, J. (2017) Working hard to know your neighbor's margins: local descriptor learning loss. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R. (Eds.) *Advances in Neural Information Processing Systems. Neural Information Processing Systems (NIPS 2017)*, 30, 1–12.

Moayedi, H., Mosallanezhad, M., Rashid, A.S.A., Jusoh, W.A.W. & Muazu, M.A. (2020) A systematic review and meta-analysis of artificial neural network application in geotechnical engineering: theory and applications. *Neural Computing and Applications*, 32(2), 495–518. Available from: https://doi.org/10.1007/s00521-019-04109-9

Mohan, M., Gálvez-López, D., Monteleoni, C. & Sibley, G. (2015) Environment selection and hierarchical place recognition. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5487–5494. Available from: https://doi.org/10.1109/ICRA.2015.7139966

Mongeon, P. & Paul-Hus, A. (2016) The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1), 213–228. Available from: https://doi.org/10.1007/s11192-015-1765-5

Mühlfellner, P., Bürki, M., Bosse, M., Derendarz, W., Philippsen, R. & Furgale, P. (2016) Summary maps for lifelong visual localization. *Journal of Field Robotics*, 33(5), 561–590. Available from: https://doi.org/10.1002/rob.21595

Mur-Artal, R., Montiel, J.M.M. & Tardós, J.D. (2015) ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5), 1147–1163. Available from: https://doi.org/10.1109/TRO.2015.2463671

Mur-Artal, R. & Tardós, 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. Available from: https://doi.org/10.1109/TRO.2017.2705103

Murphy, L. & Sibley, G. (2014) Incremental unsupervised topological place discovery. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1312–1318. Available from: https://doi.org/10.1109/ICRA.2014.6907022

Nanni, L., Ghidoni, S. & Brahnam, S. (2017) Handcrafted vs non-handcrafted features for computer vision classification. *Pattern Recognition*, 7110, 158–172. Available from: https://doi.org/10.1016/j.patcog.2017.05.025

Naseer, T., Burgard, W. & Stachniss, C. (2018) Robust visual localization across seasons. *IEEE Transactions on Robotics*, 34(2), 289–302. Available from: https://doi.org/10.1109/TRO.2017.2788045

Naseer, T., Oliveira, G.L., Brox, T. & Burgard, W. (2017) Semantics-aware visual localization under challenging perceptual conditions. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2614–2620. Available from: https://doi.org/10.1109/ICRA.2017.7989305

Naseer, T., Suger, B., Ruhnke, M. & Burgard, W. (2015)Vision-based Markov localization across large perceptual changes. In: *2015 European Conference on Mobile Robots (ECMR)*, pp. 1–6. Available from: https://doi.org/10.1109/ECMR.2015.7324181

Neubert, P., Sünderhauf, N. & Protzel, P. (2015) Superpixel-based appearance change prediction for long-term navigation across seasons. *Robotics and Autonomous Systems*, 69, 15–27. Available from: https://doi.org/10.1016/j.robot.2014.08.005

Nguyen, T.-M., Cao, M., Yuan, S., Lyu, Y., Nguyen, T.H. & Xie, L. (2022) VIRAL-Fusion: a visual-inertial-ranging-Lidar sensor fusion approach. *IEEE Transactions on Robotics*, 38(2), 958–977. Available from: https://doi.org/10.1109/TRO.2021.3094157

Nguyen, T.-M., Yuan, S., Cao, M., Lyu, Y., Nguyen, T.H. & Xie, L. (2021) NTU VIRAL: a visual-inertial-ranging-lidar dataset, from an aerial vehicle viewpoint. *International Journal of Robotics Research*, 41(3), 270–280. Available from: https://doi.org/10.1177/0278364921105231 2

Nguyen, V.A., Starzyk, J.A. & Goh, W.-B. (2013) A spatio-temporal long-term memory approach for visual place recognition in mobile robotic navigation. *Robotics and Autonomous Systems*, 61(12), 1744–1758; Online probabilistic change detection in feature-based maps. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3661–3668. Available from: https://doi.org/10.1016/j.robot.2012.12.004

Nobre, F., Heckman, C., Ozog, P., Wolcott, R.W. & Walls, J.M. (2018) Online probabilistic change detection in feature-based maps. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 3661–3668. Available from: https://doi.org/10.1109/ICRA.2018.8461111

Nuske, S., Roberts, J. & Wyeth, G.F. (2009) Robust outdoor visual localization using a three-dimensional-edge map. *Journal of Field Robotics*, 26(9), 728–756. Available from: https://doi.org/10.1002/rob.20306

Oberländer, J., Roennau, A. & Dillmann, R. (2013) Hierarchical SLAM using spectral submap matching with opportunities for long-term operation. In: 2013 International Conference on Advanced Robotics (ICAR), pp. 1–7. Available from: https://doi.org/10.1109/ICAR.2013.6766479

Oh, J. & Eoh, G. (2021) Variational Bayesian approach to condition-invariant feature extraction for visual place recognition. *Applied Sciences*, 11(19), 8976. Available from: https://doi.org/10.3390/app11198976

Ojala, T., Pietikäinen, M. & Harwood, D. (1996) A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1), 51–59. Available from: https://doi.org/10.1016/0031-3203(95)00067-4

Oliva, A. & Torralba, A. (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3), 145–175. Available from: https://doi.org/10.1023/A:1011139631724

Oliveira, G.L., Burgard, W. & Brox, T. (2016) Efficient deep models for monocular road segmentation. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4885–4891. Available from: https://doi.org/10.1109/IROS.2016.7759717

Opdenbosch, D.V., Aykut, T., Oelsch, M., Alt, N. & Steinbach, E. (2018) Efficient map compression for collaborative visual SLAM. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 992–1000. Available from: https://doi.org/10.1109/WACV.2018.00114

Ouerghi, S., Boutteau, R., Savatier, X. & Tlili, F. (2018) Visual odometry and place recognition fusion for vehicle position tracking in urban environments. *Sensors*, 18(4), 939. Available from: https://doi.org/10.3390/s18040939

Ozog, P., Carlevaris-Bianco, N., Kim, A. & Eustice, R. M. (2016) Long-term mapping techniques for ship hull inspection and surveillance using an autonomous underwater vehicle. *Journal of Field Robotics*, 33(3), 265–289. Available from: https://doi.org/10.1002/rob.21582

Page, M.J., Moher, D., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D. et al. (2021) PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372(160), 372. Available from: https://doi.org/10.1136/bmj.n160

Palazzolo, E., Behley, J., Lottes, P., Giguère, P. & Stachniss, C. (2019) ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7855–7862. Available from: https://doi.org/10.1109/IROS40897.2019.8967590

Pan, Z., Chen, H., Li, S. & Liu, Y. (2019) Clustermap building and relocalization in urban environments for unmanned vehicles. *Sensors*, 19(19), 4252. Available from: https://doi.org/10.3390/s19194252

Pandey, G., McBride, J.R. & Eustice, R.M. (2011) Ford campus vision and lidar data set. *International Journal of Robotics Research*, 30(13), 1543–1552. Available from: https://doi.org/10.1177/0278364911400640

Paul, R. & Newman, P. (2013) Self-help: seeking out perplexing images for ever improving topological mapping. *International Journal of Robotics Research*, 32(14), 1742–1766. Available from: https://doi.org/10.1177/0278364913509859

Paull, L., Saeedi, S., Seto, M. & Li, H. (2014) AUV navigation and localization: a review. *IEEE Journal of Oceanic Engineering*, 39(1), 131–149. Available from: https://doi.org/10.1109/JOE.2013.2278891

Pérez, J., Caballero, F. & Merino, L. (2015) Enhanced Monte Carlo localization with visual place recognition for robust robot localization. *Journal of Intelligent and Robotic Systems: Theory and Applications*, 80(3–4), 641–656. Available from: https://doi.org/10.1007/s10846-015-0198-y

Phuan, A. & Prakash, S. (2002) K-means fast learning artificial neural network, an alternative network for classification. *International Conference on Neural Information Processing (ICONIP 2002)*, 2, 925–929. Available from: https://doi.org/10.1109/ICONIP.2002.1198196

Piasco, N., Sidibé, D., Gouet-Brunet, V. & Demonceaux, C. (2021) Improving image description with auxiliary modality for visual localization in challenging conditions. *International Journal of Computer Vision*, 129(1), 185–202. Available from: https://doi.org/10.1007/s11263-020-01363-6

