Collaborative Mobile Robotics for Semantic Mapping: A Survey

Abdessalem Achour, Hiba Al-Assaad, Yohan Dupuis, Madeleine El Zaher

applied sciences

MDPI

Review

Collaborative Mobile Robotics for Semantic Mapping: A Survey

Abdessalem Achour [1,2,*], Hiba Al-Assaad [1], Yohan Dupuis [3] and Madeleine El Zaher [1]

[1] LINEACT CESI, Campus of Toulouse, 31670 Labège, France
[2] SMI Doctoral School, HESAM University, 75013 Paris, France
[3] LINEACT CESI, Paris La Défense, 92074 Paris, France
* Correspondence: aachour@cesi.fr

Abstract: Ensuring safety in human–robot collaboration is one of the main challenges in mobile robotics today. Semantic maps are a potential solution because they provide semantic knowledge in addition to the geometric representation of the environment. They allow robots to perform their basic tasks using geometric representation, mainly localization, path planning and navigation, and additionally allow them to maintain a cognitive interpretation of the environment in order to reason and make decisions based on the context. The goal of this paper is to briefly review semantic mapping for a single mobile robot in indoor environments, and then focus on collaborative mobile semantic mapping. In both contexts, the semantic mapping process is divided into modules/tasks, and recent solutions for each module are discussed. Possible system architectures are also discussed for collaborative semantic mapping. Finally, future directions are highlighted.

Keywords: semantic scene understanding; cooperating robots; data fusion; cognitive human–robot interaction

Citation: Achour, A.; Al-Assaad, H.; Dupuis, Y.; El Zaher, M. Collaborative Mobile Robotics for Semantic Mapping: A Survey. *Appl. Sci.* **2022**, *12*, 10316. https://doi.org/10.3390/app122010316

Academic Editor: DaeEun Kim

Received: 27 July 2022
Accepted: 11 October 2022
Published: 13 October 2022

1. Introduction

Mapping is an important task in mobile robotics on which many other tasks depend. Most of the existing mapping approaches aim at building either a metric map or a topological map of the robot environment. The difference between these two representations is that a metric map provides a geometric representation of the objects in the environment in a global reference system, while a topological map represents the environment as a graph, where nodes represent locations and edges represent relationships between them [1]. These two representations are mainly used by the robot to localize itself, plan its trajectory, avoid obstacles and navigate in the environment. However, they do not allow it to understand its environment as a human does. For example, it is unable to distinguish the kitchen from the living room, or to know which objects make up a specific room.

Over the past few years, mobile robots are increasingly sharing the same space as humans in various applications, especially service robots. Therefore, the challenge of endowing robots with cognitive interpretation capabilities of the environment has increased and many researchers have been addressing it. Some of them considered that the most appropriate way to achieve this is through semantic mapping, which consists of integrating semantic attributes about the objects and places encountered into a map, namely, a semantic map (SM). Consequently, the semantic map is an enhanced representation of the environment, which includes both geometric information and high-level features [2]. Thus, while a geometric map stores the geometric features that the robot needs for its basic tasks (localization and navigation), the high-level features represent common knowledge concepts about shapes, places, objects and even the relationships between them. The deployment of this map in mobile robots facilitates their interaction with humans and extends their capabilities, especially navigation [3,4] and task-planning capabilities [5–7].

There are a few previous works in the literature that have reviewed topics related to semantic mapping. The study [8] is one of the first reviews on semantic mapping. It focuses on highlighting its trends and main applications in indoor and outdoor scenes. Next, Qiang et aI. [9] proposed a review focusing on visual semantic mapping. They focus on semantic-information extraction techniques based on visual data through feature extraction, object/place recognition and semantic representation methods. Subsequently, semantic mapping is discussed as one of many map representations in the more general review [10], which focuses on the history and trends of the simultaneous localization and mapping (SLAM) problem. In the recent review [11], the authors reviewed works on semantic mapping and focused on the application to semantic navigation, which involves performing navigation using high-level commands such as "*go to the bathroom*". Finally, the most recent study proposed in [12], reviewed the process of semantic mapping and focused on indoor scenes. It mainly presents semantic-mapping modules and different techniques to implement them, as well as current challenges and future directions.

This review has four main differences from previous works. First, the main difference is that all previous works focus on reviewing single-robot semantic mapping, where a single robot is used for semantic mapping. However, recently, some works have started to address this problem in a collaborative mode using a fleet of robots. To the best of our knowledge, no previous work has examined collaborative semantic mapping. Indeed, there are works that have addressed collaborative geometric mapping, including collaborative SLAM [13], but not collaborative semantic mapping. Second, for the single-robot mode, this work focuses only on recent work on semantic mapping in indoor environments. Indeed, the perspective of this study is to use semantic mapping for successful human–robot interaction (HRI) in indoor environments, namely, domestic and industrial environments, so works interested in semantic mapping in outdoor environments [14–16] are not addressed in this paper, but there are some previous papers that have examined this topic [2,8,17]. For the collaborative mode, since there is much less work available, both indoor and outdoor semantic mapping are reviewed to give a better overview of what is carried out in the literature. Third, this work shows the main modules of the semantic-mapping process in single-robot and collaborative mode. It highlights the differences between them and the different solutions to implement these modules. The main contribution of this work is to explore the multi-robot data association and fusion module for the collaborative mode. Thus, it studies and classifies the different solutions available to deal with the problem of integrating data from different robots into a single semantic map. Fourth, this paper highlights the main remaining challenges in the field of single-robot and collaborative semantic mapping.

This paper is organized as follows: Section 2 is reserved for single-robot semantic mapping. We first present this process and its main modules/tasks. Then, we present solutions to implement each module, mainly geometric mapping, semantic data acquisition and map representation. In Section 3, we study collaborative semantic mapping. First, we present the mapping process, then we study the possible architectures to implement it in a collaborative setting. Finally, we focus on the different solutions proposed for the multi-robot data association and fusion module. In Section 4, we detail the remaining challenges in semantic mapping for future directions. Lastly, we conclude with the main contributions of this paper.

2. Single-Robot Semantic Mapping

Single-robot semantic mapping is the process in which a single robot explores the environment and builds its semantic map. As shown in Figure 1, this process is mainly composed of two major tasks: data acquisition and processing, and data representation. Regarding the first task, the robot collects two types of data: geometric data and semantic data. Geometric data is usually acquired by the robot's sensors and is processed to generate a geometric map of the environment. Semantic data can be acquired from different sources and is used to add semantic meanings to the mapped features. The three main sources of

semantic data are the extraction of interesting features from sensor data, the incorporation of data through HRI, and the extraction of data from a knowledge database, which is a model of the concepts involved in the explored environment and the relationships between them. As for the representation task, it consists of organizing those geometric and semantic data in a structured representation, namely, the semantic map, so that the robot can exploit them to understand its environment. In what follows, we study in depth the solutions proposed in the literature to solve these two tasks. It is important to mention that this work does not cover semantic SLAM systems [18–20], which benefit from object recognition to tightly integrate metric and semantic information in map localization and estimation [21]. However, it focuses on approaches that extract metric and semantic information in two separate processes and then merge the results into a semantic map.

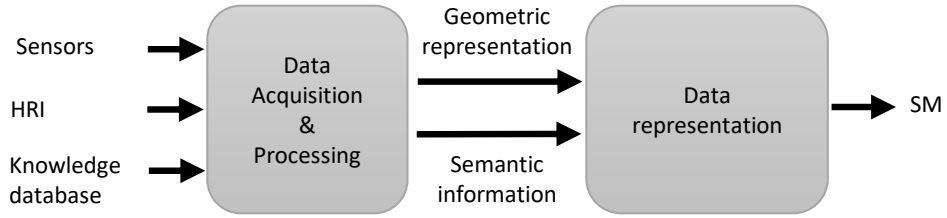

Figure 1. The single-robot semantic-mapping process.

2.1. Data Acquisition and Processing

In the single-robot semantic-mapping process, there are mainly three data-acquisition methods: perception-based data acquisition, HRI-based data acquisition and reasoning-based acquisition.

2.1.1. Perception-Based Data Acquisition

Sensor observations are used to represent the geometry of the environment and extract its semantics autonomously. The geometric representation is usually obtained using state-of-the-art SLAM methods, which consist of simultaneously estimating the state of a robot and building a model of the environment perceived by its sensors [10]. In contrast, semantic information is mainly extracted using object-detection and semantic-segmentation techniques.

Regarding systems integrating object detection, Qi et al. [4] proposed an approach to create a semantic occupancy map of a home environment. The sonar-based SLAM method proposed in [22] is used to create an occupancy grid map. Then, using odometry and a stereo camera, a method based on object detection and triangulation was used to add the object's topological area with its category to the occupancy map. In addition, Zender et al. [23] proposed a system to create a multi-layer map of an indoor environment. The laser-based EKF-SLAM algorithm [24] was used to create the low-level layers, and then receptive field cooccurrence histograms (RFCH) [25], an RGB camera-based object-recognition algorithm, was used to collect the physical feature categories and establish links to the high-level layer.

A recent trend in object detection is semantic segmentation, a paradigm that assigns a class label to pixels in an input image [26]. The semantic-mapping approach proposed in [27] incorporates semantic segmentation using a video stream from a monocular camera for 3D reconstruction of indoor and outdoor environments. The monocular semi-dense SLAM algorithm [28], which is effective in indoor and outdoor environments, is used to represent the geometric structure of the environment. In parallel, a pre-trained convolutional neural network (CNN) is used for 2D semantic segmentation. In addition, Niko et al. [29] proposed an approach that uses an RGB-D camera to map a 3D environment and reconstruct models of the detected objects on the fly. ORB-SLAM2 [30] is used to perform mapping and camera localization on each RGB-D image. In parallel, objects are detected on

the RGB images using a CNN and the associated point cloud is segmented in 3D. In [27,29], image-based deep-learning techniques were used for semantic information acquisition due to their great progress in object detection and segmentation based on 2D images.

In addition to these systems that perform object recognition and classification from visual data, there are other works that use geometric data from laser sensors to extract high-level information. Pronobis and Jensfelt [31] proposed a system that combines visual cues and laser range data using SVM classifiers to distinguish indoor areas (hallway, kitchen, meeting room, etc.). Geometric feature extraction was used to determine the shape and size properties of the environment, and visual feature extraction was used to obtain the appearance and objects present in the environment.

Although exploring object-detection and segmentation techniques is not the main focus of this work, they are necessary to understand how semantic mapping works. For interested readers, there are recent reviews of deep-learning-based semantic segmentation that provide a comprehensive survey on the topic. These reviews cover almost all popular image datasets and semantic segmentation methods, and for all modalities, such as RGB [32,33], RGB-D [34] and 3D data [35,36].

