remote sensing

MDPI

Review

A Survey on Visual Navigation and Positioning for Autonomous UUVs

Jiangying Qin [1], Ming Li [1,2,*], Deren Li [1], Jiageng Zhong [1] and Ke Yang [3]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[2] Institute of Theoretical Physics, ETH Zurich, 8039 Zurich, Switzerland
[3] School of Water Resources and Hydropower Engineering, Wuhan University, Wuhan 430079, China
* Correspondence: mingli39@ethz.ch

Abstract: Autonomous navigation and positioning are key to the successful performance of unmanned underwater vehicles (UUVs) in environmental monitoring, oceanographic mapping, and critical marine infrastructure inspections in the sea. Cameras have been at the center of attention as an underwater sensor due to the advantages of low costs and rich content information in high visibility ocean waters, especially in the fields of underwater target recognition, navigation, and positioning. This paper is not only a literature overview of the vision-based navigation and positioning of autonomous UUVs but also critically evaluates the methodologies which have been developed and that directly affect such UUVs. In this paper, the visual navigation and positioning algorithms are divided into two categories: geometry-based methods and deep learning-based. In this paper, the two types of SOTA methods are compared experimentally and quantitatively using a public underwater dataset and their potentials and shortcomings are analyzed, providing a panoramic theoretical reference and technical scheme comparison for UUV visual navigation and positioning research in the highly dynamic and three-dimensional ocean environments.

Keywords: unmanned underwater vehicles (UUVs); simultaneous localization and mapping; visual navigation; positioning; deep learning

Citation: Qin, J.; Li, M.; Li, D.; Zhong, J.; Yang, K. A Survey on Visual Navigation and Positioning for Autonomous UUVs. *Remote Sens.* **2022**, *14*, 3794. https://doi.org/10.3390/rs14153794

Academic Editor: Erica Nocerino

Received: 4 June 2022
Accepted: 4 August 2022
Published: 6 August 2022

1. Introduction

Unmanned underwater vehicles (UUVs) or underwater drones have become an attractive alternative for underwater observation and exploration, since they are cheaper than manned underwater vehicles [1]. These underwater vehicles or robotics can be divided into two categories: remotely operated underwater vehicles (ROVs) and autonomous underwater vehicles (AUVs) [2–6]. A typical ROV is connected to a ship by a neutrally buoyant cable. These cables transmit command and control signals between the operator and the ROV, allowing the remote navigation of the vehicle. Most ROVs are equipped with at least a video camera and lights. Additional equipment is commonly added to expand the vehicle's capabilities. These may include sonar, inertial navigation systems, magnetometers, etc., and often a robotic arm. Typical AUVs explore the ocean without any attached cables. An AUV can conduct its survey mission without operator intervention. Researchers drop an AUV in the ocean and pick it up at a pre-selected position. Today, AUVs can also carry equipment, such as video cameras, lights, sonar, etc., and even robotic arms. The underwater robotics market is evolving, and designs are now following commercial requirements rather than being purely developmental. The boundaries between ROVs and AUVs are becoming more and more blurred. For example, hybrid AUV/ROV designs designed by the Woods Hole Oceanographic Institution (WHOI) have merged the functions and shapes of the two in order to complete complex underwater tasks more autonomously and freely [7].

Over the past decades, there have been abundant attempts to develop UUVs to meet the challenge of diversified applications in the oceans, though the underwater environment

is one of the most challenging for robotics. With the development of artificial intelligence; photogrammetric computer vision; and underwater sensors, battery and communication technologies in recent years, UUVs have played a very significant role in ocean-related monitoring and mapping tasks and become a research hotspot in ocean engineering, marine science, and marine robotics [8]. A common requirement for these missions is that the UUV needs to know its own position and attitude in order to complete them. So, it is a crucial task for an autonomous UUV to build a map of the environment and to estimate its own location within a map or in unknown environments while it is navigating. What is technically referred to as simultaneous localization and mapping (SLAM) is now a de facto standard for autonomous vehicles, not only in underwater environments but also in ground and air environments [9–11]. Moreover, different from large-scale deep-sea scientific research, marine scientific research, such as coastal coral reef conservation, fishery management, and biological connectivity, expects new UUVs to be miniaturized, intelligent, low-cost, and easy to maintain, so that they can be deployed in large quantities and for long-term use. Therefore, how to realize the autonomous navigation and positioning of these new generation UUVs has become the primary problem to be solved. Today, most operators must still be within range of acoustic telemetry systems to keep a close eye on their underwater robotic assets. This is not always possible. This is mainly because radio waves cannot penetrate water very far, so once a UUV dives, it loses its GPS signal. A standard way for UUVs to navigate underwater is dead-reckoning based on inertial navigation systems, even though the accuracy of positioning and navigation can be improved by adding Doppler velocity logs (DVL), the error will increase over time [12,13]. Methods to improve the accuracy and robustness of UUV navigation and positioning by using hydroacoustic positioning systems, such as long-baseline (LBL), ultra-short baseline (USBL), or short baseline (SBL), are also heavily used [14]. This way has worked well in the past, but it is prone to errors driven by changes in the speed of sound underwater due to changes in temperature and salinity. This is very important when navigating close to the seafloor, where sensors, such as down-looking sonar and DVL, can help but can also be bulky and tax a UUV's limited power. So, the underwater SLAM for accurate navigation and positioning remains an unsolved problem. At the same time, the cost of these devices is high.

Since good video footage or images are core components of most ocean scientific research, research UUVs tend to be outfitted with high-output lighting systems and broadcast-quality cameras. This provides excellent conditions for us to use vision for precise positioning and navigation of underwater robots. This also coincides with the thinking of many scholars and industry professionals. If we look to terrestrial applications for inspiration, self-driving cars and aerial drones as an example, we see that their solutions to comparable problems all leverage imaging sensors and computer vision to deliver accurate navigational solutions [15]. By detecting environmental features, the platform's relative position can be tracked while simultaneously mapping the environment as it is perceived. At the same time, vision can also be coupled with other sensors to achieve more robust and comprehensive precise positioning and navigation. For these reasons, with the development of deep learning, after a period of stagnation, UUV underwater navigation and positioning based on vision can provide richer information, cheaper prices, smaller sensor sizes, and lower energy consumption. It has become a research hotspot within the scientific and industrial circles [16–19]. On the other hand, vision is a passive perception method, which can reduce the impact on the fragile ecological environment [20]. Although underwater visual navigation and positioning is a promising method and can achieve good results, especially in coastal zones with good lighting and clear water (visible distance up to 20 m), it is also limited by key factors such as underwater lighting conditions, turbidity, and feature richness. Therefore, analyzing and putting forward ideas to solve these problems in detail through a literature review and using advanced technology to achieve underwater visual navigation and positioning capabilities with a wider application range and higher accuracy is the purpose of this paper.

This paper is organized as follows. Section 2 introduces the background of visual navigation and positioning technology. Section 3 presents and compares some SOTA geometry-based visual navigation and positioning methods. Section 4 introduces and compares some SOTA deep learning-based methods. Section 5 compares representative algorithms of geometry-based and deep learning-based on the underwater public dataset and analyzes the availability and potential of different algorithms in underwater visual navigation and positioning applications. Finally, the closing section is the conclusion. Overall, the paper provides a comprehensive review of visual navigation and positioning, including the latest research progress, development trends, and possible future application directions.

2. Background

Visual navigation and positioning is a technology that realizes its own pose estimation and path recognition under the input of external perception and proprioceptive sensors, makes navigation decisions, and finally outputs continuous control to drive the robot to reach the target position. It is achieved through technologies such as positioning and mapping, obstacle avoidance, and path planning [21]. Considering the prior information of visual navigation, visual navigation, and positioning systems can be mainly divided into three categories: map-less systems, map-based systems, and map-building systems [22]. In map-less systems, no prior environment map is provided and the robot navigates simply by detecting features in the environment. Optical flow methods, appearance-based matching methods, and object recognition methods are common methods in such systems [23–25]. Map-based systems predefine environment models, such as octree maps and occupancy grid maps, which provide varying levels of detail. Robots achieve navigation tasks through obstacle avoidance and path-planning capabilities [26,27]. However, the environment map is not available in advance in all cases, and one possible solution for the robot is to map the environment while it works. In a first step, the robot explores the map until enough information is gathered and the navigation is started using the autonomously generated map. This is how map-building systems work. With the development of visual simultaneous localization and mapping (vSLAM) technology as important technical support, the application of map-building systems has been greatly expanded [28,29]. Autonomous unmanned underwater vehicles need to exhibit a high degree of autonomy in any complex environment that may exist to ensure robust navigation and positioning capabilities, and to build a map of the environment while implementing the navigation task [30,31]. Therefore, the research focus of this paper is mainly on map-building navigation, especially on its main technology, vSLAM, because it is a core technical means to realize navigation and positioning in unknown environments [32,33].