Pirker, K. (2010) Histogram of oriented cameras—a new descriptor for visual SLAM in dynamic environments. In: *British Machine Vision Conference (BMVC 2010)*, pp. 76.1–76.12. Available from: https://doi.org/10.5244/C.24.76

Pirker, K., Rüther, M. & Bischof, H. (2011) CD SLAM—continuous localization and mapping in a dynamic world. In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3990–3997. Available from: https://doi.org/10.1109/IROS.2011.6048253

Plastiras, G., Kyrkou, C. & Theocharides, T. (2019) EdgeNet: Balancing accuracy and performance for edge-based convolutional neural network object detectors. In: *International Conference on Distributed Smart Cameras (ICDSC 2019)*, pp. 1–6. Available from: https://doi.org/10.1145/3349801.3349809

Pomerleau, F., Krüsi, P., Colas, F., Furgale, P. & Siegwart, R. (2014) Long-term 3D map maintenance in dynamic environments. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3712–3719. Available from: https://doi.org/10.1109/ICRA.2014.6907397

Pronobis, A. & Caputo, B. (2009) COLD: COsy localization database. *International Journal of Robotics Research*, 28(5), 588–594. https://doi.org/10.1177/0278364909103912

Qin, C., Zhang, Y., Liu, Y., Coleman, S., Kerr, D. & Lv, G. (2020) Appearance-invariant place recognition by adversarially learning disentangled representation. *Robotics and Autonomous Systems*, 131, 103561. Available from: https://doi.org/10.1016/j.robot.2020.103561

Qin, T., Chen, T., Chen, Y. & Su, Q. (2020) AVP-SLAM: Semantic visual mapping and localization for autonomous vehicles in the parking lot. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5939–5945. Available from: https://doi.org/10.1109/IROS45743.2020.9340939

Queralta, J.P., Taipalmaa, J., Can Pullinen, B., Sarker, V.K., Nguyen Gia, T., Tenhunen, H. et al. (2020) Collaborative multi-robot search and rescue: planning, coordination, perception, and active vision. *IEEE Access*, 8, 191617–191643. Available from: https://doi.org/10.1109/ACCESS.2020.3030190

Qureshi, M.I., Khan, N., Hassan Gillani, S.M.A. & Raza, H. (2020) A systematic review of past decade of mobile learning: what we learned and where to go. *International Journal of Interactive Mobile Technologies*, 14(6), 67–81. Available from: https://doi.org/10.3991/ijim.v14i06.13479

Rabiner, L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286. Available from: https://doi.org/10.1109/5.18626

Rapp, M., Hahn, M., Thom, M., Dickmann, J. & Dietmayer, K. (2015) Semi-Markov process based localization using radar in dynamic environments. In: *2015 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 423–429. Available from: https://doi.org/10.1109/ITSC.2015.77

Redmon, J. & Farhadi, A. (2018) YOLOv3: an incremental improvement [preprint]. Available from: https://doi.org/10.48550/arXiv.1804.02767.

Ren, S., He, K., Girshick, R. & Sun, J. (2015) Faster R-CNN: towards real-time object detection with region proposal networks. *Advances in*

*Neural Information Processing Systems. Neural Information Processing Systems (NIPS 2015)*, 28, 1–9.

Ronneberger, O., Fischer, P. & Brox, T. (2015) U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, pp. 234–241. Available from: https://doi.org/10.1007/978-3-319-24574-4_28

Rosenfeld, A. (1969) Picture processing by computer. *ACM Computing Surveys*, 1(3), 147–176. Available from: https://doi.org/10.1145/356551.356554

Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. (2011) ORB: an efficient alternative to SIFT or SURF. In: *2011 IEEE International Conference on Computer Vision (ICCV)*, pp. 2564–2571. Available from: https://doi.org/10.1109/ICCV.2011.6126544

Saarinen, J.P., Andreasson, H., Stoyanov, T. & Lilienthal, A.J. (2013) 3D normal distributions transform occupancy maps: an efficient representation for mapping in dynamic environments. *International Journal of Robotics Research*, 32(14), 1627–1644. Available from: https://doi.org/10.1177/0278364913499415

Saeedi, S., Trentini, M., Seto, M. & Li, H. (2016) Multiple-robot simultaneous localization and mapping: a review. *Journal of Field Robotics*, 33(1), 3–46. Available from: https://doi.org/10.1002/rob.21620

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. (2018) MobileNetV2: Inverted residuals and linear bottlenecks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520. Available from: https://doi.org/10.1109/CVPR.2018.00474

Santos, J.M., Krajník, T., Fentanes, J.P. & Duckett, T. (2016) Lifelong information-driven exploration to complete and refine 4-D spatio-temporal maps. *IEEE Robotics and Automation Letters*, 1(2), 684–691. Available from: https://doi.org/10.1109/LRA.2016.2516594

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E. et al. (2018) Benchmarking 6DOF outdoor visual localization in changing conditions. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8601–8610. Available from: https://doi.org/10.1109/CVPR.2018.00897

Scaramuzza, D. & Fraundorfer, F. (2011) Visual odometry. part I: the first 30 years and fundamentals. *IEEE Robotics & Automation Magazine*, 18(4), 80–92. Available from: https://doi.org/10.1109/MRA.2011.943233

Schaefer, A., Büscher, D., Vertens, J., Luft, L. & Burgard, W. (2021) Long-term vehicle localization in urban environments based on pole landmarks extracted from 3-D lidar scans. *Robotics and Autonomous Systems*, 136, 103709. Available from: https://doi.org/10.1016/j.robot.2020.103709

Schmuck, P. & Chli, M. (2019) On the redundancy detection in keyframe-based SLAM. In: *2019 International Conference on 3D Vision (3DV)*, pp. 594–603. Available from: https://doi.org/10.1109/3DV.2019.00071

*Scopus*. (n.d.) Available at: https://www.scopus.com/search/form.uri [Accessed 6th May 2022].

Sheeny, M., De Pellegrin, E., Mukherjee, S., Ahrabian, A., Wang, S. & Wallace, A. (2021) RADIATE: a radar dataset for automotive perception in bad weather. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 1–7. Available from: https://doi.org/10.1109/ICRA48506.2021.9562089

Simonyan, K. & Zisserman, A. (2015) Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations (ICLR 2015)*, pp. 1–14. Available from: https://doi.org/10.48550/arXiv.1409.1556

Singh, G., Wu, M., Lam, S.-K. & Minh, D.V. (2021) Hierarchical loop closure detection for long-term visual SLAM with semantic-geometric descriptors. In: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 2909–2916. Available from: https://doi.org/10.1109/ITSC48978.2021.9564866

Singh, V. K., Singh, P., Karmakar, M., Leta, J. & Mayr, P. (2021) The journal coverage of Web of Science, Scopus and dimensions: a comparative analysis. *Scientometrics*, 126(6), 5113–5142. Available from: https://doi.org/10.1007/s11192-021-03948-5.

Siva, S., Nahman, Z. & Zhang, H. (2020) Voxel-based representation learning for place recognition based on 3D point clouds. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8351–8357. Available from: https://doi.org/10.1109/IROS45743.2020.9340992

Siva, S. & Zhang, H. (2018) Omnidirectional multisensory perception fusion for long-term place recognition. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5175–5181. Available from: https://doi.org/10.1109/ICRA.2018.8461042

Sivic, J. & Zisserman, A. (2003) Video Google: a text retrieval approach to object matching in videos. In: *2003 IEEE International Conference on Computer Vision (ICCV)*, pp. 1470–1477. Available from: https://doi.org/10.1109/ICCV.2003.1238663

Skrede, S. (2013) *Nordlandsbanen: minute by minute, season by season*.- Norwegian Broadcasting Corporation.

Smith, M., Baldwin, I., Churchill, W., Paul, R. & Newman, P. (2009) The New College vision and laser data set. *International Journal of Robotics Research*, 28(5), 595–599. Available from: https://doi.org/10.1177/0278364909103911

Song, Y., Zhu, D., Li, J., Tian, Y. & Li, M. (2019) Learning local feature descriptor with motion attribute for vision-based localization. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3794–801. https://doi.org/10.1109/IROS40897.2019.8967749

Stachniss, C. (2003a) *Freiburg building 079 (FR079)*. [Dataset: http://www.ipb.uni-bonn.de/datasets/]. Albert-Ludwigs-Universität Freiburg.

Stachniss, C. (2003b) *Freiburg building 101 (FR101)*. [Dataset: http://www.ipb.uni-bonn.de/datasets/]. Albert-Ludwigs-Universität Freiburg.