Since perceptions are used in semantic mapping to extract both geometric and semantic information, the choice of appropriate sensors, therefore, depends mainly on the type of geometric representation to be obtained (a 2D/3D map, a point cloud, etc.) and the level of the semantic description of the environment. The most popular sensors for geometric mapping are 2D/3D lasers, and, recently, RGB-D cameras are increasingly used [12]. These cameras are relatively inexpensive and can provide an RGB image and a depth map simultaneously. The depth map provides precise information on the distance of the detected elements. The visual sensors are frequently chosen for the acquisition of semantic information because they allow a perception of the environment close to that perceived by the human eye. They also allow the collection of low-level data, such as points, and high-level data, such as object categories [9].

2.1.2. HRI-Based Data Acquisition

According to [3], the automatic extraction of semantic information from sensors is limited and not robust enough. Indeed, object and place recognition are complex tasks for robots. That is why some techniques integrate the human in the process of identifying objects or places in order to create an augmented map [11,12].

In [3], the authors guided the robot through a tour of the environment to map it. The RBPF-SLAM algorithm [37,38] based on laser data and odometry data was used to generate an occupancy map in real time. At the same time, the user has an IFLYTEK-based voice application on their cell phone to label the visited places on the map. In order to avoid errors in the recognition of voice commands given by the operator, a confirmation step is applied. In the same way, in [23], the robot has a prior knowledge of spatial concepts, and the user's role is to assist it in the process of labeling places. Indeed, while walking with the robot, the user expresses what they considers relevant, for example: "*This is the corridor*" or "*This is the charging station*".

In another approach [39], as shown in Figure 2, the robot is guided by the user in a tour of the operational environment. It perceives its environment and detects the object that the user is pointing to with a commercial laser. Then, this object is segmented in the image and the robot estimates its position and orientation. Finally, it receives its description from the user via a speech-recognition module and associates it with its position to build the object representation to be added to the map. In [40], Bastianelli et al. improved this approach by implementing a system that enables adding new objects to the representation through continuous and online interaction with the user after the initial semantic mapping. Indeed, knowledge is acquired as needed by the robot and added progressively to the environment representation. For example, if the user gives a voice command indicating an unknown location for the robot, an integrated Petri net acquisition process is launched. During this process, the user can guide the system with voice commands such as "*Turn*

right", "*Follow me*" or "*Go to the kitchen*". When the robot is in front of the new location or object, the user can point to the object with the laser then tell the robot its label, e.g., "*This is the emergency door*".

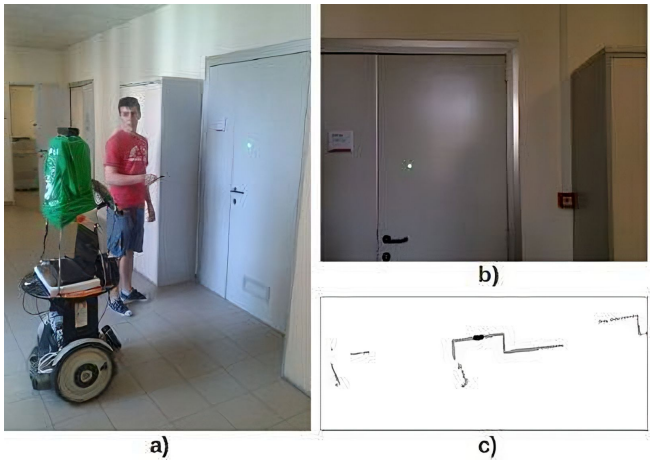

Figure 2. Adding semantic data to the map by tagging objects with a comercial laser pointer and giving a natural language description [40]. (**a**) The laser pointer is used to mark the object. (**b**) The point is detected in HSV color space. (**c**) The object is located in reference to the robot laser scan (represented by black pixels).

Pronobis and Jensfelt [31] proposed an algorithm that combines information about the presence of objects with the semantic properties of places, such as size and appearance, to classify rooms. During the mapping process, the user can input additional information about objects in the room using a graphical interface on the computer used by the robot. For example, if the user provides information about the existence of an object, the robot treats it as an additional source of information. The work of Crespo et al. [41] implemented natural language dialogues between the user and the robot through a voice interface and the keyboard. Therefore, it can add object categories and their semantic relations to the map. For example, the robot can ask the user about the possible uses of an object or about the possibilities of interaction with objects in the environment.

In conclusion, using HRI-based methods, the map representation can be enriched with various semantic information. Thus, by interacting with the robot via a mobile application, voice commands, or a user interface, an operator can include information about the objects in the scene (the categories of objects, the space they occupy, and their presence in a room), information about the locations (the categories of rooms and their characteristics), and information about the semantic relationships between the objects (the utility of an object). Moreover, while most works focus only on mapping static scenes, few works propose HRI-based data acquisition techniques to update the representation of dynamic features [40].

2.1.3. Reasoning-Based Semantic Data Acquisition

Reasoning is a way that allows the robot to acquire significant knowledge of its environment. It consists of using data already obtained by other methods and a knowledge database to infer new information about the environment. In general, the knowledge database contains common-sense knowledge, that is, knowledge that all humans possess and that they have acquired since their birth without even being aware of it [42]. This database is usually represented as a structure composed of concepts describing the environment linked by rules. For example, Galindo et al. [43] proposed a hierarchical structure of conceptual information modeled by the NeoClassic language [44]. The lowest level of this structure is composed of symbols related to physical elements detected in the environment. The conceptual structure is composed of laws that link these symbols to concepts and link concepts together to allow reasoning about symbols of known categories. For example,

if a detected physical element is related to the symbol "*oven-1*", and the symbol is related to the concept "*oven*", and the concept "*oven*" is related to the concept "*kitchen*" by a link "*contains*", then the robot's current location can be concluded as the kitchen.

In the case of [23], common-sense knowledge about an indoor office environment is modeled in the ontology shown in Figure 3a. An ontology is a data model representing a set of concepts in a domain and relationships between them. Based on this description, the system uses a reasoning software to derive more specific categories for known topological areas. Indeed, as shown in Figure 3b, if an area is classified as a room ("*area1*") and it contains a coffe machine, thus, using the conceptual knowledge given by the ontology, this area can be classified as a kitchen instance. Similarly, a semantic ontology is used in [45] to create a task scheduling strategy for service robots. In this ontology, the concept "*objects*" is separated into "*static objects*" and "*dynamic objects*". Then, the static objects are linked to the dynamic objects by probabilistic relations which allow to obtain the approximate positions of the dynamic objects by reasoning.

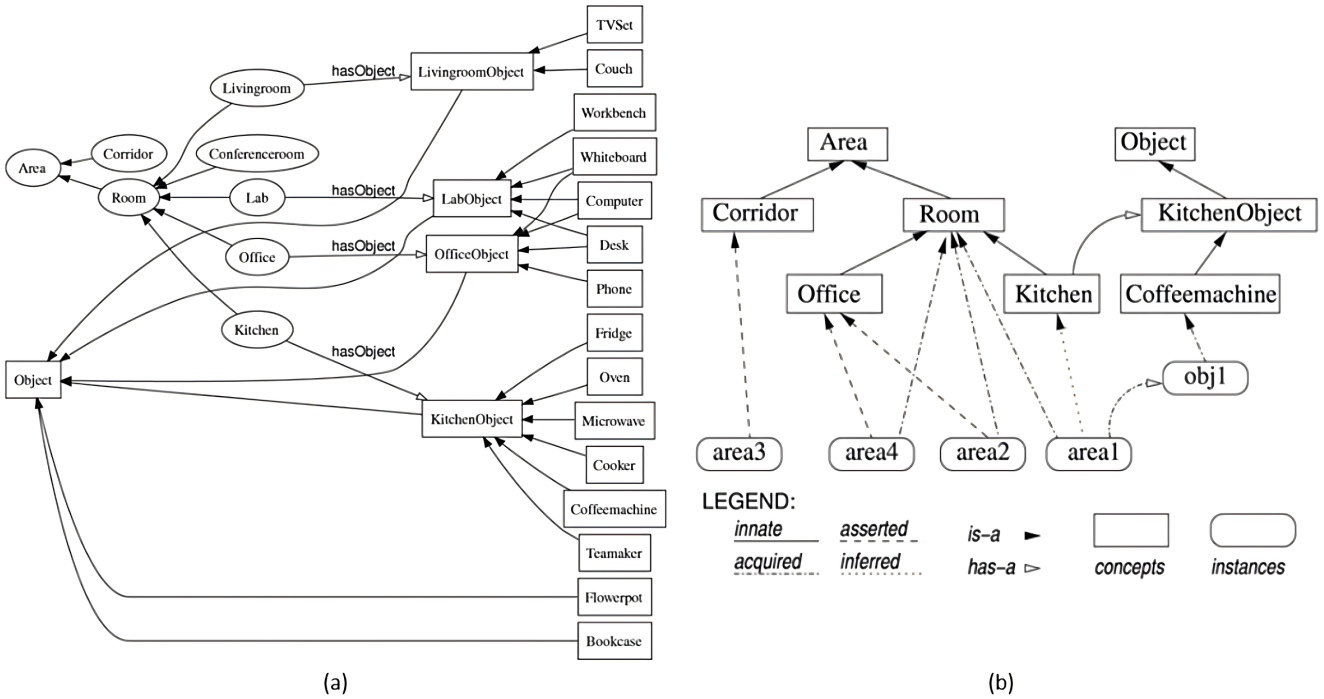

Figure 3. The example of reasoning-based data acquisition in [23]. (**a**) Illustration of a part of the common-sense ontology of an indoor office environment, (**b**) combining different types of knowledge in the conceptual map.

Furthermore, the authors of [31] propose a probabilistic conceptual structure where the rules linking the concepts can be predefined, acquired or inferred and can be probabilistic or not. The categories and properties of elements in the environment are estimated together using a probabilistic chain graph model. Room properties such as objects, appearance, area, and shape are extracted from sensors, then used with room–object relationships provided by the Open Mind Indoor Common Sense knowledge database to infer room categories. The authors of [41] proposed a relational model that allows to easily store information about the environment in tables managed by a reasoning engine. The design of this model implicitly defines the relationships between entities. Therefore, it is not necessary to manually define rules between them as for other structures. Using this model, if the robot identifies objects in the same room that have a common utility, it is possible to infer that these objects are located in a room where the robot could find other objects with related utilities. Thus, new room categories are created autonomously. In addition, the identification of objects and rooms allows the detection of possible inconsistencies.

