robotics

MDPI

*Systematic Review*

# Keyframe Selection for Visual Localization and Mapping Tasks: A Systematic Literature Review

**Nigel Joseph Bandeira Dias** *,†,‡, **Gustavo Teodoro Laureano** ‡ and **Ronaldo Martins Da Costa** ‡

Institute of Informatics, Federal University of Goias (UFG), Goiânia 74690-900, Brazil; gustavo@inf.ufg.br (G.T.L.); ronaldocosta@inf.ufg.br (R.M.D.C.)

* Correspondence: nigeldias@hotmail.com

† Current address: Pixel Lab, Universidade Federal de Goiás Campus Samambaia, Alameda Palmeiras, s/n, Chácaras Califórnia, Goiânia 74690-900, Brazil.

‡ These authors contributed equally to this work.

**Abstract:** Visual localization and mapping algorithms attempt to estimate, from images, geometrical models that explain ego motion and the positions of objects in a real scene. The success of these tasks depends directly on the quality and availability of visual data, since the information is recovered from visual changes in images. Keyframe selection is a commonly used approach to reduce the amount of data to be processed as well as to prevent useless or wrong information to be considered during the optimization. This study aims to identify, analyze, and summarize the methods present in the literature for keyframe selection within the context of visual localization and mapping. We adopt a systematic literature review (SLR) as the basis of our work, built on top of a well-defined methodology. To the best of our knowledge, this is the first review related to this topic. The results show that there is a lack of studies present in the literature that directly address the keyframe selection problem in this application context and a deficiency in the testing and validation of the proposed methods. In addition to these findings, we also propose an updated categorization of the proposed methods on top of the well-discussed categories present in the literature. We believe that this SLR is a step toward developing a body of knowledge in keyframe selection within the context of visual localization and mapping tasks by encouraging the development of more theoretical and less heuristic methods and a systematic testing and validation process.

**Keywords:** keyframe selection; visual localization; visual mapping; visual SLAM; robotics; computer vision

check for updates

**Citation:** Dias, N.J.B.; Laureano, G.T.; Da Costa, R.M. Keyframe Selection for Visual Localization and Mapping Tasks: A Systematic Literature Review. *Robotics* **2023**, *12*, 88. https://doi.org/10.3390/robotics12030088

Academic Editor: Genci Capi

Received: 2 May 2023
Revised: 23 May 2023
Accepted: 27 May 2023
Published: 15 June 2023

## 1. Introduction

Visual localization and mapping can be defined as the tasks of recovering the position and attitude of objects from images and building a virtual representation of the world from multiple scenes relative to a reference frame [1].

Common localization methods are dead reckoning-based, usually supported by inertial sensors, lasers, and tachometers, or rely on Global Positioning Systems (GPS). These methods suffer from classic problems of incremental error accumulation, drift, low accuracy, sparseness or lack of information, and the high cost of better-quality sensors. On the other hand, vision-based methods use cameras, which are highly available and low-cost, have low power consumption, can be easily mounted on other devices, and provide a large amount of data of the scene. Due to these advantages, vision-based methods have become a hot topic in recent years, especially for applications of virtual and augmented reality, robotics, and autonomous driving [1–3].

Mapping and localization are essential in any application that requires dealing with environment perception and position estimation relative to its surroundings [4]. Despite being considered independent tasks, they are closely related. For an agent (e.g., smartphone, human, vehicle, or robot) to be able to self-localize in an environment, a map of its surroundings should be provided, while to create a map of the environment, its pose should be known. When the agent does not have access to a map and does not know its own pose, the problem is addressed by Simultaneous Localization and Mapping (SLAM), where both localization and mapping are executed simultaneously. SLAM provides better results when compared to solving both localization and mapping independently.

Vision-based methods are based on geometric estimation models of the world that explain its changes from perceived visual data [5]. These models are usually solved as a nonlinear optimization problem whose complexity depends on the number of poses and perceived data to be fitted. As a result of the abundant information provided by the images, its processing can be computationally expensive, which makes the execution of these applications in real time quite challenging. One strategy to attenuate the computation load is by not processing all the frames, since consecutive frames from a video sequence are usually captured with a high degree of overlap; Figure 1.

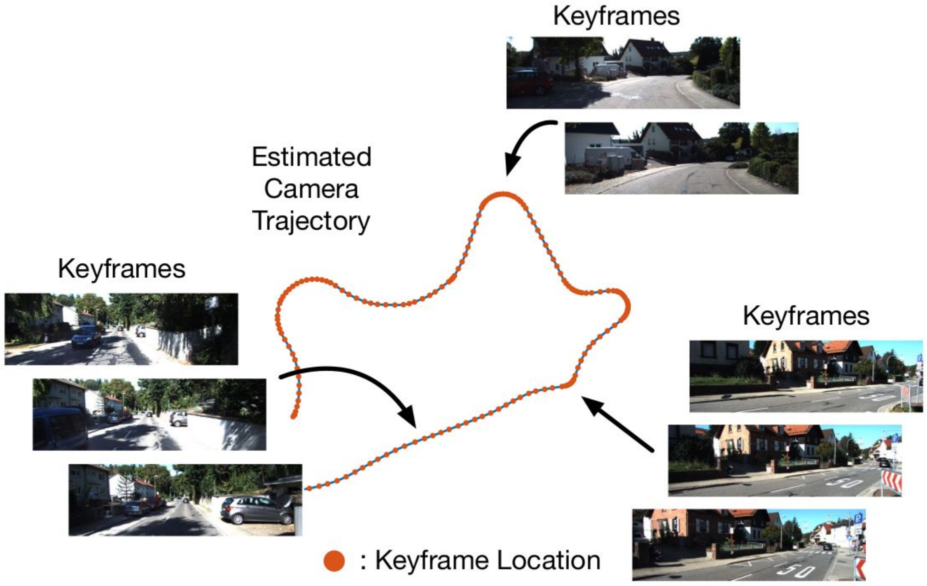

**Figure 1.** Illustration of selected keyframes from a vehicle's captured video [6]. The keyframes are uniformly distributed when the car drives straight along the street, and more frequently at the turning corners because the consecutive frames tend to be less similar.

The subset of a video sequence that can represent its visual content as closely as possible with a reduced amount of frame information is usually called a keyframe set. The keyframe selection problem has been extensively studied in video analysis and image retrieval applications, where videos are segmented into shots to aid its indexing, annotation, compression or retrieval from a database given a similar query image [7]. In vision-based localization and mapping methods, keyframes can be selected to achieve sufficient visual coverage of the environment while keeping its representation simple for computational efficiency. Moreover, by carefully selecting this subset of keyframes, we can prevent useless or wrong information from being considered during the optimization, thus avoiding degenerative configurations from ill-conditioned systems; Figure 2.

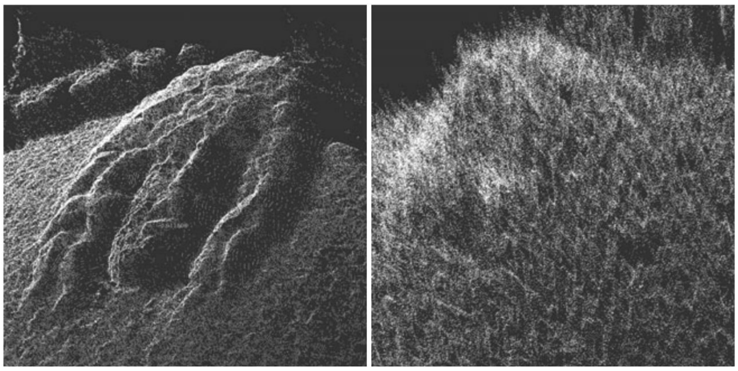

**Figure 2.** This figure illustrates the result of dense pairwise reconstruction. The **left image** shows the result obtained from a carefully selected pair of keyframes, while the **right image** shows the result from a degenerate pair [8].

Despite being considered an essential factor for the visual localization and mapping algorithm performance [3], keyframe selection does not receive as much attention as other key techniques such as feature detection and matching [9], uncertainty modeling [1], and map modeling [10]. Executing pilot searches on the most popular digital libraries in the area of computer science and robotics, we could not find any comprehensive reviews related to this topic. There are few studies that completely cover the problem of keyframe selection for these applications, even though several studies employ the keyframe selection strategy in their algorithms. Furthermore, the design of the methods is still much less theoretical and more heuristic. The lack of studies about keyframe selection for visual localization and mapping tasks motivated us to summarize all existing information about this topic in a careful and unbiased manner.

In this paper, we present a systematic literature review (SLR) aiming to identify and analyze the commonly used approaches for keyframe selection within the context of visual localization and mapping. The results of this SLR are a step toward developing a deep understanding of this topic and assisting in the formulation of new hypotheses, derived from a well-defined methodology. Moreover, bring up a discussion about the importance of keyframe selection in visual localization and mapping tasks and how this area can be deeply investigated.

The remainder of this paper is organized as follows: Section 2 presents some background to visual localization and mapping and related work. Section 3 presents the research methodology adopted to conduct this SLR. The results and discussions related to the research questions are presented in Section 4. Finally, in Section 5, we present the conclusions and findings of this review.

## 2. Background and Related Work

To the best of our knowledge, this is the first review with a specific focus on keyframe selection for visual localization and mapping tasks. In this section, we first present a brief overview of visual localization and mapping algorithms, then present closely related work on keyframe selection present in the literature.

### *2.1. Visual Localization and Mapping*

Mapping and localization are essential problems in any application that requires an agent (e.g., a smartphone, human, vehicle, or robot) to localize itself in an unknown environment [4]. Mapping is the process of generating a globally consistent model of the environment from local observations of it, while localization is the process of estimating the pose of the agent within the map according to the sensor data perceived from the environment. Even though they can be executed as independent tasks, they are closely related. In order to build a map, the agent and the structure poses need to be known in order to compute the 3D position of the objects in the scene, while in localization, a map of the environment must be available so that the pose can be computed against the map's reference. To solve this egg-and-chicken problem, simultaneous localization and mapping (SLAM) techniques have been proposed. In SLAM, both localization and mapping are executed simultaneously, where a map of an unknown environment is created incrementally while localizing the agent within the map.

In computer vision, the problem of recovering relative camera poses and the 3D structure of the environment from images is known as structure From motion (SFM). Usually, SFM is formulated as a Perspective-n-Point problem, where both the 3D reconstruction of the scene and the 3D motion of the camera are estimated from sequentially ordered or unordered image sets [5]. The final scene structure and camera poses are refined with an offline optimization called bundle adjustment, in which the camera parameters and 3D points are adjusted to minimize every error term in the process, as much as possible, in a batch manner [5]. In general, all the image-based localization and mapping algorithms have their principles derived from SFM applications.

Visual odometry (VO) is a particular case of SFM that focuses on estimating the egomotion of an agent through examination of the changes that motion induces on images captured by a single or multiple cameras attached to it. Differently from SFM, in VO, we are not concerned with the 3D structure; the camera's 3D motion is incrementally estimated (i.e., as new frames arrive), and it is usually required to be executed in real time [11]. The term VO was coined by Nister [12] due to its similarity to wheel odometry, which estimates the position of a wheeled agent by integrating the number of turns of its wheels over time. The VO process has the advantage of not being affected by wheel slip, and it is independent of the agent kinematics. VO algorithms can be classified, based on the input data, into direct [13] and feature-based [14] methods. Direct methods use image pixel intensity for motion estimation by photometric error minimization, whereas the feature-based approach relies on the geometric consistency of features extracted from the images, such as SIFT [15], ORB [16], and SURF [17].

VO is a dead reckoning process, since the camera's 3D motion is incrementally estimated by integrating the current estimation with the previous estimation. Consequently, the VO process is prone to drift accumulation due to errors introduced by each frame-to-frame motion estimation. A straightforward solution to keep drift as small as possible is through a local optimization over the last $n$ camera poses, known as sliding window bundle adjustment or windowed bundle adjustment. Even though the drift could be considerably reduced using this approach, we still obtain a local consistency of the camera trajectory. To obtain a global and consistent estimate of the camera trajectory, a global map optimization is normally performed using a SLAM strategy. The visual SLAM (VSLAM) approach is a particular case of SLAM that has only visual information as input. By keeping the track of the map, the system can detect when the agent has returned to a previously visited area and the accumulated error from the first to the current frame can be computed. This process is defined as loop closure, in which the current frame is matched with the previously acquired images, and if a closed loop is detected (i.e., the camera captures one of the previously observed views), loop constraints are used to suppress the error in the global optimization in a global bundle adjustment procedure [2].

