A Systematic Literature Review on Long-Term Localization and Mapping for Mobile Robots*

Ricardo B. Sousa Héber M. Sobreira António Paulo Moreira

September 12, 2022

Abstract

Keywords: simultaneous localization and mapping (SLAM), lifelong SLAM, long-term autonomy, mobile robots.

1 Introduction

2 Purpose of the study

2.1 Limitations of current studies

Table 1: Existent Literature Reviews, Surveys, and Tutorials on SLAM.

Topic	Reference
Probabilistic approaches and data association	Bailey and Durrant-Whyte 2006; Durrant-Whyte and Bailey 2006
SLAM back end	Grisetti et al. 2010
Multi-robot SLAM	Saeedi et al. 2016
Visual odometry	Fraundorfer and Scaramuzza 2012; Scaramuzza and Fraundorfer 2011
Overview of challenges in SLAM	Cadena et al. 2016
Trends in SLAM for autonomous vehicles	Bresson et al. 2017
Completar tabela!	

2.2 Motivations and goals

Research question: What is the current state of the art of long-term localization and mapping using mobile robots?

Goals of this review:

- which are the main strategies for accomplishing long-term operations with mobile robots;
- how to deal with varying conditions of the environment;
- how do autonomous robots deal with the dynamics of the environment;
- which are the main strategies to deal with the limited computational resources of a mobile robot on long-term operations;
- how the methods evaluate their results;
- which are the public datasets more used for evaluating long-term localization and mapping.

PICO framework (Population–Intervention–Comparison–Outcome) helps to frame the research questions of this systematic review into a more structured framework:

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

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 Preferred Reporting Items for Systematic reviews and Meta-Analysis (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. 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 decide which studies to include in the review. Next, the search strategy details the information sources considered in the review and the base string and search fields used for inquiring these sources. Furthermore, the selection process focuses on describing its stages and the quality evaluation criteria used to select works for the synthesis and analysis phase of the review. Lastly, the data extraction process details the relevant 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 a public GitHub repository[1].

*This work is financed by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia, within scholarship 2021.04591.BD.

Ricardo B. Sousa[1, 2]
up201503004@edu.fe.up.pt

Héber M. Sobreira[2]
heber.m.sobreira@inesctec.pt

António Paulo Moreira[1, 2]
amoreira@fe.up.pt

[1] Faculty of Engineering of the University of Porto, Electrical Engineering Department, Porto, 4200-465 Porto, Portugal
[2] INESC TEC – Institute for Systems and Computer Engineering, Technology and Science, CRIIS – Centre for Robotics in Industry and Intelligent Systems, Porto, 4200-465 Porto, Portugal

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

3.1 Eligibility criteria

Table 2 presents the exclusion criteria used to determine the eligible studies for the selection process. These eligibility criteria focus mainly 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 simultaneously guaranteeing peer-revision of the works. In terms of language, only considering studies with English full-texts increases the scope and visibility of the review. Similarly, the eligibility criteria reject studies not available in digital libraries for reproducibility and accessibility reasons.

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	Dataset	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

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 artificial intelligence, 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 dataset criterion rejects all works that focus only on sharing a data collection. Although these works are important for the evolution of localization and mapping algorithms in providing a benchmark for comparison and reference purposes, their scientific contribution is not directly comparable to research articles. 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 does not consider eligible for selection papers 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 for this literature review, and defining the base string and which search fields considered to obtain the results for the 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 and Paul-Hus 2016; V. K. Singh et al. 2021). Thus, the data sources 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 theirs initial one. As for inquiring 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 to the desired population. These 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 for 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, presented in Section 2. The reason for having both `"life long"` and `lifelong` terms is the existing confusion in which term is grammatically the correct one.

Furthermore, the Title, Abstract, and Keywords are the fields considered for obtaining the search results. The third one includes the author keywords, the indexed terms by the databases, and the uncontrolled ones if they are available. The selection of these search fields for this review improves the relevance of the results compared to using all fields and the full text by focusing 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 the author keywords. The indexed terms also help in obtaining records only related to the base string used

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

in this review. However, not all data sources have available the search fields considered in the review or some of them require an adaptation when performing the search. Although the *ACM Digital Library* allows searching within multiple search fields, including the ones considered in this review, the advanced search query on this library sets by default an AND operator between the different fields. This 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 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. Given the limitation of *IEEE Xplore* to 7 wildcards, the search results of this digital library using the base string for the inquiry are the grouping of different searches considering only a search field at a time, importing each search results to Parsifal and removing 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 to improve the discussion. Indeed, to best of the authors knowledge, there is not available a systematic review on long-term localization and mapping for mobile robots 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.

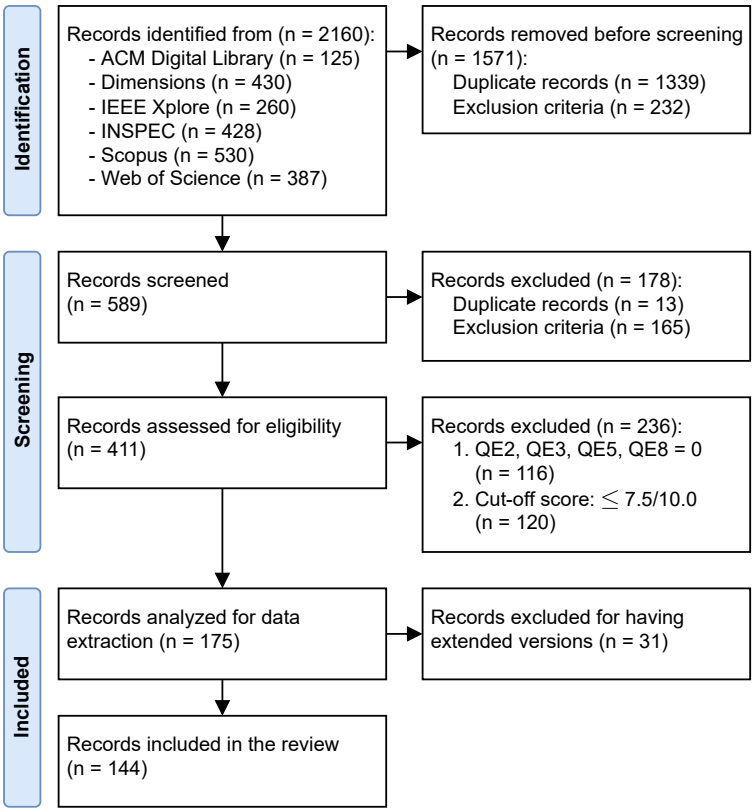


Figure 1: PRISMA flow diagram for the selection process.

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 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*, *Dimensions*, *INSPEC*, *Scopus*, and *Web of Science* data sources only require a single inquiry to obtain the search results. Given the limitation of the *IEEE Xplore* for using wildcards mentioned in Section 3.2, the number of records for this source presented in Figure 1 represents the results of 7 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 exclusion is possible due to *INSPEC*, *Scopus*, or *Web of Science* 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 of this review searches studies with a number of pages lower or equal to 4. Even though short papers have a maximum number of 3 pages, the papers with 4 pages do not usually present a detailed methodology. As for the E7 exclusion criterion, some works are possible to remove from the review by searching in their title for the term "dataset". All excluded articles of this review are double-checked to certify if the exclusion criteria are correctly applied. 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 in 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. The analysis of the results and conclusions of these publications considering the exclusion criteria either confirms the exclusion decision or reverses it to eligible works for quality assessment. As a result of the screening phase, 178 studies are rejected from the initial 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* database.

3.3.3 Quality assessment

The quality evaluation in this review of the selected works from screening 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, in which the minimum, intermediate, and maximum values correspond to none, partial, and full compliance, respectively. Furthermore, QE1, QE2, QE4, and QE8 focus on the details provided in the papers, specifically, if the discussion of the related work, the proposed methodology, the experimental setup, and the results are detailed and thoroughly analyzed in the publication, respectively. The possible scores for QE3 are twice the value of QE1, QE2, QE4, and QE8 due to this criterion being directly related to the topic of the 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 of them, the scores will be 1.0 or 0.0, i.e., partial or no compliance, respectively. QE5 evaluates the long-term results of the eligible studies and is either 2.0 (full) or 0.0 (no compliance). This criterion 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. The former criterion intends to highlight works that compare themselves to the state of the art and/or ground-truth data. The latter emphasizes the importance of having available either the implementation of the proposed methodology or the data used in the experiments for other works to be able to compare 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.

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}

After evaluating the 411 eligible works accordingly to the previously discussed QE criteria (the scores of each record are available in the GitHub repository), the first conclusion of the authors is that works with a non-detailed 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, given 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:

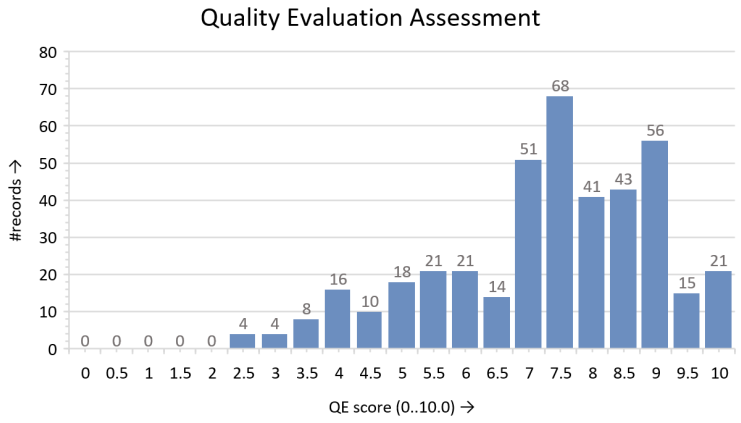
1. QE2, QE3, QE5, QE8: reject works with a 0.0 (no compliance) score;
2. cut-off score: reject works with a score lower or equal to 7.5/10.0.

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 Figures 2b and 2a, 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, while 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:

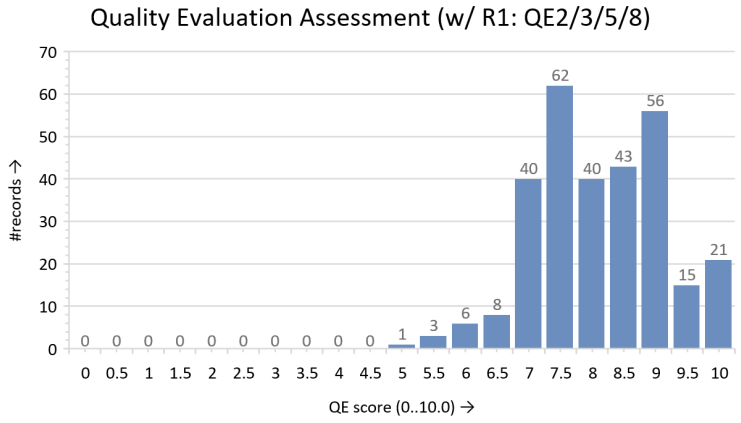
- 7.5/10.0 — 120 records (40.7%) — 175 records
- 8.0/10.0 — reject → 160 records (54.2%) — include → 135 records
- 8.5/10.0 — 203 records (68.8%) — 92 records

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, or, 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 rejection criteria previously mentioned leads to rejecting a total of 236 (57%) records. As a result, the remaining 175 records will be analyzed for data extraction.