Stachniss, C. (2010) *albert-b-laser-vision*. [Dataset: http://hdl.handle.net/1721.1/62291]. Albert-Ludwigs-Universität Freiburg.

Sturm, J., Engelhard, N., Endres, F., Burgard, W. & Cremers, D. (2012) A benchmark for the evaluation of RGB-D SLAM systems. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 573–580. Available from: https://doi.org/10.1109/IROS.2012.6385773

Sun, L., Yan, Z., Zaganidis, A., Zhao, C. & Duckett, T. (2018) Recurrent-OctoMap: learning state-based map refinement for long-term semantic mapping with 3-D-Lidar data. *IEEE Robotics and Automation Letters*, 3(4), 3749–3756. Available from: https://doi.org/10.1109/LRA.2018.2856268

Sun, L., Taher, M., Wild, C., Zhao, C., Zhang, Y., Majer, F. et al. (2021) Robust and long-term monocular teach and repeat navigation using a single-experience map. In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2635–2642. Available from: https://doi.org/10.1109/IROS51168.2021.9635886

Taisho, T. & Kanji, T. (2016) Mining DCNN landmarks for long-term visual SLAM. In: *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 570–576. Available from: https://doi.org/10.1109/ROBIO.2016.7866383

Tang, L. (2017) *YQ21: Dataset for long-term localization* [Dataset]. Available at: https://tangli.site/projects/academic/yq21/

Tang, L., Wang, Y., Ding, X., Yin, H., Xiong, R. & Huang, S. (2019) Topological local-metric framework for mobile robots navigation: a long term perspective. *Autonomous Robots*, 43(1), 197–211. Available from: https://doi.org/10.1007/s10514-018-9724-7

Tang, L., Wang, Y., Tan, Q. & Xiong, R. (2021) Explicit feature disentanglement for visual place recognition across appearance changes. *International Journal of Advanced Robotic Systems*, 18(6). Available from: https://doi.org/10.1177/17298814211037497

Thomas, H., Agro, B., Gridseth, M., Zhang, J. & Barfoot, T.D. (2021) Self-supervised learning of Lidar segmentation for autonomous indoor

navigation. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14047–14053, Available from: https://doi.org/10.1109/ICRA48506.2021.9561701

Thomas, H., Qi, C.R., Deschaud, J.-E., Marcotegui, B., Goulette, F. & Guibas, L. (2019) KPConv: flexible and deformable convolution for point clouds. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6410–6419. Available from: https://doi.org/10.1109/ICCV.2019.00651

Thrun, S. (2008) Simultaneous localization and mapping. In: Jefferies, M. E. & Yeap W.-K. (Eds.) *Robotics and cognitive approaches to spatial mapping*, pp. 13–41. Available from: https://doi.org/10.1007/978-3-540-75388-9_3

Thrun, S., Burgard, W. & Fox, D. (2005) *Probabilistic robotics*. Cambridge, Massachusetts: The MIT Press.

Tipaldi, G.D., Meyer-Delius, D. & Burgard, W. (2013) Lifelong localization in changing environments. *International Journal of Robotics Research*, 32(14), 1662–1678. Available from: https://doi.org/10.1177/0278364913502830

Tomono, M. (2004) A scan matching method using Euclidean invariant signature for global localization and map building. In: *2004 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 866–871. Available from: https://doi.org/10.1109/ROBOT.2004.1307258

Tsintotas, K.A., Bampis, L. & Gasteratos, A. (2021) Modest-vocabulary loop-closure detection with incremental bag of tracked words. *Robotics and Autonomous Systems*, 141, 103782. Available from: https://doi.org/10.1016/j.robot.2021.103782

Uy, M.A. & Lee, G.H. (2018) PointNetVLAD: deep point cloud based retrieval for large-scale place recognition. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4470–4479. Available from: https://doi.org/10.1109/CVPR.2018.00470

van Eck, N. J. & Waltman, L. (2010) Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. Available from: https://doi.org/10.1007/s11192-009-0146-3

van Eck, N. J. & Waltman, L. (2014) Visualizing bibliometric networks. In: Ding, Y., Rousseau, R. & Wolfram, D. (Eds.) *Measuring scholarly impact: Methods and practice*, Cham: Springer, pp. 285–320. Available from: https://doi.org/10.1007/978-3-319-10377-8_13

Vysotska, O., Naseer, T., Spinello, L., Burgard, W. & Stachniss, C. (2015) Efficient and effective matching of image sequences under substantial appearance changes exploiting GPS priors. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2774–2779. Available from: https://doi.org/10.1109/ICRA.2015.7139576

Walcott-Bryant, A., Kaess, M., Johannsson, H. & Leonard, J.J. (2012) Dynamic pose graph SLAM: long-term mapping in low dynamic environments. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1871–1878. Available from: https://doi.org/10.1109/IROS.2012.6385561

Wang, K., Lin, Y., Wang, L., Han, L., Hua, M., Wang, X., et al. (2019) A unified framework for mutual improvement of SLAM and semantic segmentation. In: *2019 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5224–5230. Available from: https://doi.org/10.1109/ICRA.2019.8793499

Wang, L., Chen, W. & Wang, J. (2020) Long-term localization with time series map prediction for mobile robots in dynamic environments. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8587–8593. Available from: https://doi.org/10.1109/IROS45743.2020.9468884

Wang, Z., Li, S., Cao, M., Chen, H. & Liu, Y. (2021) Pole-like objects mapping and long-term robot localization in dynamic urban scenarios. In: *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 998–1003. Available from: https://doi.org/10.1109/ROBIO54168.2021.9739599

*Web of Science*. (n.d.) https://www.webofscience.com/wos/woscc/basic-search [Accessed 6th May 2022].

Williams, S., Indelman, V., Kaess, M., Roberts, R., Leonard, J.J. & Dellaert, F. (2014) Concurrent filtering and smoothing: A parallel architecture for real-time navigation and full smoothing. *International Journal of Robotics Research*, 33(12), 1544–1568. Available from: https://doi.org/10.1177/0278364914531056

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B. & Wesslén, A. (2012) *Experimentation in software engineering*. Springer Berlin, Heidelberg. Available from: https://doi.org/10.1007/978-3-642-29044-2

Wu, L. & Wu, Y. (2019) Deep supervised hashing with similar hierarchy for place recognition. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3781–3786. Available from: https://doi.org/10.1109/IROS40897.2019.8968599

Xin, Z., Cui, X., Zhang, J., Yang, Y. & Wang, Y. (2017) Visual place recognition with CNNs: from global to partial. In: *2017 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6. Available from: https://doi.org/10.1109/IPTA.2017.8310121

Xing, Z., Zhu, X. & Dong, D. (2022) DE-SLAM: SLAM for highly dynamic environment. *Journal of Field Robotics*, 39(5), 528–542. Available from: https://doi.org/10.1002/rob.22062

Xu, X., Yin, H., Chen, Z., Li, Y., Wang, Y. & Xiong, R. (2021) DiSCO: differentiable scan context with orientation. *IEEE Robotics and Automation Letters*, 6(2), 2791–2798. Available from: https://doi.org/10.1109/LRA.2021.3060741

Yang, S., Fan, G., Bai, L., Zhao, C. & Li, D. (2020) SGC-VSLAM: a semantic and geometric constraints VSLAM for dynamic indoor environments. *Sensors*, 20(8), 2432. Available from: https://doi.org/10.3390/s20082432

Yang, X. & Cheng, K.-T.T. (2014) Local difference binary for ultrafast and distinctive feature description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1), 188–194. Available from: https://doi.org/10.1109/TPAMI.2013.150

Yang, Z., Pan, Y., Deng, L., Xie, Y. & Huan, R. (2021) PLSAV: parallel loop searching and verifying for loop closure detection. *IET Intelligent Transport Systems*, 15(5), 683–698. Available from: https://doi.org/10.1049/itr2.12054

Yin, H., Wang, Y., Ding, X., Tang, L., Huang, S. & Xiong, R. (2020) 3D LiDAR-based global localization using siamese neural network. *IEEE Transactions on Intelligent Transportation Systems*, 21(4), 1380–1392. Available from: https://doi.org/10.1109/TITS.2019.2905046

Yin, H., Xu, X., Wang, Y. & Xiong, R. (2021) Radar-to-Lidar: heterogeneous place recognition via joint learning. *Frontiers in Robotics and AI*, 8, 661199. Available from: https://doi.org/10.3389/frobt.2021.661199