The VSLAM algorithms can be divided into two main blocks: the front-end module, which normally implements a VO strategy to recover the incremental motion of the camera; and the back-end module, which executes the global optimization step whenever a loop closing is detected to obtain a globally consistent map. The front end is more related to the computer vision research fields, while the back end is essentially a state estimation problem that can be implemented using filter-based approaches such as the extended Kalman filter (EKF) or nonlinear optimization methods such as bundle adjustment (BA) [3]. The topics addressed in this review are more related to the BA-based VSLAM methods, also known as keyframe-based VSLAM. These methods formulate the VSLAM as a graph optimization problem, where the nodes represent all the poses and all the features in the map, and edges correspond to nonlinear constraints. The graph leads to a sum of nonlinear quadratic constraints that, when optimized, yields to a maximum likelihood map and a corresponding set of camera poses [4].

### *2.2. Related Work*

The use of visual sensors for localization and mapping tasks has become an active research topic due to their high accuracy, low cost, and abundant data information. Moreover, it has attracted much attention from both the academic and industrial communities due to its potential applications and problem challenges, as well as because cameras are sensors suitable to be mounted on devices such as tablets, smartphones, and robots. Therefore, image-based localization and mapping algorithms have great and widespread applications, such as virtual reality, augmented reality, robotics, and autonomous driving [2].

Due to the amount of data provided by images, their processing can be computationally expensive. On the other hand, frames from a video sequence are usually captured continuously with a high degree of overlap; thus, there is no need to process all of them. The subset of a video sequence that can represent its visual content as closely as possible with a reduced number of frame information is usually called keyframes. The keyframe selection problem has been extensively studied in video analysis and content-based image retrieval (CBIR) applications [7]. The former comprises methods that are concerned with the problem of segmenting a video into shots for its indexing, annotation, or compression, whereas the latter consists of methods that address the issue of retrieving images from a database that are similar to a query image based on an analysis of their contents. In general, keyframe selection techniques can be broadly classified into global [18] and sequential [19] data analysis. Usually, global analysis employs cluster or energy-based methods to examine the entire video sequence; therefore, they are not applicable to online applications such as visual localization and mapping. Contrarily, sequential approaches perform the analysis considering the video frames one at a time, which is more appropriate for online applications.

The complexity of the algorithms for visual localization and mapping that relies on bundle adjustment such as nonlinear optimization depends on the number of poses and features to be optimized. Thus, keyframe detection plays an important role in order to achieve a sufficient visual coverage of the robot's environment while keeping the representation simple for computational efficiency. Moreover, keyframe selection can be used as a filter to prevent useless or wrong information from being considered in the optimization, avoiding an ill-conditioned system [20].

Even though keyframe selection is considered one of the critical techniques in VSLAM algorithms [3], no reviews or surveys about this topic were found in automated searches. A low number of articles were found that review the most commonly used methods; some even proposed some categorizations, but without giving an in-depth analysis of all methods present in the literature [20–23]. Zhang et al. [20] presented one of the first and more complete studies concerned with the problem of keyframe selection for appearance-based visual SLAM. They investigated the applicability of keyframe selection techniques for CBIR applications in SLAM and a systematic comparison of the proposed methods. Nonetheless, they focused in only one category of keyframe detection methods.

## 3. Systematic Literature Review Methodology

This research is based on the guidelines proposed by Kitchenham and Charters [24] for undertaking systematic literature reviews (SLR). The proposed methodology instructs on how to perform an SLR that is appropriate to the needs of software engineering researchers, which the authors argue is not as generally rigorous as those used by medical researchers.

According to the guidelines proposed by Kitchenham and Charters [24] the stages in a systematic review can be summarized into three main phases: planning the review, conducting the review, and reporting the review. The stages associated with each phase will be addressed in the following sections. Although they will be addressed sequentially, many of the stages involve iteration, e.g., activities initiated during the planning phase could be refined during the conduction phase.

### *3.1. Planning the Review*

The main stage of the planning is the development of the review protocol, which specifies the methods that will be used to undertake the systematic review. The motivation for developing a review protocol is to reduce the possibility of research bias. For this research, the review protocol consists of the definition of the objectives, research questions, keywords and synonyms, selection sources, search strings, the inclusion and exclusion criteria, a quality assessment checklist, and data extraction forms.

The purpose of this systematic review is to identify and analyze the approaches used for keyframe selection within the context of visual localization and mapping. Although keyframe selection is a common step in those tasks, the design of the methods is still much less theoretical and more heuristic. Furthermore, there is not a clear and consistent definition of the keyframe in those contexts. Executing pilot searches on the most popular digital libraries in the areas of computer science and robotics, we could not find any SLRs or surveys related to this topic. This was a motivation to summarize all existing information about keyframe selection techniques for visual localization and mapping tasks in a careful and unbiased manner. With the collected information, we aim to answer the research questions described in Table 1.

**Table 1.** Research Questions and Motivation.

| Research Question | Motivation |
|---|---|
| **RQ1.** Which techniques are being used most for keyframe detection in visual localization and mapping tasks? | To identify and analyze the primary keyframe selection methods proposed in the literature. |
| **RQ2.** Which kinds of properties are used in keyframe classification? | To identify which kind of information the methods use to classify a frame as a keyframe. |
| **RQ3.** What is the role of keyframes in visual localization and mapping tasks? | To analyze the benefits of the keyframe selection stage on the pipeline of those tasks. |
| **RQ4.** How can we evaluate the keyframe selection method? | To identify methods and metrics used to quantitatively or qualitatively evaluate the keyframe selection methods. |

From the pilot searches, we could identify some primary studies that addressed most of the research questions. These studies were used to extract some keywords, synonyms, and alternative spellings that were used to define the search string. The search method consisted of web searches in digital libraries; the selected resources chosen were ACM Digital Library, IEEE Digital Library, Science Direct, and Springer Link.

Given the similarity with content-based image retrieval and video summarization studies, we also conducted pilot searches to refine the search string iteratively. Moreover, we realized that there could be relatively few studies that directly address the topic, so we decided not to restrict the search too much. The process consisted of excluding keywords whose inclusion did not return additional papers or the primary studies previously found while trying to retain a reasonable number of returned papers. After several iterations, we defined the following search string used to search within keywords, titles, abstracts, and full text of publications:

(“keyframe selection” OR “keyframe extraction” OR “keyframe detection”) AND (visual OR localization OR SLAM) NOT “video summarization”).

From the pilot’s search it was observed that selection, detection, and extraction are commonly used as synonyms. Studies related to video summarization applications were excluded because the extraction goals are different from the tasks of localization and mapping.

To identify the primary studies that provide direct evidence about the research questions, we defined some study selection criteria. They were defined during the protocol definition and refined during the search process. In summary, we decided to include any study that directly or indirectly addresses the problem of keyframe selection for visual localization and mapping tasks, as well as related issues, e.g., structure from motion, 3D reconstruction, and VR/AR. For exclusion, the following criteria were used:

- Studies that do not address the problem of keyframe selection;
- Studies that do not describe the methodology used for keyframe selection;
- Studies whose methods are not visual-based;
- Redundant papers of the same authorship.

In addition to the selection criteria, a quality checklist was created to assess each included article, in which the quality of the paper was measured based on the questions listed in Table 2. Following the guidelines proposed by Kitchenham and Charter [24], the questionnaire was elaborated with the aim of investigating whether quality differences explain differences in study results, as a means of weighting the importance of individual studies and to guide recommendations for further research.

**Table 2.** Study quality assessment questions.

| ID | Questions |
|---|---|
| QA1 | *Does the study directly address the problem of keyframe selection?* |
| QA2 | *Does the study answer our research questions?* |
| QA3 | *Are the methods clearly defined?* |
| QA4 | *Do they use any metric to validate the method?* |
| QA5 | *Do they use a specific datasets for validation?* |
| QA6 | *Does the study compare the proposed method against other methods?* |
| QA7 | *Does the study discuss the results obtained with the proposed method?* |

To address the quality assessment, we designed a digital extraction form to collect all the information needed. Moreover, the data extracted grant that we accurately record any information needed to answer the research questions. The fields of the data extraction form and their description are presented in Table 3.

**Table 3.** Data extraction form.

| Data Field | Description |
|---|---|
| Title, year, authors | Main information about the publication. |
| Source | Name of the source where the article was published, e.g., journal, conference, magazine, etc. |
| Key terms | Terms that are used to synthesize the main ideas of the study. |
| Objective of the study | The main purpose of the work, what kind of problem they are trying to solve. |
| Method description | Detailed description of the proposed keyframe selection method to differentiate studies that use the same approach. |
| Motivation | The motivation behind the proposed keyframe selection method to identify the role of the proposed method. |
| Materials | List of the materials used in the experiments, e.g., datasets, sensors, metrics, etc. |
| Experiment description | Detailed description about the experiments' conduction and validation. |

### *3.2. Conducting the Review*

We adopted the Parsifal online tool [25] to support the protocol definition and SLR conduction, which was designed within the context of software engineering. As pointed out earlier, most of the stages in the conducting phase were defined and piloted in the planning phase and refined during the SLR conduction. This phase consisted of the following stages: study selection, data extraction and synthesis, and study quality assessment.

The pilot searches executed during the search string definition indicate that there could be relatively few studies that directly address the problem of keyframe selection for visual localization and mapping tasks. Therefore, the search string was defined to retrieve a reasonable number of articles that could be filtered in a multistage process. The number of articles included in each stage is presented in Table 4.

**Table 4.** Number of articles selected per stage.

| Library | Stage 1 | Stage 2 | Stage 3 |
|---|---|---|---|
| ACM Digital Library | 131 | 16 | 6 |
| IEEE Digital Library | 64 | 29 | 26 |
| ScienceDirect | 93 | 18 | 11 |
| Springer Link | 210 | 31 | 18 |
| **Total** | **498** | **94** | **61** |

The first stage consists of the result obtained from the automatic search in each digital library; no filters were applied in this stage. With the Parsifal tool, we can import *bibtex* files, which allow information about the article publication and the abstract to be retrieved without having to access a full copy of the article. Some libraries do not support *bibtex* files; thus, it was necessary to manually import that information into Parsifal.

The filtering process begins in the second stage, where duplicated articles were removed, and by reading the titles and abstracts, we applied the inclusion and exclusion criteria. For a fraction of the imported studies, the abstracts provided sufficient information to decide whether to include or not. For those articles that we could not apply the selection criteria to by reading the abstracts, we downloaded a full copy for more detailed analysis in the next stage.

During the final stage, we also applied the inclusion and exclusion criteria after reading the introduction, the methodologies, and conclusions sections of the articles. At the end of this stage, 61 primary studies were selected for this review, in which we executed the data extraction and analysis, as well as the quality assessment. This process is detailed in Figure A1, accordingly to the PRISMA flow diagram [26].

The definitions of the data to be extracted that are listed in Table 3 were based on the research questions and other requirements for this review. As already mentioned, the article's metadata were automatically extracted from the Parsifal online tool. The metadata consist of the title of the article, abstract, keywords, the source where it was published, the year of publication, the names of the authors, and the DOI. Since the data were automatically detected, they were subject to errors; thus, each included article was individually revised, and the needed corrections were made.

A full reading and detailed extraction were performed to complete the remaining fields of the data extraction form. The *key terms* were used to synthesize the main ideas extracted from the reading; in contrast to the author's keywords, they are exclusively related to the purpose of this review. The *objective of the study* is to provide context to where the keyframe selection was applied; as pointed out earlier, some primary studies do not directly address the problem of keyframe selection, but they address it within a given context as a step to solve a specific problem. The *method description* contains details regarding the proposed keyframe selection method, describing its main characteristics and structure, with enough information to differentiate between studies with the same approach. The *motivation* is related to the objective of the study; it is what motivated the authors to apply the keyframe selection step as means of identifying the role of the proposed method in the pipeline. The *materials* field was used to list everything used in the validation of the proposed method, e.g., datasets used, metrics, methods used for comparison, and the kind of sensor used. The *experimental description* field contained detailed information about the experiment's conduction and validation. It is where we described how the materials were used to validate the proposed method, which tasks were executed, relevant information about the setup, and the key results obtained with the proposed method.

Besides the data extraction, we assessed the quality of the primary studies based on the quality assessment (QA) checklist presented in Table 2, to assist the data analysis and synthesis. The QA questions were judged against three possible answers with a defined weight score: "Yes" with a weight of 1, "Partially" with a weight of 0.5, and "No" with a weight of 0. These weights were defined during the planning phase, and the quality score for a particular study was computed by taking the sum of the scores of each answer. The main goal of the quality assessment was to evaluate the credibility, completeness, and relevance of the study regarding the research questions. The articles with higher scores are more relevant to this review.

QA1, "*Does the study directly address the problem of keyframe selection?*" aimed to identify the articles that we classify as core studies that have the potential to answer most of the research questions. As already mentioned, keyframe selection is commonly addressed as a step in the main task (e.g., visual odometry, visual SLAM, SfM, etc.), and consequently, it could not be a direct evaluation of the proposed method. Additionally, QA2 "*Does the study answer our research questions?*" was defined as a direct measurement of the relevance of the study for this review.

Following the same idea, QA3, "*Are the methods clearly defined?*" aimed to assess the rationales for the proposed method. Studies with a higher score on this question were those in which the authors explained or justified the motivation for the proposed keyframe selection method. This subset of study allows for an understanding of the role of keyframe selection, the techniques that are being used most, and the properties that are used to classify a frame as a keyframe within the context of those main tasks.