(a)



(b)

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).

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, the authors 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;
  - multi-session: 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**] **Multi-robot** – 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;
- [**DE8**] **Non-public experiments** – if the authors performed experiments or tests with non-public data;
- [**DE9**] **Ground-truth** – how ground-truth for non-public data is obtained or its type, if available;
- [**DE10**] **Distance and time characteristics** – relative to the non-public experiments if available, as follows:
  - total distance (km) of the non-public 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.
- [**DE11**] **Datasets** – if and which public datasets are used in the experiments;
- [**DE12**] **Evaluation metrics** – which metrics are used for evaluation.

In Section 5, a comparison table of the public datasets identified by the DE11 will contain the sensory setup, ground-truth data availability from the datasets, and the distance and time characteristics, similar to the data extraction items for non-public data, among other aspects. As a result, the distinction between public and non-public data availability represented in DE8, DE9, and DE10 allows to understand the distance and time characteristics of non-public data independently from the public datasets.

Although the data extraction phase in a systematic literature review usually does not remove any records, 31 of the analyzed 179 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, including their bibliographic references.

Consequently, 144 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 **??** and also in the repository. The included works represent 35% 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 not in terms of their scientific contribution but in terms of their bibliographic data for presenting an overview of the included records in the review. First, statistic results of the data sources in which the 144 included records could be identified in the methodology allow the evaluation of the coverage between the sources. Next, the tool VOSviewer (van Eck and Waltman 2010, 2014) is used to obtain the co-occurrence analysis for the keywords and the authors. The former focus on the keywords recency and their occurrence in the sources, while the latter discusses the research networks between the authors, and the ones with more publications in long-term localization and mapping. Lastly, two analysis are presented relative to the evolution of the 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`[3] Python library, the BibTeX files are processed to identify uncompleted 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 144 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 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 144 included ones in the review are the following ones:

- *ACM Digital Library*: 25 records (17.4%);
- *Dimensions*: 85 records (59.0%);
- *IEEE Xplore*: 68 records (47.2%);
- *INSPEC*: 104 records (72.2%);
- *Scopus*: 122 records (84.7%);
- *Web of Science*: 105 records (72.9%).

The database *Scopus* is the source that identified the greatest number of included records. This result was expected given that *Scopus* is considered as one of the largest curated databases (V. K. Singh et al. 2021), indexing more than 25000 active titles (e.g., conferences proceedings, journals) and 7000 publishers[4]. Two other sources with more than 70% of identified records are *INSPEC* and *Web of Science*. Similarly to *Scopus*, these two databases index also records from thousands of journals, conferences, and publishers[5,6]. Although *Dimensions* 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*, *Scopus*, and *Web of Science*) and could be a factor to why it obtained a lower percentage (59.0%) than the other three databases. Another possible reason is that *Scopus* and *Web of Science* 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* has better coverage in Social Sciences and Arts & Humanities (V. K. Singh et al. 2021). Even though *IEEE Xplore* 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, subject areas related to long-term localization and mapping[7]. Finally, the *ACM Digital Library* using *The ACM Guide to Computing Literature* collection only finds published records by ACM and possible links to other records focused exclusively on computing[8] and not directly related to the Computer Science or Engineering subject areas, explaining why this source obtained a lower coverage percentage of the included results than the other sources for this review.

Furthermore, Table 4 presents a coverage analysis of the identified results from each data source for the 144 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 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, i.e., 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 144 included records: $\#\{A \cup B\}/144$, 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 144 included records.

Table 4: Pairwise coverage analysis of the data sources considered in the review over the 144 included records: (a) identification only on both pairwise sources ($\#\{A \cap B\}/\min\{\#A, \#B\}$); (b) on either ones ($\#\{A \cup B\}/\#$records). Legend: dim – *Dimensions*, ieee – *IEEE Xplore*, insp – *INSPEC*, scop – *Scopus*, wos – *Web of Science*.

(a)

$A \cap B$	**acm**	**dim**	**ieee**	**insp**	**scop**	**wos**
acm	–	96.0%	44.0%	88.0%	96.0%	96.0%
dim	–	–	69.1%	77.6%	97.6%	95.3%
ieee	–	–	–	89.7%	91.2%	73.5%
insp	–	–	–	–	87.5%	68.3%
scop	–	–	–	–	–	89.5%
wos	–	–	–	–	–	–

(b)

$A \cup B$	**acm**	**dim**	**ieee**	**insp**	**scop**	**wos**
acm	–	59.7%	56.9%	74.3%	85.4%	73.6%
dim	–	–	73.6%	85.4%	86.1%	75.7%
ieee	–	–	–	77.1%	88.9%	85.4%
insp	–	–	–	–	93.8%	95.8%
scop	–	–	–	–	–	92.4%
wos	–	–	–	–	–	-

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 discussed in Section 3 to consider several data sources in the identification phase of a review. Moreover, the pairwise union coverage of *INSPEC*, *Scopus*, and *Web of Science* is greater than 90% of the included records. When evaluating the joint coverage of these three databases, they identify all 144 of the included records, i.e., 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* with the other sources, which is greater than 85%. This overlap indicates that *Scopus* covers results not only on the topic of this review but also the results obtained by the other sources considered in the methodology. Lastly, *IN-*

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

SPEC and *Web of Science* achieve a pairwise overlap percentage of 68.3% 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* identifies 33/144 records not found by *Web of Science*, and vice-versa for *Web of Science*, with 34/144 unique records.

4.2 Keywords co-occurrence

Next, VOSviewer (van Eck and 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[9] converts this processed BibTeX to a RIS file. Even though VOSviewer supports file types directly exported from *Dimensions*, *Scopus*, or *Web of Science* as input, none of these data sources obtained all the 144 included records of the review in the identification phase. Given that VOSviewer does not support BibTeX files, the conversion to RIS file is required for using as 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*, *Scopus*, or *Web of Science* 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 144 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 35 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, respectively, distances the terms more from each other than using the values recommended in the VOSviewer manual[10] (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) is used to join similar terms: spelling differences (e.g., localization – localisation), full terms versus abbreviations (simultaneous localization and mapping – SLAM), while also allowing the concatenation of long keywords for visualization reasons.

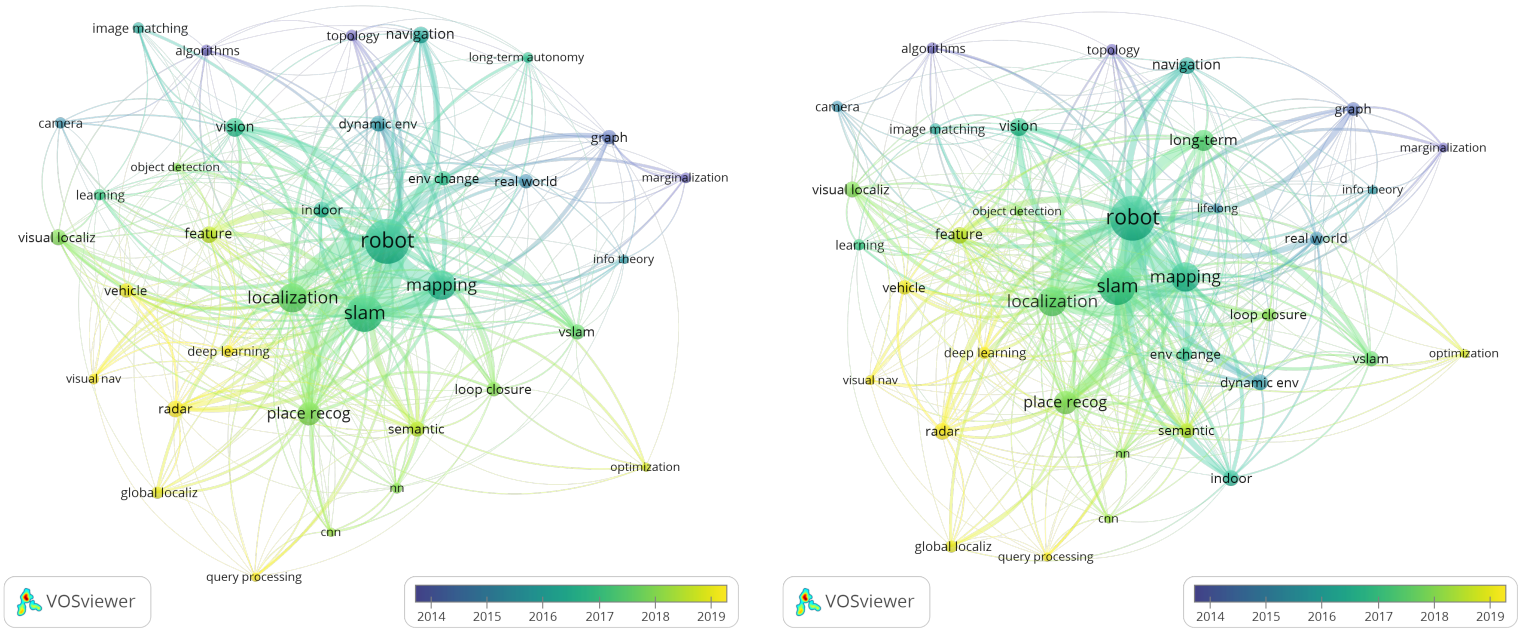
Overall, the keyword `robot` is the one that appears more times in the included records: 111 occurrences, links with 34 other terms, and has a total link strength of 403 (sum of co-occurrences of all of its links). This result is expected due to the relation of this review's topic to robotics. Similarly, 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` (75, 34, and 288), `mapping` (48, 33, and 204), and `localization` (47, 32, and 194, 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 the all the phases executed in this review's methodology, the 144 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, `long-term 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 `long-term autonomy` is not usually used by the authors nor indexed by the databases. However, the specific term of `long-term 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 `long-term` and `lifelong` into the terms themselves, obtaining 36 keywords with a minimum of 5 occurrences in the 144 included records. In terms of occurrences, number of links, and total link strength, the impact of the thesaurus keyword `long-term` is 25, 28, and 105, and for lifelong 6, 17, and 31, respectively. These values are much higher than the ones respective only to `long-term autonomy` from Figure 3a. The reason is that `long-term` in Figure 3b compiles the occurrences of keywords such as `long-term autonomy`, `long-term mapping`, and `long-term localization` (6, 2, and 2 occurrences, respectively), and `lifelong` sum up, for example, three different versions of `lifelong learning` (using `lifelong`, `life-long` 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 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

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

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



(a) (b)

Figure 3: Keywords co-occurrence analysis on the 144 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. Legend: `cnn` – Convolutional Neural Networks, `env` – environment, `localiz` – localization, `nav` – navigation, `nn` – Neural Networks, `recog` – recognition, `vslam` – visual SLAM.