In conclusion, a reasoning system can be implemented in the semantic-mapping process to infer additional semantic information. This system includes a common-sense knowledge database describing the application environment and a reasoning engine to exploit this knowledge. In general, the knowledge database is composed of predefined data and can be completed with data acquired by the robot during the mapping process.

2.2. Semantic-Map Representation

In order to exploit the semantic data extracted by the above methods, it is necessary to organize them in a structure suitable for the application, which is the output semantic map. There are two main options for the representation: either to attach the semantic information to the features detected in the geometric map, or to organize the geometric and semantic information in a hierarchical structure with different levels of abstraction. These two options are described in detail below.

2.2.1. Visual Geometric-Semantic Representations

This simply consists of attaching semantic information directly to the associated physical features in the geometric map. For example, Sunderhauf et al. [29] use ORB-SLAM2 to create a point cloud of the environment, and the points associated with each object are modeled by a different color. In [4], a semantic occupancy map is created by adding minimum bounding rectangles (MBRs), which represent objects' occupied spaces, to their topological areas in the map. The different object categories are represented by different colors. Zhao et al. [3] also used ROS's Gmapping module to create an occupancy map of the environment. During the process, as the robot receives semantic information, it combines it with its position and adds a new semantic node to the map.

These methods provide maps that enable direct interpretation of semantic information in positions. Therefore, they can be easily deployed in mobile robot navigation systems to perform semantic navigation tasks. However, such simple representations are insufficient to allow robots to perform complex tasks such as mission reasoning.

2.2.2. Hierarchical Representations

In order to allow robots to perform a reasoning process, more interesting representations have been proposed, namely, hierarchical representations, where specific information is placed in the lower levels and abstract information in the higher levels. Pronobis et al. [31] proposed a representation, shown in Figure 4, that gives the robot the ability to infer the categories of places using the properties of the space, including appearance, surface, shape and objects. They represent the spatial knowledge by a four-layer hierarchical structure: the sensory layer, the place layer, the categorical layer and the conceptual layer. The lowest level of the structure contains sensor perception data and the fourth level contains abstract conceptual knowledge. The sensory layer contains a metric map of the environment. The place layer contains a topological graph with nodes representing locations and edges encoding the path to other nodes. The categorical layer contains objects, landmarks, and both geometric and visual models. Finally, the conceptual layer contains a static common-sense knowledge ontology and the links between the concepts in this ontology and the low-level knowledge obtained from the other three layers. This structure not only allows the robot to classify rooms, but also to predict the existence of objects, properties of the space and unexplored spaces.

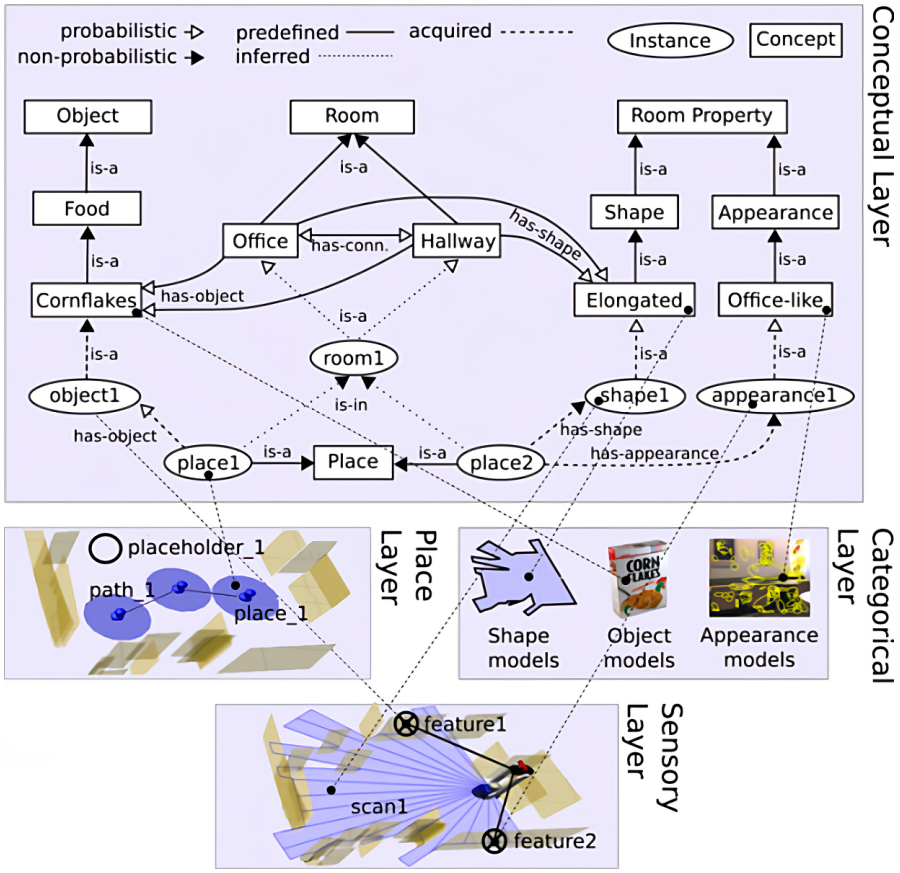

Figure 4. The layered structure of the spatial representation and a visualization of an extract of the conceptual-layer ontology proposed in [31].

While the environment in [31] is represented by a single hierarchical structure, in [43], a representation with two hierarchical structures was proposed: the first to represent spatial knowledge and the second to represent conceptual knowledge. On one side, the spatial hierarchy is composed of three levels. The first and the lowest level contains local occupancy maps, locations and images, stored by the robot. The second level contains a topological graph of the environment. In addition, the third level contains an abstract node that represents the whole environment. On the other hand, the conceptual hierarchy is composed of four levels. It is hand-coded in advance with the NeoClassic language. The top level contains a node called "*Thing*" from which two branches, "*Parts*" and "*Objects*", emerge. The third level contains object and room categories, and the lowest level contains object and room instances recognized by the robot. These two hierarchies are linked by the anchoring process, which consists of linking the entities recognized at different levels of the spatial hierarchy to the corresponding symbols that represent them in the conceptual hierarchy. These links allow the robot to exploit semantic information to determine localization errors by reasoning of the objects' expected locations and to perform semantic navigation tasks.

Similarly, in [40], the robot knowledge is divided into two parts: world knowledge, which represents the specific knowledge of a certain environment, and domain knowledge, which is the general knowledge of the application domain. The world knowledge is composed of the following elements: an occupancy grid, cell maps to represent local locations, a topological graph of the global environment and instances of recognized physical elements with their categories and properties. On the other hand, the domain knowledge is composed, as in [43], of the concepts involved in the environment as well as their properties and relationships. For this work, the main purpose of adding domain knowledge is to associate each perceived object with its spatial and functional properties in order to allow the robot to reason about the tasks related to these objects.

In conclusion [12], the representations that consist of visualizing the semantic information by different colors on the map allow to show results of semantic knowledge acquisition, but they are not easily implemented in the robots' tasks. On the other hand, the representations based on hierarchical structures do not allow to visualize this knowledge on a map, but they allow to structure and organize knowledge by levels of abstraction to make a robot able to reason. Table 1 provides a summary of all the previous sections, including the different semantic-mapping approaches used in works in the context of single-robot semantic mapping, the different methods used for semantic data acquisition, and the main data collected on objects and places.

Table 1. Summary of the reviewd works in relation to the single-robot semantic-mapping framework.

Ref.	Year	Semantic Data Sources	Approach	Collected Semantic Data	
				About Objects	About Places
[23]	2008	Perception - HRI Reasoning	Vision-based SLAM, object recognition	Categories, instances	Categories, instances
[31]	2012	Perception - HRI Reasoning	EKF-SLAM , recognition and instance reasoning, properties classification	Categories, relationships with other concepts	Appearance, surface, shape
[40]	2013	HRI	Incremental approach by HRI	Categories, positions, sizes, properties	Categories, robot positions
[3]	2015	HRI	RBPF-SLAM, voice recognition	_	Categories
[27]	2016	Perception	Monocular LSD-SLAM, 2D segmentation with CNN	Categories	_
[41]	2017	Reasoning	_	Instances, categories, utilities, characteristics, relationships with places	Instances, categories
[29]	2017	Perception	RGBD-based ORB-SLAM2, Object Detection, 3D Segmentation with CNN	Categories, instances	_
[4]	2020	Perception	Sonar-based SLAM, object detection, triangulation	Categories	Categories
[45]	2022	Perception–reasoning	Monocular SLAM, object detection	Categories, relationships with other concepts	Categories

3. Collaborative Semantic Mapping

Recently, collaborative robotics—which involves multiple agents, including robots, smart machines, and operators, working together to accomplish a task, as presented in Figure 5—has gained increasing interest in the research community. This collaboration provides a great opportunity to improve the performance of a fleet of robots. Indeed, at the individual level, it allows to increase the accuracy and robustness of a robot's estimations by integrating data collected by other agents and also to reduce its computing resources by distributing and parallelizing tasks. At the second level, it improves the efficiency of the fleet by task parallelization and by allowing the robots to substitute each other in case of a robot failure. However, on the other hand, this introduces some complexity in robotic systems and presents new challenges, especially in terms of inter-robot communication

and multi-robot data processing [46]. For example: *data exchange, data association and data representation in a global map.*

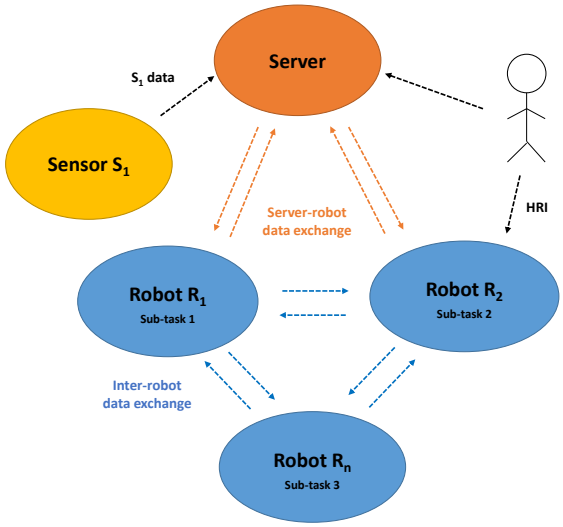

Figure 5. An example of a collaborative robotics setting.

In this context, over the past few years, a few researchers have started working on the multi-robot semantic-mapping problem, where a fleet of robots work together, while exchanging data, to create a global semantic map of the environment. These works are based on the progress made in various previous works on single-robot semantic mapping and take advantage of the advances made in the field of collaborative robotics, such as inter-robot communication methods and different data-allocation architectures to propose new consistent collaborative semantic-mapping approaches [47]. It is worth mentioning that all the works found and reviewed in this paper focus on multi-robot solutions since they only use mobile robots for mapping. Thus, in the rest of the paper, the use of the term "*collaborative*" refers to "*multi-robot*".