The classification of a frame as a keyframe can be a subjective process that depends on the application context and its goals, which makes the evaluation of the proposed method quite challenging. The remaining questions aimed to assess the studies regarding how they are addressing this problem. QA4 "*Do they use any process to validate the method?*" aimed to assess the means used to evaluate the proposed method. Studies that answered this question positively were those that presented a process to directly evaluate the implementation of the proposed method for keyframe selection. Complementarily, QA5, "*Do they use a specific dataset to validate the proposed method?*" allows for identification of whether the evaluation process can be replicated given the metrics used.

QA6, "*Does the study compare the proposed method against other methods?*" aimed to identify whether the results were compared to others. The two last QAs are related to the tests and the validations of the implemented method, showing their experimental measurements. Nonetheless, it is important to define a standardized comparison process to identify the most suited solution for each application. Finally, QA7, "*Does the study discuss the results obtained with the proposed method?*" aimed to identify whether the authors qualitatively or quantitatively explain the meaning of the presented measurements and results.

Table A1 contains the checklist used for the quality assessment of the included articles and also the meaning of the three possible answers for each questions.

### *3.3. Risk of Bias*

In contrast to traditional reviews, the systematic literature review (SLR) aims to provide scientific rigor, as it is built upon a well-defined methodology. However, despite following a predefined review protocol, there is still a risk of introducing bias into our conclusions.

When selecting libraries for searching, we chose those considered the most popular in the fields of computer science and robotics. It is highly likely that relevant articles have been published in various sources. Nevertheless, the described protocol contains all the necessary information to expand this review by incorporating new libraries and aggregating the results.

Although pilots' research was conducted to define and refine the search string, ensuring its accuracy remains a challenge. Similarly, while the criteria decisions were designed to minimize subjectivity, applying them objectively proved difficult due to the lack of direct attention given to the keyframe selection problem in most studies. To mitigate these biases, the entire research protocol underwent validation, with consensus among the authors.

## 4. Results and Discussion

This section presents the results of the proposed methodology and a discussion of the research questions presented in Section 3.1. The selection process in the course of the conduction phase of this review resulted in 61 studies that attend the inclusion criteria. To the best of our knowledge, combined with the results of the selection process, there is no systematic literature review or conventional reviews related to this topic.

We categorized the included studies into four primary research areas: visual SLAM, localization, visual odometry, and 3D reconstruction. Most of the principles of these areas are derived from the structure from motion area; in that sense, we decided to not include it in this categorization. The localization research area comprises studies that do not perform the dead reckoning or mapping processes, such as appearance-based localization methods.

Analyzing the extracted data (Figure 3), we can see that 37 papers belong to the visual SLAM research area, which represents 61% of the included articles. This can be explained by the fact that keyframe selection is a crucial step in keyframe-based SLAM approaches, since it decides when to introduce a new node into the topological map [20]. Combining this information with the temporal view of the included articles presented in Figure 4, we can infer some patterns. As we can see, the included articles were published between 2004 and 2021, and there was an increase in the number of studies on this topic from 2010. In that year, the work of [27] was published, which received the best paper award at the ICRA conference. In that study, the authors performed a series of experiments and a rigorous analysis comparing filter-based and keyframe-based SLAM approaches, where they pointed out that keyframe-based approaches give the most accuracy per unit of computing time. Since then, keyframe-based SLAM approaches have received considerable attention, notably after 2016, when ORB-SLAM2 [28] was published, which is considered one of the state-of-the-art methods for stereo visual SLAM.

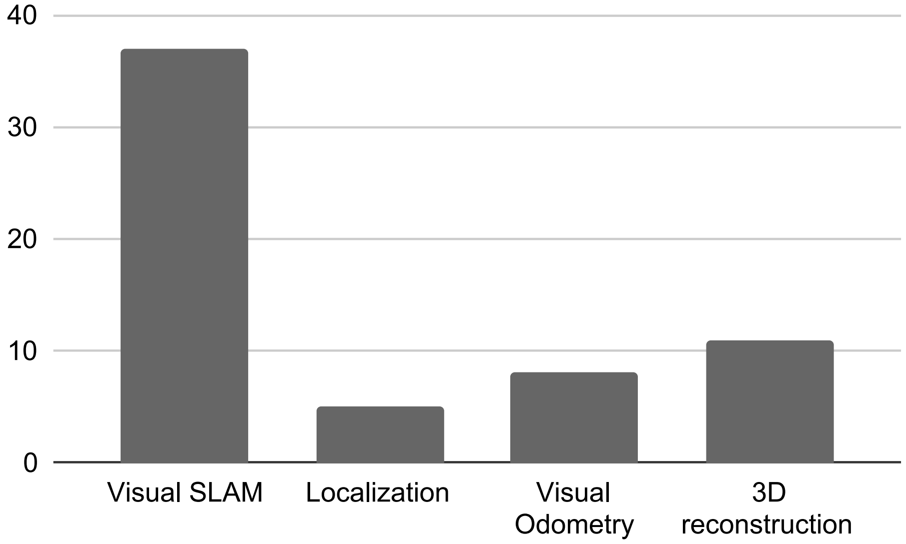

**Figure 3.** Number of included articles per research area.

The quality assessment helped to weigh the importance of individual studies to assist the data analysis and synthesis of the included articles, regarding the objectives of this SLR. The results are presented in Table 5 according to the assessment questions described in Table A1. It was noticed that most of the answers were negative, which means that information provided by authors about the keyframe selection methods and their evaluation were poorly reported or did not directly address the problem. This is explicitly observed by QA2, which is "*Does the study directly address the problem of keyframe selection*", with only 16% of positive answers. Most of the included articles did not focus on the keyframe selection method; it is addressed as an intermediary stage that is not directly evaluated.

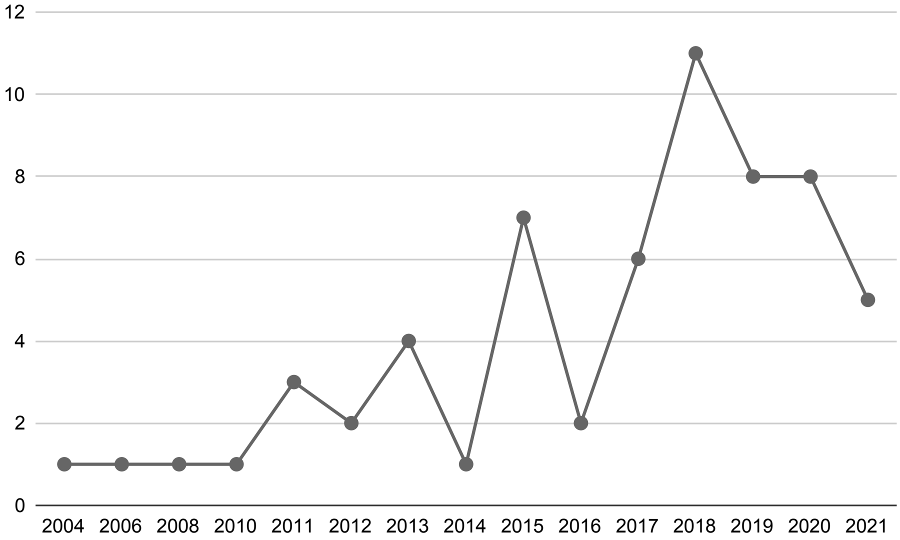

**Figure 4.** Number of included articles per year of publication.

QA5, "*Do they use specific datasets for validation*" received the lowest number of positive answers, 7%; and the highest number of negative answers, 77%. This is because most of the included papers evaluated their methods on commonly used datasets for other tasks, such as visual odometry, visual SLAM, etc. This is also related to the fact that in most of the studies, keyframe selection method is not the main focus. A minority of the studies used or proposed datasets designed specifically for the evaluation of keyframe selection methods.

The question with the second-lowest number of positive answers, with only 10%, was QA2, "*Does the study answer our research questions?*", which was expected, since the research questions were defined based on the goals of this review. Even though these goals should be related to the article's objectives, they are not expected to be precisely the same. The only question positively answered was QA3, "*Are the methods clearly defined?*", with 72% of positive answers, a considerable number. This means that the majority of the included articles presented and explained the motivation for the proposed method of keyframe selection.

In the following sections, we will present and discuss the results of each research question.

**Table 5.** Study quality assessment results.

| QA ID | Yes (%) | Partially (%) | No (%) |
|---|---|---|---|
| QA1 | 10 (16%) | **28 (46%)** | 23 (38%) |
| QA2 | 6 (10%) | 26 (43%) | **29 (48%)** |
| QA3 | **44 (72%)** | 14 (23%) | 3 (5%) |
| QA4 | 10 (16%) | 25 (41%) | **26 (43%)** |
| QA5 | 4 (7%) | 10 (16%) | **47 (77%)** |
| QA6 | 11 (18%) | 17 (28%) | **33 (54%)** |
| QA7 | 20 (33%) | **22 (36%)** | 19 (31%) |

### *4.1. RQ1: Which Techniques Are Being Used Most for Keyframe Detection in Visual Localization and Mapping Tasks?*

Keyframe detection is a known problem in the research areas of video analysis and summarization, where it is used as a mechanism for generating a short summary of a video [7]. A variety of methods have been proposed based on the measure of similarity between images. However, not all of them are directly applicable to visual localization and mapping problems, since some of these approaches require the examination of the entire video sequence, which is not feasible in online applications.

The main objective of keyframe detection in visual localization and mapping tasks is to achieve sufficient visual coverage of the robot's environment while decreasing the amount of redundant information. This means that the selected keyframe should be sufficiently different from the previously detected one, but with enough overlap to guarantee the tracking. Due to their success in video analysis applications, image similarity as a metric of keyframe detection has become a standard approach for visual localization and mapping tasks [20]. However, the notion of image similarity can be subjective and context-dependent; thus, most of the approaches are heuristics-based. In this SLR, we propose the following categorization: heuristics-based methods, probabilistic-based methods, and learning-based methods. In Figure 5, we present a diagram illustrating the new categorization proposed in this work. As shown in Table 6, 83.61% of the included articles are heuristic-based, 14.75% are probabilistic-based methods, and only one article uses learning-based approaches to detect keyframes. The following subsections present the included studies and the proposed approaches for each category.

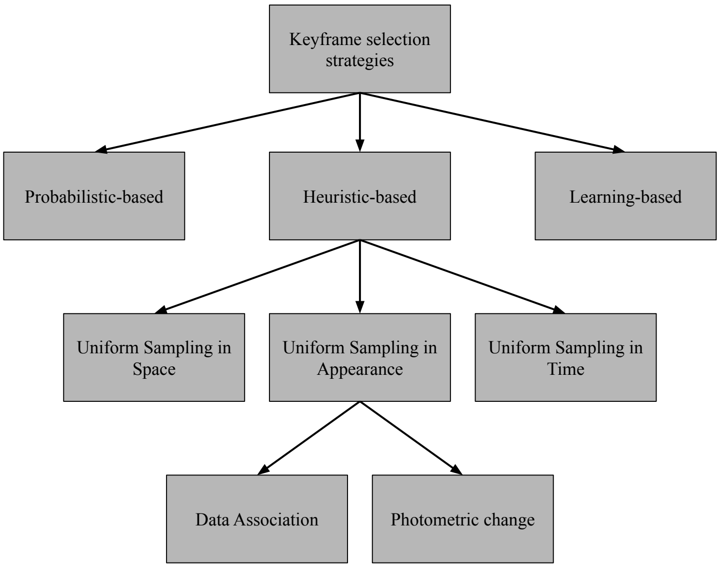

**Figure 5.** Keyframe selection strategies for visual localization and mapping tasks.

**Table 6.** Number of included articles per method.

| **Method** | **Number (#)** | **Percentage (%)** |
|---|---|---|
| Heuristic-based | 51 | 83.61 |
| Probabilistic-based | 9 | 14.75 |
| Learning-based | 1 | 1.64 |

#### 4.1.1. Heuristic-Based Methods

Most of the studies argue that a keyframe should be sufficiently different from the previously detected one in order to achieve an ample visual representation of the environment while reducing the amount of redundant information between frames. Therefore, the methods are based on heuristics that usually rely on certain assumptions about the camera view change and the environment. As pointed out in [20], the commonly used approaches can be divided into uniform sampling in space, uniform sampling in time, and uniform sampling in appearance.

Uniform sampling in time approaches aim to select a keyframe at every *n*th frame captured by the camera. These methods assume an acceptable correlation between the time interval between frames and the appearance change of the environment. Usually, they are not concerned with the selection itself, but reducing the number of frames to be processed without increasing the system's complexity. For instance, in [29,30], the proposed methods focus on the data structure regardless of the subsampling mechanism. The authors argue that the robustness of loop closure detection depends on the representation of the candidate's keyframes, and an efficient image query process is better than sophisticated heuristics for data sampling.