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 (5 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 keywork learning occurs 7 times with an average publication year of 2017.00. Neural Netowrks (`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 another 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`, `marginalization`, and information theory (`info theory`) while having similar average publication years (2014.36, 2014.00, and 2015.50, respectively). Even though the number of occurrences of these terms is low (11, 6, and 5 for `graph`, `marginalization`, and `info theory`, respectively), their map proximity could indicate a focus in the past on the topic of graph sparsity, i.e., 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 information theory, the respective works focus on removing uninformative data from the map (Kretzschmar and 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 (multi-session 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 multi-session 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 aver-

age publication year. Instead, the main goal of the co-authorship analysis in this review is to present possible research networks detected in the 144 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 29 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 144 included records using the original RIS file used in Section 4.2 compared to 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.

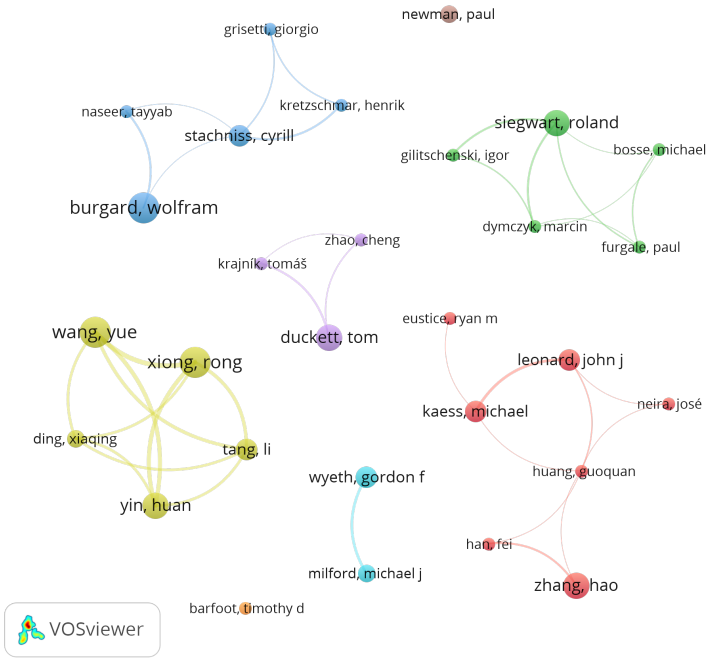


Figure 4: Co-authorship analysis on the 144 included records generated by VOSviewer. Parameters used for generating the co-occurrence network: minimum number of occurrences = 3, attraction = 2, repulsion = -3, scale = 1.49, circles size variation = 1.0, lines size validation = 1.0.

Analyzing Figure 4, the co-authorship network presents 8 clusters. These clusters are separated from each other, i.e., 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 miminum of 3 co-authored documents there is not a connection between these 8 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 144 records included in this review. As a results, the following enumeration presents the authors that belong to each cluster in the format of author (number of co-authored documents):

1. Rong Xiong (7), Yue Wang (7), Huan Yin (6), Li Tang (5), and Xiaqing Ding (4);
2. Hao Zhang (6), John J. Leonard (5), Michael Kaess (5), Fei Han (3), Guoquan Huang (3), José Neira (3), and Ryan M. Eustice (3);
3. Wolfram Burgard (7), Cyrill Stachniss (5), Giorgio Grisetti (3), Henrik Kretzschmar (3), and Tayyab Naseer (3);
4. Roland Siegwart (6), Igor Gilitschenski (3), Marcin Dymczyk (3), Michael Bosse (3), and Paul Furgale (3);
5. Tom Duckett (6), Cheng Zhao (3), and Tomáš Krajník (3);
6. Gordon F. Wyeth (5) and Michael J. Milford (4);
7. Paul Newman (4);
8. Timothy D. Barfoot (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 (H. Yin, Y. Wang, 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. 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 Universitario de Investigación en Ingeniería de Aragón (I3A) at the Universidad de Zaragoza in Spain. Although there are 5 different affiliations to which the 7 authors stated in the respective records, 4 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 by 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

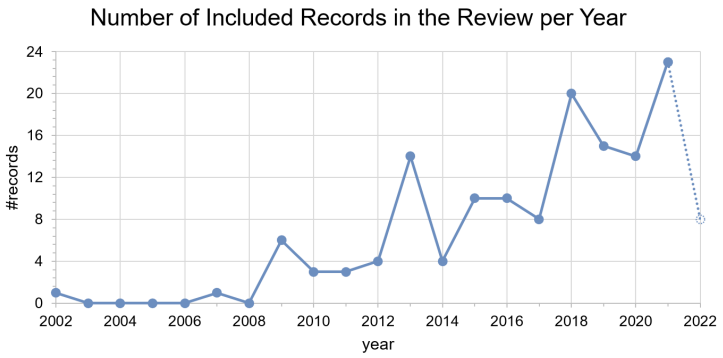


Figure 5: Evolution of published records per year considering the 144 included records in this review

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 144 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 biase 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 6 works, compared to 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 8 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 144 included records in the review is relative to the publication venue. Table 5 presents the venues with more than 1 publication, separating the journals and conferences in two different tables (Tables 5a and 5b, respectively). The columns μ 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 (μ), the third column (σ) 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 144 records included in the review for discussion.

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

(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	3.8	2021	18
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

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[11]. With more than 5 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 3 records published in them, compared to 22 and 18 papers in ICRA and IROS, respectively. When considering that 62 of the 144 included records are published in conferences, ICRA and IROS with a total of 40 published works related to this review's topic represent 65% of works published in conferences and 27.8% 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 144 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 for removing redundant data from the map or identifying novelty data to keep the map size constrained to the environment size. Section 5.4 discusses how methods handle multi-session 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. However, the discussion should also focus on how the included works evaluated their results in long-term operations. Thus, Section 5.6 presents the evaluation metrics used in the experiments, and Section 5.8 analyzes the experimental data and datasets used for evaluating the proposed methodologies.

5.1 Appearance variance

Next, the discussion focuses on included works categorized in DE1 as appearance. 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. In order to improve the discussion, the analysis of the proposed techniques related with appearance invariance is organized into the following topics: experience maps for treating different appearances as multiples experiences, handcrafted features, features extracted using Convolutional Neural Networks (CNN), assessment of feature stability, multi-modal features, leverage of temporal coherence by image sequence matching, and 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, where each experience is a view of the environment at a certain position and wheel odometry provides the relative pose for the links. 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 combines the mapping of RatSLAM with the place recognition of FAB-MAP (Cummins and Newman 2008b). The latter improves the loop closure detection of the original RatSLAM due to FAB-MAP having light invariant characteristics for data association by learning a generative model for the Bag of Words (BoW) model (Sivic and Zisserman 2003). Both RatSLAM and the hybrid RatSLAM+FAB-MAP systems uses visual data to retrieve information from the environment. Although Martini et al. 2020 uses also experience-based mapping, the main sensor is a radar, where an experience is represented by a point cloud from the sensor and the point descriptors retrieved from it. 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).

The concept of adding the environment changes to the map identified by the degradation in localization is also employed by Konolige and Bowman 2009 and Tang, Y. Wang, Ding, et al. 2019. The former implements a keyframe SLAM system created from the Visual Odometry (VO) module, where each keyframe represents a view of the environment, while a place recognition module tries to match the current frame to similar views already in the map for loop closure. The latter 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, i.e., no restriction for global consistency. New nodes are trigered either from localization failure or after a certain length is traveled by the robot. 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 defines it as a whole sequence of the saved poses and related features directly obtained from VO. In this case, the topological mapping links experiences not geometrically but instead if two experiences observe the same space. However, the method does not implement a specific place recognition module for loop closure, assuming that the robot will subsequently return to a place that can have successful localization. Gadd and Newman 2016 builds on the work of Churchill and Newman 2013 for multi-robot systems. This method adds FAB-MAP for place recognition in the existing map maintained by a centralized versioning framework. The selection of the most relevant experiences by the centralized framework for localizing multiple agents in the system 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 N. Zhang et al. 2018. Similar to

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

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 multi-experience 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 navigation methods try to generate new experiences if the environment changes, expecting 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. However, these approaches are not scalable in the long-term timeframe nor to deal with dynamic elements, even using central servers as in Gadd and Newman 2016 with more computational resources than the robots. Pruning algorithms would be required to remove redundant or outdated information, as in Konolige and Bowman 2009 or Tang, Y. Wang, Ding, 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 to color images for increasing 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 3 color channels into an one-dimensional space with a weighting parameter conditioned by the peak spectral responses of each channel, 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 Z. Yang et al. 2021 uses this transformation for preprocessing the color images into grayscale ones demonstrating the robustness of the illumination invariant space when lighting changes appear.

An alternative to using predefined illumination invariant transformations is to learn them. Clement et al. 2020 learns a nonlinear transformation mapping function from the RGB color space to grayscale also combining the three-channel log-responses, but relaxing the constraints of the one-dimensional space due to the original weighting parameter used in Arroyo et al. 2018 and Z. Yang et al. 2021. Instead of using the same parameters independently of the image content, Clement et al. 2020 trains 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 the applying the optimized transformation to the color images.

Even though the Gamma correction does not transform an image to an invariant color space, this transformation can be used to strengthen low-illumination changes. Li Sun et al. 2021 uses 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 Li Sun et al. 2021 to improve day-to-night matching.

5.1.3 Handcrafted features

Many localization and mapping algorithms rely on detection and extraction of features. The designation of handcrafted features refers to properties derived from the sensors data as a two-step process: a keypoint detector to locate the features and their characterization by computing a descriptor capable of distinguishing each feature from the others (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 defines 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) or Speeded Up Robust Features (SURF) selecting the stable ones. Using an incremental tree-like PIRF (with inverted index as in BoW) dictionary, the method has shown robustness to viewpoint variance and unstable features. Also, PIRF-based localization improved the recall over FAB-MAP in the experiments.

Moreover, Histogram of Oriented Gradients (HOG) features have been used in different works to improve robustness to appearance variance, given that HOG descriptors capture local gradient information robust to seasonal changes (Naseer, Suger, et al. 2015). Li et al. 2015 computes local HOG descriptors from visually-salient image patch features in an underwater environment. Using a trained Support-Vector Machine (SVM) to classify the matching between corresponding patches, the method achieved approximately 80% accuracy with dramatic appearance changes. Although Naseer, Suger, et al. 2015 computes HOG descriptors from each cell of a partitioned image, a global descriptor for the whole image joins all the cell ones. The global descriptor proved to be robust to foliage color changes, occlusions, and seasonal changes. Vysotska et al. 2015 uses the same global HOG descriptor as in Naseer, Suger, et al. 2015, but applied to image sequence matching requiring a rough global pose estimation for the images (e.g., GPS) for efficient matching.