Yin, P., Xu, J., Zhang, J. & Choset, H. (2021) FusionVLAD: a multi-view deep fusion networks for viewpoint-free 3D place recognition. *IEEE Robotics and Automation Letters*, 6(2), 2304–2310. Available from: https://doi.org/10.1109/LRA.2021.3061375

Yin, P., Xu, L., Liu, Z., Li, L., Salman, H., He, Y., et al. (2018) Stabilize an unsupervised feature learning for LiDAR-based place recognition. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1162–1167. Available from: https://doi.org/10.1109/IROS.2018.8593562

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. Available from: https://doi.org/10.1007/s40903-015-0032-7

Yu, C., Liu, Z., Liu, X.-J., Qiao, F., Wang, Y., Xie, F. et al. (2019) A DenseNet feature-based loop closure method for visual SLAM system. In: *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 258–265. Available from: https://doi.org/10.1109/ROBIO49542.2019.8961714

Yue, Y., Yang, C., Zhang, J., Wen, M., Wu, Z., Zhang, H. et al. (2020) Day and night collaborative dynamic mapping in unstructured environment based on multimodal sensors. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2981–2987. Available from: https://doi.org/10.1109/ICRA40945.2020.9197072

Zaffar, M., Ehsan, S., Stolkin, R. & Maier, K.M. (2018) Sensors, SLAM and long-term autonomy: a review. In: *2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, pp. 285–290. Available from: https://doi.org/10.1109/AHS.2018.8541483

Zeng, T. & Si, B. (2021) A brain-inspired compact cognitive mapping system. *Cognitive Neurodynamics*, 15(1), 91–101. Available from: https://doi.org/10.1007/s11571-020-09621-6.

Zhang, G., Yan, X. & Ye, Y. (2019) Loop closure detection via maximization of mutual information. *IEEE Access*, 7, 124217–124232. Available from: https://doi.org/10.1109/ACCESS.2019.2937967

Zhang, H., Chen, X., Lu, H. & Xiao, J. (2018) Distributed and collaborative monocular simultaneous localization and mapping for multi-robot systems in large-scale environments. *International Journal of Advanced Robotic Systems*, 15(3), 1–20. Available from: https://doi.org/10.1177/1729881418780178

Zhang, K., Jiang, X. & Ma, J. (2022) Appearance-based loop closure detection via locality-driven accurate motion field learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(3), 2350–2365. Available from: https://doi.org/10.1109/TITS.2021.3086822

Zhang, M., Chen, Y. & Li, M. (2019) SDF-Loc: signed distance field based 2D relocalization and map update in dynamic environments. In: *2019 American Control Conference (ACC), 2019-July*, pp. 1997–2004. Available from: https://doi.org/10.23919/acc.2019.8814347

Zhang, N., Warren, M. & Barfoot, T.D. (2018) Learning place-and-time-dependent binary descriptors for long-term visual localization. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 828–835. Available from: https://doi.org/10.1109/ICRA.2018.8460674

Zhou, W., Berrio, J.S., De Alvis, C., Shan, M., Worrall, S., Ward, J. et al. (2020) Developing and testing robust autonomy: the University of Sydney Campus data set. *IEEE Intelligent Transportation Systems Magazine*, 12(4), 23–40. Available from: https://doi.org/10.1109/MITS.2020.2990183

Zhu, J., Ai, Y., Tian, B., Cao, D. & Scherer, S. (2018) Visual place recognition in long-term and large-scale environment based on CNN feature. In: *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1679–1685. Available from: https://doi.org/10.1109/IVS.2018.8500686

Zhu, J.-Y., Park, T., Isola, P. & Efros, A.A. (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251. Available from: https://doi.org/10.1109/ICCV.2017.244

Zhu, S., Zhang, X., Guo, S., Li, J. & Liu, H. (2021) Lifelong localization in semi-dynamic environment. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14389–14395. Available from: https://doi.org/10.1109/ICRA48506.2021.9561584

**How to cite this article:** Sousa, R.B., Sobreira, H.M. & Moreira, A.P. (2023) A systematic literature review on long-term localization and mapping for mobile robots. *Journal of Field Robotics*, 1–78. https://doi.org/10.1002/rob.22170

## APPENDIX A: DATA EXTRACTION RESULTS OF THE INCLUDED RECORDS IN THE SYSTEMATIC LITERATURE REVIEW ON LONG-TERM LOCALIZATION AND MAPPING FOR MOBILE ROBOTS

Table A1

**TABLE A1** Data extraction items retrieved from the included records in the review.

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Davison and Murray (2002) | | | x | | | EKF (2D, 3DoF) | Feature (Harris corner detector) | – | | x | x | | | x | | Wheel odometry, camera (gray, stereo) | x | Manual | – | – | – | – | – | Innovation covariance, pose error |
| Filliat (2007) | x | | | | | Image classification (location) | Dictionary (BoW, location category) | – | x | x | x | | | x | | Camera (color, mono) | x | Manual | – | – | – | 1d | – | Confusion matrix, localization rate |
| Konolige and Bowman (2009) | x | | x | | | Visual odometry (3D, 6DoF), vocabulary tree (location) | Keyframe (graph, 6DoF edges) | – | | x | x | | | x | | Camera (stereo) | x | – | – | – | – | 4d | – | Execution time, localization rate, memory |
| Bosse and Zlot (2009) | x | | | | x | Point cloud matching (2D, 3DoF) | Submap (2D point cloud, graph, 3DoF edges) | – | | x | | x | | x | | Laser (2D) | x | SLAM-based | 245.-9 | – | 6.8 | 5d | – | Pose error, ROC curves |
| Biber and Duckett (2009) | | x | | | | Point cloud matching (2D, 3DoF) | Submap (2D point cloud, graph, 3DoF edges) | – | x | x | x | | | x | | Wheel odometry, laser (2D) | x | – | 9.6 | – | – | 5w | – | Average point cloud likelihood, covariance eigenvalues, memory |
| Hochdorfer and Schlegel (2009) | | | x | | | EKF (2D, 3DoF) | Feature (SURF) | – | | x | x | | | x | | Wheel odometry, camera (omni) | x | Position | 0.11-5 | – | – | – | – | Position error, #map points |
| Hochdorfer et al. (2009) | | | x | | | EKF (2D, 3DoF) | Feature (SURF) | – | | x | x | | | x | | Wheel odometry, camera (omni) | x | Position | 0.15 | – | – | – | – | Covariance eigenvalues, position error, #map points |

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Nuske et al. (2009) | x | | | | | Particle filter (2D, 3DoF) | Feature (building edges) | – | | x | | x | | x | | Wheel odometry, camera (mono) | x | Laser-based | 3.92 | – | 10.5 | 1d | – | Execution time, localization rate, pose error |
| Glover et al. (2010) | x | | | | | Visual odometry (2D, 3DoF), Bayesian (location) | Experience (graph, pose + local views, 3DoF edges) | – | x | x | | x | | x | | Camera (mono) | – | – | – | – | – | – | St Lucia 07 | Confusion matrix, precision-recall, #nodes |
| Kretzschmar et al. (2010) | | | x | | | Point cloud matching (2D, 3DoF) | Pose graph (graph, 2D point clouds, 3DoF edges) | – | | x | x | | | x | | Laser (2D) | – | – | – | – | – | – | FR079, Intel 2003 | Execution time, graph connectivity, #edges, #nodes |
| Ikeda and Kanji (2010) | | | | | x | Particle filter (location) | Dictionary (semantic hashing) | – | x | x | | x | | x | | Camera (mono) | x | GPS | 40 | 20 | – | – | – | Execution time, localization rate, memory |
| Dayoub et al. (2011) | | x | x | | | Feature matching (location) | Keyframe (graph, 6DoF edges) | – | | x | x | | | x | | Camera (omni) | x | Initial position, laser-based | – | – | – | 3d | – | Orientation error, similarity score, #localization failures |
| Pirker et al. (2011) | | | x | | | Feature matching (3D, 6DoF) | Keyframe (graph) | – | | x | x | x | | x | | Camera (gray, mono) | x | – | 1.2 | – | – | 2w | – | Position error, #map points |
| Walcott-Bryant et al. (2012) | | x | x | | | Point cloud matching (2D, 3DoF) | Pose graph (graph, 2D point clouds, | – | | x | x | | | x | | Laser (2D) | x | – | 8.4 | – | – | 5w | – | Execution time, position error, |

(Continues)