Distance-based sampling approaches comprise the methods that select a keyframe at every unit of linear or angular distance traveled by the camera. These methods assume a good correlation between appearance change and spatial change, which is highly sensitive to the unknown geometry of the scene [20]. The distance can be inferred directly from the camera egomotion or from any external source, for example, an odometer. In [31], the authors proposed to combine the robot's linear and angular velocity with the camera frame rate to define a time interval to extract keyframes, while in [32], the authors decided to use the IMU preintegration data to infer the parallax between frames. Other methods, such as that in [33], proposed to use the robot control feedback as a constraint for rough positioning of the keyframes, and then an accurate positioning was obtained based on several criteria. The commonly used approaches infer the distance traveled between frames based on the estimated camera motion [34–44]. In these cases, the distance is computed as a Euclidean distance between the estimated projection matrix, it can be either considering the camera's center or the rotation matrix and translation vector.

The appearance-based sampling approaches are expected to be more practical, considering that they directly measure the view change of camera observation with respect to previously detected keyframes. This requires a means of assessing image similarity, which can be subjective and context-dependent, since the images could not be exactly the same. Therefore, most of the studies proposed engineering heuristics that seek a reasonable tradeoff between precision, recall, and computation time. The most popular methods used two main factors to infer the view change of camera observation: data association and photometric change.

The methods based on data association are widely used due to the success of the feature-based methods for visual SLAM and visual Odometry. It is worth mentioning that these methods are considered appearance-based methods by the fact that the appearance change between frames is associated with the change in data association. In general, the data association is constructed by means of feature matching or feature tracking; however, the criteria used for keyframe selection are distinctive. The majority used the number of features matched/tracked or the inlier ratio between camera observation and previous frames, or the existing map, for view change thresholding [45–51]. Other methods used feature correspondence to construct auxiliary graph structures to determine the degree of overlap between the camera observation and the previously detected keyframes [21,52,53]. Even though the feature correspondence is able to measure the degree of similarity or overlap between frames, it is highly sensitive to the distribution of the detected features. Therefore, others argue that feature flow-based metrics for similarity are more suitable to detect the changes in the scene [54]. Likewise, other studies claim that only points with enough flow magnitude provide information about both translation and rotation motions; thus, the number of points with sufficient displacement can be used as a criterion for keyframe selection, as proposed in [55].

On the other hand, methods based on photometric change are more popular in direct-based methods for visual SLAM and visual odometry. These methods rely on the assumption that the similarity between frames can be measured from the photometric change between the images. Moreover, by analyzing the pixel intensities, it is possible to determine the degree of informativeness and the quality of the frames. This is important, considering that changes in brightness due to camera exposure time, lighting conditions, and fast motion blur can make tracking against the previous keyframe difficult [56,57]. The simplest—and a fairly naive—approach is to use a pixelwise difference as means of measuring the similarity between frames. However, this could not be effective for keyframe selection, due to image noise and pixel correlation carelessness [20]. A more sophisticated approach is to use the structural similarity (SSIM) metric, which relies on the assumption that pixels have strong interdependencies, especially when they are spatially close, as proposed in [58]. Others characterize the images in terms of their histogram, and the similarity is inferred according to some metric of the distance between the histograms. In this case, it could be a global histogram that captures the global intensity, or a local histogram that computes the histogram of subregions of the image [20,59]. The global histogram loses the spatial information, within which the difference between two frames may exist, and this is partially alleviated through the use of local histograms. One advantage of the photometric change approach over the data association approach is that the similarity or the informativeness of the frame can be inferred without the extraction of features, which is a time-consuming process. In [57], the authors proposed to use the difference-of-Gaussian (DoG) filter as a distinctiveness detector to represent the informativeness of each frame. Experimental results show that the proposed keyframe measure is positively proportional to the number of SIFT features.

All the methods above rely on the assumption that if the similarity between the camera observation and the previous keyframe drops below some constant threshold, then the frame can be declared as a new keyframe. This requires careful tuning of the threshold to achieve an optimal result in different scenarios; thus, some studies have proposed methods that define the threshold adaptively. In [32], the authors proposed a model to adjust the threshold according to the situation. The proposed model comprises a weight factor of the velocity, position, and rotation information from IMU preintegrated data, which are compared to the threshold and updated in an iterative process. Alternatively, in [23], the authors proposed to use the feedback of a designed PD controller to define to threshold dynamically. The PD controller input error is computed as the difference between a defined ideal view change and an estimated view change with regard to the build map.

Alternatively to the employment of individual heuristics, there are studies that propose to combine different metrics for keyframe decisions. For instance, in [60], the authors proposed to use a multidimensional weighted cost function to select the keyframe with minimum value.

#### 4.1.2. Probabilistic-Based Methods

Probabilistic-based methods are those that propose the design of the keyframe selection mechanisms that are much less heuristic and more theoretical. In these cases, there are no heuristic rules or assumptions about the scene; the decision is made based on the informativeness of the frame regarding the geometry of the scene in a probabilistic fashion.

Besides the benefit of reducing the amount of data to be processed, the keyframe selection can be applied to discard frames that could potentially lead to degenerated configurations. There are several situations that could result in degenerate camera poses, such as numerical errors in triangulation, bad correspondence, and the absence of translation in the camera movement, which can lead to an ill-conditioned solution. On the other hand, degenerate camera poses can introduce errors in the reconstruction process. Therefore, probabilistic approaches seek to apply geometric and statistical analysis to infer the frame quality for pose and structure estimation.

Correspondence goodness is the most straightforward approach for finding good image pairs for pose estimation, since it is a commonly used source of data for geometric model estimation. The quality of the correspondence can be obtained using information criteria such as the geometric robust information criterion (GRIC) [61]. The GRIC is a function that scores a model based on the number of correspondences, the residuals, the error standard deviation, and the complexity of the model, i.e., the number of parameters and dimensions. It returns the lowest score for the model that best fits the data; thus, given a model and the correspondence between two scene views, the quality of the correspondence can be inferred. In [62], the authors proposed to use keyframes to delimit coherent subsequences for visual search tasks, and the selection procedure was built upon the GRIC function. The fundamental matrix and planar homography models were used to evaluate the opportunity to instantiate a new keyframe by checking the consistency of the subject's *point-of-regard* between the current observation and the last keyframe. Similarly, in [8], the authors proposed the *relGRIC* metric that computes the relative comparison of residual errors between the fundamental matrix and the homography matrix. They argued that the fundamental matrix has a high number of errors when there is a small baseline, and contrarily, the homography has a high number of errors with wider baselines; thus, the *relGRIC* metric allows for the search for the sweet spot in baseline size based on the camera and correspondence goodness.

Instead of inferring the camera pose goodness from the correspondences between the views, some studies proposed the use of information-theoretic approaches to measure the informativeness of the estimated pose. In general, camera motion estimation has a probabilistic formulation that results in a nonlinear least-squares problem that can be solved by maximum likelihood estimators such as Gauss–Newton. Given a nonlinear and differentiable measurement process, the first-order approximation of the information matrix can be obtained from the Jacobian matrix. The diagonal of the information matrix represents the amount of certainty in the pose parameters, and can be used as a criterion for keyframe selection [5].

Based on the epipolar constraint, in [63], the authors defined the error function as the Euclidean distance between the matched points and their corresponding epipolar line. The residuals were used to compute the Jacobian matrix, and consequently, the information matrix was obtained. They proposed to compare the latest keyframe against two consecutive candidate frames by analyzing the eigenvalues of the information matrix. They argue that the eigenvalues indicate the amount of certainty there is in the corresponding eigenvector; thus, the frame with the highest least uncertain pose estimate can be selected as a keyframe. Differently, the authors in [64] proposed to use the determinant of the information matrix for selection thresholding. To be computationally feasible, the covariance matrix is estimated from a pose graph optimization considering the landmarks in the current observation and relative keyframes. The proposed thresholding process reduces the number of user-defined parameters, and naturally, it adaptively selects the keyframes based on the surrounding environment. Alternatively, some studies [13,22,65] propose to transform the covariance matrix into a scalar based on the concept of differential entropy for multivariate Gaussian distribution. The absolute value of the entropy abstracts the uncertainty encoded in the covariance matrix into a scalar value. However, its value varies along with the trajectory and between different scenes; thus, simple thresholding for keyframe selection is not feasible. Therefore, in [13,65], the entropy ratio between the pose estimated from the last keyframe to the current frame and the estimated pose from the last keyframe and its first consecutive frame is used for thresholding. Meanwhile, in [22], they proposed to apply an average filter on the measured entropy value, and current observation is defined as a keyframe if its entropy value is below a certain percentage of the tracked average.

Beyond the pose uncertainty, there are studies that seek to select keyframes based on the expected improvement of the map. While other methods strive to maintain the map integrity, this method aims to reduce the uncertainty on the map. In [66], the authors proposed two approaches that attempt to reduce the map point entropy using information-theoretic concepts. *Cumulative point entropy reduction (CPER)* seeks to select keyframes that are expected to maximize the point entropy reduction in the existing map, while *point pixel flow discrepancy (PPFD)* seeks to select keyframes that best initialize new features for the camera to track in the future. The CPER approach attempts to predict the map point covariance if the observing keyframe was inserted into the bundle adjustment process; it applies the Shannon entropy function to assess the uncertainty reduction across all map points. On the other hand, the PPFD approach uses the point flow to construct two discrete probability distribution functions: *existing point flow PDF (E-PFP)*, which provides the areas with a high probability of inserting new points; and *future point flow PDF (F-PFP)*, which predicts which of the candidate's keyframes will add new image points in high-probability areas of the E-PFP. They compute the relative entropy between the probability distributions, and the candidate frame that F-PFP minimizes the relative entropy of is selected as the keyframe.

In addition to the map uncertainty reduction, it is important to guarantee that the ever-increasing size of the map does not become a bottleneck. In [67], the authors proposed to apply information-theoretic metrics to remove redundant scene measurements in graph-based maps with a minimum decrease in accuracy. The main idea is to calculate the mutual information (MI) between a keyframe and its neighbors to quantify how much information the reference keyframe adds to the estimated SLAM map.

#### 4.1.3. Learning-Based Methods

Over the last few years, there has been an increasing interest in integrating machine learning-based methods into the visual localization and mapping process to increase the robustness of specific modules in challenging environments [68], mostly improving the depth estimation, feature extraction, and data association modules considering the environment dynamics to better estimate the localization and mapping. Nevertheless, despite being a hot topic in many applications, such as video classification, action recognition, and video summarization [69], the result of this SLR shows that machine learning-based keyframe detection is not an active research area in visual localization and mapping tasks.

Of the included articles, only one proposed a complete machine learning-based approach for keyframe selection. Other studies, such as [23,32], proposed learning models to define the thresholds adaptively to the environment. However, the selection process is still based on heuristics rules; thus, they have not been classified as learning-based methods. Similarly, in [56], the authors have proposed a semantic content criterion to score the frames based on their relevance for high-level tasks to be performed; yet, the final score is still based on heuristics and assumptions about the scene.

The learning-based strategy proposed in [6] aims to overcome the limitations in state-of-the-art learning-based visual odometry methods to capture large geometric changes, since they are learned from short-length consecutive frames. According to the authors, a keyframe should include considerably geometric and visual changes with regard to to the last keyframe to augment the geometry description of the visual odometry process. On the other hand, the visual odometry process can provide valuable geometry clues by capturing challenging motion patterns and by checking the geometric consistency between the target image and the keyframes. Hence, they proposed an unsupervised collaborative learning framework to jointly learn the keyframe selection and visual odometry tasks in an end-to-end fashion. The keyframe selector network consists of a similarity regression function that fuses visual features and geometric features derived from the visual odometry network, used by a loss function that intensively applies the triplet losses to measure the similarity between frames. The output is a set of representative keyframes that are managed and updated based on inserting and merging operations during the training phase. The described process was not so clear, we tried to reach the authors by email but without success.

### *4.2. RQ2: Which Kinds of Properties Are Used in Keyframe Classification?*

This question aims to synthesize the most used properties to classify keyframes. It was specifically motivated by the heuristics-based approaches that classify keyframes based on some attributes that hold their assumptions about the camera view change and the environment. It is important to note that usually, a frame is not classified as a keyframe by itself; these attributes are defined regarding their relationship with the previously detected keyframes. Generally speaking, a frame could be assessed by its informativeness, considering that changes in brightness, fast motion blur, and the lack of distinctiveness information make pairwise image registration difficult. However, it is worthless to assess a frame by its informativeness if it can not maintain a spatial and temporal relationship with the previously detected keyframe.