Local Difference Binary (LDB) 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 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 introduces 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 and clusters the descriptors into seasonal-specific vocabularies using hierarchi-

cal 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 performance compared to not comparing with the predicted image, although 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 builds on SIFT Flow 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, while the GPS, the feature tracks from a visual SLAM, and the compass measurements bias the image registration. The proposed method was 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 F. Cao, Zhuang, et al. 2018 and F. Cao, Yan, 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. F. Cao, Zhuang, et al. 2018 transforms 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 ORB 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 LiDAR in the experiments was its less sensitivity to lighting conditions relative to visual sensors while not being incapacitated in dark environments. The proposed method outperformed M2DP (He et al. 2016) – global descriptor for point clouds –, given that M2DP could not deal in situations where the point clouds distributions were centralized and similar to each other. As for F. Cao, Yan, et al. 2021, the proposed method accepts also 2D laser data by accumulating a sequence of scans. The 2D representation used differs from F. Cao, Zhuang, 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 to detect and describe the contours of the images, F. Cao, Yan, et al. 2021 generates Binary Robust Independent Elementary Features (BRIEF) 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 and Gordon. F. Wyeth 2012b) (sequential place recognition) and PointNetVLAD (Uy and 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 extracts visual features used for tracking using a blob detector based on a Hessian matrix. These features are extracted from a 2D cartesian image transformed from the polar image representation of radar, while also compensating the distortion from the vehicle's motion. 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, while also outperforming ORB-SLAM2 (Mur-Artal and 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 2013a extracts 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 extracts 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 formulates 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) framework.

Furthermore, Meng et al. 2021 extracts 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 while the histogram cross-correlation of the Normal Distribution Transform (NDT) that computes local probability density functions of the surface smoothness identifies the loop closures. The proposed methods 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 or Hough transform peaks due to the moment grid descriptor includes higher order of moments relative to other descriptors.

Poles are structures also used for long-term localization. Schaefer et al. 2021 retrieves 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, Ward, et al. 2019 extracts 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 formulates an aerial image registration algorithm based on the positions of the crops and the gaps between them remaining the same over time. The method computes a vegetation mask by exploiting the Excess Green Index (ExG) of RGB images. Using the Hough transform to find lines between vegetation, the center of the crops are the peaks on vegetation histograms perpendicular to the rows. The testing results demonstrated invariance of the registration algorithm to changing conditions caused by weather and crop growth over one 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 discriminate power compared to handcrafted features while

being able to be more robust in challenging environments (Taisho and 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, L. Liu, 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, the place recognition is formulated as a classification problem (Chen, L. Liu, et al. 2018). In Taisho and Kanji 2016, the layer fc6 (fully connected) of AlexNet 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 defines 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 and FAB-MAP on a cross-season experiment in precision-recall metrics. Chen, L. Liu, et al. 2018 also formulates a classification task for place recognition, using a VGG16 network for generating local features, while adding a convolutional a fully-connected, and a softmax layer to learn the correct label output for classification. The proposed architecture outperformed FABMAP and SeqSLAM 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; B. Liu et al. 2021; Xin et al. 2017), while local features are used for obtaining a more accurate estimation based on spatial matching (Camara et al. 2020; Xin et al. 2017) or geometrical verification (B. Liu et al. 2021). Xin et al. 2017 extracts both global and local features using a convolutional layer (conv3) of the AlexNet network, where 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 uses layers from VGG16 for feature extraction, specifically, conv5-2 and conv4-2 layers for global and local features, respectively. As for B. Liu et al. 2021, the MobileNetV2 network is selected for global feature extraction due to its computational efficiency. However, their work uses grid-based motion statistics with Oriented FAST and Rotated BRIEF (ORB) local features instead of CNN features.

Deep features can be combined with handcrafted features and preprocessing techniques to facilitate learning and further enhance their discriminative properties. K. Zhang et al. 2022 uses the Key.Net network for keypoint generation, given that combines handcrafted and learned filters to detect keypoints at different scale levels, helping reduce the number of learnable parameters. Combined with HardNet for descriptor extraction, the method outperformed a BoW approach in viewpoint and illumination changing conditions. H. Yin, Y. Wang, et al. 2020 proposes a handcrafted rotational invariant feature to be the input of a LocNet 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, Li Sun et al. 2021 uses a the Gamma transform 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 keypoint detector. DarkPoint achieved approximately 1.7x 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, J. Zhu et al. 2018 extracts features from three different 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 Z. Yang et al. 2021. The conv5-3 achieved higher accuracy than conv4-4 and conv3, indicating that the spatial information increases in deeper layers improving the place recognition. J. 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 chooses DenseNet for feature extraction due to this network reusing feature maps, i.e., connecting all layers with the same map sizes directly with each other. Then, Yu et al. 2019 uses 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) and to a BoW place recognition method.

The included works also focus on 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). P. Yin, L. Xu, et al. 2018 encodes directly the BEV of a LiDAR into a low dimensional global feature using a bidirectional Generative Adversarial Network (GAN). Using the extracted features within the SeqSLAM framework, the proposed method improved the precision-recall metrics over the original SeqSLAM in changing conditions. Similarly, Martini et al. 2020 extracts a global descriptor from the BEV using NetVLAD but with the point cloud from a radar sensor. Kim, B. Park, et al. 2019 formulates 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, B. Park, et al. 2019 uses a LeNet network for feature extraction and place classification. The proposed architecture outperforms PointNetVLAD (Uy and Lee 2018) and the handcrafted point cloud feature M2DP in precision-recall. Based on SCI (Kim, B. Park, et al. 2019), X. Xu et al. 2021 proposes 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 (X. Xu et al. 2021), H. Yin, X. Xu, et al. 2021 also uses SCI and FFT for feature extraction of point clouds. The difference is the use of a shared U-Net architecture to extract features of LiDAR and radar data, training simultane-

ously 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, P. Yin, J. Xu, et al. 2021 also uses 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 and M2DP on the recall metric in appearance variant conditions.

Lastly, a trend found in the included works to improve the discriminative power of CNN features is the use of triplets (B. Liu et al. 2021; Martini et al. 2020; Piasco et al. 2021; Li Sun et al. 2021; H. Yin, X. Xu, et al. 2021; P. Yin, J. Xu, et al. 2021) in training. A triplet consists in an anchor image, a positive corresponding match, 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) (Li Sun et al. 2021). Additionally, Piasco et al. 2021 uses also depth information during training, given that depth maps and their geometric information remain more stable across time than visual ones. 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 during 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, Oliveira, et al. 2017 uses the FastNet network for extracting saliency 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 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 modifies an U-Net 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, Worrall, et al. 2021 also segments an image with pixel-wise labels, discriminating 12 classes: pole, building, road, vegetation, undrivable road, pedestrian, rider, sky, fence, vehicle, and unknown. 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 IMU data for localization and mapping. The long-term evaluation of the map corrections showed a decrease over time demonstrating the stability of these features in outdoor environments. In addition to pixel-wise segmentation, G. Singh et al. 2021 connects 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, FAB-MAP, and a BoW-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 relationships between features for improving the robustness of place recognition. Han, Beleidy, et al. 2018 proposes 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, where the nodes are images or the semantic classes, and 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, and color and AlexNet-based descriptors in changing conditions. As for Gao and H. Zhang 2020, the proposed method formulated the place recognition task into 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 that relate the features represent theirs spatial distance and angular relations, and their appearance similarity (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, H. 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 other sensors such as LiDAR. Z. Wang et al. 2021 uses the RangeNet++ network for inferring semantic labels of 3D point clouds from LiDAR data. Even though the network can label 10 different 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, B. Park, et al. 2019) in an outdoor experiment with moving elements and dense vegetation.

Appearance-content disentanglement A location has different representations due to weather or seasonal changing conditions in the long-term perspective, among other factors. In terms of image data, the information retrieved from these representations could be separated in terms of its contents and appearance. The included works studied this possibility by learning the appearance-content disentanglement for feature representation that assumes the decomposition of the images latent space into apperance and content spaces (C. Qin et al. 2020).

Oh and Eoh 2021 adopts a Variational AutoEncoders (VAE) architecture that 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 proposes 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 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 indicated that viewpoint change is considered to be content in the proposed algorithm. With a similar architecture to C. Qin et al. 2020, Tang, Y. Wang, Tan, et al. 2021 also considers a place domain discriminator to ensure that the content discriminator only contains the place information and not also its appearance, while also using data augmentation in training to increase robustness against viewpoint changes. In the experiments, all images generated from a zero-appearance feature vector looked similar, while their place information remains conserved indicating that the proposed method can disentangle the input image across appearance changes.

Even though Hu et al. 2022 does 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 adopts a multi-domain image-to-image architecture that expands the CycleGAN architecture from two 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, while also outperforming FAB-MAP.

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 et al. 2016 proposes a CNN architecture based on AlexNet 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), where the feature label is binary - stable or unstable - computed for training by assessing the number of the feature observations over multiple sessions. In the experiments of over 15 months and changing conditions, the proposed method outperformed random selection of features for localization in terms of f-score, while the addition of depth information improved the method's performance.

Other approaches in the included works define predictor functions for evaluating the feature stability. Berrio, Ward, et al. 2019 defines 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, Worrall, et al. 2021 also uses the concentration ratio and maximum driven length as predictors, their approach simplifies the selection by including features that have been observed for more than 1m and conserving the ones in sparse density areas to avoid localization failures. Berrio, Worrall, et al. 2021 also defines a visibility measure related to the maximum range from where the feature is detected at a particular angle and the respective probability of detection to only the feature metrics when it is a match or if both not detected and not occluded.

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 defines 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 that measures 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 does not add features after creating the visual database used as a map, the 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, i.e., reprojecting the keypoints of the database on the query image using the transformation and re-compute the respective descriptors. The descriptors similarity, a spatial and temporal constraints, and the number of successful matches determine if the environment changed to update the feature weights 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 models 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, following the intuition 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 proposes 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. 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 Multi-modal features

Another type of approach to feature-based localization and mapping is the use of multi-modal features, given that these features can be more discriminative than only considering a single feature

space (Latif et al. 2017). Filliat 2007 proposes a two-stage voting scheme for localization integrating 3 different feature spaces: SIFT, local color histograms, and local normalized grey 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 multi-modal features – GIST and feature maps from a CNN – by concatenating their descriptors into a single vector. In both Filliat 2007 and Latif et al. 2017, the use of multiple feature spaces improved the localization performance over considering only a single feature space.

The included works also cover a more specific approach to multi-modal 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, H. Wang, et al. 2018; Han, X. Yang, et al. 2017; Siva, Nahman, et al. 2020; Siva and H. Zhang 2018). Han, X. Yang, et al. 2017 formulates the Shared Representative Appearance Learning (SRAL) for fusing multi-modal visual features from 6 different spaces applied on downsampled images as scene descriptors: color histograms, GIST, HOG, Local Binary Patterns (LBP), SURF, and AlexNet (conv3). SRAL outperformed the individual feature spaces and also the concatenation of the 6 spaces into a single descriptor. Han, H. Wang, et al. 2018 proposes the RObust Multimodal Sequence-based loop closure detection (ROMS), that is the adaptation of the regularized optimization to image sequence matching. The modalities considered are LDB (Arroyo et al. 2018), GIST, Faster R-CNN, and ORB. ROMS outperformed both FAB-MAP and SeqSLAM in appearance changing conditions, while improving the performance over considering a single feature space. In addition to learn discriminative modalities, Siva and H. Zhang 2018 formulates 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, Nahman, 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 while exploring the temporal sequential relationships of the images (V. A. Nguyen et al. 2013; Ouerghi et al. 2018). Ouerghi et al. 2018 builds on SeqSLAM (Milford and Gordon. F. Wyeth 2012b) 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, H. Wang, et al. 2018 compared frame-to-frame matching to the proposed ROMS algorithm that 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, while outperforming SeqSLAM and FAB-MAP in changing conditions.