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| | | | | | | | 3DoF edges) | | | | | | | | | | | | | | | | | #edges, #nodes |
| Kretzschmar and Stachniss (2012) | | | x | | | Point cloud matching (2D, 3DoF) | Pose graph (graph, 2D point clouds, 3DoF edges) | – | | x | x | | | x | | Laser (2D) | – | – | – | – | – | – | FHW, FR079, FR101, Intel 2003 | Execution time, #edges, #nodes |
| Maddern et al. (2012) | | | x | | | Wheel odometry (2D, 3DoF), particle filter (location) | Pose graph (graph, 3DoF edges) | – | | x | | x | | x | | Wheel odometry, camera (color, mono) | – | – | – | – | – | – | New College (FAB-MAP) | Execution time, memory, precision-recall |
| Latif et al. (2012) | | | | x | | – | Pose graph (graph) | – | | x | x | x | | x | | Odometry | – | – | – | – | – | – | Bicocca (indoor), Intel 2003, New College | ATE, execution time |
| Kawewong et al. (2013) | x | | | | | Vocabulary tree (location) | Dictionary (BoW, hierarchical tree) | – | | x | x | x | | x | | Camera | – | – | – | – | – | – | City Center, New College (FAB-MAP) | Execution time, precision-recall |
| Bacca et al. (2013) | | x | x | | | Bayesian (2D, 3DoF) | Keyframe (graph) | – | | x | x | | | x | | Camera (omni), laser (2D) | x | No pruning | 1.63-5 | – | – | 1y | – | Pose error, precision-recall, #map points |
| Ball et al. (2013) | x | | | | | Visual odometry (2D, 3DoF), feature matching (location) | Experience (graph, pose + local views, 3DoF edges) | – | | x | | x | | x | | Camera (mono) | – | – | – | – | – | – | New College, St Lucia 07 | Pose error, #nodes |
| Einhorn and Gross (2013) | | x | x | | | Odometry | Pose graph (graph, 2D/3D NDT) | – | | x | x | | | x | | Camera (mono, RGBD), laser (2D) | x | – | 7 | – | 3 | 2d | – | Execution time, #nodes |

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Tipaldi et al. (2013) | | x | | | | Particle filter (2D, 3DoF) | Grid (occupancy, 2D) | – | x | x | x | | | x | | laser (2D) | x | Manual, SLAM-based | – | – | – | 1d | – | Computational complexity, localization rate, pose error |
| Huang et al. (2013) | | | x | | | Odometry, point cloud matching (2D, 3DoF) | Pose graph (graph, 2D point clouds, 3DoF edges) | – | | x | x | | | x | | Wheel odometry, laser (2D) | x | Simulation | – | – | – | – | Intel 2003, MIT Kilian Court | Pose error, #nodes |
| Johannsson et al. (2013) | | | x | | | Odometry (3D, 6DoF), BoW (location) | Keyframe (graph, 6DoF edges) | – | | x | x | | | x | | Wheel odometry, camera (stereo, RGBD), IMU | – | Manual, no pruning | – | – | – | – | MIT Stata Center | Execution time, #localization failures, #nodes |
| Oberländer et al. (2013) | | | | x | x | Fourier-Mellin transform matching (2D, 3DoF) | Submap (2D occupancy grid, graph, 3DoF edges) | – | | x | x | | | x | | Wheel odometry, laser (2D) | – | SLAM-based | – | – | – | – | Albert-b-laser-vision, FR079, Intel 2003 | Execution time, pose error, precision-recall |
| Saarinen et al. (2013) | | x | | | | – | 3D NDT, grid (occupancy, 3D) | – | | x | x | | | x | | Camera (RGBD), laser (3D) | x | – | 5 | – | 17 | – | TUM RGBD | Execution time, map similarity |
| Biswas and Veloso (2013) | x | | | | | Particle filter (2D, 3DoF) | Feature (2D line segments) | – | | x | x | | | x | | Wheel odometry, camera (RGBD), laser (2D) | – | Manual | – | – | – | – | CoBots | Pose error, #localization failures |
| Paul and Newman (2013) | | | x | | | Image classification (location) | Database (images, semantic visual topics) | – | x | x | | x | | x | | Camera (color, mono) | x | GPS, manual | 28 | – | – | – | City Center, New College (FAB-MAP) | Execution time, f-beta, precision-recall |

(Continues)

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Nguyen et al. (2013) | x | | | | | Feature matching (location) | Pose graph (graph) | – | | x | x | | | x | | Camera (color, mono) | – | – | – | – | – | – | COLD | Computational complexity, execution time, localization rate |
| Churchill and Newman (2013) | x | | | | | Visual odometry (3D, 6DoF) | Experience (graph, local views, observability edges) | – | | x | x | | | x | | Camera (color, stereo) | x | RTK-GPS | 37 | 0.7 | – | 3m | – | Execution time, #localization failures |
| Pomerleau et al. (2014) | | x | | | | Point cloud matching (3D, 3DoF) | Point cloud (laser, 3D) | – | | x | | x | | x | | Wheel odometry, laser (3D) | x | Map, targeted speed | 3.9 | 1.3 | – | 7m | – | Execution time, velocity error |
| Murphy and Sibley (2014) | | | x | | | Image classification (location) | Keyframe (graph) | – | | x | | x | | x | | Camera (color, mono) | x | – | – | – | – | 1w | New College | Execution time, confusion matrix, precision-recall, #nodes |
| Carlevaris-Bianco et al. (2014) | | | x | | | Odometry, point cloud matching (2/3D, 3/6DoF) | Pose graph (graph) | – | | x | x | x | | x | x | Camera (mono), laser (2D/3D) | – | No pruning | – | – | – | – | Intel 2003, MIT Kilian Court, NCLT | Execution time, KLD, pose error, #nodes |
| Williams et al. (2014) | | | | | x | Odometry (3D, 6DoF) | Pose graph (graph) | – | | x | | x | x | x | | Wheel odometry, camera (color, stereo), IMU | x | RTK-GPS, simulation | – | – | – | – | KITTI | Execution time, KLD |
| Einhorn and Gross (2015) | | x | x | | | Odometry | Pose graph (graph, 2D/3D NDT) | – | | x | x | | | x | | Camera (mono, RGBD), laser (2D) | x | – | 7 | – | 3 | 2d | – | Execution time, #nodes |

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Pérez et al. (2015) | x | | | | | Particle filter (3D, 6DoF) | Grid (2D, occupancy), dictionary (BoW) | – | x | x | | x | | x | | Wheel odometry, camera (gray, stereo), laser (2D) | x | SLAM-based | 11.5 | – | 3.5 | – | – | Pose error |
| Li et al. (2015) | x | | | | | Feature matching (location) | Pose graph (graph) | – | – | – | | x | | | x | Camera (gray, mono) | x | Manual | – | – | – | 3y | – | Confusion matrix |
| Mohan et al. (2015) | | | | | x | BoW (location) | Dictionary (BoW) | – | x | x | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | Bicocca (indoor), Ford Campus, Malaga 09, New College, Nordland, St Lucia 07 | Confusion matrix, execution time, precision-recall |
| Dymczyk et al. (2015) | | | x | | | Feature matching (location) | Keyframe (graph) | – | x | x | | x | | x | | Camera (mono) | x | No pruning | 1.03-4 | – | – | 10d | – | f-score, localization rate, #nodes |
| Rapp et al. (2015) | | x | | | | Particle filter (2D, 3DoF) | Grid (occupancy, 2D) | – | – | – | | x | | x | | Odometry, radar | x | – | – | – | – | – | – | Pose error |
| Vysotska et al. (2015) | x | | | | | Image sequence matching (location) | Database (images, sequence) | – | – | – | | x | | x | | Camera (color, mono) | x | Manual | 3 | – | – | – | – | Computational complexity, confusion matrix, precision-recall |
| Neubert et al. (2015) | x | | | | | Feature matching (location) | Dictionary (translation, winter - summer) | – | – | – | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | Nordland | Precision-recall |

(Continues)