Images are informative; they provide rich information about the environment, and consequently, their processing is computationally expensive. On the other hand, adjacent frames are usually captured at the exact location with a high degree of overlap; thus, there is no need to process all of them. Therefore, regarding its appearance, a candidate keyframe should be sufficiently different from the previously detected keyframe. In other words, the selected keyframe should provide a considerable amount of new information while ensuring that the covisibility with the last keyframe is maintained. In general, most of the proposed approaches aim to infer the amount of view change of the camera observation with respect to the existing keyframes, and the decision to select a keyframe is made by thresholding the view change value. However, the notion of image similarity is quite subjective and context-dependent; hence, most of the approaches seek to establish a linear relationship between the change in the appearance of frames and the actual agent's view change, to provide a proper sampling of the environment. So, regarding the appearance, what differs between the different techniques is not the intrinsic properties of the keyframe, but the way in which the change in appearance is measured. Usually, it is heuristically inferred by combining one or more factors such as camera motion, data association, and photometric change that reflect the camera view change.

Keyframe assessment by the degree of view change is a worthy approach to reduce the amount of redundant information processed; however, for visual localization and mapping tasks, the quality of the results depends heavily on the quality and consistency of the data to be processed. Therefore, besides their similarity, it is important to classify the keyframes by their geometric relationships, which basically means identifying frames that lead to ill-conditioned solutions. In that sense, adjacent frames tend to have a higher number of matches, but there could be an insufficient amount of translation information in which the epipolar geometry estimation can be ill-posed or undefined. Similarly, small baselines in relation to the depth of the viewed scene can introduce numerical errors in triangulation computation. On the other hand, a wider baseline can result in bad correspondences due to occlusion, which can lead to degenerate camera pose estimation. Consequently, degenerate camera poses seriously affect the quality of reconstruction. Thereby, for visual localization and mapping tasks, it is essential to establish a good interkeyframe geometric relationship.

In summary, keyframes are commonly classified not only by their informativeness, but also by their similarity and how they are temporally, spatially, and geometrically related to the other keyframes. There are numerous strategies that allow this intrakeyframe assessment to be carried out, and there is no consensus on which is the best strategy. In practice, the criteria are defined in a convenient way, aiming at some aspects of the application, such as the type of data used, i.e., a feature-based or direct approach; correspondence strategy, i.e., feature-matching or feature-tracking, and so on. Likewise, these criteria can be defined according to the application's goals. For example, in visual odometry tasks, it is preferable to a generous policy to guarantee the tracking with regard to the last keyframe, while for visual SLAM tasks, the frame should be sufficiently different visually from the previous keyframe in the map graph in order to decide whether a new node should be created. Similarly, in 3D reconstruction applications, a keyframe could be inserted in order to maintain the map integrity, while in visual SLAM, it is important to reduce the uncertainty in the map to improve the accuracy of the localization. To conclude, the keyframe classification problem is more related to the classic problem of choosing frame pairs with an optimal baseline. Adjacent frames share a high amount of redundant information and tend to lead to a poor motion estimate, i.e., the translation between them is ill-conditioned. On the other hand, as the baseline gets wider, image registration becomes harder and subject to bad correspondence due to occlusion.

### *4.3. RQ3: What Is the Role of Keyframes in Visual Localization and Mapping Tasks?*

This research question aims to identify and analyze the purposes of applying keyframe selection and relate them to the different application domains. By analyzing the included articles, we found distinct motivations for applying keyframe selection, but we identified four primary goals that are commonly addressed: reducing the redundancy, improving the tracking robustness, improving the map quality, and improving loop closure detection.

Figure 6 shows the number of included articles per target; they are not mutually exclusive, which means that a study may have multiple targets. It could be noted that most of the included articles seek to reduce the amount of data being processed and/or improve the robustness of the tracking process. Methods that seek to reduce the redundancy select the keyframe based on sufficient content change to avoid redundancy between adjacent frames. Differently, methods that attempt to improve the tracking robustness are not necessarily concerned with the amount of redundant information between the frames; they aim to select the keyframes based on their geometric relation, to avoid frames that could lead to an ill-conditioned solution.

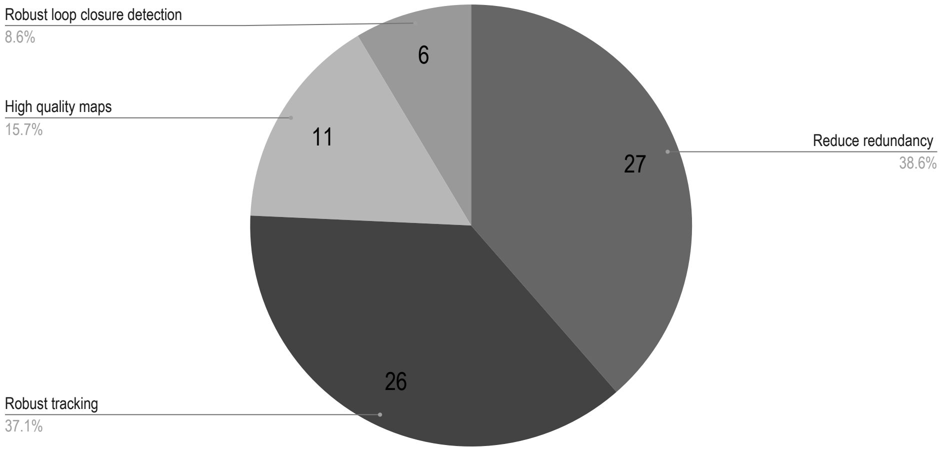

**Figure 6.** Number of included articles classified by keyframe selection target.

Regarding the map quality, we found that among the included articles, there are two main targets: maintaining the map integrity or minimizing the uncertainty in the map. Most of the methods define the target based on the map representation and/or the application domain. Methods that attempt to maintain map integrity are more likely to be applied in 3D reconstruction applications, where maps are commonly represented by sparse geometric shapes such as points and straight lines. Basically, these methods seek to reduce errors in multiview reconstructions by discarding frames that can lead to structure degeneracy due to degenerate camera poses or numerical errors in the triangulation. Differently, applications such as VSLAM are more concerned with the geometric relations between places or landmarks; thus, they aim to select keyframes that will directly improve the system's ability to localize. Usually, especially in graph-based SLAM techniques, the keyframes are selected in order to reduce the uncertainty and the problem size of the bundle adjustment optimization.

Loop closure is one of the key components of every visual SLAM algorithm, as it defines error constraints to be optimized in order to reduce the accumulated errors during long-term and large-scale operations. That explains the fact that most of the included articles that proposed keyframe selection to improve the loop closure were applied in visual SLAM applications. The loop closure process is closely related to scene recognition, where given the current observation, one wants to query a database to retrieve the most similar stored image. This would require a comparison between the current frame and all past frames, which is computationally infeasible. Thus, some studies overcome this problem by selecting only a subset of keyframes to be compared. Even though this sounds to be a worthy solution, as can be observed in Figure 7, among the included articles related to VSLAM applications, only 5 (13%) proposed keyframe selection methods to improve loop closure detection. This can be explained by the fact that the loop closure is essentially a large-scale image retrieval algorithm; consequently, most algorithms are more concerned with designing hierarchical data structures that help speed up the matching process than increasing the complexity by applying a keyframe selection mechanism.

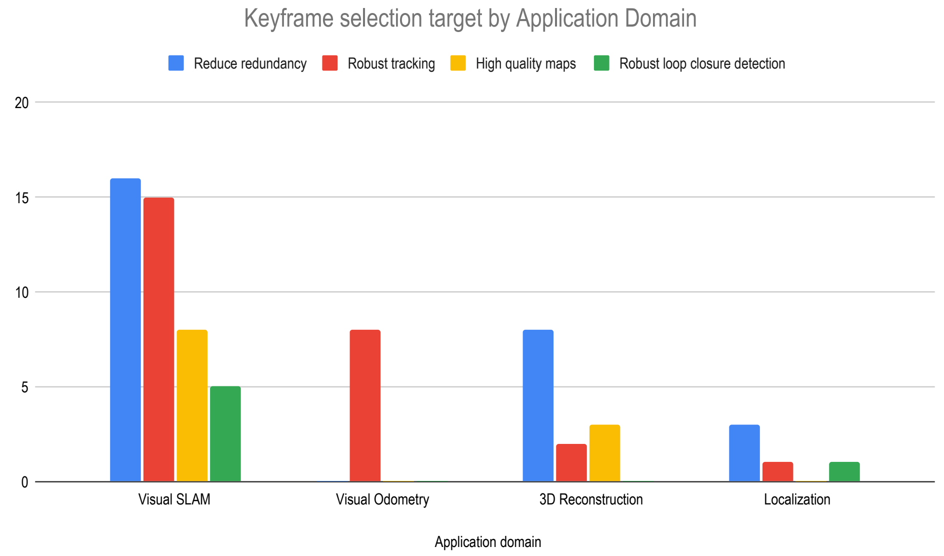

**Figure 7.** Number of article that address each target by application domain.

### 4.4. RQ4: How Can We Evaluate the Keyframe Selection Method?

As already mentioned in previous sections, the keyframe concept is quite subjective and application-dependent, which makes the validation process challenging. Therefore, in this research question, we show some of the methods and metrics used to validate the proposed keyframe selection methods for visual localization and mapping tasks.

Figure 8 shows the most common validation methods found in this review. Note that they are not mutually exclusive; one article could use more than one validation method. This result shows the deficit in the literature with respect to the validation and testing process. Note that 21% of the included articles did not use any method to validate the proposed approach or did not even mention their impact on the results. This information corroborates the motivation of this review, showing the lack of consideration regarding the keyframe selection procedure despite being a stage of the pipelines with a high impact on the final result, as we previously showed. Additionally, Figure 9 shows the most used measurement metrics; generally, they are related to the validation method adopted.

For most of the included articles, 51.6%, the application-level comparison was used as a validation method. This means that most of the studies did not directly evaluate the proposed method for keyframe selection; instead, they evaluated the impact of the method on the pipeline they were applied to. Usually, the proposed pipeline was compared against others present in the literature based on application domain metrics such as trajectory error, reprojection error, reconstruction quality, etc. Likewise, some studies adopted a self-comparison approach, where the proposed pipeline was evaluated with and without the proposed keyframe selection module.

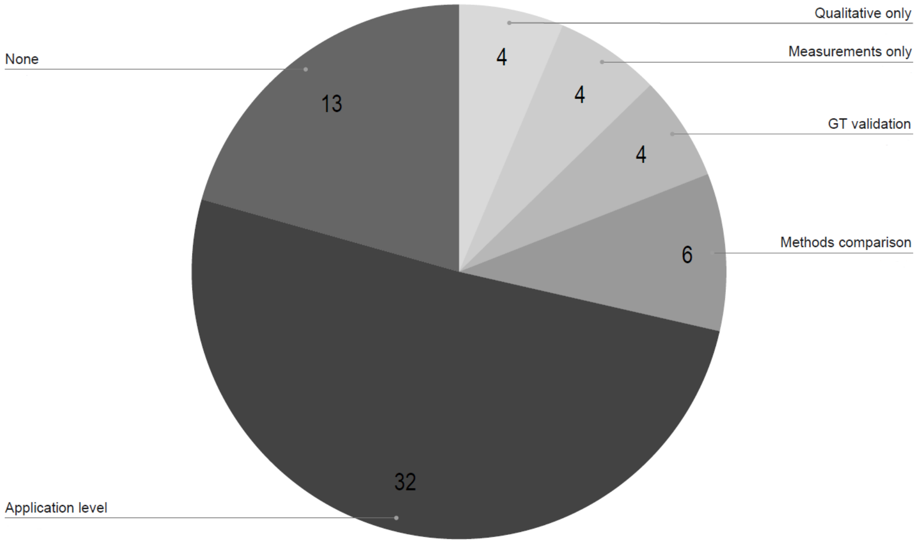

**Figure 8.** Number of selected articles per validation method.

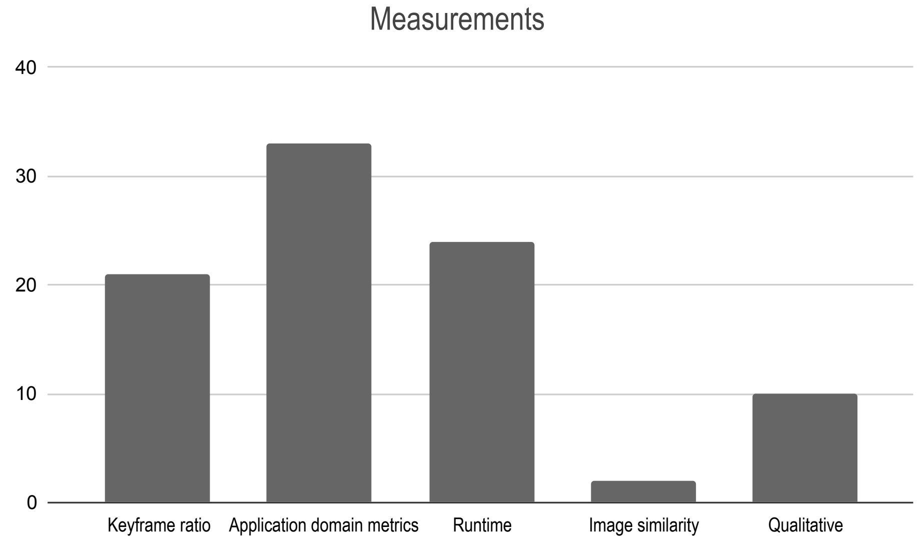

**Figure 9.** Number of selected articles per validation measurements.

Among the included articles, 9.7% compared the proposed method against other keyframe selection strategies. When the comparison was not made against methods presented in the literature, the uniform sampling in time strategy (i.e., selecting a keyframe at every *n*th frame captured by the camera) was used as the baseline. The application run time and the keyframe ratio (i.e., the total number of selected keyframes over the total number of frames in the sequence) were the most common metrics used to directly compare different strategies for keyframe selection. This is in compliance with the results in Section 4.3, where we showed that most of the included articles sought to reduce the amount of data to be processed.