In the following sections, we first present a description of the multi-robot semantic-mapping system, then we detail the different tasks/data allocation architectures and existing strategies for multi-robot data association and fusion. It is important to mention that in this work, the focus is not on the works that treat the multi-robot semantic SLAM problem where object detection and recognition techniques are used to solve a SLAM problem, the focus is only on works that solve the collaborative semantic-mapping problem.

3.1. Multi-Robot Semantic-Mapping Pipeline

The multi-robot semantic-mapping system is mainly composed of the same main modules as the single-robot semantic-mapping system (Section 2). First, the geometric mapping module is in charge of creating the geometric representation of the environment. Then, the semantic segmentation or object-recognition module is in charge of understanding the environment and collecting semantic data. Last, the map-representation module is in charge of creating and updating the model of the environment.

Nevertheless, the collaborative semantic-mapping system includes an additional module, which can be called the data association and fusion module. It treats the problem of associating and fusing multi-robot data into a coherent global model. In addition, this system is generally characterized by three main features: the task and data distribution scheme, the communication policy and the multi-robot data association and fusion strategy.

The task and data distribution scheme: defines the system architecture, its components, the connections between them and the tasks assigned to each component.

The communication policy:

- Defines the topology of the exchanges: the different agents with which each robot has the physical possibility and the authorization to communicate;
- Defines the type of communicated data: sensor data or partial semantic maps;
- Plans the exchanges: which information to send, to which robot and at which time.

The multi-robot data association and fusion strategy: is the solution used to integrate the multi-robot exchanged data into a global semantic map.

3.2. Tasks and Data Allocation Architectures

Regarding the tasks and data distribution scheme, there are two main architectures that can be used to implement a collaborative semantic-mapping system: a centralized or a distributed architecture. They are different in several aspects, including the system components, the distribution of modules and the topology of exchanges between the different components.

3.2.1. Centralized Semantic-Mapping Architecture

The centralized architecture, presented in Figure 6, is composed of two main agents: the robots and the server. Indeed, the tasks are distributed and the data is exchanged between these two components in order to establish the semantic mapping. It is mainly used when the robots in the fleet have insufficient processing power to run the deployed system and when the communication infrastructure is always available and can support the necessary bandwidth [48]. In this case, the communication topology between the robots and the server can be a unidirectional or bidirectional communication. The first one consists in sending data in one direction only, from the robots to the server, and it is mainly used when the semantic map created by the server will not be used by the fleet of robots for their tasks and will be directly exploited for other server-related tasks. On the other hand, bidirectional communication is used to exchange data in both directions between the server and the robots. It is usually used to send the global semantic map created to the robot fleet so that they can use it in their tasks such as semantic navigation or location classification.

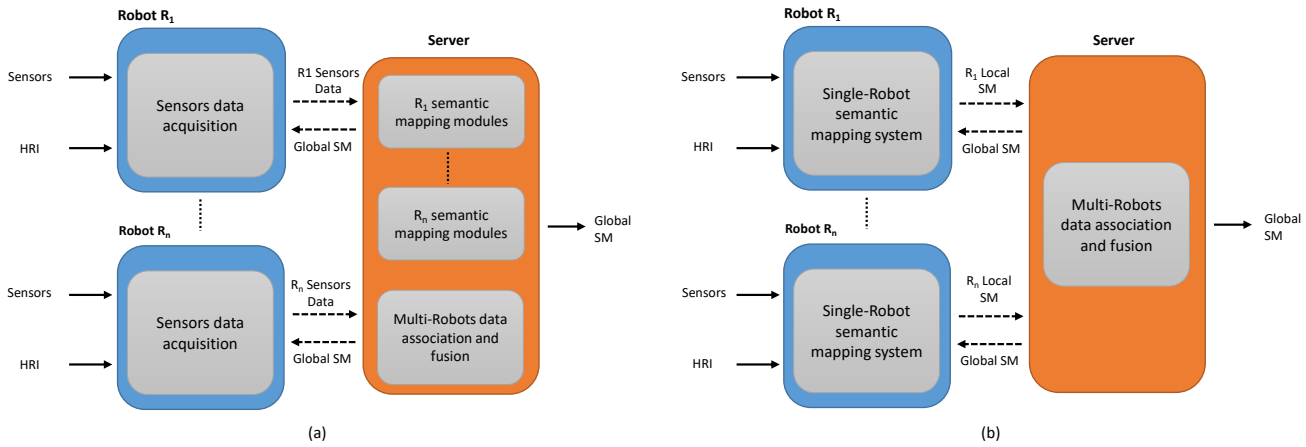

Figure 6. Representation of the distribution of a semantic-mapping system in a centralized architecture. (**a**) Robots play the role of sensors and the whole process of semantic mapping is performed in the server, (**b**) each robot implements a complete semantic-mapping system and the server manages the fusion and association of the multi-robot local maps.

Figure 6 presents two potential centralized formalizations of the collaborative semantic-mapping problem. They mainly depend on the degree of autonomy granted to the mobile robot and the degree of task integration in the server. For the first case, the robots are considered as mobile sensors. Their only task is to acquire data from the environment

using their sensors, and then send them to the server for processing and estimation of the global semantic map. More specifically, all the single-robot semantic-mapping modules, of each robot, are implemented in the server, mainly the SLAM task and the semantic data acquisition task. In addition, the server determines the correspondences between the data of different robots and fuses them in order to maintain a consistent global semantic map. In this case, the robots are very dependent on the server and have a low degree of autonomy. If the server or communication breaks down, the robot loses contact with the other robots.

In the second case, a complete semantic-mapping system is implemented in each robot, as described in Section 2. Then, the robots communicate their individual semantic maps to the server where only the inter-robot tasks are performed, namely, the multi-robot semantic-map association and fusion. In this case, the knowledge database is removed from the inputs of the single-robot semantic-mapping system and generally implemented in the server to be shared among all robots. This formalization makes the robots more autonomous, as they always have their own representations in case of server or communication failures. However, in some cases, the robot does not build a complete individual semantic map. Indeed, it runs basic modules such as the SLAM algorithm to collect geometric data and an object-recognition or semantic-segmentation algorithm to collect semantic data, but it does not create a local semantic map. Then, the data exchanged here is coupled with semantic and geometric information and not local maps. Therefore, the multi-robot data association and fusion module, implemented in the server, must process the correspondences and inconsistencies of the incoming data from several robots and integrate them into a single representation of the environment (Section 3.3.2).

Considering these formalizations, we can conclude that three types of data can be communicated by the robots to the server: sensor row data, separate geometric and semantic data, and local semantic maps. Therefore, each case has its specific requirements in terms of communication bandwidth latency and robot processing power.

In the literature, there are a few works that use a centralized architecture for semantic mapping, which covers the second case presented in Figure 6b. In this work [49], the authors proposed a server–client architecture for mapping a domestic environment. On the one hand, the client implements four main components: an object-recognition module, an object information-processing module, a robot-localization module and a communication module. Indeed, the object-recognition module is used to detect the robot's environment and the objects within it. The output of this component is used as input of the object information-processing module, which collects and packages information such as the object category, its size, its 3D position in the robot frame, its orientation and its confidence score. Meanwhile, the localization module is in charge of localizing the robot on a predefined map and collecting its position information. Then, a communication module transmits the resulting semantic and geometric informations to the server. On the other hand, the server is composed of several elements, the main ones being: the general manager, the object manager, the ontology manager and the graphical user interface. The general manager handles the data flow by creating connections with the clients. The object manager processes the received object detections and creates or updates objects in a virtual environment. The ontology manager manages the model containing the semantic informations. Finally, the graphical user interface allows an operator to adjust and modify the resulting map.

3.2.2. Distributed Semantic-Mapping Architecture

While a centralized architecture is composed of agents and a server, the distributed architecture, shown in Figure 7, is composed only of a set of robots that work together to estimate the global map. Indeed, each robot runs its own semantic-mapping system. Then, the multi-robot data association and fusion process is performed by a single robot or by each member of the fleet depending on the application, the processing power of the robots and the communication constraints, in order to fuse the resulting individual semantic maps.

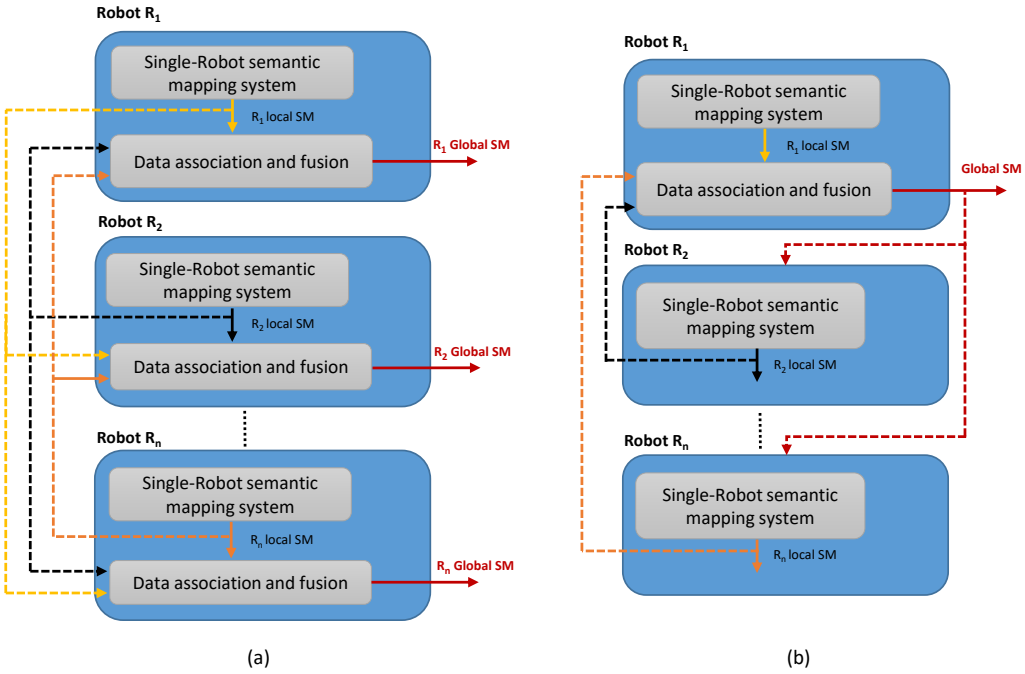

Figure 7. Representation of the distribution of a semantic-mapping system in a distributed architecture. (**a**) Each robot collects data from neighboring robots and establishes data association and fusion to create its global semantic map, (**b**) a single robot collects maps from neighboring robots, creates the global semantic map and distributes it to other robots.