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Mur-Artal et al. (2015) | | | x | | | Bundle adjustment (3D, 6DoF) | Keyframe (graph) | – | | x | | x | | x | | Camera (gray, mono) | – | – | – | – | – | – | KITTI, New College, TUM RGBD | ATE, execution time, pose error, recall, #nodes |
| Naseer et al. (2015) | x | | | | | Image sequence matching (location) | – | – | x | | | x | | x | | Camera (color, mono) | x | GPS | – | – | – | – | New College (FAB-MAP) | f-beta, precision-recall |
| Karaoğuz and Bozma (2016) | x | | | | | Feature matching (location) | Pose graph (graph, similarity edges) | – | | x | x | x | | x | | Camera (color, mono) | x | – | 0.32-5 | – | – | – | COLD, New College | Execution time, precision-recall |
| Santos et al. (2016) | | x | | | | – | Grid (occupancy, 3D) | – | | x | x | | | x | | Camera (RGBD) | x | Simulation | – | – | – | 5d | – | Environment model error, execution time |
| Dymczyk, Stumm, et al. (2016) | x | | | | | Feature matching (location) | – | – | | x | x | x | | x | | Camera (gray, mono), IMU | x | Feature labels | 4.05 | 0.15 | – | 3m | NCLT | Execution time, f-score |
| Dymczyk, Schneider, et al. (2016) | | | x | | | – | Keyframe (graph) | – | | x | x | | | x | | Camera, IMU | x | SLAM-based | – | 0.15 | – | – | – | f-score, #map points |
| Gadd and Newman (2016) | x | | x | | | Visual odometry (3D, 6DoF) | Experience (graph, local views, 6DoF edges) | x | | x | | x | | x | | Camera (grey, mono) | x | – | 100 | – | – | 1m | – | memory, #localization failures |
| Mazuran et al. (2016) | | | x | | | – | Pose graph (graph) | – | | x | x | | | x | | – | – | – | – | – | – | – | Intel 2003, MIT Kilian Court | KLD, #nodes |
| Ozog et al. (2016) | | | x | x | | Particle filter (3D, 6DoF) | Pose graph (graph, planar | – | | x | | x | | | x | Camera (gray, mono), IMU, DVL | x | Map model | 10.1-59 | | | | | – |

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| | | | | | | | segments, 6DoF edges) | | | | | | | | | | | | | | | | | |
| _ | 3- | | | | | y | – | K- | | | | | | | | LD, pose error, #nodes | | | | | | | | |
| Mühlfellner et al. (2016) | | | x | x | | Reprojection minimization (3D, 6DoF) | Keyframe (graph) | – | x | x | | x | | x | | Wheel odometry, camera (gray, mono) | x | RTK-GPS | 22 | – | – | 1y | – | Computational complexity, localization rate, pose error, #map points |
| An et al. (2016) | x | x | | | | EKF (2D, 3DoF) | Pose graph (graph) | – | | x | x | | | x | | Wheel odometry, camera (gray, mono), laser (2D) | x | Manual, simulation | 0.25-4 | – | 0.33 | – | – | Pose error, #edges, #nodes |
| Taisho and Kanji (2016) | x | | | | x | Image classification (location) | Database (images, image features) | – | – | – | | x | | x | | Camera (color, mono) | x | Manual | – | – | – | – | – | Localization rate |
| Han et al. (2017) | x | | | | | Feature matching (location) | – | – | x | x | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | CMU-VL, Nordland, St Lucia 07 | Execution time, precision-recall |
| Biswas and Veloso (2017) | | x | | | | Bayesian (2D, 3DoF) | Feature (2D line segments) | – | | x | x | x | | x | | Wheel odometry, laser (2D), camera (RGBD) | – | Manual, SLAM-based | – | – | – | – | CoBots | Pose error, #localization failures |
| Griffith and Pradalier (2017) | x | | | | | SIFT flow (3D, 6DoF) | Pose graph (graph, 6DoF edges) | – | x | | | x | | | x | Camera (color, mono), GPS, IMU | x | Manual | 100 | – | – | 1y2m | – | Absolute alignment error |

(Continues)

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Naseer et al. (2017) | x | | | | x | Feature matching (location) | – | – | | x | | x | | x | | Camera (color, mono) | x | Manual | 100 | – | – | 3y | – | f-score, precision-recall |
| Krajník et al. (2017) | | x | | | | – | Grid (occupancy, 3D) | – | | x | x | x | | x | | Camera (RGBD) | x | External tracking system | – | – | – | 112d | NCLT, Witham Wharf RGB-D | Computational complexity, memory, pose error |
| Ila et al. (2017) | | | x | | | Odometry (3D, 6DoF) | Pose graph (graph) | – | | x | x | x | | x | | – | – | Simulation | – | – | – | – | KITTI | Pose error, #nodes |
| Latif et al. (2017) | x | | | | x | Dictionary search (location) | Dictionary (incremental) | – | | x | | x | | x | | Camera (mono) | – | – | – | – | – | – | Bicocca (indoor), KITTI, New College | Confusion matrix, execution time, precision-recall |
| Xin et al. (2017) | x | | | | x | Feature matching (location) | – | – | – | – | | x | | x | | Camera (mono) | – | – | – | – | – | – | CMU-VL, Gardens Point Campus | Computational complexity, f-score, precision-recall |
| Bescos et al. (2018) | | x | | | | Bundle adjustment (3D, 6DoF) | Keyframe (graph) | – | x | | x | | | x | | Camera (color, mono, stereo, RGBD) | – | – | – | – | – | – | KITTI, TUM RGBD | ATE, execution time, pose error |
| Opdenbosch et al. (2018) | | | x | | | – | Keyframe (graph, Hamming distance edges) | – | – | – | x | x | x | | | Camera (mono, stereo, RGBD) | – | – | – | – | – | – | EuRoC | Memory |
| Han, Wang, et al. (2018) | x | | | | | Image sequence matching (location) | – | – | – | – | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | CMU-VL, Nordland, St Lucia 07 | Precision-recall |
| Han, Beleidy, et al. (2018) | x | | | | | Feature matching (location) | – | – | – | – | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | CMU-VL, Nordland | Precision-recall |

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Cao et al. (2018) | x | | | | | Odometry (3D, 6DoF), vocabulary tree (location) | Pose graph (graph, 2D point clouds, 3DoF edges) | – | | x | | x | | x | | Wheel odometry, laser (2D, 3D), IMU | x | – | – | – | – | 1d | – | Execution time, precision-recall |
| Nobre et al. (2018) | x | | | | | Mahalanobis distance minimization (2D, 3DoF) | Feature | – | | x | x | | | x | | Wheel odometry, camera (color, mono) | x | Simulation | – | – | – | – | UTIAS Multi-Robot | Precision-recall |
| Zhang, Chen, et al. (2018) | | | x | | | Feature matching (3D, 6DoF) | Keyframe (graph) | x | | x | x | x | | x | | Camera (mono) | x | – | – | – | – | – | KITTI | Localization rate |
| Zhu et al. (2018) | x | | | | | Image sequence matching (location) | – | – | x | | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | City Center, Nordland | Precision-recall |
| MacTavish et al. (2018) | x | | | | | visual odometry (3D, 6DoF) | Keyframe (graph) | – | | x | | x | | x | | Camera (stereo) | x | External tracking system | 26.0-8 | 0.16 | – | 4m | – | Execution time, localization rate, pose covariance, pose error |
| Sun et al. (2018) | | x | | | | Point cloud matching (3D, 6DoF) | Grid (occupancy, 3D) | – | x | x | | x | | x | | Laser (3D) | x | Manual | – | – | – | 2w | KITTI | map accuracy |
| Lázaro et al. (2018) | | x | x | x | | Point cloud matching (2D, 3DoF) | Submap (2D point cloud, graph, 3DoF edges) | – | | x | x | | | x | | Wheel odometry, laser (2D) | – | – | – | – | – | – | MIT Stata Center, Witham Wharf RGB-D | Execution time, #edges, #nodes |
| Zhang, Warren, et al. (2018) | x | | | | | Visual odometry (3D, 6DoF) | Keyframe (graph) | – | | x | | x | | x | | Camera (stereo) | x | – | 25 | 0.25 | – | 4m | – | Matching accuracy |

(Continues)

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Chebrolu et al. (2018) | x | | | | | Feature matching (location) | – | – | – | – | | x | x | | | Camera (color, mono), GPS | x | Manual | – | – | – | 1m | – | Matching accuracy |
| Yin et al. (2018) | x | | | | | Feature matching (location) | Grid (occupancy, 3D) | – | – | – | | x | | x | | Laser (3D) | – | – | – | – | – | – | KITTI, NCLT | Precision-recall |
| Egger et al. (2018) | x | | x | | | Feature matching (3D, 6DoF), odometry (3D, 6DoF) | Submap (surfel, graph) | – | | x | | x | | x | | Wheel odometry, laser (3D), IMU | x | RTK-GPS | – | – | – | 1y6m | – | Execution time, memory |
| Arroyo et al. (2018) | x | | | | | Image sequence matching (location) | – | – | – | – | | x | | x | | Camera (mono, stereo) | – | – | – | – | – | – | KITTI, New College, Nordland | Confusion matrix, execution time, precision-recall |
| Ouerghi et al. (2018) | x | | | | | Image sequence matching (location), visual odometry (3D, 2DoF), EKF (2D, 3DoF) | Keyframe (graph) | – | | x | | x | | x | | Wheel odometry, camera (mono) | – | – | – | – | – | – | KITTI | Execution time, pose error |
| Siva and Zhang (2018) | x | | | | | Feature matching (location) | – | – | – | – | | x | | x | | Camera (omni, RGBD) | x | GPS | 19.1-5 | – | – | 1y | – | Precision-recall |
| Luthardt et al. (2018) | x | | x | | | Visual odometry (3D, 6DoF) | Pose graph (graph) | – | x | x | | x | | x | | Camera (gray, mono) | x | GPS | – | – | – | – | – | Pose error |
| Chen et al. (2018) | x | | | | | Image classification (location) | – | – | – | – | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | Nordland, St Lucia 07 | Precision-recall |