The measurements-only validation method represents the studies that present a quantitative evaluation of the proposed method without comparing it with other methods, usually using the keyframe ratio or the application runtime as metrics. On the other hand, the qualitative evaluation mainly relies on the author's opinion, expressing how good their methods are without any metric information. Generally, this evaluation is made by visually inspecting how similar the extracted keyframes are, or how the proposed method samples the data in pure translation or rotation camera motions. In the case of reconstruction application, it could be made by visually inspecting the quality of the reconstruction.

To fairly compare different approaches, it is important to establish a dataset so they can be evaluated based on their consistency with the ground truth. In Figure 8, it can be observed that only 6.3% of the included articles used a ground truth in their validation process. Additionally, in Table 7, it can be seen that most of the studies used a public dataset to validate their pipelines; however, most of them are datasets and benchmarks made specifically for application-level validation. Among these, we can highlight the KITTI dataset [70], TUM datasets [71], and the Euroc dataset [72]. These results ratify that there is a gap in the literature regarding the methodologies used to compare the performance of different keyframe detectors. Nevertheless, we found a study [20] that could be considered the first to propose a methodology to compare keyframe selection techniques in the field of visual localization and mapping. Accordingly to the authors, a good keyframe detector should sample the environment fairly, in such a way that it does not oversample the environment and at the same time does not cause an unnecessary computational burden. Therefore, the authors proposed an accurate similarity measure, which, according to them, truthfully reflects the degree of change in the camera view, and in an ideal case, establishes a linear relationship with the actual change in the camera view. It is important to point out that this methodology is only viable in pure translation camera motions so that visible objects do not get distorted by nonlinear transformations such as rotation and scale. In that sense, the methods should be evaluated according to their ability to measure the similarity between two consecutively acquired frames and compared against the proposed ground truth similarity value. They also reported that the methods should be evaluated according to their stability detection when the similarity thresholds are changed. They argue that the threshold should work equally well independently of the environment.

**Table 7.** Number of included articles per dataset category.

| Category | Number (#) | Percentage (%) |
|---|---|---|
| Public with real data | 43 | 70.49 |
| Public with synthetic data | 4 | 6.56 |
| Custom with real data | 16 | 26.23 |
| Custom with synthetic data | 6 | 9.84 |

Later on, in their second study [54], the authors proposed two scenarios based on synthetic data, for which ground truth similarity can be inferred. The first scenario consists of the camera moving parallel to a rectangular wall with the optical axis perpendicular to the surface. The ideal similarity measure is defined based on the width of the overlapping region and the distance moved by the camera from its starting position, which decreases as the robot moves in a straight line parallel to the wall. The second scenario consists of the camera moving away from a large square wall, with the camera initial view being one-eighth of the width and one-eighth of the height of the wall. In this scenario, they proposed to calculate the ideal similarity as the fraction of the image where the original view is visible. In both scenarios, they assume that the camera is an ideal projective camera, and the synthesized features were assigned with unique IDs, so the methods did not get affected by the matching strategy used to detect keyframes.

Other articles have proposed pseudo-ground truth to validate their methods. In [6], the authors defined a set of snippets whose starting frame is a reference keyframe and a pseudo-ground-truth keyframe is located in the middle of the snippet. They proposed to use the ground-truth camera motion and interpolated depth maps to quantify the overlap between the frames. The pseudo-ground-truth keyframe is detected if the ratio of the overlapping area with regard to the reference keyframe is just below 50%. The keyframe detectors are quantitatively evaluated by the ratio of the detected keyframes within a fixed window around the defined pseudo-ground-truth keyframe. Similarly, in [62], the authors proposed to manually annotate the starting keyframe of each subsequence to produce pseudo-ground-truth data. Based on the application goals, they requested experts to select coherent subsequences and annotate the starting keyframe for each one by using their human pattern recognition skills. The performance measure was defined as the ratio between the number of keyframes recognized by the keyframe detectors over the number of keyframes identified by the experts.

## 5. Conclusions

In this article, we present a systematic literature review aiming to synthesize the existing knowledge about keyframe selection within the context of visual localization and mapping. The SLR draws on 61 articles selected out of 498 from the most popular digital libraries in computer science and robotics, filtered through a multistage process. An important feature of the review is that it does not restrict itself to a specific application domain or algorithm. This broad scope allows for a deeper understanding of the problem modeling and its peculiarities regarding the application context. Despite being considered an essential factor for the performance of the algorithms, the keyframe selection problem has not received as much attention as other key techniques. Therefore, we believe that this SLR is a step toward developing a deep understanding of this topic and bringing up a discussion about its importance to encourage the development of new methods. Additionally, by analyzing the commonly used approaches in the literature, a new categorization of the methods has been proposed. In the following sections, we present the most relevant findings and their implications for further research.

### *5.1. Lack of a Unified Classification of Keyframe Selection Methods*

Through the examination of the included articles, we could observe a lack of standardization in the classification of the methods present in the literature. Method categorization or classification is an important means to provide an intuitive and comprehensive way to present and summarize the existing approaches and a effective way to present the results in a more structured manner. Few articles presented or mentioned the groups or types of methods present in the literature and which of them their methods fit. Moreover, analyzing the methods that have presented such categorization, we could not find a standard nomenclature. This motivated us to propose a new and updated classification of the existing methods in the literature for keyframe selection within the context of visual localization and mapping tasks.

Despite the increasing interest over the last few years, especially for depth estimation, feature extraction, and data association, learning-based methods are still not a hot topic for keyframe selection in visual localization and mapping tasks. Of the included articles, only one proposed a complete machine learning-based approach; the proposed method fuses visual and geometric features to train a deep neural network in a self-supervised manner. Even though there was only a single study, we decided to include this category because learning-based methods for keyframe selection have been intensively studied in other applications, such as video classification, action recognition, and video summarization. Moreover, considering the results presented in the mentioned article, we believe that this is a promising method that should be explored in future research.

### *5.2. High Dependency of Heuristics-Based Solutions*

The majority of the included articles proposed heuristics-based approaches to selected keyframes. Although heuristic techniques could provide a solution in reasonable time and complexity, close to a exact solution, they do not give an optimal or rational result, which can lead to inaccurate judgments about how commonly things occur. In visual localization and mapping tasks, a frame is not classified as a keyframe by itself; it should be performed regarding its relationship with the previously detected keyframe. Most of the methods rely on the assumption that if the similarity between the current frame and the previous detected keyframe drops below some constant threshold, then the frame can be declared as a new keyframe. In other words, they seek to establish a linear relationship between the change in the appearance of the frames and the actual camera view change, which is not always true. Furthermore, we could observe that what differs between these methods is the way in which the change in appearance is measured. In practice, the criteria are defined in a convenient way according to the feature detector or the correspondence strategy adopted, which makes it difficult to say which is the best strategy.

### *5.3. High Application Dependency*

As already mentioned, this review was not restricted to a specific application domain (i.e., visual odometry, visual SLAM, 3D reconstruction, localization) to obtain a broad understanding of the problem modeling. In general, vision-based localization and mapping algorithms have strictly related pipelines, with their basis in geometric modeling of the world from visual information. Nonetheless, the results of this review show that most of the keyframe selection strategies have been proposed for visual SLAM problems. This can be explained by the fact that keyframe selection is a key component of keyframe-based SLAM approaches because it defines when to introduce a new node into the topological map. In that sense, usually, these methods seek to select the keyframes to achieve sufficient visual coverage of the environment while keeping its representation simple for computational efficiency. On the order hand, for applications such as visual odometry that require real-time execution, it is preferable to have a generous policy or to not apply keyframe selection to guarantee the tracking with respect to the last frame. In general, when applied, they are more concerned with tracking robustness by discarding frames that could lead to ill-conditioned solutions. Likewise, regarding the map data, 3D reconstruction applications are more concerned with map integrity, while in SLAM applications, they are more likely to select keyframes to minimize the uncertainty in the map representation. Therefore, keyframe selection methods are commonly modeled regarding the application goals.

### *5.4. Lack of Validation Methodologies*

The results of this SLR show a deficiency in the literature regarding the validation and testing process. A considerable number of studies did not use any methodology to validate their methods. Furthermore, the majority of the included articles did not directly evaluate the proposed method; instead, they evaluated the impact of the method in the pipeline they were applied to, based on the application domain metrics. There are no public benchmarks available for validation; a few articles have proposed some kind of methodology to compare keyframe selection techniques. The lack of rigorous tests, validation procedures, and public benchmarks prevents a fair comparison of different approaches. To assist and improve the development of new methods, better solutions for testing and validation should be developed.

### *5.5. Further Research*

The findings of this systematic literature review (SLR) represent a significant stride towards the development of a comprehensive knowledge base concerning keyframe selection in the domain of visual localization and mapping. However, despite this progress, numerous unresolved questions persist, as a considerable number of studies do not directly confront this issue. In light of these observations, we present the following comments as potential research directions that we believe merit further exploration.

The most commonly employed approaches for keyframe selection were based on heuristics. These methods often assume that a frame should be considered a keyframe if its similarity to the previous keyframe falls below a fixed threshold. However, fine-tuning is necessary to achieve optimal results with these approaches. Only a few studies have proposed adaptive thresholding methods.

Probabilistic methods evaluate frames based on their informativeness with respect to the scene's geometry, disregarding frames that may potentially degrade pose estimation and scene structure. We believe that these approaches hold the potential for greater effectiveness compared to assessing frames solely based on appearance changes. However, further exploration is needed in this area, as only a limited number of studies have addressed these strategies.

Assessing a frame based on more than just visual similarity involves establishing a meaningful relationship between camera observations, scene geometry, and the dynamics of the agent to which the camera is attached. This is a challenging task that requires capturing intricate patterns, relationships, and dependencies within the data. Deep learning approaches have shown remarkable effectiveness in tackling this type of task by automatically learning these complex representations from data, rather than relying on handcrafted features. Consequently, we find this area particularly intriguing, and believe it deserves further exploration.

**Author Contributions:** Conceptualization, N.J.B.D., G.T.L. and R.M.D.C.; methodology, N.J.B.D., G.T.L. and R.M.D.C.; validation, G.T.L. and R.M.D.C.; formal analysis, N.J.B.D., G.T.L. and R.M.D.C.; investigation, N.J.B.D.; resources, N.J.B.D., G.T.L. and R.M.D.C.; data curation, N.J.B.D. and G.T.L.; writing—original draft preparation, N.J.B.D.; writing—review and editing, N.J.B.D., G.T.L. and R.M.D.C.; visualization, N.J.B.D., G.T.L. and R.M.D.C.; supervision, G.T.L. and R.M.D.C.; project administration, G.T.L. and R.M.D.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** The APC was funded by Centro de Excelência em Inteligência Artificial (CEIA).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Quality Assessment Checklist

**Table A1.** Quality assessment checklist and the meaning of each possible answer.

| Question | Yes | Partially | No |
|---|---|---|---|
| **QA1.** *Does the study directly address the problem of keyframe selection?* | When the work explicitly addresses the problem of keyframe selection | When the work addresses the keyframe selection problem for a specific context. There is not a direct evaluation of the the proposed method | When the keyframe selection is used to solve a specific problem in a convenient way without justification. |
| **QA2.** *Does the study answer our research questions?* | When the work addresses all the questions | When the work addresses most of the questions | When the work only addresses one or none of the questions |
| **QA3.** *Are the methods clearly defined?* | The work presents and explains the motivation for the proposed method | The work presents the method but there is not a clear justification | The work just presents the method without justification |
| **QA4.** *Do they use any process to validate the method?* | When the work directly evaluates the impact of the proposed method and its parameterization in a specific task | When the work just evaluates the final result of a specific task (e.g., reconstruction error, tracking error, etc.) | When the proposed method is not cited in the results |
| **QA5.** *Do they use a specific dataset to validate the proposed method?* | When the work presents a specific dataset to evaluate KF selection methods | When the method uses commonly used datasets but defines specific metrics to evaluate the method | When the results are presented in an unrepeatable way |
| **QA6.** *Does the study compare the proposed method against other methods?* | Directly compares different approaches for keyframe selection | Just compares the same system/framework with and without the proposed KF selection method | It only presents experimental measurements without any comparison |
| **QA7.** *Does the study discuss the results obtained with the proposed method?* | Qualitative and quantitative discussion about the proposed method | Qualitative discussion only | There is not a discussion about the proposed method |