This architecture is considered the appropriate technique when the bandwidth is limited and the robots have heterogeneous needs and capabilities [48], or when the application requires to maintain local autonomy for decision making (security, robustness, etc.). However, it is more challenging for inter-robot communication than the centralized architecture because the topology becomes unstructured, it can change dynamically depending on whether the other robots are in proximity or not, and also because of the data correlation problem that may occur in the case of a cyclic communication between robots. The main advantage of this architecture is that it ensures flexibility and autonomy of the robot fleet. Thus, in the case of a centralized architecture, when the server fails, all robots fail, whereas in a decentralized architecture, even if one robot fails, the other robots can continue their task using their own semantic maps and existing data from the other robots.

Figure 7 shows the two possible formalizations of the semantic-mapping problem in a distributed architecture. In Figure 7a, each robot builds its own local semantic map, and then receives the maps of its neighbors. Indeed, it receives data from robots with which communication is physically possible and within communication range at the time of multi-robot data gathering, to associate and merge them with its own map. In contrast, in the second scheme of Figure 7b, each robot builds its own local semantic map, but only one robot collects all the maps of the neighboring robots and runs the data association and fusion algorithm to create the global map, and then sends it to the robots that are within its range and needed for their task. This case is very similar to the second formalization of the centralized architecture, where each robot creates its local semantic map, and then the fusion is executed in the server. The only difference here is that the task of data association and fusion is performed by a robot instead of the server. In general, if the robots in the fleet are not identical, the robot with the highest computational power performs this task. It is also possible to have a third option by combining the two schemes of the Figure 7, where each robot maintains a global map and the global maps are exchanged between the robots.

In the literature, the majority of works that are interested in proposing a collaborative semantic-mapping system used a distributed architecture. For example, Yue et al. [47,50]

used the Husky Clearpath mobile robots, equipped with powerful controllers. Each robot estimates its local map and updates it incrementally using the partial maps received from the other robots. In [51], 12 marine robots were used, where each robot creates its local semantic model of an ocean environment, and then a fusion algorithm is run by any robot in the fleet that has collected the maps from the other robots in order to fuse them into a global map. The authors specify in this paper that this algorithm can be also implemented in a server and executed manually by an operator each time a new match is needed, but it is computationally light enough to be implemented by any robot in the fleet.

3.3. Multi-Robot Data Association and Semantic Maps Fusion

The data association and fusion process can be performed in two modes, either a "*real-time incremental mode*" or a "*one-shot mode*". In the first, the individual robots incrementally create their partial semantic maps and communicate them in real time to a fusion module to progressively build the global semantic map. In the second mode, the semantic maps of individual robots are created in advance and then fed to the fusion module to create the global semantic map in one step. Table 2 specifies the fusion mode for the reviewed works and provides an overview of the input data type of the fusion module, the methods used to solve this task, and the resulting map representation.

Table 2. Different data association and fusion techniques used in collaborative semantic mapping.

Ref.	Year	Fusion Mode	Input Data	Data Association and Fusion Methods		Map Representation
				Data association	Fusion	
[52]	2019	Incremental	Local grid-based multi-modal maps	Point-to-point matching	Map alignment	Global grid-based multi-modal map
[53]	2019	One-shot	Local OSTM [54]	Graph matching	Graph merging	Global OSTM
[47]	2020	Incremental	3D Octree local maps [55]	Voxel-to-voxel matching	Bayesian fusion	3D Octree global map
[51]	2021	One-shot	BNP-ROST unsupervised learning models [56]	Matching topics developed in individual robot models	Associating a single label to similar topics	Global semantic occupancy model
[49]	2021	Incremental	Coupled geometric and semantic informations	Grouping labels associated with the same physical element under a common virtual object	Assigning a parent label to each virtual object and linking it to the ontology	Labeled 3D virtual environment + Ontology

This process can be viewed as two main sub-problems: the data-association problem and the fusion and optimization problem. Thus, the data-association problem consists of establishing correspondences between local semantic maps using geometric and semantic data. In addition, the fusion and optimization problem consists in integrating multi-robot data into a global representation by using data associations and dealing with data dissimilarities. In what follows, we review the techniques proposed to address these two problems.

3.3.1. Data Association

The data-association techniques proposed in the reviewed papers are designed according to the type of semantic map generated by the single robot. Indeed, the local map type defines the kind of correspondences to be established between these maps. Moreover, these techniques depend on the characteristics of the mapped environment, such as whether it is dense or flat, static or dynamic, and feature-rich or feature-less. For example, it is easier to match data when the environment contains distinctive features than when it contains similar repeating patterns. Below, we classified the techniques used for semantic-map association into three main categories: metric–semantic maps matching, topological–semantic maps matching, and grouping data referring to the same element.

Metric–semantic maps matching: This category includes methods where the data-association problem is to match semantic-occupancy maps, specifically by determining point-to-point correspondences in the case of 2D local maps and voxel-to-voxel correspondences in the case of 3D local maps. For example, Yue et al. [47] focused on matching 3D semantic-occupancy grid maps to construct a global map of an indoor–outdoor environment. They consider that existing works using only geometric data, such as planes, lines and points, to establish data association between local maps are not very effective in feature-less environments. Therefore, they propose to match local maps using both semantic and geometric data. To this end, as shown in Figure 8, each robot follows a trajectory and performs a semantic-mapping system proposed by [55] in a part of the environment. Then, the generated local semantic maps, which are 3D occupancy maps divided into voxels (3D Octree maps), are merged. Indeed, each of these voxels has an occupancy probability and a semantic-class probability computed by applying the Bayes rule. In order to determine the data association, each robot receives the maps of the neighboring robots according to a certain permutation and calculates the relative transformation matrix between the received map and its own map using the matching algorithm proposed in [50]. Then, the expectation-maximization (EM) algorithm [57] is used to infer the hidden data association. This algorithm is divided into two steps: the E-step and M-step. The E-step establishes the correspondences between the robot map and the received map by computing the minimum relative distances using geometric and semantic data, and the M-step uses these correspondences to update the occupancy and the semantic-class probabilities of the robot global map voxels. In this work, the main advantage is that the EM algorithm can assign a probabilistic data association and update it iteratively instead of assigning hard decisions.

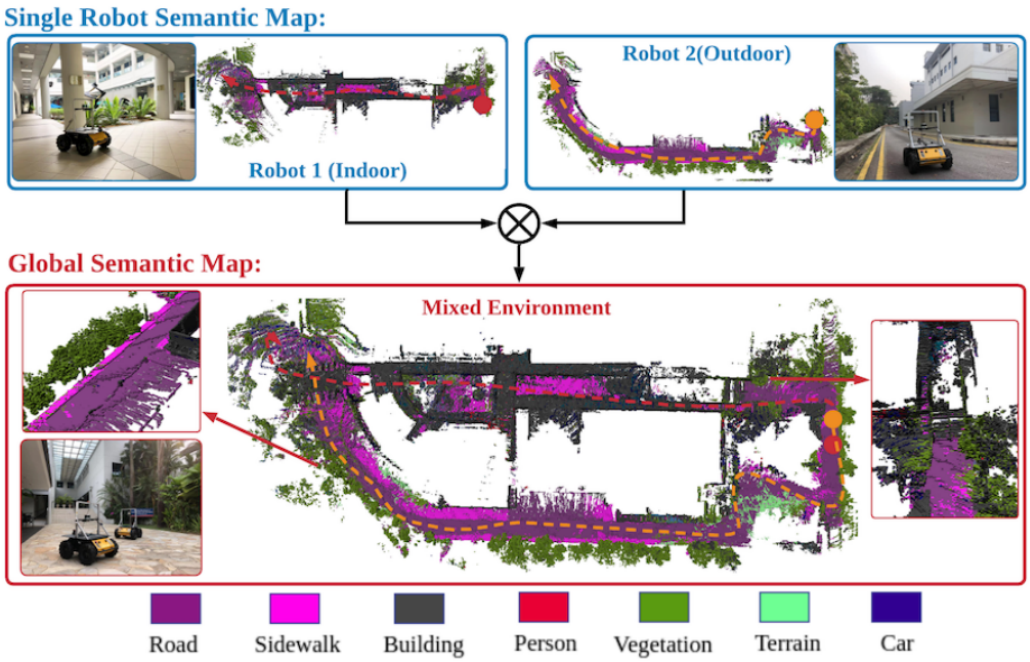

Figure 8. An example of collaborative semantic mapping in mixed indoor–outdoor environment. Red trajectory is robot 1, and orange trajectory is robot 2. Top: local semantic maps generated by two robots. Bottom: collaboratively generated global semantic map [47].

In [47], voxel-to-voxel correspondences were established to associate local 3D semantic maps; in [52], point-to-point correspondences were established to associate local 2D semantic maps. This work is different from other works because it proposes a method to establish the association of data coming from heterogeneous robots, namely, an aerial vehicle map and a ground vehicle map. This method is specifically designed for the agricultural context, where the generated maps are typically composed of similar and repetitive patterns. Specifically, the aerial vehicle can quickly provide a coarse reconstruction of a large area,

which can be updated with portions of more detailed, higher resolution maps generated by the ground vehicle visiting selected areas. First, a 2D grid map is created for each robot using its colored point cloud generated from the collected GPS/IMU-tagged RGB images. This map is composed of a multi-modal semantic representation of the environment, where each cell stores the vegetation index and height information of the local surface. Then, the idea to match these two maps is to find a dense flow of point-to-point correspondences between them using the large diplacement dense optical flow (LDOF) system [58]. The key intuition for using this method is that points belonging to one cloud locally share similar displacement vectors that associate them with points in the other cloud. The LDOF system is modified to fit the environment specificity by including the vegetation index and the geometric properties of the surface in the calculation of the cost function.

Topological–semantic maps matching: In this method, the data-association problem is defined in a different way. Indeed, the type of local maps is a graph augmented with semantic information. The problem is, therefore, to match the nodes of these graphs using semantic and geometric data. This issue is addressed in [53], where the authors propose a different technique for matching topological–semantic maps. They did not use the state-of-the-art sub-graph fusion techniques based on combining geometric and topological information [59,60], but instead proposed a new technique based solely on semantic and topological information. It does not require the global or relative pose of the robots to merge the maps. Each robot builds its partial oriented semantic topological map (OSTM) using the exploration algorithm [54]. This map is a semantically structured topological graph of the environment, where vertices represent locations in a building (room, hallway) and edges connect locations with a direction to traverse them. Then, to merge the sub-graphs, a bio-inspired algorithm [61], called inverse Warrington's object recognition model (IWORM), is used. It makes the correspondences between the two subgraphs using two comparison layers. The first layer, called semantic categorization, compares the two sub-graphs in terms of node orientation and semantic label. The second layer, called perceptual categorization, compares the two sub-graphs in terms of local structure, such as the number of neighbors of a vertex and the connections between those neighbors, in order to address labeling and orientation errors that may occur during the single-robot semantic-mapping process.