In the field of UUV visual navigation and positioning, some papers have been investigated and reviewed. Ref. [34] gave a relatively complete review of vSLAM, but it was published before some of the latest vSLAM algorithms could be included. Ref. [35] gave a broad overview of vSLAM but mainly focused on metric and semantic SLAM. Ref. [36] reviewed visual–inertial simultaneous localization and mapping (viSLAM) mainly from a filtering-based and optimization-based perspective. The latest review [37], achieved a structured overview of existing vSLAM and viSLAM designs and proposed its own classification. Ref. [38] reviewed the most representative vSLAM algorithms and compared the main advantages and disadvantages of each method. These reviews investigate the existing vSLAM and viSLAM from different aspects but are limited to the research of vSLAM and viSLAM in conventional ground environments. Ref. [39] contained some related content to underwater vSLAM. Ref. [40] introduced active and passive techniques for UUV positioning, and Ref. [41] defined complex and constrained underwater environments and introduced vSLAM techniques from the perspective applicable to the defined environment. However, these two reviews focused on multi-sensor positioning rather than vision positioning. Based on this, in order to enable readers to quickly establish an understanding of the research and development of UUV visual navigation and positioning, this paper introduces excellent state-of-the-art (SOTA) vSLAM and vSLAM algorithms fused with

other sensors and undertakes a comparative study using a public underwater dataset. This paper will be the first paper focusing on the panoramic review of underwater vSLAM and conducting experimental comparison and analysis through an underwater public dataset.

3. Geometry-Based Methods

As a key technology of visual navigation and positioning, vSLAM technology has achieved rich results and effective applications in the fields of indoor, aerial, and planetary exploration. However, due to the unstructured characteristics and the complexity of the underwater environment, research on underwater vSLAM is still in a very early stage, and most is focused on the improvement of local algorithms, such as keypoint extraction in the traditional vSLAM framework. At present, there is no unified algorithm and framework, and the improvement of performance, such as accuracy and robustness, still needs to be further studied [42–45].

Geometry-based methods are classic solutions of vSLAM, which realize real-time positioning and mapping by restoring the mapping relationship between the camera's geometric model and the photo [46,47]. Geometry-based methods are mainly divided into two categories: vision-only methods and vision-based multi-sensor fusion methods [48]. Vision-only methods only realize positioning through photos, and it is the first choice for autonomous positioning and environment perception without prior knowledge of the environment and external auxiliary information sources. Vision-based multi-sensor fusion methods achieve more accurate and robust vSLAM by combining vision cameras with other sensors, such as inertial navigation, sonar, depth meters, etc. The rest of this section will introduce geometry-based methods. Figure 1 is the pipeline of geometry-based methods and Figure 2 is an overview of SOTA geometry-based methods.

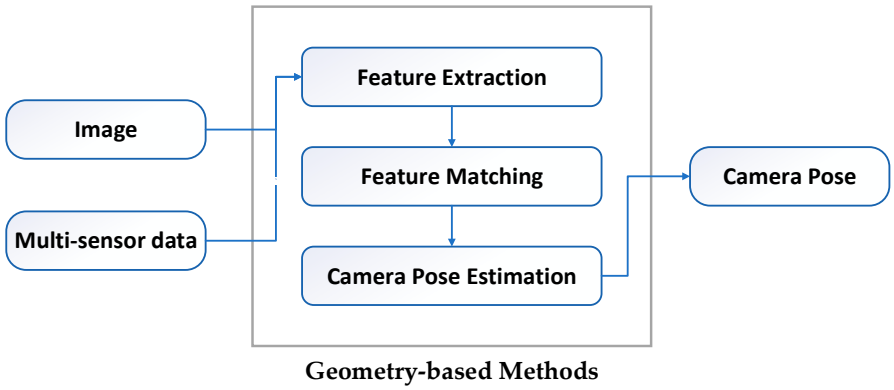


Figure 1. Pipeline of geometry-based methods.

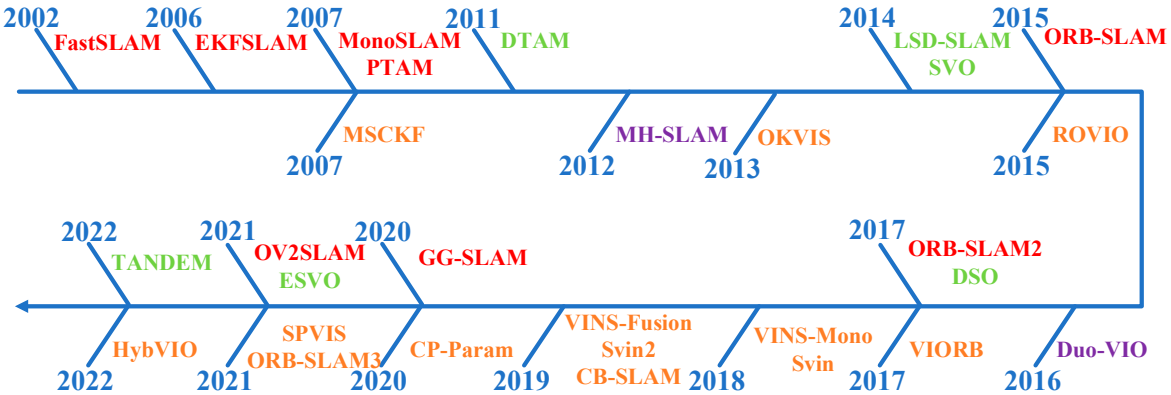


Figure 2. Overview of SOTA geometry-based methods. The ones above the timeline represent vision-only methods, with feature-based methods in red and direct methods in green. Those below the timeline represents vision-based multi-sensor fusion methods, where purple represents loosely-coupled and orange represents tightly-coupled.

3.1. Vision-Only Methods

Traditional geometry-based vSLAM methods can be divided into two main categories, feature-based vSLAM methods and direct vSLAM methods.

Feature-based vSLAM method mainly detects the feature points in adjacent images and matches them by comparing the feature descriptors, and then solves the camera pose according to the matching relationship [49]. Among the early vSLAM, feature-based vSLAM methods, especially extended Kalman filter (EKF) SLAM, dominated for a long time [50,51]. MonoSLAM [52] was the first real-time monocular vSLAM system, which used EKF as the back-end to track very sparse feature points at the front-end and used a single thread to update the camera pose and map frame by frame. The problem is that the computational cost increases proportionally to the size of the environment and even the minimal complexity FastSLAM [53] severely limited the application of vSLAM. PTAM, proposed by [54] in 2007, is an important breakthrough in the field of vSLAM, which innovatively proposed a dual-thread architecture for tracking and mapping. The tracking thread was used to update the camera pose frame by frame in real time, while the mapping thread did not need to be updated frame by frame, thus having longer processing times. Therefore, the bundle adjustment (BA) algorithm that could previously only be applied to offline structure from motion (SfM) is also applied to vSLAM. Since then, almost all vSLAM algorithms have followed this multi-threaded architecture. Following the two-thread structure of PTAM, the three-thread structure (tracking, local mapping, loop closing) of the ORB-SLAM2 [55] also achieved very good tracking and mapping effects, which can ensure the global consistency of the trajectory and the map. ORB-SLAM2 was a further extension to the monocular version ORB-SLAM [56]. It supported monocular cameras, binocular cameras, and RGB-D cameras and was the representative of feature-based vSLAM methods.