## Appendix B. Selected Studies

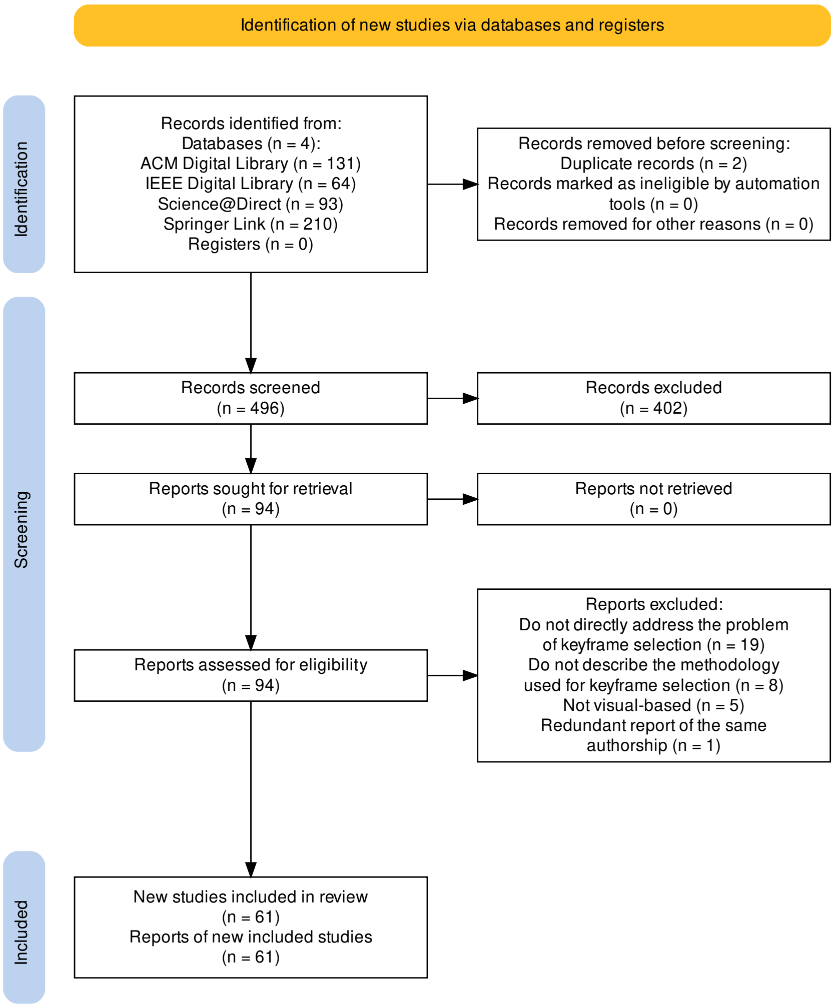
Identification of new studies via databases and registers
Identification
Screening
Included
Records identified from:
Databases (n = 4):
ACM Digital Library (n = 131)
IEEE Digital Library (n = 64)
Science@Direct (n = 93)
Springer Link (n = 210)
Registers (n = 0)
Records removed before screening:
Duplicate records (n = 2)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Records screened
(n = 496)
Records excluded
(n = 402)
Reports sought for retrieval
(n = 94)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 94)
Reports excluded:
Do not directly address the problem of keyframe selection (n = 19)
Do not describe the methodology used for keyframe selection (n = 8)
Not visual-based (n = 5)
Redundant report of the same authorship (n = 1)
New studies included in review
(n = 61)
Reports of new included studies
(n = 61)

**Figure A1.** PRISMA flow diagram illustrating the process of applying inclusion and exclusion criteria of the selected studies.

**Table A2.** A list of the selected studies grouped by the library that they were imported from.

| Library | Selected Studies |
|---|---|
| ACM Digital Library | [30,36,38,45,47,60] |
| IEEE Digital Library | [6,13,20,22,23,31–33,44,46,48–52,54,56,57,65–67,73–77] |
| ScienceDirect | [21,35,40,41,43,53,58,62,78–80] |
| Springer Link | [8,29,34,37,39,42,55,59,63,64,81–88] |

## References

1. Rosen, D.; Doherty, K.; Espinoza, A.; Leonard, J. Advances in Inference and Representation for Simultaneous Localization and Mapping. *Annu. Rev. Control Robot. Auton. Syst.* **2021**, *4*, 215–242. [CrossRef]
2. Wu, Y.; Tang, F.; Li, H. Image-based camera localization: An overview. *Vis. Comput. Ind. Biomed. Art* **2018**, *1*, 8. [CrossRef]
3. Jia, G.; Li, X.; Zhang, D.; Xu, W.; Lv, H.; Shi, Y.; Cai, M. Visual-SLAM Classical Framework and Key Techniques: A Review. *Sensors* **2022**, *22*, 4582. [CrossRef] [PubMed]
4. Thrun, S.; Burgard, W.; Fox, D. *Probabilistic Robotics*; MIT Press: Cambridge, MA, USA; London, UK, 2005.
5. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2003.

6. Sheng, L.; Xu, D.; Ouyang, W.; Wang, X. Unsupervised Collaborative Learning of Keyframe Detection and Visual Odometry Towards Monocular Deep SLAM. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: New York, NY, USA, 2019; pp. 4301–4310. [CrossRef]
7. Truong, B.T.; Venkatesh, S. Video Abstraction: A Systematic Review and Classification. *ACM Trans. Multimedia Comput. Commun. Appl.* **2007**, *3*, 3-es. [CrossRef]
8. Knoblauch, D.; Hess-Flores, M.; Duchaineau, M.A.; Joy, K.I.; Kuester, F. Non-Parametric Sequential Frame Decimation for Scene Reconstruction in Low-Memory Streaming Environments. In Proceedings of the Advances in Visual Computing, Las Vegas, NV, USA, 26–28 September 2011; Bebis, G., Boyle, R., Parvin, B., Koracin, D., Wang, S., Kyungnam, K., Benes, B., Moreland, K., Borst, C., DiVerdi, S., et al., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 359–370.
9. Gauglitz, S.; Höllerer, T.; Turk, M. Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking. *Int. J. Comput. Vis.* **2011**, *94*, 335–360. [CrossRef]
10. Garcia-Fidalgo, E.; Ortiz, A. Vision-based topological mapping and localization methods: A survey. *Robot. Auton. Syst.* **2015**, *64*, 1–20. [CrossRef]
11. Scaramuzza, D.; Fraundorfer, F. Visual Odometry [Tutorial]. *IEEE Robot. Autom. Mag.* **2011**, *18*, 80–92. [CrossRef]
12. Nister, D.; Naroditsky, O.; Bergen, J. Visual odometry. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington, DC, USA, 27 June–2 July 2004; Volume 1, p. I. [CrossRef]
13. Kerl, C.; Sturm, J.; Cremers, D. Dense visual SLAM for RGB-D cameras. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; IEEE: New York, NY, USA, 2013; pp. 2100–2106. [CrossRef]
14. Dias, N.; Laureano, G. Accurate Stereo Visual Odometry Based on Keypoint Selection. In Proceedings of the 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE), Rio Grande, Brazil, 23–25 October 2019; IEEE: New York, NY, USA, 2019; pp. 74–79. [CrossRef]
15. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
16. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; IEEE: New York, NY, USA, 2011; pp. 2564–2571. [CrossRef]
17. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded Up Robust Features. In Proceedings of the Computer Vision—ECCV 2006, Graz, Austria, 7–13 May 2006; Leonardis, A., Bischof, H., Pinz, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 404–417.
18. Chatzigiorgaki, M.; Skodras, A.N. Real-time keyframe extraction towards video content identification. In Proceedings of the 2009 16th International Conference on Digital Signal Processing, Santorini, Greece, 5–7 July 2009; IEEE: New York, NY, USA, 2009; pp. 1–6. [CrossRef]
19. Almeida, J.; Torres, R.D.S.; Leite, N.J. Rapid Video Summarization on Compressed Video. In Proceedings of the 2010 IEEE International Symposium on Multimedia, Taichung, Taiwan, 13–15 December 2010; IEEE: New York, NY, USA, 2010; pp. 113–120. [CrossRef]
20. Zhang, H.; Li, B.; Yang, D. Keyframe detection for appearance-based visual SLAM. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; IEEE: New York, NY, USA, 2010; pp. 2071–2076. [CrossRef]
21. Vázquez-Martín, R.; Bandera, A. Spatio-temporal feature-based keyframe detection from video shots using spectral clustering. *Pattern Recognit. Lett.* **2013**, *34*, 770–779. [CrossRef]
22. Kuo, J.; Muglikar, M.; Zhang, Z.; Scaramuzza, D. Redesigning SLAM for Arbitrary Multi-Camera Systems. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; IEEE: New York, NY, USA, 2020; pp. 2116–2122. [CrossRef]
23. Chen, W.; Zhu, L.; Lin, X.; He, L.; Guan, Y.; Zhang, H. Dynamic Strategy of Keyframe Selection with PD Controller for VSLAM Systems. *IEEE/ASME Trans. Mechatron.* **2022**, *27*, 115–125. [CrossRef]
24. Kitchenham, B.; Charters, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering; Technical Report EBSE 2007-001; Keele University and Durham University Joint Report. 2007. Available online: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf (accessed on 23 July 2021).
25. Parsifal. Online Tool Designed to Support Researchers to Perform Systematic Literature Reviews within the Context of Software Engineering. 2013. Available online: https://parsif.al/ (accessed on 23 July 2021).
26. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
27. Strasdat, H.; Montiel, J.M.M.; Davison, A.J. Real-time monocular SLAM: Why filter? In Proceedings of the 2010 IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 3–8 May 2010; IEEE: New York, NY, USA, 2010; pp. 2657–2664. [CrossRef]
28. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]