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Yu et al. (2019) | x | | | | x | Vocabulary hashing (location) | – | – | – | – | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | City Center, New College (FAB-MAP) | Confusion matrix, precision-recall |
| Boniardi et al. (2019) | | x | x | | | Point cloud matching (2D, 3DoF) | Pose graph (graph, 2D point clouds, 3DoF edges) | – | | x | x | | | x | | Laser (2D) | x | External tracking system, map model | 4.65-7 | – | 2.85 | 4d | – | Execution time, pose error, #nodes |
| Kim et al. (2019) | x | | | | | Feature matching (location) | Grid (location, 2D) | – | – | – | | x | | x | | Laser (3D) | – | – | – | – | – | – | NCLT, Oxford RobotCar | Precision-recall |
| Berrio et al. (2019) | x | | x | | | Point cloud matching (2D, 3DoF) | Feature (pole, corners) | – | x | | | x | | x | | Laser (3D) | x | Manual | – | 0.5 | – | 6m | – | Pose covarianc-e, #map points |
| Wang et al. (2019) | | x | | | | Bundle adjustment (3D, 6DoF) | Keyframe (graph) | – | – | – | x | x | | x | | Camera (RGBD) | x | Simulation | – | – | – | – | TUM RGBD | ATE, precision |
| Wu and Wu (2019) | | | | | x | Image classifica-tion (location) | – | – | | x | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | Gardens Point Campus, Nordland | Execution time, f-score, precision-recall |
| Tang et al. (2019) | x | | x | | | BoW (location), point cloud matching (3D, 6DoF) | Submap (graph, manifold) | – | | x | | x | | x | | Camera (color, stereo) | – | SLAM-based | – | – | – | – | YQ21 | Execution time, localiza-tion rate, pose error, #nodes |
| Bürki et al. (2019) | x | | | x | | Feature matching (3D, 6DoF) | Keyframe (graph) | – | x | x | | x | | x | | Wheel odometry, sensor (gray, mono) | x | – | – | – | – | 1y | NCLT | Execution time, pose error |

(Continues)

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Labbé and Michaud (2019) | | | | x | x | BoW (location), odometry (2/3D, 3/6DoF) | Pose graph (graph) | – | | x | x | x | x | x | | Wheel odometry, camera (stereo, RGBD), laser (2D/3D) | – | – | – | – | – | – | EuRoC, KITTI, MIT Stata Center, TUM RGBD | ATE, execution time, #nodes |
| Zhang, Chen, et al. (2019) | | x | | | | Point cloud matching (2D, 3DoF) | Grid (signed distance field, 2D) | – | | x | x | | | x | | Laser (2D) | x | SLAM-based | – | – | – | – | – | Execution time, pose error |
| Schmuck and Chli (2019) | | | x | | | Odometry (3D, 6DoF) | Keyframe (graph) | – | | x | x | | x | | | Camera, IMU | – | – | – | – | – | – | EuRoC | Pose error |
| Ganti and Waslander (2019) | | x | | | | Bundle adjustment (3D, 6DoF) | Keyframe (graph) | – | – | – | | x | | x | | Camera (stereo) | – | – | – | – | – | – | KITTI | Pose error |
| Ding et al. (2019) | | | x | | x | EKF (3D, 6DoF) | Keyframe (graph) | – | | x | | x | | x | | Camera (stereo), IMU | x | Laser-based | – | – | 1.32 | 1y | – | ATE, communication constraints |
| Song et al. (2019) | | x | | | | Odometry (3D, 6DoF) | Keyframe (graph) | – | | x | | x | | x | | Camera, IMU | x | RTK-GPS | – | – | – | – | – | pose error |
| Pan et al. (2019) | | x | | | | Odometry, reprojection minimization (3D, 6DoF) | Feature (point clusters) | – | x | x | | x | | x | | Camera (mono), laser (3D) | x | – | – | – | – | 3m | KITTI | Execution time, localization rate, pose error, reprojection error |
| Ali et al. (2020) | | | | | x | Visual odometry (3D, 6DoF) | Keyframe (graph) | – | | x | x | | | x | | Camera (RGBD) | x | Laser-based | – | – | 0.5 | – | TUM RGBD | Communication constraints, execution time, |

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| | | | | | | | | | | | | | | | | | | | | | | | | localization rate, memory, pose error |
| C. Qin et al. (2020) | x | | | | | Feature matching (location) | – | – | | x | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | Alderley, FAS, Nordland, Oxford RobotCar, St Lucia 07 | Confusion matrix, execution time, precision-recall |
| Martini et al. (2020) | x | | | | | Feature matching (location), point cloud matching (2D, 3DoF) | Experience (graph) | – | x | x | | x | | x | | Radar | – | – | – | – | – | – | Oxford Radar RobotCar | Confusion matrix, localization rate, pose error, precision-recall |
| Karaoğuz and Bozma (2020) | | | x | | | – | Pose graph (graph) | x | | x | | x | | x | | Camera (mono) | x | – | – | – | – | – | COLD | Computational complexity, precision-recall |
| Yin et al. (2020) | x | | | | | Particle filter (2D, 3DoF), point cloud matching (3D, 6DoF) | Pose graph (graph, 6DoF edges) | – | | x | | x | | x | | Laser (3D) | – | – | – | – | – | – | KITTI, YQ21 | Execution time, f-score, pose error, precision-recall |
| Clement et al. (2020) | x | | | | | Feature matching (3D, 6DoF) | – | – | – | – | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | Oxford RobotCar | Confusion matrix, matching accuracy |
| Wang et al. (2020) | | x | | | | Particle filter (2D, 3DoF) | Grid (occupancy, 2D) | – | | x | x | | | x | | Wheel odometry, laser (2D) | x | Simulation, SLAM-based | – | – | – | – | – | Execution time, memory, pose error |

(Continues)

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Camara et al. (2020) | x | | | | x | Feature matching (location) | – | – | | x | | x | | x | | Camera | – | – | – | – | – | – | Berlin Kudamm, Gardens Point Campus, Nordland | Execution time, memory, precision-recall |
| Gao and Zhang (2020) | x | | | | | Feature matching (location) | – | – | x | | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | CMU-VL, St Lucia 07 | Precision-recall |
| Yang et al. (2020) | | x | | | | Visual odometry (3D, 6DoF) | Keyframe (graph) | – | | x | x | | | x | | Camera (RGBD) | – | – | – | – | – | – | TUM RGBD | ATE, execution time, pose error |
| Siva et al. (2020) | x | | | | | Feature matching (location) | – | – | – | – | | x | | x | | Laser (3D) | x | Simulation | – | – | – | – | NCLT | Precision-recall |
| T. Qin et al. (2020) | x | | | | | EKF (2D, 3DoF) | Feature (semantic) | – | | x | x | x | | x | | Wheel odometry, camera (color, mono), IMU | x | RTK-GPS | 0.32-4 | – | – | 1m | – | ATE, memory, recall |
| Ding et al. (2020) | x | x | | | | Bundle adjustment (3D, 6DoF) | Point cloud (3D) | – | x | x | | x | | x | | Camera (color, stereo), laser (3D) | – | – | – | – | – | – | KITTI, YQ21 | ATE, execution time, pose error |
| Yue et al. (2020) | | x | | | | – | Point cloud (3D) | x | | x | | x | | x | | Camera (color, mono, thermal), laser (3D) | x | – | – | – | – | – | – | ATE, memory |
| Schaefer et al. (2021) | x | | x | | | Particle filter (2D, 3DoF) | Feature (poles) | – | x | x | | x | | x | | Laser (3D) | – | – | – | – | – | – | KITTI, NCLT | Pose error |
| Liu et al. (2021) | x | | | | | EKF (3D, 6DoF), feature matching (location) | Keyframe (graph) | – | | x | x | x | | x | | Camera (color, mono), IMU | – | – | – | – | – | – | City Center, KITTI, New College (FAB-MAP) | Execution time, memory, precision-recall |