The direct vSLAM method does not calculate keypoints or descriptors but directly calculates the camera motion according to the pixel information of the image. DTAM [57] was the first attempt of direct vSLAM, which computed keyframes to build a dense depth map by minimizing the global spatial canonical energy function. The camera pose was computed using the depth map through direct image matching. This method had good robustness to missing features and blurred images, but it required a large amount of computation and GPU parallel computing. Additionally, it was not robust enough for global illumination processing. LSD-SLAM [58] marked the successful application of direct methods in vSLAM. Its core contribution was the application of the direct method to semi-dense monocular vSLAM, which did not need to calculate feature points. It generated a map with global consistency by directly registering image photometrics and using a probabilistic model to represent semi-dense depth maps. However, this method was very sensitive to camera intrinsics and exposure. DSO [59] proposed a complete photometric calibration method by combining the effects of exposure time, lens vignetting, and nonlinear response functions, thereby enhancing the robustness to exposure. In addition, there are also some semi-direct vSLAM methods that combine the advantages of parallel tracking and mapping of the feature-based method with the advantages of the direct method, which is fast and robust [60]. In recent years, the field of vSLAM has been extensively studied, many excellent algorithms have been proposed and their accuracy and robustness have also been greatly improved. GG-SLAM [61] proposed an efficient optimization algorithm, Good Graph, which can efficiently select well-preserved size reduction graphs in local BA. OV2SLAM [62] proposed a complete vSLAM framework that integrates the most up-to-date visual positioning results, which can solve monocular and binocular images of different scales and frame rates. ESVO [63] proposed a binocular event camera-based visual odometry solution that takes only raw events from a calibrated camera as input, while estimating the pose of the binocular event camera and reconstructing the environment using a semi-dense depth map. TANDEM [64] seamlessly combined classic direct visual odometry and learning-based MVS reconstruction and utilized the depth map rendered by the global TSDF model to achieve monocular dense tracking front-end.

3.2. Vision-Based Multi-Sensor Fusion Methods

With the deepening of research on UUV and the continuous improvement of application requirements, it has become a realistic choice to use a combination of two or more sensors to complete autonomous navigation and positioning functions. Especially in the complex and unstructured underwater environment, the fusion positioning of multiple sensors can give full play to the advantages of each sensor to achieve more accurate and robust UUV autonomous navigation and positioning [65].

viSLAM is a common solution for multi-sensor fusion positioning, which couples inertial measurements with vision methods to estimate the sensor pose. By adding an inertial measurement unit (IMU), the information richness of the environment can be increased, providing high-frequency robust short-term precise positioning in areas that are difficult for vision cameras, such as texture-missing and image blur. On the other hand, the vision system can solve the cumulative error and significant drift of the IMU, enabling the fused vSLAM algorithm to achieve higher accuracy and robustness [66]. In addition, the miniaturization, low power consumption, and low cost of IMUs also provide a richer selection of sensors for many lightweight applications. According to the type of sensor coupling, viSLAM algorithm can be divided into two categories: loosely-coupled methods and tightly-coupled methods. In loosely-coupled methods, the poses calculated by the vision sensor and the IMU are directly fused. The fusion process does not affect the two themselves and is output as a post-processing method. Tightly-coupled methods combine the data of the vision sensor and the IMU through optimization and filtering to jointly construct the motion equation, so as to obtain the final pose information by considering the state estimation of the two kinds of data. At present, there are few studies based on loosely-coupled optimization, and they mainly fuse the estimation results of two sensors through extended Kalman filtering [67,68]. Many excellent algorithms have emerged in the tightly-coupled field, which can be roughly divided into two categories: filter-based methods and optimization-based methods according to the type of back-end optimization. MSCKF [69] was a typical representative of filter-based methods. It integrated IMU and visual information under the EKF framework. Compared with traditional visual methods, it can adapt to more violent motion, texture loss for a certain period of time, etc. Similarly, ROVIO [70,71] realized coupling based on the EKF filtering of sparse image patches. It used IMU data for state propagation. In terms of vision, it extracted multi-level patches around features, and obtained innovation terms based on these patches to update the filter state. OKVIS [72,73] and VIORB [74] were important optimization-based algorithms that jointly optimize the reprojection error of matching points and IMU error data. The algorithm flow of OKVIS predicts the current state through the IMU measurement value and perform feature extraction and feature matching according to the prediction. The 3D point feature and the 2D image feature constitute the optimal reprojection, while the predicted IMU state quantity and the optimized parameter constitute the IMU measurement error. These two errors are combined for optimization. The VINS series (VINS-Mono, Vins–Fusion) [75,76] used a tightly coupled, nonlinear optimization-based approach to obtain high-accuracy visual-inertial odometry by fusing pre-integrated IMU measurements. ORB-SLAM3 [77] was a vSLAM system that supported vision, vision inertial, and hybrid maps, which only relied on maximum a posteriori estimation; applied visual and inertial estimation separately; and then jointly optimized the two. In addition, there were some algorithms based on coplanar point-line parameter constraints [78] and pose graph constraints [79] to achieve optimization-based coupling of visual–inertial data. HybVIO [80] proposed a novel hybrid method that combined filter-based methods visual-inertial odometry with optimization-based vSLAM. Its core was a highly robust, self-contained VIO with improved IMU bias modeling, outlier suppression, stationarity detection, and feature trajectory selection that can be tuned to run on embedded hardware.

Furthermore, the fusion of vision and various sensors, such as sonar, depth gauge, etc., which may be carried out by diversified and personalized UUV operations, has also attracted the attention of some scholars. The Svin series [81,82] proposed image enhancement

techniques for the underwater domain. In addition, combining binocular camera vision data, IMU angular velocity and linear acceleration data, mechanical scanning sonar sensor distance data, and depth measurement data enables robust and accurate sensor trajectory estimation. CB-SLAM [83] proposed a systematic approach to the real-time reconstruction of underwater environments using sonar, visual, inertial, and depth data. It utilized well-defined edges between well-lit areas and dark areas to provide additional features, resulting in denser 3D point clouds for higher-resolution 3D environment reconstruction and more robust UUV autonomous navigation.

3.3. Summary

Geometry-based visual navigation and positioning is a relatively mature field in the non-underwater field, and many classic frameworks and methods have emerged. Vision-only methods can rely only on vision to achieve navigation and positioning without any other information aids. However, considering the unstructured and complex underwater environment, coupling vision and other sensors, such as IMU and sonar, is a more complex but effective solution, which can combine the advantages of different sensors to achieve real-time positioning and mapping with higher accuracy and stronger robustness. Under ideal conditions, geometry-based navigation and positioning can accurately estimate the pose of the camera with high positioning accuracy. However, limited by various complex environments that may actually exist, such as dynamic targets, missing textures, complex light, possible inaccurate camera calibration, and inaccurate system modeling, it may lead to low positioning accuracy or even positioning failure.

4. Deep Learning-based Methods

Deep Learning-based visual navigation and positioning technology has been highly expected in recent years. Different from relying on geometric models and mapping relationships to achieve navigation and positioning, deep learning-based methods provide a data-driven alternative. Deep learning-based methods can automatically discover task-relevant features using highly expressive neural networks, which also makes them better suited to more scenarios. Benefiting from the ever-increasing amount of data and computing power, these methods are rapidly advancing to a new stage [84,85]. At present, deep learning-based methods are mainly divided into two categories. One is to use deep learning to replace one or several modules in traditional vSLAM, such as feature extraction [86], depth estimation [87,88], loop closure detection [89], and so on. The second is to integrate semantic information into vSLAM to improve the positioning accuracy of vSLAM and build a semantic map [90]. The neural networks currently used in vSLAM can be mainly divided into convolutional neural networks (CNN), recurrent neural networks (RNN), generative adversarial networks (GAN), graph neural network (GNN), spiking neural network (SNN), and so on. Subsequent sections will introduce the application of these neural networks in navigation and positioning fields. Figure 3 is the pipeline of deep learning-based methods and Figure 4 is an overview of SOTA deep learning-based methods.

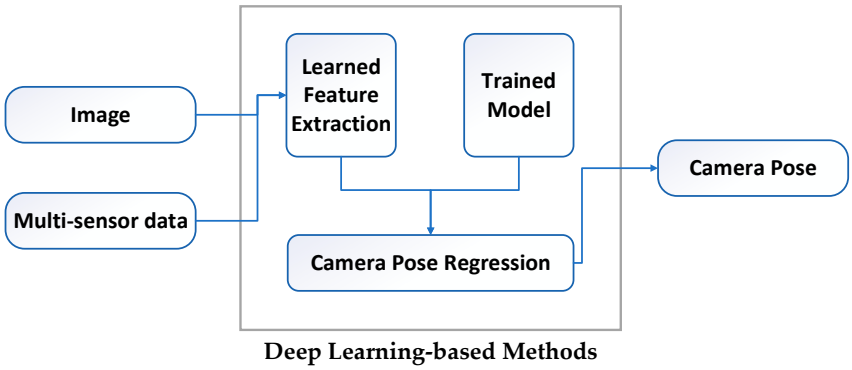


Figure 3. Pipeline of deep learning-based methods.