Grouping labels referring to the same element: In this third category, the main objective of the data association is to group together the labels develeped by different robots and referring to the same physical element of the environment. For example, this method is used in [49], where the authors proposed a server–client system to build and maintain a semantic map of a dense domestic environment. This system generates a 3D environment reconstructed under Unity 3D, where each detection is added as a label to the virtual environment and linked to a semantic concept in an ontology. In effect, each robot processes its detections to determine object categories, 3D positions, dimensions and confidence scores. Then, it sends these data with its position to the server. The latter first determines the global positions of all detected objects in the reference system of the virtual environment. Then, it establishes multi-robot data correspondences using the physics provided by Unity 3D. It actually detects when virtual objects are close to each other or even in contact with each other with respect to a distance threshold. In the case of a collision, if they also share the same labels, then they are considered to refer to the same object. This method of data association is quite simple, but its main difficulty lies in the definition of the threshold distance allowing to consider the existence of a collision or not, because the objects in the environment have variable sizes. Thus, in some cases, the distance between two labels can be large enough to consider that they refer to two distinct objects, while in reality they refer to the same large object.

In the same spirit, the data-association method proposed in [51], consists of estimating the number of unique phenomena developed by different marine robots after completely exploring their area of the environment. As shown in Figure 9, each robot develops its individual semantic model of the environment, and then the correspondences between them are resolved into a global semantic map. While in [49], a pre-trained deep-learning

model is used for semantic mapping, so that the labels are known; in this work, the online unsupervised BNP-ROST model [56] is used, so that the detected features are developed as topics without associated labels (Topic 1, Topic 2, . . .) and are different from one robot to another. To establish the data association, the designed algorithm uses the final models of all robots to compute a topic similarity matrix to identify when the robots have developed semantically equivalent topics. Then, noisy topic similarities are removed in a two-step process. In the first step, similarity weights below a defined threshold are removed to obtain the noisy association graph. Then, in the second step, the obtained noisy association graph is rectified using the CLEAR algorithm [62] to generate the final cluster graph, where the topics belonging to the same cluster represent a unique phenomena. This method is very interesting for collaborative exploration of unknown environments, but it is highly dependent on the performance of individual robot models, which do not require pre-training, but need to be well-tuned by many varying parameters depending on the mission, which is not such a simple task.

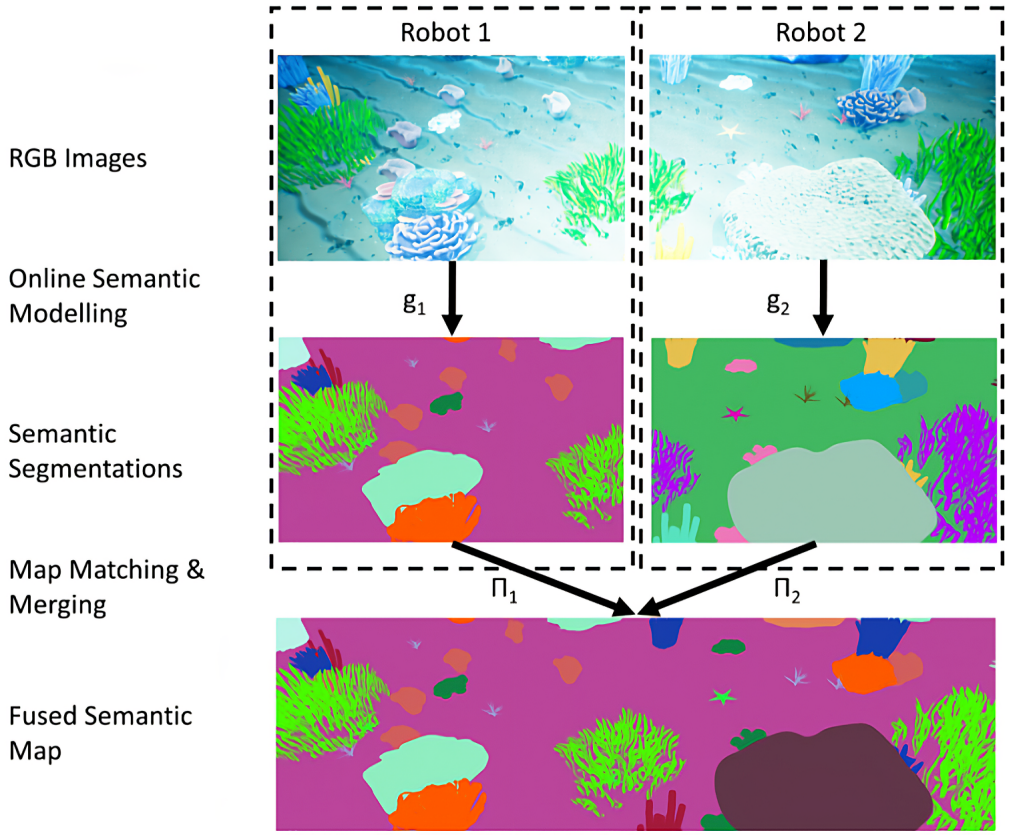

Figure 9. The fusion of two different semantic models developed by two robots in the same environment [51].

3.3.2. Fusion and Optimization

In order to integrate multi-robot local semantic maps into a global representation, it is important to have a module that handles inconsistencies and dissimilarities between them. The fusion and optimization process is responsible for this task. Indeed, it is performed after the data-association process and builds the global semantic map using the established correspondences. The solutions proposed in the literature address several types of dissimilarities, mainly dissimilarities caused by multiple view-point errors, annotation errors, and localization errors, as presented in Table 3.

Table 3. Sources of dissimilarity addressed in the reviewed works in relation to the collaborative semantic-mapping framework.

Ref.	Year	Sources of Dissimilarity		
		Multiple View-Points Errors	Annotation Errors	Localization Errors
[52]	2019	X	X	X
[53]	2019	-	X	-
[47]	2020	X	-	-
[51]	2021	-	-	-
[49]	2021	X	-	-

The most addressed type is dissimilarities caused by sensing the environment from multiple viewpoints. In that case, the detected features may have different semantic labels and some variation in their positions in the corresponding maps. For example, in [47], the authors consider that the same object can be observed from different viewpoints by different robots, and voxels representing the same object can have different semantic classes. Therefore, the fusion method they propose is capable of correcting false labels and improving true labels. After computing the exact transformation matrix between the neighboring robot's 3D Octree map and its own map, the robot uses it to perform pair-wise voxel fusion. Since each voxel is considered as a Gaussian distribution, the problem of pair-wise voxel fusion consists of integrating two Gaussian distributions. Therefore, the fusion of 3D Octree maps consists of integrating multi-dimensional Gaussian distributions into a global representation. For this purpose, the authors used the 3D statistical fusion method proposed in [63]. The experiments conducted in this paper demonstrate that this fusion method can obtain a high-quality 3D Octree map in dense and large environments.

This same problem is addressed in [49], where identical mobile robots explore the environment at different times following the same trajectories, but the only difference between them is that the camera's point of view is changed. Each robot sends its detections to the server. Then, the server checks whether this object has been detected before or not, as explained in Section 3.3.1. If the object is detected for the first time, it is added to the virtual environment as a 3D bounding box with its respective dimension and is represented in the ontology as an instance of its associated category or concept. Nevertheless, if it has been detected before, the properties of all detections are merged to obtain a virtual parent object that encapsulates: the category of the item, the average confidence score, the number of items that have been merged and the union of the child bounding boxes. After that, the corresponding instance of the object in the ontology is updated with the new properties of the virtual parent. However, a record of all previous detections is maintained by entering each of them as an instance of the parent object. It is important to mention that this work is one of the few papers that address the problem of merging semantic-map representations containing a conceptual part (the ontology), as other works only treat the merging of annotated spatial representations.

Regarding dissimilarities caused by localization and annotation errors, the proposed method in [52] addresses various related problems such as local inconsistencies, global deformation and relatively large initial misalignment. For this purpose, after computing the dense set of point-to-point correspondences between an air-vehicle map and a ground-vehicle map, the largest set of correspondences with consistent and similar flows is identified and used to infer a preliminary alignment transformation between the maps. This step deletes dissimilarities resulting from local inconsistencies that may be caused by detection or communication errors. Then, in order to deal with the global deformation resulting from the inaccuracy of the location and orientation data provided by GPS (Global Positioning System) and AHRS (Attitude and Heading Reference System), a non-rigid point-set registration algorithm is used to estimate an affine transformation. Finally, the global semantic map is obtained by performing robust point-to-point registration on the

input point clouds using only the points belonging to the vegetation in the local semantic maps. In another work [53], the global semantic map is a topological graph connecting many robot sub-graphs. In this particular work, only semantic data is used to match the maps, so there are no errors caused by metric data, but there are dissimilarities caused by labeling errors, such as the association of a wrong label or the absence of a node.

Finally, in addition to the automatic fusion systems described above, some works have proposed a system based on human intervention to optimize and maintain the resulting global representation. Indeed, in [49], since the robots map the environment at different times, the authors considered that the location of an object mapped by the first robot can be modified before the other robot maps the environment. This case is not solved by the automatic system designed to merge dissimilarities caused by multiple viewpoints. Therefore, they included the option for the operator to modify the final reconstructed labeled 3D map by replacing or deleting erroneous labels. It is also important to mention that some works do not focus on the fusion and optimization part but only on solving the data-association problem. For example, in [51], after clustering the topics associated with the unique phenomena explored in the ocean, the positions of the marine robots are considered error-free, so the fusion part consists only of concatenating the local semantic maps and segmenting the topics referring to the same phenomena with the same color.

In conclusion, the fusion and optimization methods proposed in the literature mainly deal with multiple viewpoint errors and annotation errors in order to integrate local semantic maps into a global map. The majority of works consider that robot positions exist and attempt to optimize the detection positions by resolving inconsistencies in the corresponding maps. Meanwhile, few works take into account dynamic objects in the fusion process.

4. Open Problems and Ongoing Trends

Although many semantic-mapping systems have been proposed, there are still challenges and possible improvements in the state-of-the-art solutions for both single-robot systems and collaborative systems. In this section, we highlight some of the potential improvements and remaining challenges.