Moreover, Vysotska et al. 2015 defines 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 descriptor (Naseer, Suger, et al. 2015). Instead of formulating the sequence matching as a network flow optimization problem, Vysotska et al. 2015 estimates 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 image matches. Naseer, Suger, et al. 2015 leverages 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. The 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, Suger, et al. 2015 outperformed SeqSLAM 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 J. 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 uses the LDB 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 and SeqSLAM in terms of precision-recall metrics. As for J. Zhu et al. 2018, the feature maps from VGG16 are normalized into a binary descriptor. The method outperformed FAB-MAP, SeqSLAM, and ABLE (Arroyo et al. 2018) in a cross-season experiment.

Lastly, V. A. Nguyen et al. 2013 proposes an approach to identify topological places based on an image stream. The method uses a clustering scheme K-iteration Fast Learning Neural Network (FLANN) 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, in which the topological locations defined as image sequences are 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 sensors as well as ranging-based ones 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 (H. Yin, X. Xu, 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 to appearance changes. In addition to the works already discussed previously,Pérez et al. 2015, Coulin et al. 2022, and T.-M. Nguyen, M. Cao, et al. 2022 also focus on appearance invariance upon changing environments while using more than one modality. Pérez et al. 2015 introduces an appearance-based particle injection in the Monte Carlo Localization (MCL) framework to account the visual place recognition of FAB-MAP Cummins and Newman 2008b. The BoW 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 proposes 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 2.4m to 0.033m compared to using vision-only in the MSCKF. As for T.-M. Nguyen, M. Cao, et al. 2022, the proposed Visual-Inertial-Ranging-Lidar (VIRAL) sensor fusion algorithm includes an IMU, LiDAR, a camera, and Ultra-Wide-Band (UWB) data for localizing an aerial vehicle in indoor environments, with the first three sensor modalities for odometry and UWB for absolute positioning 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 the experiments performed with an aerial vehicle 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. 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: specific map representations used to model or deal with dynamic elements in the scene, identification of dynamic elements matching the current observation to the current map, future prediction of dynamic properties of scene elements, and 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 multi-store 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 implements 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 decreases the localization failure rate compared to a static view. Instead of using a state machine, Bacca et al. 2013 implements 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 pedestrian dynamics from the 2D laser and camera data while also achieving a more accurate representation of the environment compared to a static approach.

Although Biber and Duckett 2009 does 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 5 different timescales, where the smaller one ($\sim$3.1s) represents an STM map updated at every instant and the other 4 are LTM submaps ranging from $\sim$0.43session to $\sim$13.5day, updated after each season or daily. Instead of only localizing on the LTM maps as inDayoub et al. 2011 and Bacca et al. 2013, Biber and Duckett 2009 selects 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, while the 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, including 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. K. Wang et al. 2019 uses a tracking map with short-term static points and a long-term map only containing long-term static points identified by a semantic segmentation module with ORB-SLAM2 (Mur-Artal and Tardós

2017). Also, S. Zhu et al. 2021 creates offline a semi-dynamic map and a static one, where the former has semi-dynamic objects (parked cars in a parking lot environment) and the latter 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 K. 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, although it was affected by false positives and negatives, and by the blur effect in the 2D grid map. K. Wang et al. 2019 improved the ATE in a dynamic environment over ORB-SLAM2 and DynaSLAM (Bescos et al. 2018). As for S. 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 determines 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, M. Zhang et al. 2019 maintains a Signed Distance Field (SDF) 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 detects semi-static changes leveraging the ICP 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 minimizes the Gibbs energy defined on the proposed Long-term Consistent Conditional Random Field (LC-CRF) for detecting dynamic points, considering that these points have often a large reprojection error in frame-to-map matching and points tend to have the same dynamic properties as the neighbor ones. In a dynamic scene, LC-CRF achieved lower ATE than ORB-SLAM.

Furthermore, Pan et al. 2019 and Ding et al. 2020 leverage clustering properties of the observations evaluating the observations count. Pan et al. 2019 segments the points of a LiDAR point cloud into different clusters assuming 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 uses 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), allowing 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 uses 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 and indeed, in the experiments, Pomerleau et al. 2014 had a more precise and cleaner map of the environment than using a standard ICP matching. Instead of weighting the map points, An et al. 2016 proposes 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, decrease gradually the weight until not detecting the obstacle. Integrating DEL in an exploration scheme, nodes with a edge weight average lower than a certain threshold, meaning frequent moving obstacles or changed structure near that node, are not considered for exploration due to the robot may be unable to move to that position.

Although the standard NDT 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 and adopt exponential weighted moving average and covariance for new measurements having an higher influence than old ones. Einhorn and Gross 2015 proposes 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 depends 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, where the first 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 uses a dynamic occupancy grid and exploits the stationary distribution and the state holding time associated with Hidden Markov Mod-

els (HMM) 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 implements a semi-Markov process extended by a Levy process to model a time dependency on the state holding time of Markov processes, also predicting as Tipaldi et al. 2013 the expected retention time for each cell being in a specific state. 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, Fentanes, Santos, et al. 2017 proposes 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 and Newman 2013) in terms of localization error by selecting the most likely visible features at each location for localization. Santos et al. 2016 adopts the FreMEn within an exploration scheme, where the 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, L. Wang et al. 2020 models both aperiodic and periodic changes by an Auto-regressive Moving Average Model (ARMA). 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 uses a KPConv 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 for obtaining the training annotation of 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 proposes a collaborative dynamic mapping for detecting humans using visual and thermal images and a 3D LiDAR. The YOLOv3 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. S. Zhu et al. 2021 also uses YOLOv3 to extract bounding boxes of dynamic classes (parked cars) from a RGB image, and the projection of LiDAR points allows the creation of the static and semi-dynamic maps required for localization. Even though Zhu does not discard the dynamic objects, the localization module reduce the importance of weight of the corresponding observations. Instead of using visual data to detect object classes, L. Sun et al. 2018 adapts the PointNet for object recognition (pedestrian, cyclist, car, or background) to classify the scan points of a LiDAR. The proposed Recurrent-OctoMap maintains the occupancy and semantic information, 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). 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. Instead of identifying object classes, Song et al. 2019 proposes 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 proposes the Semantically Informed Visual Odometry (SIVO) to improve the performance of ORB-SLAM2 by using the Bayesian neural network SegNet 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 immediate 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 proposes DynaSLAM as a front end for ORB-SLAM2 to segment potential dynamic classes using a Mask R-CNN. The semantic labeling is improved using a multi-view 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, K. Wang et al. 2019 implements a front end for ORB-SLAM2 to identify movable objects with a ResNet-based network for segmentation. The segmentation of the previous frame and a geometric verification 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 Constraints Visual SLAM (SGC-VSLAM) (S. 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 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 proposes the DE-SLAM to deal with Short-Term Dynamics (STD) and Long-Term Dynamics (LTD) at the same time. A MobileNetv2 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 features extracted from ORB 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

berrio-et-al:2019:8814289 Using the pole and corner features extracted from a 3D laser, Berrio uses the following predictor variables to evaluate a feature: number of detections, maximum detected spanning angle, maximum length driven while observing the landmark, maximum area of detection, maximum possible spanned angle, and concentration ratio. Then, a cross-validated elastic net regularized regression algorithm using training data (based on the number of observations across multiple sessions) adjusts the coefficients for identified predictors. The concentration ratio and the maximum angle predictors seem to have much lower coefficients relative to the other ones.

berrio-et-al:2021:3094485 Using the pole and corner features extracted from the fusion of image segmentation and 3D laser data, Berrio use two predictors to assess the quality of features for including in the map or not: concentration ratio and the maximum length driven while observing the landmark. Instead of using a regression algorithm ,the method selects landmarks that have been observed for more than 1m of distance. The concentration ratio avoids the exclusion of landmarks when the density is very sparse (so eliminating one landmark could lead to localization failures), and excludes the ones in high density areas. The prior map contains the positions and attributes of the features including a visibility measure of each map feature – visibility volumetric volume to assess the stability of the features: maximum range from where the feature was detected at a particular angle and the probability at detecting at that particular angle. Allows to only updating the feature metrics when it is a match or not detected + not occluded.

5.4 Multi-session

5.5 Computational

taisho-kanji:2016:7866383 Principal Component Analysis (PCA) can be used for dimensionality reduction (4096 to 128-dimensional features), defining the proposed descriptor PCA-NBNN.

xin-et-al:2017:8310121 Studies the possibility of using a random selection technique for feature dimension reduction, given that requires no need for further training nor significant loss in efficiency and effectiveness compared to Local Sensitive Hushing (LSH) and Principal Component Analysis (PCA). Cosine distance for evaluate query image compared to database (global descriptor).

yu-et-al:2019:8961714 The 4 max-pooling by channel proposed to reduce the descriptors dimensions with minimal accuracy reduction. 1024-dimension divided into 256 groups and maximum of ech group used as final descriptor. Compared to PCA, 4 max-pooling by channel has less computational complexity but similar performance.

camara-et-al:2020:9196967 PCA to reduce each vector of the 16 cubes to 100 dimensions.

piasco-et-al:2021:6 Reduce dimension of the descriptors by applying PCA and whitening.

yang-et-al:2021:12054 select max-pooling layer instead of convulutional one

naseer-et-al:2017:7989305 Uses Sparse Random Projections for embedding high-dimensional feature vectors into lower dimensions. The precision-recall metrics is very similar to the full dimensional descriptor.

5.6 Evaluation metrics

5.7 Long-term experimental data

Table 6: Datasets used in the 144 included records for long-term localization and / or mapping experiments. Legend: odo – odometry (wheeled, laser, visual, inertial, or a combination of odometry sources, dist. – total distance length of the dataset, path – total path distance if repeated several times, time – total operation time, int. – time interval between the start and end acquisition dates / time instants (d/w/m/y equivalent to day/week/month/year, 0 if only 1 run), and seq. – number of sequences of the dataset.