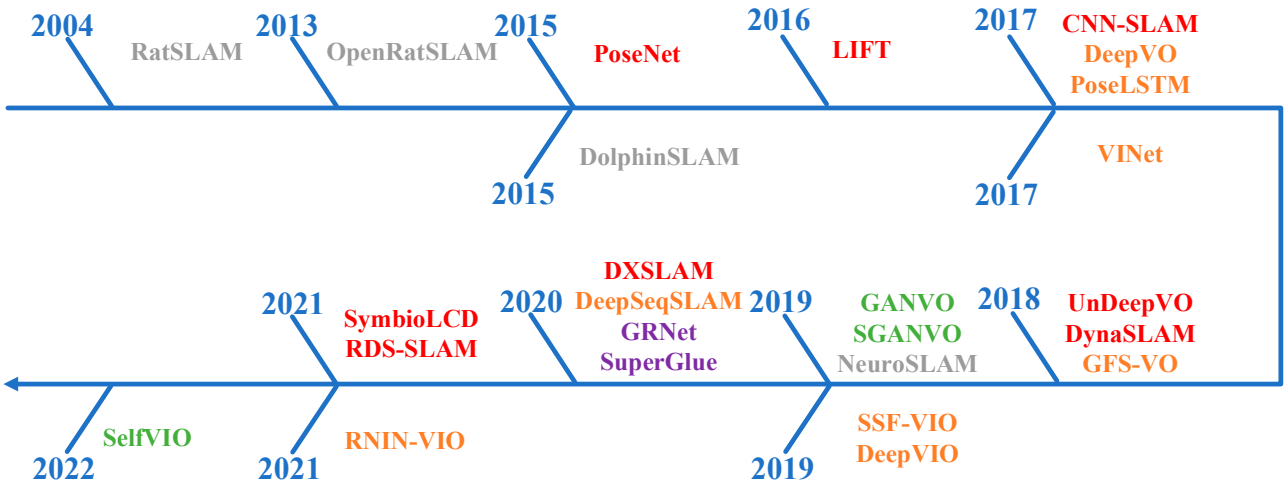


Figure 4. Overview of SOTA deep learning-based methods. The methods above the timeline represent vision-only methods, while those below the timeline represents vision-based multi-sensor fusion methods. Among them, red represents CNN-based methods, orange represents RNN-based methods, green represents GAN-based methods, purple represents GNN-based methods, and gray represents SNN-based methods.

4.1. CNN-Based Methods

The convolutional neural network is a kind of feedforward neural network with a deep structure, including convolution calculation. Its convolution kernel parameter sharing and the sparsity of inter-layer connections make it possible to learn grid-like topology features with a small amount of computation. It has stable effects and no additional feature engineering requirements of the data. CNN mainly includes convolutional layers, pooling layers, and fully connected layers. The convolutional layer performs feature extraction on the input data, and the output feature map is passed to the pooling layer for feature selection. The layer-by-layer abstraction of features is achieved through the alternate stacking of the two, and the results are output through the fully connected layer at the end. Convolutional neural networks are commonly used in the processing and analysis of image data and have also been successfully applied to other forms of data, such as speech [91,92].

Convolutional neural networks are widely used in image data processing, including visual navigation and positioning due to their superior performance in image data dimensionality reduction and image feature retention. PoseNet [93] innovatively applied convolutional neural network to the task of camera pose regression to achieve an end-to-end regression of 6 degrees of freedom (DoF) camera pose from RGB images. LIFT [94] used convolutional neural networks to implement a complete feature point processing pipeline, including feature point extraction, orientation estimation, and feature description. Similarly, DXSLAM [95] applied convolutional neural networks to feature point extraction and description. CNN-SLAM [96] fused the dense depth map predicted by CNN with the depth measurements obtained by direct monocular vSLAM, enabling the vSLAM system to have both an accurate absolute scale and robust tracking capability. UnDeepVO [97] used convolutional neural networks for pose estimation and depth estimation, respectively, and realized real-scale monocular visual odometry in an unsupervised manner by exploiting spatial and temporal geometric constraints. SymbioLCD [98] utilized object and vBoW features extracted by CNN for loop closure detection (LCD) candidate prediction, creating a more robust and symbiotic LCD system by adding object semantic and spatial awareness elements. DynaSLAM [99] and RDS-SLAM [100] used convolutional neural networks for semantic segmentation, distinguished dynamic and static regions, and achieved real-time robust tracking and mapping in dynamic environments through semantic-based optimization threads.

4.2. RNN-Based Methods

The recurrent neural network is a sequence-to-sequence model with memory capability, which can better capture temporal features in sequence data, such as videos. RNN consists of three parts: input unit, hidden unit, and output unit, where the input of the hidden unit consists of the current input unit and the output of the previous hidden unit. Traditional RNN is prone to the problems of gradient disappearance and gradient explosion, so LSTM and its variants are generally used at present [101,102].

The recurrent neural network has attracted much attention in the field of visual navigation and positioning due to its good ability to capture temporal states in video sequences. DeepVO [103] proposed an end-to-end visual odometry based on deep recurrent convolutional neural networks. On the one hand, it automatically learned the effective feature representation of VO through convolutional neural networks. On the other hand, it implicitly modeled the time series model (motion model) and the data association model (image sequence) through the recurrent neural networks. PoseLSTM [104] proposed a new CNN + LSTM architecture to solve the visual positioning problem. It used CNN to learn robust feature representations for various complex environments, such as motion blur, illumination changes, etc. In addition to this, it used LSTM units to perform structured dimensionality reduction on feature vectors to further improve positioning performance. GFS-VO [105], DeepSeqSLAM [106] proposed a trainable CNN + RNN framework for feature learning and image sequence inference, respectively, to jointly learn visual and position representations from a single image sequence.

In addition, RNN has a good ability to model the temporal state evolution of inertial data and is also widely used in viSLAM. VINet [107] was the first attempt to solve VIO problem using a deep learning framework. It obtained the optical flow motion features between adjacent frames through FlowNet, used a small LSTM network to process the original data of the IMU to obtain the motion features under the IMU data, and finally inputted them into the core LSTM network for feature fusion and pose estimation. SSF-VIO [108], DeepVIO [109], and RNIN-VIO [110] also use the LSTM-style IMU pre-integrated network and fusion network, learned by minimizing the loss function of self-motion constraints to achieve robust visual–inertial odometry.

4.3. GAN-Based Methods

The generative adversarial network is one of the most promising deep learning models for unsupervised learning on complex distributions in recent years. It produces good output results through mutual game learning between the generative model and the discriminative model. The large number of datasets required for supervised learning can cause large costs if is collected and annotated manually, while GAN can automatically generate datasets to provide low-cost training data [111].

The generative adversarial network plays an important role in navigation, positioning, and mapping, especially supervised depth estimation, due to its feature of automatically generating low-cost training data. On the one hand, GAN can address the difficulty of supervised deep learning in environments without richly labeled data. On the other hand, supervised depth estimation models need to learn 3D mapping and scale information from ground-truth depth maps. However, ground-truth depth maps are difficult to obtain in real scenarios. Through the introduction of the GAN model, sharper and more realistic depth maps can be generated. GANVO [112] proposed a generative unsupervised learning framework. It predicted camera motion poses with 6DoF and monocular depth maps from unlabeled RGB image sequences by using deep convolutional generative adversarial neural networks. It generated supervision signals by transforming the sequence of views and minimizing the objective function used in multi-view pose generation and single-view depth generation networks. SGANVO [113] proposed a novel unsupervised network system for visual depth and inter-frame motion estimation. It consisted of a stack of GANs, the lowest layers estimated depth and inter-frame motion while higher layers estimated spatial features. It also captured temporal dynamics due to the use of cross-layer recurrent

representations. SelfVIO [114] used adversarial training and adaptive vision-inertial sensor fusion. It learned the joint estimations of 6DoF ego-motion and scene depth maps from unlabeled monocular RGB image sequences and IMU readings. It can perform VIO without IMU internal parameters or external calibration between IMU and camera.

4.4. GNN-Based Methods

The graph neural network captures graph dependencies through messages passing between graph nodes. It updates the state of a node from its neighbors at any depth, which improves the problem of traditional neural networks, such as traditional CNNs and RNNs, which can only process structured data because the features of nodes need to be arranged in a certain order. Its excellent processing ability for unstructured data enables it to achieve good results in data analysis, the optimal combination of graphs, etc. Its variants, such as the graph convolutional networks (GCN), graph attention network (GAT), graph recurrent networks (GRN), and others have been widely used [115].