29. Fei, X.; Tsotsos, K.; Soatto, S. A Simple Hierarchical Pooling Data Structure for Loop Closure. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 321–337.
30. Chen, K.; Wu, J.; Li, Z.; Tu, R. A Robust Visual Loop-Closure Detection Method of VSLAM for Ambiguous Environment. In Proceedings of the 2020 the 3rd International Conference on Control and Computer Vision (ICCCV'20), Macau, China, 23–25 August 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 77–83. [CrossRef]
31. Yue, H.; Yu, Y.; Wu, X.; Chen, W. Keyframe extraction and loop closure detection considering robot motion. In Proceedings of the 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, China, 31 May–2 June 2018; IEEE: New York, NY, USA, 2018; pp. 847–851. [CrossRef]
32. Piao, J.C.; Kim, S.D. Real-Time Visual–Inertial SLAM Based on Adaptive Keyframe Selection for Mobile AR Applications. *IEEE Trans. Multimed.* **2019**, *21*, 2827–2836. [CrossRef]
33. Zhang, C.; Wang, H.; He, S.; Li, H.; Liu, J. Photography Constraint Aided Keyframe Selection and Matching Method for UAV 3D Reconstruction. In Proceedings of the 2018 Chinese Control And Decision Conference (CCDC), Shenyang, China, 9–11 June 2018; IEEE: New York, NY, USA, 2018; pp. 5030–5035. [CrossRef]
34. Quan, M.; Piao, S.; He, Y.; Liu, X.; Qadir, M.Z. Monocular Visual SLAM with Points and Lines for Ground Robots in Particular Scenes: Parameterization for Lines on Ground. *J. Intell. Robot. Syst.* **2021**, *101*, 72. [CrossRef]
35. Lu, J.; Fang, Z.; Gao, Y.; Chen, J. Line-based visual odometry using local gradient fitting. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103071. [CrossRef]
36. Chen, C.F.; Suma Rosenberg, E. Capture to Rendering Pipeline for Generating Dynamically Relightable Virtual Objects with Handheld RGB-D Cameras. In Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology, Virtual, 1–4 November 2020; Association for Computing Machinery: New York, NY, USA, 2020. [CrossRef]
37. Ni, T.; Shi, Y.; Sun, A.; Ju, B. Simultaneous identification of points and circles: Structure from motion system in industry scenes. *Pattern Anal. Appl.* **2021**, *24*, 333–342. [CrossRef]
38. Tang, C.; Wang, O.; Liu, F.; Tan, P. Joint Stabilization and Direction of 360° Videos. *ACM Trans. Graph.* **2019**, *38*, 1–13. [CrossRef]
39. Li, R.; Gu, D.; Liu, Q.; Long, Z.; Hu, H. Semantic Scene Mapping with Spatio-temporal Deep Neural Network for Robotic Applications. *Cogn. Comput.* **2018**, *10*, 260–271. [CrossRef]
40. Wei, X.; Xu, X.; Zhang, J.; Gong, Y. Specular highlight reduction with known surface geometry. *Comput. Vis. Image Underst.* **2018**, *168*, 132–144. [CrossRef]
41. Dhou, S.; Motai, Y. Dynamic 3D surface reconstruction and motion modeling from a pan–tilt–zoom camera. *Comput. Ind.* **2015**, *70*, 183–193. [CrossRef]
42. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-Scale Direct Monocular SLAM. In Proceedings of the Computer Vision—ECCV Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 834–849.
43. Castle, R.; Murray, D. Keyframe-based recognition and localization during video-rate parallel tracking and mapping. *Image Vis. Comput.* **2011**, *29*, 524–532. [CrossRef]
44. Rachmielowski, A.; Birkbeck, N.; Jägersand, M.; Cobzas, D. Realtime Visualization of Monocular Data for 3D Reconstruction. In Proceedings of the 2008 Canadian Conference on Computer and Robot Vision, Windsor, ON, Canada, 28–30 May 2008; IEEE: New York, NY, USA, 2008; pp. 196–202. [CrossRef]
45. Guo, D.; Teng, X.; Guo, Y.; Zhou, X.; Liu, Z. SiFi: Self-Updating of Indoor Semantic Floorplans for Annotated Objects. *ACM Trans. Internet Things* **2021**, *2*, 1–21. [CrossRef]
46. Xie, P.; Su, W.; Li, B.; Jian, R.; Huang, R.; Zhang, S.; Wei, J. Modified Keyframe Selection Algorithm and Map Visualization Based on ORB-SLAM2. In Proceedings of the 2020 4th International Conference on Robotics and Automation Sciences (ICRAS), Chengdu, China, 6–8 June 2020; IEEE: New York, NY, USA, 2020; pp. 142–147. [CrossRef]
47. Singh, D. Stereo Visual Odometry with Stixel Map Based Obstacle Detection for Autonomous Navigation. In Proceedings of the Advances in Robotics (AIR 2019), Chennai, India, 2–6 July 2019; Association for Computing Machinery: New York, NY, USA, 2019. [CrossRef]
48. Yuan, Y.; Ding, Y.; Zhao, L.; Lv, L. An Improved Method of 3D Scene Reconstruction Based on SfM. In Proceedings of the 2018 3rd International Conference on Robotics and Automation Engineering (ICRAE), Guangzhou, China, 17–19 November 2018; IEEE: New York, NY, USA, 2018; pp. 228–232. [CrossRef]
49. Chen, C.W.; Hsiao, W.Y.; Lin, T.Y.; Wang, J.; Shieh, M.D. Fast Keyframe Selection and Switching for ICP-based Camera Pose Estimation. In Proceedings of the 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 27–30 May 2018; IEEE: New York, NY, USA, 2018; pp. 1–4. [CrossRef]
50. Gan, Y.; Ye, M.; Xing, G.; Zeng, F. A new keyframe decision mechanism with translation constraint for visual slam. In Proceedings of the 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 15–17 December 2017; IEEE: New York, NY, USA, 2017; pp. 152–157. [CrossRef]
51. Kuang, H.; Zhang, K.; Li, R.; Liu, X. Monocular SLAM Algorithm Based on Improved Depth Map Estimation and Keyframe Selection. In Proceedings of the 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Changsha, China, 10–11 February 2018; IEEE: New York, NY, USA, 2018; pp. 350–353. [CrossRef]

52. Stalbaum, J.; Song, J.B. Keyframe and inlier selection for visual SLAM. In Proceedings of the 2013 10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Jeju, Republic of Korea, 30 October–2 November 2013; IEEE: New York, NY, USA, 2013; pp. 391–396. [CrossRef]
53. Li, S.; Zhang, T.; Gao, X.; Wang, D.; Xian, Y. Semi-direct monocular visual and visual-inertial SLAM with loop closure detection. *Robot. Auton. Syst.* **2019**, *112*, 201–210. [CrossRef]
54. Stewart, R.L.; Zhang, H. Image similarity from feature-flow for keyframe detection in appearance-based SLAM. In Proceedings of the 2011 IEEE International Conference on Robotics and Biomimetics, Karon Beach, Thailand, 7–11 December 2011; IEEE: New York, NY, USA, 2011; pp. 305–312. [CrossRef]
55. Bellavia, F.; Fanfani, M.; Colombo, C. Selective visual odometry for accurate AUV localization. *Auton. Robot.* **2017**, *41*, 133–143. [CrossRef]
56. Alonso, I.; Riazuelo, L.; Murillo, A.C. Enhancing V-SLAM Keyframe Selection with an Efficient ConvNet for Semantic Analysis. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; IEEE: New York, NY, USA, 2019; pp. 4717–4723. [CrossRef]
57. Hong, S.; Kim, J. Visual SLAM with keyframe selection for underwater structure inspection using an autonomous underwater vehicle. In Proceedings of the 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Xi'an, China, 19–22 August 2016; IEEE: New York, NY, USA, 2016; pp. 558–562. [CrossRef]
58. Tian, F.; Gao, Y.; Fang, Z.; Gu, J.; Yang, S. 3D reconstruction with auto-selected keyframes based on depth completion correction and pose fusion. *J. Vis. Commun. Image Represent.* **2021**, *79*, 103199. [CrossRef]
59. Vacchetti, L.; Lepetit, V.; Ponder, M.; Papagiannakis, G.; Fua, P.; Thalmann, D.; Thalmann, N.M. A Stable Real-time AR Framework for Training and Planning in Industrial Environments. In *Virtual and Augmented Reality Applications in Manufacturing*; Ong, S.K., Nee, A.Y.C., Eds.; Springer: London, UK, 2004; pp. 129–145.
60. Valentin, J.; Kowdle, A.; Barron, J.T.; Wadhwa, N.; Dzitsiuk, M.; Schoenberg, M.; Verma, V.; Csaszar, A.; Turner, E.; Dryanovski, I.; et al. Depth from Motion for Smartphone AR. *ACM Trans. Graph.* **2018**, *37*, 1–19. [CrossRef]
61. Torr, P.H.S. Geometric Motion Segmentation and Model Selection. *Phil. Trans. R. Soc. Lond. A* **1998**, *356*, 1321–1340. [CrossRef]
62. Ntouskos, V.; Pirri, F.; Pizzoli, M.; Sinha, A.; Cafaro, B. Saliency prediction in the coherence theory of attention. *Biol. Inspired Cogn. Archit.* **2013**, *5*, 10–28. [CrossRef]
63. Gan, T.S.Y.; Drummond, T.W. Vision-Based Augmented Reality Visual Guidance with Keyframes. In Proceedings of the Advances in Computer Graphics, Hangzhou, China, 26–28 June 2006; Nishita, T., Peng, Q., Seidel, H.P., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 692–701.
64. Lim, H.; Lim, J.; Kim, H.J. Online 3D Reconstruction and 6-DoF Pose Estimation for RGB-D Sensors. In Proceedings of the Computer Vision—ECCV 2014 Workshops, Zurich, Switzerland, 6–12 September 2014; Agapito, L., Bronstein, M.M., Rother, C., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 238–254.
65. Gomez-Ojeda, R.; Moreno, F.A.; Zuñiga-Noël, D.; Scaramuzza, D.; Gonzalez-Jimenez, J. PL-SLAM: A Stereo SLAM System Through the Combination of Points and Line Segments. *IEEE Trans. Robot.* **2019**, *35*, 734–746. [CrossRef]
66. Das, A.; Waslander, S.L. Entropy based keyframe selection for Multi-Camera Visual SLAM. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–3 October 2015; IEEE: New York, NY, USA, 2015; pp. 3676–3681. [CrossRef]
67. Schmuck, P.; Chli, M. On the Redundancy Detection in Keyframe-Based SLAM. In Proceedings of the 2019 International Conference on 3D Vision (3DV), Quebec City, QC, Canada, 16–19 September 2019; IEEE: New York, NY, USA, 2019; pp. 594–603. [CrossRef]
68. Beghdadi, A.; Mallem, M. A comprehensive overview of dynamic visual SLAM and deep learning: Concepts, methods and challenges. *Mach. Vis. Appl.* **2022**, *33*, 54. [CrossRef]
69. Yan, X.; Gilani, S.Z.; Feng, M.; Zhang, L.; Qin, H.; Mian, A. Self-Supervised Learning to Detect Key Frames in Videos. *Sensors* **2020**, *20*, 6941. [CrossRef] [PubMed]
70. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: New York, NY, USA, 2012; pp. 3354–3361. [CrossRef]
71. Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 7–12 October 2012; IEEE: New York, NY, USA, 2012; pp. 573–580. [CrossRef]
72. Burri, M.; Nikolic, J.; Gohl, P.; Schneider, T.; Rehder, J.; Omari, S.; Achtelik, M.W.; Siegwart, R. The EuRoC micro aerial vehicle datasets. *Int. J. Robot. Res.* **2016**, *35*, 1157–1163.
73. Yoo, W.; Kim, H.; Hong, H.; Lee, B.H. Scan Similarity-based Pose Graph Construction method for Graph SLAM. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; IEEE: New York, NY, USA, 2018; pp. 131–136. [CrossRef]
74. Fu, Z.; Quo, Y.; Lin, Z.; An, W. FSVO: Semi-direct monocular visual odometry using fixed maps. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: New York, NY, USA, 2017; pp. 2553–2557. [CrossRef]

75. Zeng, F.; Zeng, W.; Gan, Y. ORB-SLAM2 with 6DOF Motion. In Proceedings of the 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, China, 27–29 June 2018; IEEE: New York, NY, USA, 2018; pp. 556–559. [CrossRef]
76. Tu, X.; Xu, C.; Liu, S.; Xie, G.; Huang, J.; Li, R.; Yuan, J. Learning Depth for Scene Reconstruction Using an Encoder-Decoder Model. *IEEE Access* **2020**, *8*, 89300–89317. [CrossRef]
77. Soares, J.; Meggiolaro, M. Keyframe-Based RGB-D SLAM for Mobile Robots with Visual Odometry in Indoor Environments Using Graph Optimization. In Proceedings of the 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE), Joao Pessoa, Brazil, 6–10 November 2018 ; IEEE: New York, NY, USA, 2018; pp. 94–99. [CrossRef]
78. Athira, S.V.; George, M.; Jose, B.R.; Mathew, J. A Global Image Descriptor Based Navigation System for Indoor Environment. *Procedia Comput. Sci.* **2017**, *115*, 466–473. [CrossRef]
79. Gutierrez-Gomez, D.; Mayol-Cuevas, W.; Guerrero, J. Dense RGB-D visual odometry using inverse depth. *Robot. Auton. Syst.* **2016**, *75*, 571–583. [CrossRef]
80. Pire, T.; Fischer, T.; Castro, G.; De Cristóforis, P.; Civera, J.; Jacobo Berlles, J. S-PTAM: Stereo Parallel Tracking and Mapping. *Robot. Auton. Syst.* **2017**, *93*, 27–42. [CrossRef]
81. Turan, M.; Shabbir, J.; Araujo, H.; Konukoglu, E.; Sitti, M. A deep learning based fusion of RGB camera information and magnetic localization information for endoscopic capsule robots. *Int. J. Intell. Robot. Appl.* **2017**, *1*, 442–450. [CrossRef]
82. Neubert, J.; Pretlove, J.; Drummond, T. Rapidly constructed appearance models for tracking in augmented reality applications. *Mach. Vis. Appl.* **2012**, *23*, 843–856. [CrossRef]
83. Li, J.N.; Wang, L.H.; Li, Y.; Zhang, J.F.; Li, D.X.; Zhang, M. Local optimized and scalable frame-to-model SLAM. *Multimed. Tools Appl.* **2016**, *75*, 8675–8694. [CrossRef]
84. Fioraio, N.; Cerri, G.; Di Stefano, L. Towards Semantic KinectFusion. In Proceedings of the Image Analysis and Processing—ICIAP 2013, Naples, Italy, 9–13 September 2013; Petrosino, A., Ed.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 299–308.
85. Fioraio, N.; Di Stefano, L. SlamDunk: Affordable Real-Time RGB-D SLAM. In Proceedings of the Computer Vision—ECCV 2014 Workshops, Zurich, Switzerland, 6–12 September 2014; Agapito, L., Bronstein, M.M., Rother, C., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 401–414.
86. Turan, M.; Pilavci, Y.Y.; Ganiyusufoglu, I.; Araujo, H.; Konukoglu, E.; Sitti, M. Sparse-then-dense alignment-based 3D map reconstruction method for endoscopic capsule robots. *Mach. Vis. Appl.* **2018**, *29*, 345–359. [CrossRef]
87. Li, Y.; He, H.; Yang, D.; Wang, S.; Zhang, M. Geolocalization with aerial image sequence for UAVs. *Auton. Robot.* **2020**, *44*, 1199–1215. [CrossRef]
88. Lee, J.K.; Yoon, K.J. Joint Estimation of Camera Orientation and Vanishing Points from an Image Sequence in a Non-Manhattan World. *Int. J. Comput. Vis.* **2019**, *127*, 1426–1442. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.