4.1. Semantic Data Gathering Challenges

Currently, sensors such as lasers and RGB/RGBD cameras are widely used in the semantic-mapping process. In order to extract the semantic information, this input data must be processed. However, it is possible to equip the robot with additional specific sensors to directly extract the semantic information. For example, a temperature sensor to collect temperature values or a humidity sensor to collect humidity values. Regarding collaborative semantic mapping, most works collect semantic data using a fleet of mobile robots. However, it is possible to add static sensors to collect specific semantic data, for example, a camera implementing a people-detection algorithm to calculate the number of people in a scene, or a temperature sensor to provide the temperature variation over time within a scene, giving continuous data of a scene instead of a limited set of data collected when the robot explores that local area.

4.2. Map Representation Challenges

There are several challenges that can be addressed at the representation level. For example:

4.2.1. Task-Oriented Map Representation

The semantic-mapping process is generally designed with consideration of the map representation needed to accomplish the task. Today, there are many proposed representations, but they solve only a few specific tasks, mainly semantic navigation and room classification. Indeed, it is still difficult to implement these maps to solve more complex problems, such as task planning. For example, the robot uses the map to execute a command such as "*Bring me a cup*", where it must use the map to navigate to the kitchen, find

the exact location of the cup, pick it up, and bring it to the user. In addition, if it finds the kitchen door closed, it is able to understand it and open the door to continue its task. In order to perform this, the map representation must contain the data needed to plan the task, and the robot must have the appropriate reasoning system to interpret its commands and make a decision among a set of possible actions depending on the situation.

4.2.2. Context-Aware Map Representation

Some of the existing semantic-mapping systems build metric semantic maps, such as semantic occupancy grid maps, which allow the robot to interpret the positions of represented elements using their associated labels. Other works build hierarchical semantic maps, which give a context-aware representation, where the detected features are associated using their labels to a common-sense database giving a more detailed description of the context, especially about the concepts involved in the environment and the relationships between them. For single-robot semantic-mapping systems, there are many works that focus on building a context-aware map representation, but for collaborative systems, most of the existing works focus on fusing local geometric-semantic maps and do not focus on building a shared contextual database of the environment. There are only a few works that have built and maintained a common-sense database on the server that is shared by all robots. This database is populated using data collected by multiple robots. Therefore, this can cause some problems due to multi-robot data dissimilarities, but it permits to build a much detailed and extensive description of the environment to expand the knowledge database of individual robots.

4.2.3. Knowledge Database Representation

Despite the variety in the hierarchical representations proposed, most works use an ontology model to describe the contextual knowledge database. There are few works that use other structures as the relational semantic model proposed in [41]. Indeed, a possible direction in future works could be to explore other data representation models in order to optimize the reasoning system or to facilitate the reuse of semantic-mapping systems in different environments. In current works focusing on domestic and office environments, the models include concepts and connections specific to these environments, such as "*kitchen*" and "*living room*" as room categories, "*supplies*" and "*offices*" as object categories, and "*oven is in the kitchen*" as an object–place relationship. When transferring the semantic-mapping system to another environment with a different context, it is necessary to design a new ontology. For example, for an industrial environment, new concepts and relationships are needed in the model, such as "*workshop*" and "*warehouse*" as place categories, "*machine*" and "*container*" as object categories, and "*container is in the warehouse*" as an object–place relationship. In other cases [41], the semantic relational model can be directly adapted to the industrial environment because the relationships between the concepts are implicitly included in the model design. It is, therefore, sufficient to modify the information on the domestic environment by information on the industrial environment to adapt it to the new context.

4.3. Semantic Mapping of Dynamic Environments

Although many single-robot semantic-mapping systems have been proposed, most of the works focus on static environments. The majority of them propose a process to create the initial map of the environment, but they do not consider maintaining it. Only a few works designed a system to update the map to take into account moving objects, newly introduced objects, and deleted objects. To perform this process, these systems are not autonomous and require the assistance of an operator. Moreover, only a few works considered collecting semantic data about humans in the environment, which can be interesting in a collaborative context. Similarly, in collaborative semantic-mapping systems, the majority of works focus on static scenes, and only a few works considered updating the dynamic feature representation after the mapping process. Updating the map through

an autonomous process is a challenging task that could be addressed in future works on semantic mapping. Indeed, such a system is very important for mapping a dynamic environment, where the structure changes at a high frequency. For example, in an industrial environment, where the human and the robot share the same workspace and where objects can be introduced and removed every second [64], a real-time semantic map is needed to give the robot real-time information about its environment and to ensure the safety of the human–robot collaboration.

4.4. Collaborative Semantic Mapping of Indoor Environments

While many works on single-robot semantic mapping focus on indoor environments, the few recent works that address collaborative semantic mapping focus primarily on outdoor environments. Indeed, in these environments, the area to be mapped is large and using a fleet of robots has many potential advantages, such as distributing tasks to reduce work time and obtaining a more accurate representation of the environment. For indoor environments, work in the literature focuses on the semantic mapping of office and domestic environments to facilitate the integration of robots into these environments and extend their capabilities. In these settings, a single service robot is typically used, so the collaborative approach is not very interesting, but in other indoor environments that could be explored in future work, there may be benefits. For example, in industrial environments, where high precision is required, a collaborative approach can provide a more accurate representation of the environment.

4.5. Semantic-Map Fusion Challenges

While some data-fusion problems have been addressed, such as dissimilarities caused by multiple viewpoints or mislabeling, many other problems can still be explored. While currently most works address the problem of merging local semantic maps of the same type, a first possible challenge to be explored in future works is the merging of heterogeneous semantic maps. This solution can be very interesting when the robots in the fleet have different sensing capabilities. Moreover, in most works, the final global map output has the same type of local semantic map, so another challenge is to build a different representation of the output map. For example, in a centralized architecture, the map representation can benefit from the server data in addition to the local maps of the robots to build a better map representation. Finally, it is possible to consider optimization challenges such as the optomization of the data exchanged between the robots to reduce the processing power and increase the performance of the system, especially in real-time applications.

5. Conclusions

This paper reviews recent work on single-robot semantic indoor mapping and collaborative semantic mapping. In general, for single-robot semantic mapping, state-of-the-art SLAM solutions are used to generate a geometric representation of the environment. Then, several semantic-data acquisition methods are used to augment this representation. In recent works, object detection/recognition techniques and deep-learning-based segmentation techniques are widely used for semantic-information extraction from RGB or RGB-D images. Alternatively, HRI-based acquisition methods integrate the human into the semantic-mapping process in order to label the physical features perceived by the robot or to introduce additional information about the features already detected by the sensors using different human–robot interfaces. Reasoning-based acquisition methods allow the robot to infer new information useful to its mission using previously acquired data and a knowledge database about the environment. To achieve a semantic map, there are two possible representations, either by directly visualizing the semantic information on its area in the geometric representation, or by organizing the semantic and geometric information in a hierarchical structure, which is more suitable for semantic-map applications such as semantic navigation and task planning. Regarding collaborative robotics, this review shows that there are few works that address collaborative semantic mapping. Most of these

works focus on two main topics: proposing a centralized or distributed semantic-mapping architecture, and addressing the challenge of multi-robot data association and fusion. The proposed solutions mainly focus on outdoor environments. They are dependent on the environment and the characteristics of the fleet of robots. Although considerable progress has been made in the field of semantic mapping in recent years, this paper shows that there are still some challenges to be addressed, including semantic data gathering, map representation, environment and collaboration challenges.

Author Contributions: Supervision: A.A., H.A.-A., Y.D. and M.E.Z.; writing—original draft preparation, A.A.; writing—review and editing, A.A., H.A.-A., Y.D. and M.E.Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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.

References

1. Wang, F.; Zhang, C.; Tang, F.; Jiang, H.; Wu, Y.; Liu, Y. Lightweight Object-level Topological Semantic Mapping and Long-term Global Localization based on Graph Matching. *arXiv* **2022**, arXiv:2201.05977.
2. Nüchter, A.; Hertzberg, J. Towards semantic maps for mobile robots. *Robot. Auton. Syst.* **2008**, *56*, 915–926. [CrossRef]
3. Zhao, C.; Mei, W.; Pan, W. Building a grid-semantic map for the navigation of service robots through human–robot interaction. *Digit. Commun. Netw.* **2015**, *1*, 253–266. [CrossRef]
4. Qi, X.; Wang, W.; Yuan, M.; Wang, Y.; Li, M.; Xue, L.; Sun, Y. Building semantic grid maps for domestic robot navigation. *Int. J. Adv. Robot. Syst.* **2020**, *17*. [CrossRef]
5. Galindo, C.; Fernández-Madrigal, J.A.; González, J.; Saffiotti, A. Robot task planning using semantic maps. *Robot. Auton. Syst.* **2008**, *56*, 955–966. [CrossRef]
6. Kantaros, Y.; Pappas, G.J. Optimal temporal logic planning for multi-robot systems in uncertain semantic maps. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4127–4132.
7. Kantaros, Y.; Kalluraya, S.; Jin, Q.; Pappas, G.J. Perception-based temporal logic planning in uncertain semantic maps. *IEEE Trans. Robot.* **2022**. [CrossRef]
8. Kostavelis, I.; Gasteratos, A. Semantic mapping for mobile robotics tasks: A survey. *Robot. Auton. Syst.* **2015**, *66*, 86–103. [CrossRef]
9. Liu, Q.; Li, R.; Hu, H.; Gu, D. Extracting semantic information from visual data: A survey. *Robotics* **2016**, *5*, 8. [CrossRef]
10. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, D.; Neira, J.; Reid, I.; Leonard, J.J. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robot.* **2016**, *32*, 1309–1332. [CrossRef]
11. Crespo Herrero, J.; Castillo Montoya, J.C.; Martínez Mozos, Ó.; Barber Castaño, R.I. Semantic information for robot navigation: A survey. *Appl. Sci.* **2020**, *10*, 497. [CrossRef]
12. Han, X.; Li, S.; Wang, X.; Zhou, W. Semantic Mapping for Mobile Robots in Indoor Scenes: A Survey. *Information* **2021**, *12*, 92. [CrossRef]
13. Saeedi, S.; Trentini, M.; Seto, M.; Li, H. Multiple-robot simultaneous localization and mapping: A review. *J. Field Robot.* **2016**, *33*, 3–46. [CrossRef]
14. Wolf, D.F.; Sukhatme, G.S. Semantic mapping using mobile robots. *IEEE Trans. Robot.* **2008**, *24*, 245–258. [CrossRef]
15. Bernuy, F.; Ruiz del Solar, J. Semantic mapping of large-scale outdoor scenes for autonomous off-road driving. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 35–41.
16. Bai, Y.; Fan, L.; Pan, Z.; Chen, L. Monocular Outdoor Semantic Mapping with a Multi-task Network. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 1992–1997.
17. Lang, D.; Friedmann, S.; Hedrich, J.; Paulus, D. Semantic mapping for mobile outdoor robots. In Proceedings of the 2015 14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 18–22 May 2015; pp. 325–328.
18. Atanasov, N.; Zhu, M.; Daniilidis, K.; Pappas, G.J. Semantic Localization Via the Matrix Permanent. In Proceedings of the Robotics: Science and Systems, Berkeley, CA, USA, 12–16 July 2014; Volume 2, pp. 1–10. Available online: https://www.x-mol.com/paper/1477015746142314496 (accessed on 10 September 2022).
19. Reid, I. Towards semantic visual SLAM. In Proceedings of the 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), Singapore, 10–12 December 2014; p. 1.