The graph neural network has been introduced into the field of visual navigation and positioning due to its excellent processing capability for unstructured data to provide unstructured knowledge learning and reasoning, such as time-consistent constraints between discontinuous multi-view frames. GRNet [116] was the first system to solve multi-view camera relocalization using GNN, which redefined the node, edge, and embedding functions in traditional GNN. It combined CNN and redesigned GNN for feature extraction and learning the interaction relationship between different features to iteratively process multi-view high-dimensional image features at different levels. In this way, the information of the entire image sequence can be better mined, and even frames relatively far away in time can provide interrelationships, thereby enabling more accurate absolute pose estimation. SuperGlue [117] proposed a GNN feature-matching network based on the attention mechanism, which can perform feature matching and false matching point elimination at the same time. Feature matching was solved by solving the optimal transport problem, using GNNs to predict the optimized cost function, and using intra- and inter-image attention mechanisms to exploit the spatial relationship between keypoints and their visual appearances. The model can run in real-time on the GPU and can be easily integrated into vSLAM systems.

4.5. SNN-Based Methods

The spiking neural network is based on the premise that biological properties can provide higher processing power, so it uses the model that best fits the biological neuron mechanism for calculation. In fact, compared to the high computational demands of current vSLAM, even animals with very small brains excel at combining local visual cues and self-motion cues for spatial navigation in the brain, enabling the robust mapping and navigation of 3D environments. Therefore, seeking inspiration from biology to develop vSLAM systems is a promising path. SNN draws on biologically-inspired local unsupervised and global weakly-supervised biological optimization methods, so it has strong capabilities of spatiotemporal information representation, asynchronous event information processing, network self-organizing learning, and so on. The basic idea is that a neuron in a dynamic neural network is not activated in each iterative propagation, but only when its membrane potential reaches a certain value. When a neuron is activated, it generates a signal to other neurons, raising or lowering the membrane potential [118].

The piking neural network has attracted attention in the field of vSLAM due to its effective utilization of biological characteristics and its excellent representation and integration capabilities of different information dimensions. At present, there are few related studies on SNN in vSLAM. RatSLAM [119] was a real-time positioning and mapping method proposed based on the rodent hippocampal complex, which simulated the computational model of rodents and used a competitive attractor network structure to combine mileage information with landmark sensing to form a consistent representation of the environment. RatSLAM can operate with ambiguous visual input and recover from minor and major

path integration errors. OpenRatSLAM provided an open source version of RatSLAM [120]. DolphinSLAM [121] extended RatSLAM from 2D ground vehicles to 3D underwater environments and proposed a vSLAM system based on mammalian navigation. DolphinSLAM used SNNs to position and process low-resolution monocular imagery and imaging sonar data, enabling the cooperation of acoustic–optical sensors and robust positioning in ambiguous environments. NeuroSLAM [122], a neural-inspired 4DoF vSLAM system, was the first SNN vSLAM system capable of mapping and positioning in large real-world 3D environments. NeuroSLAM developed a functional computational model of connected pose units consisting of 3D grid units and multiple layers of head orientation units. In addition to this, it proposed a novel multi-layer graphical experience map that combined local view units, 3D grid units, multi-layer head orientation units, and 3D visual odometry.

4.6. Summary

Deep learning-based methods offer a data-driven alternative. CNN has been widely used in the field of visual navigation and positioning due to its superior performance in image data dimensionality reduction and image feature retention. The sequence-to-sequence mode of RNN makes it good at capturing temporal features in sequence data, such as videos. In addition, RNN can model the time evolution state of inertial data, so the fusion of CNN and RNN vSLAM is also a common solution, in which CNN is used for feature representation and RNN is used for image sequence reasoning. The fact that GANs can automatically generate features from low-cost training data opens up the possibility of supervised deep learning vSLAM within unenriched labeled data. GNN has an excellent processing ability for unstructured data and can overcome the problem of limited information expressed by RNN. It allows the learning and reasoning of unstructured knowledge, such as time consistency constraints between discontinuous multi-view frames, and solves the under-use of multi-view image data for visual ambiguity. SNN aims to bridge the gap between neuroscience and deep learning, using models that best fit biological neuron mechanisms for computation. At present, there are few related works on SNN in the field of visual navigation and positioning, but SNN's simulation of biological characteristics and its excellent expression and integration capabilities for different information dimensions are likely to bring new changes to the field of visual navigation and positioning. To facilitate the use of these methods by other researchers we have made the relevant code of our experiment publicly available [123].

5. Experiments and Discussions

5.1. Dataset

This paper compares and analyzes different algorithms by conducting experiments on the public dataset AQUALOC [124]. AQUALOC is a dataset dedicated to underwater vehicles positioned close to the seabed, equipped with monocular cameras, MEMS-IMUs, and pressure sensors, which provide these data simultaneously. This paper selects sequence 1 and sequence 7 of the habor dataset for experiments. The habor dataset is challenging. Its motion changes drastically, and there is even a temporary unavailability of visual information due to collisions. In addition, due to the low brightness of the underwater, the dataset uses a lighting system, which also causes some photos to have unstable lighting, such as overexposure. This is a big challenge to the performance of test algorithms. Among them, the overexposure and sudden movements of sequence 7 are more serious than sequence 1. The example images of the dataset are shown in Figure 5.

5.2. Experimental Analysis of Vision-Only Methods

Considering the representativeness and achievability of the above algorithms, we selected four algorithms as the representatives of vision-only methods, which are LSD-SLAM, ORB-SLAM2, PoseLSTM, and RatSLAM. We use average positioning accuracy as the measurement of positioning performance. We use different algorithms to position the habor dataset, and the positioning accuracy is shown in Table 1.

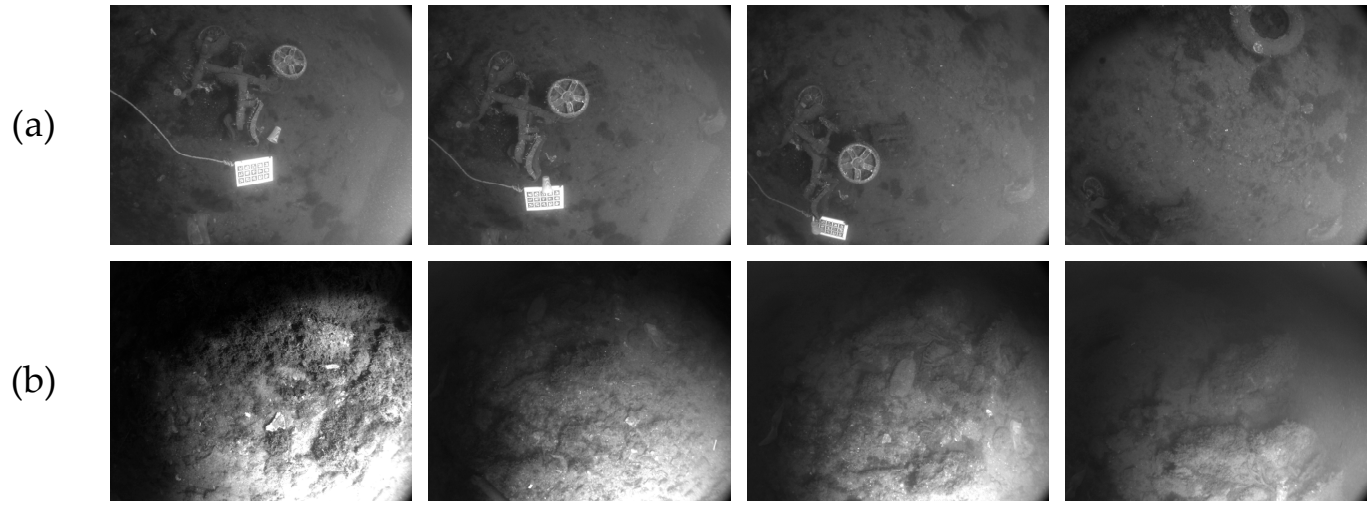


Figure 5. Example images of the public dataset AQUALOC. (**a**) is the example image of sequence 1, while (**b**) is the example image of sequence 7.

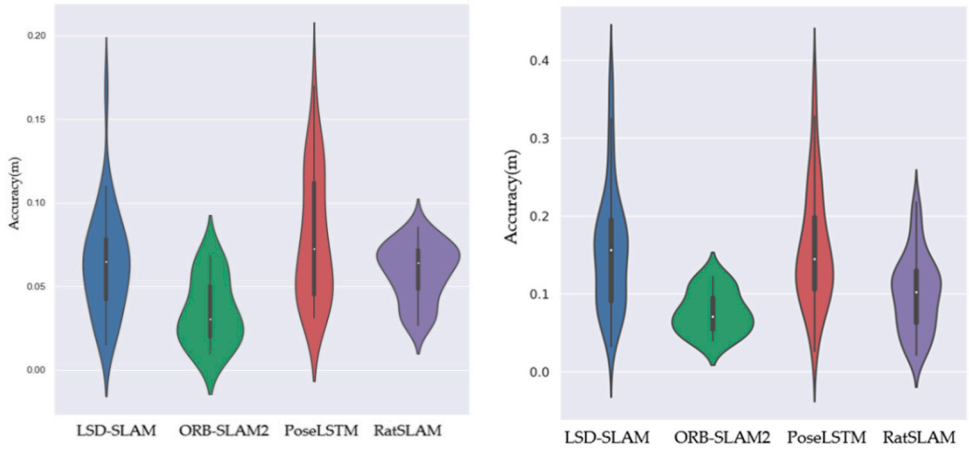
Table 1. Vision-only methods positioning accuracy comparison.