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Kim et al. (2021) | | | | | x | Particle filter, point cloud matching (3D, 6DoF) | Grid (geodetic, NDT) | – | | x | | x | | x | | Laser (3D) | x | RTK-GPS, SLAM-based | – | – | – | – | KITTI | Memory, pose error |
| Derner et al. (2021) | x | | | | | Feature matching (3D, 6DoF) | Database (images, features, pose) | – | x | x | x | | | x | | Wheel odometry, camera (RGBD) | x | Manual | 0.19-8 | – | – | – | Witham Wharf RGB-D | Execution time, localization rate, pose error |
| Cao et al. (2021) | x | | | | | Sequence matching (location) | – | – | | x | | x | | x | | Laser (2D/3D) | – | – | – | – | – | – | NCLT, Oxford RobotCar | Execution time, precision-recall |
| G. Singh et al. (2021) | x | x | | | | Feature matching (location) | Pose graph (graph, BoW) | – | | x | | x | | x | | Camera (stereo, RGBD) | – | – | – | – | – | – | CBD, KITTI | Execution time, precision-recall |
| Kurz et al. (2021) | | | x | | | – | Pose graph (graph) | – | | x | x | | | x | | Wheel odometry, laser (2D), IMU | – | No pruning | – | – | – | – | MIT Stata Center, Witham Wharf RGB-D | Execution time, pose error, #nodes |
| Yin, Xu, Wang, et al. (2021) | x | | | | | Location matching (location) | – | – | – | – | | x | | x | | Laser (3D), radar | – | – | – | – | – | – | MulRan, Oxford Radar RobotCar | Confusion matrix, precision-recall |
| Thomas et al. (2021) | | x | | | | Point cloud matching (3D, 6DoF) | Grid (occupancy, 3D) | – | x | x | x | | | x | | Wheel odometry, laser (3D) | x | Simulation | – | – | – | – | – | Confusion matrix, execution time, precision-recall |
| Berrio et al. (2021) | x | | x | | | – | Grid (feature, 2D) | – | x | | | x | | x | | Wheel odometry, camera (color, mono), | – | – | – | – | – | – | USyd Campus | Pose covariance |

(Continues)

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| | | | | | | | | | | | | | | | | laser (3D), IMU | | | | | | | | |
| Oh and Eoh (2021) | x | | | | | Feature matching (location) | – | – | – | – | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | KAIST, Nordland | Precision-recall |
| Tsintotas et al. (2021) | | | x | | | BoW (location) | Dictionary (BoTW, incremental) | – | | x | x | x | x | x | | Camera (mono) | – | – | – | – | – | – | City Center, EuRoC, KITTI, Lip6Ind, Lip6Out, Malaga 09 | Execution time, memory, precision-recall |
| Sun et al. (2021) | x | | | | | Feature matching (3D, 6DoF) | Keyframe (graph) | – | | x | | x | | x | | Camera (color, mono) | x | SLAM-based | 0.74-1 | – | – | 1d | – | ATE, execution time, matching error, pose error |
| Tang et al. (2021) | x | | | | | Feature matching (location) | – | – | – | – | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | Alderley, Nordland, Oxford RobotCar, YQ21 | Localization rate, precision-recall |
| Piasco et al. (2021) | x | | | | x | Feature matching (location) | – | – | x | x | | x | | x | | Camera (RGBD) | – | – | – | – | – | – | CMU-VL, Oxford RobotCar | Precision-recall |
| Yin, Xu, Zhang, et al. (2021) | x | | | | | Feature matching, sequence matching (location) | – | – | | x | | x | | x | | Laser (3D) | x | – | 132 | 11 | – | – | KITTI, NCLT | Execution time, memory, precision-recall |
| Meng et al. (2021) | x | | | | | Laser odometry (3D, 6DoF) | Pose graph (graph) | – | | x | | x | | x | | Laser (3D) | – | – | – | – | – | – | KITTI | ATE, execution time, pose error |
| Zhu et al. (2021) | | x | | | | Particle filter (2D, 3DoF) | Grid (occupancy, 2D) | – | x | x | x | | | x | | Wheel odometry, camera | x | Manual | – | – | – | – | – | Pose error |

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| | | | | | | | | | | | | | | | | (color, mono), laser (2D), IMU | | | | | | | | |
| Zeng and Si (2021) | | | x | | | – | Pose graph (graph) | – | | x | x | | | x | | Wheel odometry, camera (color, mono) | x | No pruning | – | – | – | – | – | #edges, #nodes |
| Ali et al. (2021) | | | x | | | Point cloud matching, visual odometry (3D, 6DoF) | Keyframe (graph), submap (image, graph) | – | | x | x | x | | x | | Camera, laser (3D) | x | – | – | – | – | – | KITTI | CPU usage, memory, pose error, precision-recall |
| Xu et al. (2021) | x | | | | | Feature matching (location) | – | – | | x | | x | | x | | Laser (3D) | – | – | – | – | – | – | MulRan, NCLT, Oxford RobotCar | Execution time, precision-recall |
| Yang et al. (2021) | x | | | | x | BoW, feature matching (location) | – | – | | x | | x | | x | | Camera (color, mono) | – | – | – | – | – | – | City Center, KITTI, Lip6Ind, Lip6Out, Malaga 09, New College | Confusion matrix, execution time, precision-recall |
| Wang et al. (2021) | x | | x | | | Point cloud matching (3D, 6DoF) | Feature (poles) | – | – | – | | x | | x | | Laser (3D) | x | GPS | 5.52 | – | – | 1m | – | Localization rate, pose error |
| Hu et al. (2022) | x | | | | | Feature matching (location) | – | – | – | – | | x | | x | | Camera (color, mono) | x | RTK-GPS | – | – | – | – | CMU-Seasons, RobotCar Seasons | Execution time, precision-recall |
| Coulin et al. (2022) | x | | | | | EKF (3D, 6DoF) | Magnetic (vector) | – | x | x | x | | | x | | Camera (stereo), IMU | x | SLAM-based | 1.66-5 | – | – | 1y | – | ATE, execution time |

(Continues)

**TABLE A1** (Continued)

| DE: | 1: | | | | | 2: | 3: | 4: | 5: | | 6: | | | | | 7: | 8: | 9: | 10: | | | | 11: | 12: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref. | Appearance | Dynamic | Sparsity | Multisession | Compute | Localization | Mapping | Multirob. | Offline | Online | Indoor | Outdoor | Air | Ground | Water | Sensor | Self-acq. | Ground-truth | Dist.(km) | Path(km) | Time(h) | Int.(d\|w\|m\|y) | Datasets | Metrics |
| Zhang et al. (2022) | x | | | | | Feature matching (location) | – | – | | x | | x | | x | | Camera (color, mono) | – | Manual | – | – | – | – | City Center, KITTI, Malaga 09, St Lucia 07 | Execution time, precision-recall |
| Nguyen et al. (2022) | x | | | | | Bundle adjustment, sensor fusion (3D, 6DoF) | – | – | | x | x | | x | | | Camera (mono), laser (3D), IMU, UWB | – | – | – | – | – | – | EuRoC, NTU VIRAL | Execution time, pose error |
| Bouaziz et al. (2022) | | | x | | | Feature matching (3D, 6DoF) | Keyframe (graph) | – | x | x | | x | | x | | Camera (gray, mono) | – | – | – | – | – | – | IPLT, Oxford RobotCar | Execution time, memory, #localization failures |
| Du et al. (2022) | | x | | | | Reprojection minimization (3D, 6DoF) | Keyframe (graph) | – | | x | x | | | x | | Camera (RGBD) | – | – | – | – | – | – | Bonn RGB-D Dynamic, TUM RGBD | ATE, execution time, pose error |
| Xing et al. (2022) | | x | | | | Feature matching (3D, 6DoF) | Keyframe (graph) | – | | x | x | x | x | x | | Camera (RGBD), IMU | x | – | – | – | – | – | EuRoC, KITTI, TUM RGBD | Execution time, localization rate, pose error |
| Hong et al. (2022) | x | | | | | Feature matching (location), point cloud matching (2D, 3DoF) | Keyframe (graph) | – | | x | | x | | x | | Radar | – | – | – | – | – | – | MulRan, Oxford RobotCar, RADIATE | ATE, execution time, pose error |