20. Kundu, A.; Li, Y.; Dellaert, F.; Li, F.; Rehg, J.M. Joint semantic segmentation and 3d reconstruction from monocular video. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 703–718. Available online: https://link.springer.com/chapter/10.1007/978-3-319-10599-4_45citeas (accessed on 10 September 2022).
21. Bowman, S.L.; Atanasov, N.; Daniilidis, K.; Pappas, G.J. Probabilistic data association for semantic slam. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 1722–1729.
22. Lee, K.; Lee, S.J.; Kölsch, M.; Chung, W.K. Enhanced maximum likelihood grid map with reprocessing incorrect sonar measurements. *Auton. Robot.* **2013**, *35*, 123–141. [CrossRef]
23. Zender, H.; Mozos, O.M.; Jensfelt, P.; Kruijff, G.J.; Burgard, W. Conceptual spatial representations for indoor mobile robots. *Robot. Auton. Syst.* **2008**, *56*, 493–502. [CrossRef]
24. Folkesson, J.; Jensfelt, P.; Christensen, H.I. Vision SLAM in the measurement subspace. In Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18–22 April 2005; pp. 30–35.
25. Ekvall, S.; Kragic, D. Receptive field cooccurrence histograms for object detection. In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 84–89.
26. Miyamoto, R.; Adachi, M.; Nakamura, Y.; Nakajima, T.; Ishida, H.; Kobayashi, S. Accuracy improvement of semantic segmentation using appropriate datasets for robot navigation. In Proceedings of the 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), Paris, France, 23–26 April 2019; pp. 1610–1615.
27. Li, X.; Belaroussi, R. Semi-dense 3d semantic mapping from monocular slam. *arXiv* **2016**, arXiv:1611.04144.
28. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-scale direct monocular SLAM. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 834–849. Available online: https://link.springer.com/chapter/10.1007/978-3-319-10605-2_54 (accessed on 10 September 2022).
29. Sünderhauf, N.; Pham, T.T.; Latif, Y.; Milford, M.; Reid, I. Meaningful maps with object-oriented semantic mapping. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 5079–5085.
30. Mur-Artal, R.; Tardós, J.D. Probabilistic Semi-Dense Mapping from Highly Accurate Feature-Based Monocular SLAM. In Proceedings of the Robotics: Science and Systems, Rome Italy, 13–17 July 2015, Volume 2015. Available online: https://www.researchgate.net/publication/282807894_Probabilistic_Semi-Dense_Mapping_from_Highly_Accurate_Feature-Based_Monocular_SLAM (accessed on 10 September 2022).
31. Pronobis, A.; Jensfelt, P. Large-scale semantic mapping and reasoning with heterogeneous modalities. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 3515–3522.
32. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857
33. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
34. Wang, C.; Wang, C.; Li, W.; Wang, H. A brief survey on RGB-D semantic segmentation using deep learning. *Displays* **2021**, *70*, 102080. [CrossRef]
35. Zhang, J.; Zhao, X.; Chen, Z.; Lu, Z. A review of deep learning-based semantic segmentation for point cloud. *IEEE Access* **2019**, *7*, 179118–179133. [CrossRef]
36. Xie, Y.; Tian, J.; Zhu, X.X. Linking points with labels in 3D: A review of point cloud semantic segmentation. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 38–59. [CrossRef]
37. Grisetti, G.; Stachniss, C.; Burgard, W. Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling. In Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18–22 April 2005; pp. 2432–2437.
38. Grisetti, G.; Stachniss, C.; Burgard, W. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE Trans. Robot.* **2007**, *23*, 34–46. [CrossRef]
39. Randelli, G.; Bonanni, T.M.; Iocchi, L.; Nardi, D. Knowledge acquisition through human–robot multimodal interaction. *Intell. Serv. Robot.* **2013**, *6*, 19–31. [CrossRef]
40. Bastianelli, E.; Bloisi, D.D.; Capobianco, R.; Cossu, F.; Gemignani, G.; Iocchi, L.; Nardi, D. On-line semantic mapping. In Proceedings of the 2013 16th International Conference on Advanced Robotics (ICAR), Montevideo, Uruguay, 25–29 November 2013; pp. 1–6.
41. Crespo, J.; Barber, R.; Mozos, O. Relational model for robotic semantic navigation in indoor environments. *J. Intell. Robot. Syst.* **2017**, *86*, 617–639. [CrossRef]
42. Darlington, K. Common Sense Knowledge, Crucial for the Success of AI Systems. *OpenMind BBVA* **2020**.
43. Galindo, C.; Saffiotti, A.; Coradeschi, S.; Buschka, P.; Fernandez-Madrigal, J.A.; González, J. Multi-hierarchical semantic maps for mobile robotics. In Proceedings of the 2005 IEEE/RSJ International Conference On Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 2278–2283.
44. Patel-Schneider, P.F.; Abrahams, M.; Resnick, L.A.; McGuinness, D.L.; Borgida, A. *Neoclassic Reference Manual: Version 1.0*; Artificial Intelligence Principles Research Department, AT&T Labs Research: 1996. Available online: http://www.bell-labs.com/project/classic/papers/NeoTut/NeoTut (accessed on 26 July 2022).

45. Wang, Z.; Tian, G. Hybrid Offline and Online Task Planning for Service Robot Using Object-Level Semantic Map and Probabilistic Inference. *Inf. Sci.* **2022**, *593*, 78–98. [CrossRef]
46. Dubois, R.; Eudes, A.; Frémont, V. Sharing visual-inertial data for collaborative decentralized simultaneous localization and mapping. *Robot. Auton. Syst.* **2022**, *148*, 103933. [CrossRef]
47. Yue, Y.; Zhao, C.; Li, R.; Yang, C.; Zhang, J.; Wen, M.; Wang, Y.; Wang, D. A hierarchical framework for collaborative probabilistic semantic mapping. In Proceedings of the 2020 IEEE international conference on robotics and automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 9659–9665.
48. Martins, G.S.; Ferreira, J.F.; Portugal, D.; Couceiro, M.S. MoDSeM: Modular framework for distributed semantic mapping. In *Poster Papers*; 2019; p. 12. Available online: https://www.ukras.org.uk/publications/ras-proceedings/UKRAS19/pp12-15 (accessed on 10 September 2022).
49. Fernandez-Chaves, D.; Ruiz-Sarmiento, J.R.; Petkov, N.; Gonzalez-Jimenez, J. ViMantic, a distributed robotic architecture for semantic mapping in indoor environments. *Knowl.-Based Syst.* **2021**, *232*, 107440. [CrossRef]
50. Yue, Y.; Wen, M.; Zhao, C.; Wang, Y.; Wang, D. COSEM: Collaborative Semantic Map Matching Framework for Autonomous Robots. *IEEE Trans. Ind. Electron.* **2021**, *69*, 3843–3853. [CrossRef]
51. Jamieson, S.; Fathian, K.; Khosoussi, K.; How, J.P.; Girdhar, Y. Multi-Robot Distributed Semantic Mapping in Unfamiliar Environments through Online Matching of Learned Representations. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 8587–8593.
52. Potena, C.; Khanna, R.; Nieto, J.; Siegwart, R.; Nardi, D.; Pretto, A. AgriColMap: Aerial-ground collaborative 3D mapping for precision farming. *IEEE Robot. Autom. Lett.* **2019**, *4*, 1085–1092. [CrossRef]
53. Rincon, J.L.S.; Carpin, S. Map Merging of Oriented Topological Semantic Maps. In Proceedings of the 2019 International Symposium on Multi-Robot and Multi-Agent Systems (MRS), New Brunswick, NJ, USA, 22–23 August 2019; pp. 202–208.
54. Rincon, J.L.S.; Carpin, S. Time-constrained exploration using toposemantic spatial models: A reproducible approach to measurable robotics. *IEEE Robot. Autom. Mag.* **2019**, *26*, 78–87. [CrossRef]
55. Hornung, A.; Wurm, K.M.; Bennewitz, M.; Stachniss, C.; Burgard, W. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Auton. Robot.* **2013**, *34*, 189–206. [CrossRef]
56. Girdhar, Y.; Dudek, G. Modeling curiosity in a mobile robot for long-term autonomous exploration and monitoring. *Auton. Robot.* **2016**, *40*, 1267–1278. [CrossRef]
57. Dellaert, F. The Expectation Maximization Algorithm. Technical Report; Georgia Institute of Technology, 2002. Available online: https://ieeexplore.ieee.org/document/543975 (accessed on 10 September 2022).
58. Hu, Y.; Song, R.; Li, Y. Efficient coarse-to-fine patchmatch for large displacement optical flow. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5704–5712. Available online: https://ieeexplore.ieee.org/document/7780984 (accessed on 10 September 2022).
59. Huang, W.H.; Beevers, K.R. Topological map merging. *Int. J. Robot. Res.* **2005**, *24*, 601–613. [CrossRef]
60. Bonanni, T.M.; Della Corte, B.; Grisetti, G. 3-d map merging on pose graphs. *IEEE Robot. Autom. Lett.* **2017**, *2*, 1031–1038. [CrossRef]
61. Warrington, E.K. The selective impairment of semantic memory. *Q. J. Exp. Psychol.* **1975**, *27*, 635–657. [CrossRef]
62. Fathian, K.; Khosoussi, K.; Tian, Y.; Lusk, P.; How, J.P. Clear: A consistent lifting, embedding, and alignment rectification algorithm for multiview data association. *IEEE Trans. Robot.* **2020**, *36*, 1686–1703. [CrossRef]
63. Yue, Y.; Senarathne, P.N.; Yang, C.; Zhang, J.; Wen, M.; Wang, D. Hierarchical probabilistic fusion framework for matching and merging of 3-d occupancy maps. *IEEE Sens. J.* **2018**, *18*, 8933–8949. [CrossRef]
64. Macenski, S.; Jambrecic, I. SLAM Toolbox: SLAM for the dynamic world. *J. Open Source Softw.* **2021**, *6*, 2783. [CrossRef]