Algorithm	Sequence 1 (m)	Sequence 7 (m)
LSD-SLAM	0.07052	0.17432
ORB-SLAM2	0.03948	0.08043
PoseLSTM	0.08706	0.17581
RatSLAM	0.0611	0.11169

From the quantification point of positioning accuracy, the best performer is ORB-SLAM2, which is an algorithm based on geometric features, and its positioning accuracy on two datasets are 0.03 m and 0.08 m, respectively. The second is the SNN-based RatSLAM, which has an accuracy of 0.06 m and 0.11 m. The geometry-based direct LSD-SLAM algorithm and RNN-based PoseLSTM algorithm have close accuracy in sequence 7, but LSD-SLAM performs better in sequence 1.

Figure 6 is the comparison chart of the positioning accuracy of the four algorithms. For each algorithm, the corresponding graph is the fitted graph of the distribution of its positioning results at different levels of accuracy. The white dots indicate the average positioning accuracy, the black line indicates the accuracy distribution range, and the width of the graph indicates the number of positioning results at this level of accuracy. It can be seen from the figure that ORB-SLAM2 shows excellent positioning performance in terms of the highest accuracy, the lowest accuracy and the average accuracy, and its positioning accuracy on two sequences is concentrated between 0.02–0.07 m and 0.04–0.13 m, respectively. The positioning accuracy of RatSLAM is second, and its positioning accuracy is concentrated between 0.03–0.08 m and 0.02–0.21 m. The positioning results of LSD-SLAM are slightly better than that of PoseLSTM, but the results of both have a large deviation from the real results. However, it is worth noting that during the operation of the dataset, ORB-SLAM2 suffered some short-term tracking loss, and it is easy to fail in positioning in the case of sudden movement and overexposure.

As can be seen from the results in the table, the geometry-based methods ORB-SLAM3 and OKVIS have higher positioning accuracy. It is worth mentioning that the geometry-based method ORB-SLAM3, which integrates visual inertial information, achieved fairly high accuracy, reaching around 0.02 m. The accuracy of OKVIS is 0.04 m and 0.11 m and the accuracy of deep learning-based algorithm VINet is relatively low at 0.05 m and 0.15 m.



(**a**) (**b**)

Figure 6. Accuracy comparison of vision-only methods. (**a**) Sequence 1. (**b**) Sequence 7.

5.3. Experimental Analysis of Vision-Inertial Methods

Similarly, we choose three methods as representatives of visual–inertial methods, that is OKVIS, ORB-SLAM3, and VINet. The accuracy comparison is shown in Table 2.

Table 2. Vision–inertial methods positioning accuracy comparison.

Algorithm	Sequence 1 (m)	Sequence 7 (m)
OKVIS	0.040636	0.11707
ORB-SLAM3	0.019821	0.02119
VINet	0.049718	0.14946

Figure 7 shows the trajectories solved by the three algorithms compared to the ground-truth. Overall, the positioning performance of sequence 1 is better than that of sequence 7, which may be related to the drastic motion changes and overexposure of sequence 7. As can be seen from the figure, ORB-SLAM3 has a good tracking result for sequence 1. Except for a small offset at the beginning, the rest of the trajectory is basically the same as the real trajectory. It is worth mentioning that for sequence 7, ORB-SLAM3 fails to track in the first half of the trajectory and does not successfully solve the pose data, which is related to the dramatic viewpoint changes in the first half of the dataset. However, in the second half of the dataset, the trajectory solved by ORB-SLAM3 is basically consistent with the ground-truth. Except for a small drift at the sudden change of direction, the results of other parts are excellent. The positioning errors of OKVIS and VINet are relatively large, and VINet is more prone to large drift. In addition, from the perspective of the comparison of vision–inertial and vision-only methods, the overall positioning results of the vision–inertial fusion algorithm show a large improvement over the vision-only method. It is worth noting that the ORB-SLAM3 vision–inertial method has significantly improved the positioning accuracy after coupling the inertial information to the visual information compared with the ORB-SLAM2 of the same series that only uses the visual information, which also confirms the effectiveness of the fusion of visual information and inertial information for positioning.

Overall, vision–inertial fusion positioning results are better. When the quality of some visual information is low or even missing, the high-frequency robust short-term precise positioning provided by the IMU can make up for the strong dependence of visual positioning on the quality of visual information. On the other hand, the high accuracy of visual positioning for long-term positioning can also significantly improve the cumulative error problem of the IMU, thereby achieving higher accuracy. The experimental results in this paper have also proved this point. In addition, the accuracy of geometry-based visual

navigation and positioning methods are generally higher than that of deep learning-based methods. However, the stronger robustness of deep learning-based methods makes them less prone to position failure, especially in some complex or even extreme situations. In this case, deep learning-based methods position more robust and stable. In addition, time is also an issue that needs special attention in the field of visual navigation and positioning, especially in applications with high real-time requirements. In this regard, deep learning-based methods show better performance, the average time for deep learning-based methods to position a frame of image is about 5 ms, while for geometry-based methods, the average positioning time reaches 30 ms.

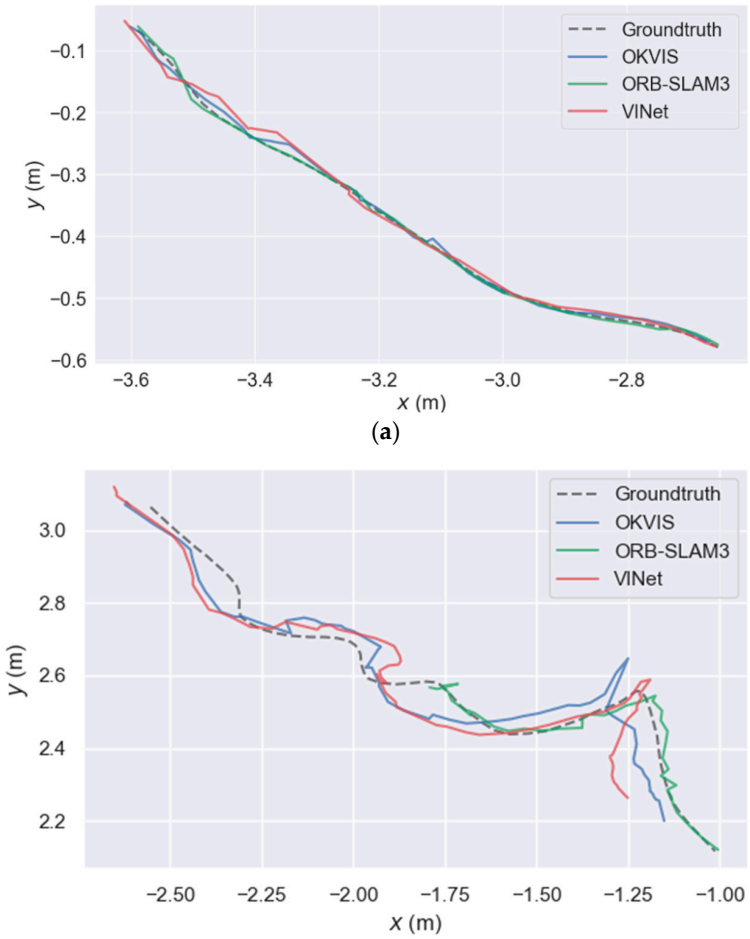


Figure 7. Trajectory comparison of vision–inertial methods. (**a**) Sequence 1. (**b**) Sequence 7.