		Long-term								Sensor												Calib.							
											camera						laser												
Dataset	**Year**	**lighting**	**day/night**	**weather**	**seasonal**	**dynamics**	**sparsity**	**Environ.**	**Domain**	**odo**	**gray**	**color**	**monocular**	**stereo**	**omni**	**RGBD**	**thermal**	**2D**	**3D**	**radar**	**sonar**	**IMU**	**GPS**	**intrinsic**	**extrinsic**	**GT data**	**Format**	**dist.** (km)	**path** (km)	**time** (h)	**int.** (d/w/m/y)	**#seq.**
FHW	2001						x	indoor (museum)	ground (TOURBOT)	x								x								–	CARMEN	–	–	1.98	–	1
FR079	2003						x	indoor (office)	ground (robot)	x								x								–	CARMEN	–	–	0.29	–	1
FR101	2003						x	indoor (office)	ground (robot)	x								x								–	CARMEN	–	–	0.29	–	1
Intel Research Lab 2003	2003						x	indoor (office)	ground (robot)	x								x								–	CARMEN	0.506	–	0.75	–	1
MIT Killian Court	2004						x	indoor (office)	ground (robot)	x								x			x					–	CARMEN	2.2	–	2.5	–	1
City Center (FAB-MAP)	2008	x				x		outdoor (urban)	ground (robot)			x	x										x	x		GPS, manual	plain text (non-image), jpg (image)	2	–	–	–	1
Lip6Indoor	2008							indoor (office)	ground (hand-held)			x	x											x		manual	ppm (images)	–	–	0.11	–	1
Lip6Outdoor	2008	x				x		outdoor (campus)	ground (hand-held)			x	x											x		manual	ppm (images)	–	–	0.3	–	1
New College (FAB-MAP)	2008	x				x		outdoor (campus)	ground (robot)			x	x										x	x		GPS, manual	plain text (non-image), jpg (image)	1.9	–	–	–	1
St Lucia Brisbane 2007	2008	x				x		outdoor (urban)	ground (car)			x		x												–	–	66	–	1.67	–	1
Bicocca (indoor)	2009	x				x	x	indoor (office)	ground (Robocom)	x		x	x	x	x			x			x	x		x	x	map model, laser-based	plain text (non-image), png (image)	–	–	2.5	3d	5
COLD	2009	x	x	x		x		indoor (office)	ground (Pioneer 3, ATRV Mini, PeopleBot)	x			x	x		x		x								laser-based, manual	plain text (non-image), jpg (image)	0.92	–	0.99	–	76
Malaga 2009	2009	x				x		outdoor (parking, campus)	ground (car)				x	x				x				x	x	x	x	RTK-GPS	Rawlog MRPT	6.358	–	–	–	6
New College	2009	x				x		outdoor (campus)	ground (Segway)	x		x	x		x	x		x				x	x	x	x	GPS	plain text (non-image), png (stereo), jpg (omni)	2.2	–	0.73	–	1
albert-b-laser-vision	2010	x					x	indoor (office)	ground (iRobot B21r)	x			x	x				x								–	CARMEN (non-image), jpg (image)	–	–	0.18	–	1
CMU-VL	2011	x			x	x		outdoor (urban)	ground (car)				x	x								x	x	x	x	GPS	–	–	8.5	–	1y	16
Ford Campus	2011	x				x		outdoor (campus, urban)	ground (car)				x		x			x	x			x	x	x	x	RTK-GPS	LCM log	–	–	–	2m	–
UTIAS Multi-Robot	2011	x						indoor (empty space)	ground (iRobot Create)	x			x	x										x		external tracking system	jpg (image), dat (non-image)	–	–	4.78	–	9
Alderley Brisbane	2012	x		x		x		outdoor (urban)	ground (car)				x	x												manual	–	16	8	–	–	2
TUM RGBD	2012	x				x		indoor (office, industrial hall)	ground (handheld, Pioneer 3)				x			x						x			x	x	external tracking system	plain text (non-image), png (image + depth), ROS bag	0.285	–	0.35	–	15
CoBots long-term	2013	x			x	x	x	indoor (office)	ground (robot)	x			x			x		x							x	–	ROS bag	131	–	260	2y3m	1082
KITTI	2013	x				x		outdoor (urban)	ground (car)			x	x	x					x			x	x	x	x	RTK-GPS	png (image), binary (laser), plain text (imu, gps)	–	–	1.18	8d	61
MIT Stata Center	2013	x			x	x	x	indoor (office)	ground (PR2)	x			x		x		x	x				x			x	map model	ROS bag	42	–	38	1y9m	84
Nordland	2013	x		x	x			outdoor (railway)	ground (train)				x	x									x				GPS	mp4 (video stream), plain text (gps)	2916	729	39.74	–	4
Gardens Point Campus of QUT	2014	x	x			x		indoor, outdoor (campus)	ground (hand-held)				x	x												ground-plane position	png (images), plain text (ground plane)	–	–	–	–	3
Witham Wharf RGB-D (LCAS STRANDS)	2014	x	x		x	x		indoor (office)	ground (SCITOS-G5)				x			x		x							x	–	ROS bag	–	–	–	1y1m	368
KAIST	2015	x	x			x		outdoor (urban)	ground (car)				x		x		x					x	x	x	x	RTK-GPS	png (images), plain text (imu, gps)	84	–	–	18d	36
EuRoC	2016	x						indoor (industrial hall, office)	air (AscTec Firefly)			x		x	x							x		x	x	external tracking system	ROS bag	0.8936	–	0.37	–	11
NCLT	2016	x		x	x	x		indoor, outdoor (campus)	ground (Segway)				x		x			x	x			x	x	x	x	RTK-GPS, SLAM-based	binary (laser), tiff (image), plain text (non-laser or image)	147.4	–	34.9	1y4m	27
Berlin Kudamm	2017	x				x		outdoor (urban)	ground (car)				x	x												manual	jpg (image)	–	–	–	–	2
Oxford RobotCar	2017	x	x	x	x	x		outdoor (urban)	ground (car)	x			x	x	x			x	x			x	x	x	x	RTK-GPS	png (image), binary (laser), plain text (imu, gps, odo)	1010.46	10	–	1y8m	133
YQ21	2017	x				x		outdoor (campus)	ground (car)				x	x	x				x			x	x	x	x	RTK-GPS	binary (laser), jpg (images), plain text (gps)	23	–	6.5	1w	21
CMU-Seasons	2018	x			x	x		outdoor (urban)	ground (car)				x	x										x	x	manual	jpg (image)	–	8.5	–	330d	17
Freiburg Across Seasons	2018	x			x	x		outdoor (urban)	ground (car)				x		x								x			GPS, manual	jpg (image)	110	–	–	3y	3
RobotCar Seasons	2018	x		x		x		outdoor (urban)	ground (car)				x	x	x									x	x	manual	jpg (image)	–	10	–	178d	10
Bonn RGB-D Dynamic	2019					x		indoor (office)	ground				x			x								x	x	external tracking system	png (images, depth), plain text (imu, gps)	–	–	–	–	26
CBD	2019	x				x		outdoor (urban)	ground				x	x	x											manual	png (images)	–	–	–	–	1
MulRan	2020					x		outdoor (urban)	ground (car)										x	x		x	x	x		x	SLAM-based	binary (laser), CSV (global pose, radar ray), png (radar polar image)	41.2	–	–	2m15d	12

Table 6: continued from previous page

		Long-term								Sensor													Calib.							
											camera								laser											
Dataset	**Year**	**lighting**	**day/night**	**weather**	**seasonal**	**dynamics**	**sparsity**	**Environ.**	**Domain**	**odo**	**gray**	**color**	**monocular**	**stereo**	**omni**	**RGBD**	**thermal**	**2D**	**3D**	**radar**	**sonar**	**IMU**	**GPS**	**intrinsic**	**extrinsic**	**GT data**	**Format**	**dist.** (km)	**path** (km)	**time** (h)	**int.** (d/w/m/y)	**#seq.**
Oxford Radar RobotCar	2020	x		x		x		outdoor (urban)	ground (car)	x		x	x	x					x	x		x	x	x	x	RTK-GPS, SLAM-based	png (image, raw laser, radar), binary (laser), plain text (imu, gps, odo)	280	10	–	1m	32
USyd Campus	2020	x			x	x	x	outdoor (campus)	ground (car)	x			x	x					x				x	x	x	GPS	ROS bag	–	–	–	1y	52
IPLT	2021	x	x	x	x	x	x	outdoor (parking)	ground (car)	x	x			x				x					x	x	x	GPS	ROS bag	–	0.2	–	2y	127
RADIATE	2021	x	x	x			x	outdoor (parking, urban)	ground (car)				x	x	x				x	x		x	x	x	x	RTK-GPS	ROS bag	–	–	4.98	–	–
NTU VIRAL	2022	x						indoor, outdoor (campus)	air (DJI M600)		x			x	x				x				x	x	x	external tracking system	ROS bag	1.845	–	0.9	–	9

5.8 Final observations

6 Challenges and Future Directions

- multi-robot long-term localization and mapping
- fusion of different sensors
- availability of datasets in continuous operation in challenging long-term scenarios (urban, industry – storage / logistics facilities)
- edge/cloud computing
- parameters of the algorithms adapting along time accordingly to changes in the environment
- specific hardware for improving computational performance (FPGA, operating systems specific for real-time usage, etc.)
- interaction with the user: map update and consequent visualization by the users not being on raw data but instead on higher levels of abstraction (CAD, automatic extract of the environment topological locations)
- hierarchical localization and mapping (topological location ¿ coarse localization ¿ millimeter level of localization or something like that)

7 Limitations of the Study

Section to discuss possible limitations of the study (timeframe, wider approach, etc.).

- only one query for discussion: e.g., for searching datasets, possibly, a different query should have been used
- overview long-term SLAM vs in-depth analysis and discussion of each type of techniques: our review synthesizes all types of techniques, if the reader wants an in-depth analysis, different reviews should be performed
- related to previous one, each category appearance dynamics sparsity multi-session and computational have all different aspects related to each other, that probably should be treated differently in different data extraction items to improve the data organization – however the main goal of this review was to overview all trends and not focus on one specifically, so this categorization should be done separately and probably in singular literature reviews.
- limited information on the experiments conditions (traveled distance, duration, etc.) given by the authors
- some works such as PointNetVLAD, FAB-MAP and SeqSLAM (+LOAM, and LEGO-LOAM) did not appeared in the identification phase. However, the ones included have improvements over all of these...
- discussion does not focus on the datasets used but rather on the methodologies, to not extend even further the review
- ORBSLAM2, ORBSLAM3 not identified

8 Conclusions

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.

ORCID

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

References

Acm digital library. (n.d.). Retrieved May 6, 2022, from https://dl.acm.org/

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. https://doi.org/10.1007/s10514-015-9453-0

Angeli, A., Filliat, D., Doncieux, S., & Meyer, J.-A. (2008a). *Lip6Indoor* (No. *5*) [http://cogrob.ensta-paris.fr/loopclosure.html]. https://doi.org/10.1109/TRO.2008.2004514

Angeli, A., Filliat, D., Doncieux, S., & Meyer, J.-A. (2008b). *Lip6Outdoor* (No. *5*) [http://cogrob.ensta-paris.fr/loopclosure.html]. 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. 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.). Academic Press. https://doi.org/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), 1539–58. https://doi.org/10.1016/j.robot.2013.07.003

Badino, H., Huber, D., & Kanade, T. (2011). *Cmu visual localization dataset* [http://3dvis.ri.cmu.edu/data-sets/localization]. Carnegie Mellon University. https://doi.org/10.1109/IVS.2011.5940504

Bailey, T., & Durrant-Whyte, H. (2006). Simultaneous localization and mapping (SLAM): Part II. *IEEE Robotics Automation Magazine*, *13*(3), 108–117. 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. 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* [https://ori.ox.ac.uk/news/the-oxford-radar-robotcar-dataset/]. University of Oxford. https://doi.org/10.1109/ICRA40945.2020.9196884

Berrio, J. S., Ward, J., Worrall, S., & Nebot, E. Identifying robust landmarks in feature-based maps. In: *2019-June*. 2019, 1166–1172. 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*. 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. https://doi.org/10.1109/LRA.2018.2860039

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

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

Biswas, J., & Veloso, M. M. (2013b). *Localization and navigation of the CoBots over long-term deployments* (No. *14*) [https://www.cs.cmu.edu/~coral/cobot/data.html]. Carnegie Mellon University. https://doi.org/10.1177/0278364913503892