In summary, the development of geometry-based visual navigation and positioning methods are relatively mature. They mainly rely on extracting geometric constraints from images to estimate motion. Since they derive from elegantly established principles with strict model constraints, these methods enable accurate cross-environment motion estimation without time constraints. However, real-world application scenarios are often not ideal, and situations such as viewpoint changes and missing textures may affect the positioning accuracy of geometry-based methods or even cause position failure. The high generalization and strong autonomous learning ability of deep learning enables it to learn the inherent laws and expression levels of sample data in a data-driven mode and discover task-related features, which allows it to better adapt to complex environments. Therefore, deep learning-based methods are expected to make up for this limitation and greatly improve the efficiency of online processing. This is critical for autonomous UUVs with limited space and energy. However, at present, there is still a gap between deep

learning-based methods and geometry-based methods in terms of positioning accuracy when the texture richness is good, which is also one of the important directions for future research on deep learning-based methods. Another point worth noting is the addition and enhancement of multi-sensor coupling to visual-only navigation and positioning capabilities. Our experimental results have demonstrated the enhancement of vision–inertial methods over vision-only methods. In fact, the coupling of other sensors, such as sonar, depth meter, etc., with vision has also achieved good results, which has been discussed in the literature. However, adding more sensors provides more observations for the optimization of navigation and positioning algorithms, but comes with equipment cost, space occupation, and energy consumption issues. This is a long-standing dilemma for large-scale, low-cost, long-term UUV underwater scientific research deployments. So, how to balance these conflicts and provide reliable and available high-precision underwater autonomous navigation and positioning UUV is our unremitting pursuit in the future.

6. Conclusions

In this work, vision-based techniques for the autonomous navigation and positioning of UUVs have been studied in depth. This paper summarizes and reviews the navigation and positioning of UUVs based on vision and the fusion of vision and other sensors and introduces the current development trend of UUVs and the main methods of visual navigation and positioning. The methods outlined can overcome some limitations and take advantage of image data more fully. In this paper, visual navigation and positioning methods are divided into two categories: geometry-based methods and deep learning-based methods. This paper not only introduces the development process of two kinds of methods but also selects typical algorithms to compare the navigation and positioning performance of the two methods through typical underwater dataset. The paper also predicts upcoming research trends and priorities for UUV visual navigation and positioning, which enables the availability and potential of this technology. From the perspective of accuracy, the positioning accuracy of different algorithms is quite different, and the accuracy gap between them can reach an order of magnitude. Different precisions have different implications for personal consumption applications and scientific applications. For the field of personal consumption, the requirements for the positioning accuracy of UUVs are relatively low, and it is mainly used for entertainment, such as underwater shooting and video sharing. However, for scientific research applications, higher precision can help us take better underwater measurements and observations. For example, in the field of coral survey and protection, coral reefs are a very fragile ecosystem, and we hope that UUVs engaged in underwater surveys have higher-precision autonomous navigation and positioning capabilities in order to avoid damage to corals. At the same time, they can also provide efficient technical means for oceanographers to conduct high-precision mapping of coral habitats and automatic periodic surveys in different periods.

Author Contributions: Conceptualization, J.Q. and M.L.; methodology, J.Q.; software, J.Z. and K.Y.; validation, J.Q. and M.L.; formal analysis, J.Z.; investigation, J.Z. and K.Y.; resources, M.L.; data curation, J.Q.; writing—original draft preparation, J.Q.; writing—review and editing, J.Q. and M.L.; visualization, J.Q.; supervision, D.L.; project administration, D.L.; funding acquisition, D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China, grant number 2018YFB0505400, the National Natural Science Foundation of China (NSFC), grant number 41901407 and the College Students' Innovative Entrepreneurial Training Plan Program, grant number S202210486307, Research on visual navigation, perception and localization algorithm of unmanned underwater vehicle/robot (UUV).

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the reviewers for their constructive suggestions.

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

References

1. Khawaja, W.; Semkin, V.; Ratyal, N.; Yaqoob, Q.; Gul, J. Threats from and Countermeasures for Unmanned Aerial and Underwater Vehicles. *Sensors* **2022**, *22*, 3896. [CrossRef]
2. Chemisky, B.; Nocerino, E.; Menna, F.; Nawaf, M.; Drap, P. A Portable Opto-Acoustic Survey Solution for Mapping of Underwater Targets. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *43*, 651–658. [CrossRef]
3. Gilson, C. The Future of Optical Sensors Will Enhance Navigation. *Ocean. News Technol. Mag.* **2021**, *12*, 12–13.
4. Petillot, Y.R.; Antonelli, G.; Casalino, G.; Ferreira, F. Underwater robots: From remotely operated vehicles to intervention-autonomous underwater vehicles. *IEEE Robot. Autom. Mag.* **2019**, *26*, 94–101. [CrossRef]
5. He, Y.; Wang, D.; Ali, Z. A Review of Different Designs and Control Models of Remotely Operated Underwater Vehicle. *Meas. Control* **2020**, *53*, 1561–1570. [CrossRef]
6. Yoerger, D.; Govindarajan, A.; Howland, J.; Llopiz, J.; Wiebe, P.; Curran, M.; Fujii, J.; Gomez-Ibanez, D.; Katija, K.; Robison, B.; et al. A hybrid underwater robot for multidisciplinary investigation of the ocean twilight zone. *Sci. Robot.* **2021**, *6*, 1901–1912. [CrossRef] [PubMed]
7. HROV Nereid Under Ice. Available online: https://www.whoi.edu/what-we-do/explore/underwater-vehicles/hybrid-vehicles/nereid-under-ice/ (accessed on 28 September 2021).
8. Vasilijević, A.; Nađ, Đ.; Mandić, F.; Mišković, N.; Vukić, Z. Coordinated navigation of surface and underwater marine robotic vehicles for ocean sampling and environmental monitoring. *IEEE/ASME Trans. Mechatron.* **2017**, *22*, 1174–1184. [CrossRef]
9. Ken Kostel. Terrain Relative Navigation: From Mars to the Deep Sea. Available online: https://oceanexplorer.noaa.gov/okeanos/explorations/ex2102/features/trn/trn.html (accessed on 11 May 2021).
10. Sun, K.; Cui, W.; Chen, C. Review of Underwater Sensing Technologies and Applications. *Sensors* **2021**, *21*, 7849. [CrossRef]
11. Burguera, B.A.; Bonin-Font, F. A Trajectory-Based Approach to Multi-Session Underwater Visual SLAM Using Global Image Signatures. *J. Mar. Sci. Eng.* **2019**, *7*, 278. [CrossRef]
12. Wu, Y.; Ta, X.; Xiao, R.; Wei, Y.; Li, D. Survey of Underwater Robot Positioning Navigation. *Appl. Ocean Res.* **2019**, *90*, 101845–101860. [CrossRef]
13. Tan, H.P.; Diamant, R.; Seah, W.K.G.; Waldmeyer, M. A survey of techniques and challenges in underwater localization. *Ocean Eng.* **2011**, *38*, 1663–1676. [CrossRef]
14. Toky, A.; Singh, R.; Das, S. Localization Schemes for Underwater Acoustic Sensor Networks—A Review. *Comput. Sci. Rev.* **2020**, *37*, 100241–100259. [CrossRef]
15. Ran, T.; Yuan, L.; Zhang, J. Scene perception based visual navigation of mobile robot in indoor environment. *ISA Trans.* **2021**, *109*, 389–400. [CrossRef] [PubMed]
16. Zhang, J.; Ila, V.; Kneip, L. Robust visual odometry in underwater environment. In Proceedings of the OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO), Kobe, Japan, 28–31 May 2018.
17. Nash, J.; Bond, J.; Case, M.; Mccarthy, I.; Teahan, W. Tracking the fine scale movements of fish using autonomous maritime robotics: A systematic state of the art review. *Ocean Eng.* **2021**, *229*, 108650–108671. [CrossRef]
18. Liu, J.; Gong, S.; Guan, W.; Guan, W.; Li, B.; Li, H.; Liu, J. Tracking and Localization based on Multi-angle Vision for Underwater Target. *Electronics* **2020**, *9*, 1871. [CrossRef]
19. Kim, J. Cooperative localization and unknown currents estimation using multiple autonomous underwater vehicles. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2365–2371. [CrossRef]
20. Plum, F.; Labisch, S.; Dirks, J.H. SAUV—A bio-inspired soft-robotic autonomous underwater vehicle. *Front. Neurorobot.* **2020**, *14*, 8. [CrossRef]
21. Lu, Y.; Xue, Z.; Xia, G.S.; Zhang, L. A survey on vision-based UAV navigation. *Geo-Spat. Inf. Sci.* **2018**, *21*, 21–32. [CrossRef]
22. Guilherme, N.D.; Avinash, C.K. Vision for mobile robot navigation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 237–267.
23. Nourani-Vatani, N.; Borges, P.V.K.; Roberts, J.M.; Srinivasan, M.V. On the use of optical flow for scene change detection and description. *J. Intell. Robot. Syst.* **2014**, *74*, 817–846. [CrossRef]
24. Zhang, X.; Wang, L.; Su, Y. Visual place recognition: A survey from deep learning perspective. *Pattern Recognit.* **2021**, *113*, 107760–107781. [CrossRef]
25. Cho, D.M.; Tsiotras, P.; Zhang, G.; Marcus, J. Robust feature detection, acquisition and tracking for relative navigation in space with a known target. In Proceedings of the AIAA Guidance, Navigation, and Control (GNC) Conference, Boston, MA, USA, 19–22 August 2013.
26. Moravec, H.; Elfes, A. High resolution maps from wide angle sonar. In Proceedings of the IEEE International Conference on Robotics and Automation, St. Louis, MO, USA, 25 March 1985.
27. Thoma, J.; Paudel, D.P.; Chhatkuli, A.; Probst, T.; Gool, L.V. Mapping, localization and path planning for image-based navigation using visual features and map. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Los Angeles, CA, USA, 16–20 June 2019.
28. Saputra, M.R.U.; Markham, A.; Trigoni, N. Visual SLAM and structure from motion in dynamic environments: A survey. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–36. [CrossRef]
29. Yasuda, Y.D.V.; Martins, L.E.G.; Cappabianco, F.A.M. Autonomous visual navigation for mobile robots: A systematic literature review. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–34. [CrossRef]