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

Blanco, J.-L., Moreno, F.-A., & González, J. (2009). *Malaga dataset 2009 with 6D ground truth* [https://www.mrpt.org/malaga_dataset_2009]. Universidad de Málaga. https://doi.org/10.1007/s10514-009-9138-7

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. https://doi.org/10.1016/j.robot.2018.11.003

Bosse, M., & Leonard, J. J. (2004). *MIT Killian Court* (No. *12*) [http://ais.informatik.uni-freiburg.de/slamevaluation/datasets.php]. MIT. 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. 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* [http://ipltuser:iplt_ro@iplt.ip.uca.fr/datasets/]. Université Clermont Auvergne. https://doi.org/10.5220/0010518303830387

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

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. https://doi.org/10.1002/rob.21870

Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M. W., & Siegwart, R. (2016). *The EuRoC micro aerial vehicle datasets* (No. *10*) [https://projects.asl.ethz.ch/datasets/doku.php?id=kmavvisualinertialdatasets]. Autonomous Systems Lab, ETH Zürich. https://doi.org/10.1177/0278364915620033

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., & Leonard, J. J. (2016). Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, *32*(6), 1309–1332. https://doi.org/10.1109/TRO.2016.2624754

Camara, L. G., Gäbert, C., & Přeučil, L. Highly robust visual place recognition through spatial matching of CNN features. In: 2020, 3748–3755. 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. 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), 563–74. 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. https://doi.org/10.1109/JSEN.2018.2815956

Carlevaris-Bianco, N., Ushani, A., & Eustice, R. M. (2016). *The University of Michigan North Campus Long-Term vision and LIDAR dataset* (No. *9*) [http://robots.engin.umich.edu/nclt/]. University of Michigan. https://doi.org/10.1177/0278364915614638

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. 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. https://doi.org/10.1109/LRA.2018.2859916

Chen, Z., Maffra, F., Sa, I., & Chli, M. (2017). *Berlin Kudamm dataset* [http://imr.ciirc.cvut.cz/Datasets/Ssm-vpr]. 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* [https://sites.google.com/site/alldaydataset/]. Korea Advanced Institute of Science and Technology. https://www.researchgate.net/publication/282147318_All-Day_Visual_Place_Recognition_Benchmark_Dataset_and_Baseline

Churchill, W., & Newman, P. (2013). Experience-based navigation for long-term localisation. *International Journal of Robotics Research*, *32*(14), 1645–1661. 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. 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. https://doi.org/10.1109/LRA.2021.3136241

Cummins, M., & Newman, P. (2008a). *City Center (FAB-MAP)* [https://www.robots.ox.ac.uk/~mobile/IJRR_2008_Dataset/data.html]. Mobile Robotics Group, University of Oxford. https://doi.org/10.1177/0278364908090961

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

Cummins, M., & Newman, P. (2008c). *New College (FAB-MAP)* [https://www.robots.ox.ac.uk/~mobile/IJRR_2008_Dataset/data.html]. Mobile Robotics Group, University of Oxford. https://doi.org/10.1177/0278364908090961

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. https://doi.org/10.1016/j.robot.2011.02.013

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. https://doi.org/10.1016/j.robot.2020.103676

Dimensions. (n.d.). Retrieved May 6, 2022, from https://app.dimensions.ai/discover/publication

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

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), 1745–57. 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., Stumm, E., Nieto, J., Siegwart, R., & Gilitschenski, I. Will it last? Learning stable features for long-term visual localization. In: 2016, 572–581. https://doi.org/10.1109/3DV.2016.66.

Egger, P., Borges, P. V. K., Catt, G., Pfrunder, A., Siegwart, R., & Dubé, R. PoseMap: Lifelong, multi-environment 3D LiDAR localization. In: 2018, 3430–3437. https://doi.org/10.1109/IROS.2018.8593854.

Einhorn, E., & Gross, H.-M. Generic 2D/3D SLAM with NDT maps for lifelong application. In: 2013, 240–247. 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. https://doi.org/10.1016/j.robot.2014.08.008

Fallon, M., Johannsson, H., Kaess, M., & Leonard, J. J. (2013). *The MIT Stata Center data set* (No. *14*) [http://projects.csail.mit.edu/stata/]. MIT. https://doi.org/10.1177/0278364913509035

Filliat, D. A visual bag of words method for interactive qualitative localization and mapping. In: 2007, 3921–3926. https://doi.org/10.1109/ROBOT.2007.364080.

Fontana, G., Matteucci, M., Sorrenti, D. G., Marzorati, D., Giusti, A., Taddei, P., Rizzi, D., & Ceriani, S. (2009). *Bicocca (indoor)* [http://www.rawseeds.org/rs/datasets/view/6]. 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. https://doi.org/10.1109/MRA.2012.2182810

Freitas, V. (2014). *Parsifal*. Retrieved May 12, 2022, from https://parsif.al/

Gadd, M., & Newman, P. Checkout my map: Version control for fleetwide visual localisation. In: *2016-November*. 2016, 5729–5736. https://doi.org/10.1109/IROS.2016.7759843.

Ganti, P., & Waslander, S. L. Network uncertainty informed semantic feature selection for visual SLAM. In: 2019, 121–128. https://doi.org/10.1109/CRV.2019.00024.

Gao, P., & Zhang, H. Long-term place recognition through worst-case graph matching to integrate landmark appearances and spatial relationships. In: 2020, 1070–1076. https://doi.org/10.1109/ICRA40945.2020.9196906.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). *Vision meets robotics: The KITTI dataset* (No. *11*) [http://www.cvlibs.net/datasets/kitti/]. https://doi.org/10.1177/0278364913491297

Glover, A. J. (2014). *Day and night with lateral pose change datasets (Gardens Point Campus of QUT)* [https://wiki.qut.edu.au/display/raq/Day+and+Night+with+Lateral+Pose+Change+Datasets]. Queensland University of Technology. https://doi.org/10.5281/zenodo.4561862

Glover, A. J., Maddern, W. P., Milford, M. J., & Wyeth, G. F. FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day. In: 2010, 3507–3512. https://doi.org/10.1109/ROBOT.2010.5509547.

Griffith, S., & Pradalier, C. (2017). Survey registration for long-term natural environment monitoring. *Journal of Field Robotics*, *34*(1), 188–208. 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. https://doi.org/10.1109/MITS.2010.939925

Hähnel, D. (2001). *Foundation of the Hellenic World (FHW, Athens)* [http://www.ipb.uni-bonn.de/datasets/]. https://doi.org/10.1023/A:1026272605502

Hähnel, D. (2003). *Intel Research Lab (Seattle)* [http://ais.informatik.uni-freiburg.de/slamevaluation/datasets.php]. 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. 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. 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. https://doi.org/10.1109/LRA.2017.2662061

He, L., Wang, X., & Zhang, H. M2DP: A novel 3D point cloud descriptor and its application in loop closure detection. In: 2016, 231–237. https://doi.org/10.1109/IROS.2016.7759060.

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*. https://doi.org/10.1177/02783649221080483

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. https://doi.org/10.1109/JAS.2021.1003907

Ieee xplore. (n.d.). Retrieved May 6, 2022, from https://ieeexplore.ieee.org/Xplore/home.jsp

Inspec. (n.d.). Retrieved May 6, 2022, from https://www.engineeringvillage.com/search/quick.url

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

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. https://doi.org/10.1080/01691864.2013.826410

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. 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* [https://sites.google.com/view/mulran-pr]. Korea Advanced Institute of Science and Technology. https://doi.org/10.1109/ICRA40945.2020.9197298

Konolige, K., & Bowman, J. Towards lifelong visual maps. In: 2009, 1156–1163. https://doi.org/10.1109/IROS.2009.5354121.

Krajník, T., Fentanes, J. P., Mozos, O. M., Duckett, T., Ekekrantz, J., & Hanheide, M. (2014). *Witham Wharf RGB-D (LCAS STRANDS)* [https://lcas.lincoln.ac.uk/nextcloud/shared/datasets/]. Lincoln Centre for Autonomous Systems, University of Lincoln. 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. https://doi.org/10.1109/TRO.2017.2665664

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

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. https://doi.org/10.1016/j.robot.2017.03.016

Lázaro, M. T., Capobianco, R., & Grisetti, G. Efficient long-term mapping in dynamic environments. In: 2018, 153–160. https://doi.org/10.1109/IROS.2018.8594310.

Leung, K. Y. K., Halpern, Y., Barfoot, T. D., & Liu, H. H. T. (2011). *The UTIAS multi-robot cooperative localization and mapping dataset* (No. *8*) [http://asrl.utias.utoronto.ca/datasets/mrclam/]. Autonomous Space Robotics Lab, University of Toronto Institute for Aerospace Studies. https://doi.org/10.1177/0278364911398404

Li, J., Eustice, R. M., & Johnson-Roberson, M. (2015). High-level visual features for underwater place recognition. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 3652–3659. 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. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 11241–7. https://doi.org/10.1109/ICRA48506.2021.9561126

Luthardt, S., Willert, V., & Adamy, J. LLama-SLAM: Learning high-quality visual landmarks for long-term mapping and localization. In: *2018-November*. 2018, 2645–2652. 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. https://doi.org/10.1002/rob.21838

Maddern, W., Pascoe, G., Linegar, C., & Newman, P. (2017). *Oxford RobotCar dataset* (No. *1*) [https://robotcar-dataset.robots.ox.ac.uk/]. Mobile Robotics Group, University of Oxford. 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

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. https://doi.org/10.1109/TMECH.2021.3062647

Milford, M. J., & Wyeth, G. F. [Gordon F.]. (2008). *Mapping St Lucia: openRatSLAM dataset (Brisbane)* [https://wiki.qut.edu.au/display/cyphy/OpenRatSLAM+datasets]. Queensland University of Technology. https://doi.org/10.4225/09/543DE62F1105C

Milford, M. J., & Wyeth, G. F. [Gordon F.]. (2012a). *Alderley Brisbane dataset* [https://wiki.qut.edu.au/display/cyphy/Michael+Milford+Datasets+and+Downloads]. Queensland University of Technology. https://doi.org/10.1109/ICRA.2012.6224623

Milford, M. J., & Wyeth, G. F. [Gordon. F.]. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In: 2012, 1643–1649. https://doi.org/10.1109/ICRA.2012.6224623.

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

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. https://doi.org/10.1109/TRO.2017.2705103

Nanni, L., Ghidoni, S., & Brahnam, S. (2017). Handcrafted vs non-handcrafted features for computer vision classification. *Pattern Recognition*, *71*, 158–172. https://doi.org/10.1016/j.patcog.2017.05.025

Naseer, T., Burgard, W., & Stachniss, C. (2018). *Freiburg Across Seasons (FAS) dataset* (No. *2*) [https://goo.gl/1Jf3kI, https://goo.gl/AvZvjc, https://goo.gl/Y2I6CI]. Albert-Ludwigs-Universität Freiburg. 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. *2017 IEEE International Con-*

ference on Robotics and Automation (ICRA), 2614–20. https://doi.org/10.1109/ICRA.2017.7989305

Naseer, T., Suger, B., Ruhnke, M., & Burgard, W. Vision-based Markov localization across large perceptual changes. In: 2015, 1–6. 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. 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. https://doi.org/10.1109/TRO.2021.3094157

Nguyen, T.-M., Yuan, S., Cao, M., Lyu, Y., Nguyen, T. H., & Xie, L. (2022). *NTU VIRAL: A visual-inertial-ranging-lidar dataset, from an aerial vehicle viewpoint* (No. *3*) [https://ntu-aris.github.io/ntu_viral_dataset/]. Nanyang Technological University. https://doi.org/10.1177/02783649211052312

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. https://doi.org/10.1016/j.robot.2012.12.004

Nobre, F., Heckman, C., Ozog, P., Wolcott, R. W., & Walls, J. M. Online probabilistic change detection in feature-based maps. In: 2018, 3661–3668. 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. https://doi.org/10.1002/rob.20306

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

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. https://doi.org/10.3390/s18040939

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n160

Palazzolo, E., Behley, J., Lottes, P., Giguère, P., & Stachniss, C. (2019). *Bonn RGB-D Dynamic dataset* [http://www.ipb.uni-bonn.de/data/rgbd-dynamic-dataset/]. Universität Bonn. 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. https://doi.org/10.3390/s19194252

Pandey, G., McBride, J. R., & Eustice, R. M. (2011). *Ford Campus vision and lidar data set* (No. *13*) [http://robots.engin.umich.edu/SoftwareData/Ford]. University of Michigan. https://doi.org/10.1177/0278364911400640

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. https://doi.org/10.1007/s10846-015-0198-y

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. https://doi.org/10.1007/s11263-020-01363-6

Pomerleau, F., Krüsi, P., Colas, F., Furgale, P., & Siegwart, R. Long-term 3D map maintenance in dynamic environments. In: 2014, 3712–3719. https://doi.org/10.1109/ICRA.2014.6907397.

Pronobis, A., & Caputo, B. (2009). *COLD: The CoSy localization database* [https://www.cas.kth.se/COLD/]. 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. https://doi.org/10.1016/j.robot.2020.103561

Qin, T., Chen, T., Chen, Y., & Su, Q. AVP-SLAM: Semantic visual mapping and localization for autonomous vehicles in the parking lot. In: 2020, 5939–5945. https://doi.org/10.1109/IROS45743.2020.9340939.

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

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), 1627–44. 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. https://doi.org/10.1002/rob.21620

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. https://doi.org/10.1109/LRA.2016.2516594

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., & Pajdla, T. (2018a). *CMU-Seasons dataset* [https://www.visuallocalization.net/datasets/]. https://doi.org/10.1109/CVPR.2018.00897

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., & Pajdla, T. (2018b). *RobotCar Seasons dataset* [https://www.visuallocalization.net/datasets/]. 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. 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. https://doi.org/10.1016/j.robot.2020.103709

Scopus. (n.d.). Retrieved May 6, 2022, from https://www.scopus.com/search/form.uri

Sheeny, M., De Pellegrin, E., Mukherjee, S., Ahrabian, A., Wang, S., & Wallace, A. (2021). *RADIATE: A radar dataset for automotive perception in bad weather* [http://pro.hw.ac.uk/radiate/]. Heriot-Watt University. https://doi.org/10.1109/ICRA48506.2021.9562089

Singh, G., Wu, M., Lam, S.-K., & Minh, D. V. Hierarchical loop closure detection for long-term visual SLAM with semantic-geometric descriptors. In: *2021-September*. 2021, 2909–2916. 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. https://doi.org/10.1007/s11192-021-03948-5

Siva, S., Nahman, Z., & Zhang, H. Voxel-based representation learning for place recognition based on 3D point clouds. In: 2020, 8351–8357. https://doi.org/10.1109/IROS45743.2020.9340992.

Siva, S., & Zhang, H. Omnidirectional multisensory perception fusion for long-term place recognition. In: 2018, 5175–5181. https://doi.org/10.1109/ICRA.2018.8461042.

Sivic, J., & Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In: 2003, 1470–1477. https://doi.org/10.1109/ICCV.2003.1238663.

Skrede, S. (2013). *Nordlandsbanen: Minute by minute, season by season* [https://nrkbeta.no/2013/01/15/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* (No. *5*) [https://academictorrents.com/details/9e738f5ef5f1412974ab793f315450bc8da76e73]. 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. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3794–801. https://doi.org/10.1109/IROS40897.2019.8967749

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

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

Stachniss, C. (2010). *Albert-b-laser-vision* [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 (TUM RGBD)* [https://vision.in.tum.de/data/datasets/rgbd-dataset]. Technische Universität München. https://doi.org/10.1109/IROS.2012.6385773

Sun, L. [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. https://doi.org/10.1109/LRA.2018.2856268

Sun, L. [Li], Taher, M., Wild, C., Zhao, C., Zhang, Y., Majer, F., Yan, Z., Krajník, T., Prescott, T., & Duckett, T. Robust and long-term monocular teach and repeat navigation using a single-experience map. In: 2021, 2635–2642. https://doi.org/10.1109/IROS51168.2021.9635886.

Taisho, T., & Kanji, T. Mining dcnn landmarks for long-term visual SLAM. In: 2016, 570–576. https://doi.org/10.1109/ROBIO.2016.7866383.

Tang, L. (2017). *YQ21: Dataset for long-term localization* [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. 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). https://doi.org/10.1177/17298814211037497

Thomas, H., Agro, B., Gridseth, M., Zhang, J., & Barfoot, T. D. Self-supervised learning of Lidar segmentation for autonomous indoor navigation. In: *2021-May*. 2021, 14047–14053. https://doi.org/10.1109/ICRA48506.2021.9561701.

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

Uy, M. A., & Lee, G. H. PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition. In: 2018, 4470–4479. 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. https://doi.org/10.1007/s11192-009-0146-3

van Eck, N. J., & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 285–320). Springer. 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. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2774–9. https://doi.org/10.1109/ICRA.2015.7139576

Walcott-Bryant, A., Kaess, M., Johannsson, H., & Leonard, J. J. Dynamic pose graph SLAM: Long-term mapping in low dynamic environments. In: 2012, 1871–1878. https://doi.org/10.1109/IROS.2012.6385561.

Wang, K., Lin, Y., Wang, L., Han, L., Hua, M., Wang, X., Lian, S., & Huang, B. A unified framework for mutual improvement of SLAM and semantic segmentation. In: *2019-May*. 2019, 5224–5230. https://doi.org/10.1109/ICRA.2019.8793499.

Wang, L., Chen, W., & Wang, J. Long-term localization with time series map prediction for mobile robots in dynamic environments. In: *2020-January*. 2020, 8587–8593. https://doi.org/10.1109/IROS45743.2020.9468884.

Wang, Z., Li, S., Cao, M., Chen, H., & Liu, Y. Pole-like objects mapping and long-term robot localization in dynamic urban scenarios. In: 2021, 998–1003. https://doi.org/10.1109/ROBIO54168.2021.9739599.

Web of science. (n.d.). Retrieved May 6, 2022, from https://www.webofscience.com/wos/woscc/basic-search

Xin, Z., Cui, X., Zhang, J., Yang, Y., & Wang, Y. (2017). Visual place recognition with CNNs: From global to partial. *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 6pp.–. 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*. 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. 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. https://doi.org/10.3390/s20082432

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. 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. 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. 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. https://doi.org/10.1109/LRA.2021.3061375

Yin, P., Xu, L., Liu, Z., Li, L., Salman, H., He, Y., Xu, W., Wang, H., & Choset, H. Stabilize an unsupervised feature learning for LiDAR-based place recognition. In: 2018, 1162–1167. https://doi.org/10.1109/IROS.2018.8593562.

Yu, C., Liu, Z., Liu, X.-J., Qiao, F., Wang, Y., Xie, F., Wei, Q., & Yang, Y. A DenseNet feature-based loop closure method for visual SLAM system. In: 2019, 258–265. https://doi.org/10.1109/ROBIO49542.2019.8961714.

Yue, Y., Yang, C., Zhang, J., Wen, M., Wu, Z., Zhang, H., & Wang, D. (2020). Day and night collaborative dynamic mapping in unstructured environment based on multimodal sensors. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2981–7. https://doi.org/10.1109/ICRA40945.2020.9197072

Zhang, G., Yan, X., & Ye, Y. (2019). *Dynamic scenes dataset for visual SLAM (CBD)* [https://doi.org/10.7910/DVN/NZETVT]. Zhengzhou University. https://doi.org/10.1109/ACCESS.2019.2937967

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. https://doi.org/10.1109/TITS.2021.3086822

Zhang, M., Chen, Y., & Li, M. SDF-Loc: Signed distance field based 2D relocalization and map update in dynamic environments. In: *2019-July*. 2019, 1997–2004. https://doi.org/10.23919/acc.2019.8814347.

Zhang, N., Warren, M., & Barfoot, T. D. Learning place-and-time-dependent binary descriptors for long-term visual localization. In: 2018, 828–835. https://doi.org/10.1109/ICRA.2018.8460674.

Zhou, W., Berrio, J. S., De Alvis, C., Shan, M., Worrall, S., Ward, J., & Nebot, E. (2020). *Developing and testing robust autonomy: The University of Sydney Campus data set* (No. *4*) [https://ieee-dataport.org/open-access/usyd-campus-dataset]. University of Sydney. 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. *2018 IEEE Intelligent Vehicles Symposium (IV)*, 1679–85. https://doi.org/10.1109/IVS.2018.8500686

Zhu, S., Zhang, X., Guo, S., Li, J., & Liu, H. Lifelong localization in semi-dynamic environment. In: *2021-May*. 2021, 14389–14395. https://doi.org/10.1109/ICRA48506.2021.9561584.

Ricardo B. Sousa obtained a Master of Science (M.Sc.) degree in Electric and Computers Engineering (ECE) at Faculty of Engineering of the University of Porto (FEUP), in 2020. He is currently working towards the Ph.D. degree in electrical and computer engineering with FEUP, and he has a graduate research scholarship from FCT – Fundação para a Ciência e a Tecnologia at the Centre for Robotics in Industry and Intelligent Systems from INESC TEC. Also, he is an invited assistant lecturing the courses Software Design and Industrial Informatics from the M.Sc. in ECE at FEUP. His research interests include robotics, sensor fusion, and localization and mapping for autonomous robots.

Héber M. Sobreira was born in Leiria, Portugal, in July 1985. He graduated with an M.Sc. degree (2009) and a Ph.D. degree (2017) in Electrical Engineering from the University of Porto. Since 2009, he has been developing his research within the Centre for Robotics in Industry and Intelligent Systems at INESC TEC. His main research areas are navigation and control of indoor autonomous vehicles.

António Paulo Moreira graduated with a degree in electrical engineering at the University of Oporto, in 1986. Then, he pursued graduate studies at University of Porto, obtaining a M.Sc. degree in electrical engineering – systems in 1991 and a Ph.D. degree in electrical engineering in 1998. Presently, he is Associate Professor with tenure at the Faculty of Engineering of the University of Porto and researcher and head of the Centre for Robotics in Industry and Intelligent Systems at INESC TEC. His main research interests are process control and robotics.