30. Paull, L.; Saeedi, S.; Seto, M.; Li, H. AUV navigation and localization: A review. *IEEE J. Ocean. Eng.* **2013**, *39*, 131–149. [CrossRef]
31. Orpheus Explores the Ocean's Greatest Depths. Available online: https://www.whoi.edu/multimedia/orpheus-explores-the-oceans-greatest-depths/ (accessed on 18 September 2019).
32. Zhou, Z.; Liu, J.; Yu, J. A Survey of Underwater Multi-Robot Systems. *IEEE/CAA J. Autom. Sin.* **2021**, *9*, 1–18. [CrossRef]
33. Qin, J.; Yang, K.; Li, M.; Zhong, J.; Zhang, H. Real-time Positioning and Tracking for Vision-based Unmanned Underwater Vehicles. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *46*, 163–168. [CrossRef]
34. Fuentes-Pacheco, J.; Ruiz-Ascencio, J.; Rendón-Mancha, J.M. Visual simultaneous localization and mapping: A survey. *Artif. Intell. Rev.* **2015**, *43*, 55–81. [CrossRef]
35. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, S.; 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]
36. Chen, C.; Zhu, H.; Li, M.; You, S. A review of visual-inertial simultaneous localization and mapping from filtering-based and optimization-based perspectives. *Robotics* **2018**, *7*, 45. [CrossRef]
37. Servières, M.; Renaudin, V.; Dupuis, A.; Antigny, N. Visual and visual-inertial SLAM: State of the art, classification, and experimental benchmarking. *J. Sens.* **2021**, *2021*, 2054828. [CrossRef]
38. Macario, B.A.; Michel, M.; Moline, Y.; Corre, G.; Carrel, F. A Comprehensive Survey of Visual SLAM Algorithms. *Robotics* **2022**, *11*, 24. [CrossRef]
39. González-García, J.; Gómez-Espinosa, A.; Cuan-Urquizo, E.; Garcia-Valdovinos, L.G.; Salgado-Jimenez, T.; Cabello, J.A.E. Autonomous underwater vehicles: Localization, navigation, and communication for collaborative missions. *Appl. Sci.* **2020**, *10*, 1256. [CrossRef]
40. Maurelli, F.; Krupiski, S.; Xiang, X.; Petillot, Y. AUV localisation: A review of passive and active techniques. *Int. J. Intell. Robot. Appl.* **2021**, *6*, 246–269. [CrossRef]
41. Watson, S.; Duecker, D.A.; Groves, K. Localisation of unmanned underwater vehicles (UUVs) in complex and confined environments: A review. *Sensors* **2020**, *20*, 6203. [CrossRef]
42. Wirth, S.; Carrasco, P.L.N.; Codina, G.O. Visual odometry for autonomous underwater vehicles. In Proceedings of the MTS/IEEE OCEANS-Bergen, Bergen, Norway, 10–14 June 2013.
43. Bellavia, F.; Fanfani, M.; Colombo, C. Selective visual odometry for accurate AUV localization. *Auton. Robot.* **2017**, *41*, 133–143. [CrossRef]
44. Choi, J.; Lee, Y.; Kim, T.; Jung, J.; Choi, H. Development of a ROV for visual inspection of harbor structures. In Proceedings of the IEEE Underwater Technology (UT), Busan, Korea, 21–24 February 2017.
45. Xu, Z.; Haroutunian, M.; Murphy, A.J.; Neasham, J.; Norman, R. An Underwater Visual Navigation Method Based on Multiple ArUco Markers. *J. Mar. Sci. Eng.* **2021**, *9*, 1432. [CrossRef]
46. Li, M.; Qin, J.; Li, D.; Chen, R.; Liao, X.; Guo, B. VNLSTM-PoseNet: A novel deep ConvNet for real-time 6-DOF camera relocalization in urban streets. *Geo-Spatial Inf. Sci.* **2021**, *24*, 422–437. [CrossRef]
47. Ferrera, M.; Moras, J.; Trouvé-Peloux, P.; Creuze, V. Real-time monocular visual odometry for turbid and dynamic underwater environments. *Sensors* **2019**, *19*, 687. [CrossRef] [PubMed]
48. Taketomi, T.; Uchiyama, H.; Ikeda, S. Visual SLAM algorithms: A survey from 2010 to 2016. *IPSJ Trans. Comput. Vis. Appl.* **2017**, *9*, 16. [CrossRef]
49. Azzam, R.; Taha, T.; Huang, S.; Zweiri, Y. Feature-based visual simultaneous localization and mapping: A survey. *SN Appl. Sci.* **2020**, *2*, 224. [CrossRef]
50. Bailey, T.; Nieto, J.; Guivant, J.; Stevens, M.; Nebot, E. Consistency of the EKF-SLAM algorithm. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 9–15 October 2006.
51. Yan, J.; Guorong, L.; Shenghua, L.; Zhou, L. A review on localization and mapping algorithm based on extended kalman filtering. *Int. Forum Inf. Technol. Appl.* **2009**, *2*, 435–440.
52. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef] [PubMed]
53. Stentz, A.; Fox, D.; Montemerlo, M. FastSLAM: A factored solution to the simultaneous localization and mapping problem with unknown data association. In Proceedings of the AAAI National Conference on Artificial Intelligence, Edmonton, AB, Canada, 28 July–1 August 2002.
54. Klein, G.; Murray, D. Parallel tracking and mapping for small AR workspaces. In Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, Japan, 13–16 November 2007.
55. 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]
56. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [CrossRef]
57. Newcombe, R.A.; Lovegrove, S.J.; Davison, A.J. DTAM: Dense tracking and mapping in real-time. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
58. 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.
59. Engel, J.; Koltun, V.; Cremers, D. Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 611–625. [CrossRef]

60. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–5 June 2014.
61. Zhao, Y.; Smith, J.S.; Vela, P.A. Good graph to optimize: Cost-effective, budget-aware bundle adjustment in visual SLAM. *arXiv* **2020**, arXiv:2008.10123.
62. Ferrera, M.; Eudes, A.; Moras, J.; Sanfourche, M.; Besnerais, G.L. OV2SLAM: A Fully Online and Versatile Visual SLAM for Real-Time Applications. *IEEE Robot. Autom. Lett.* **2021**, *6*, 1399–1406. [CrossRef]
63. Zhou, Y.; Gallego, G.; Shen, S. Event-based stereo visual odometry. *IEEE Trans. Robot.* **2021**, *37*, 1433–1450. [CrossRef]
64. Koestler, L.; Yang, N.; Zeller, N.; Cremers, D. TANDEM: Tracking and Dense Mapping in Real-time using Deep Multi-view Stereo. In Proceedings of the Conference on Robot Learning, Auckland, New Zealand, 14–18 December 2022.
65. Xu, Z.; Haroutunian, M.; Murphy, A.J.; Neasham, J.; Norman, R. An Integrated Visual Odometry System for Underwater Vehicles. *IEEE J. Ocean. Eng.* **2020**, *46*, 848–863. [CrossRef]
66. Jinyu, L.; Bangbang, Y.; Danpeng, C.; Wang, N.; Zhang, G.; Bao, H. Survey and evaluation of monocular visual-inertial SLAM algorithms for augmented reality. *Virtual Real. Intell. Hardw.* **2019**, *1*, 386–410. [CrossRef]
67. Weiss, S.M. Vision Based Navigation for Micro Helicopters. Ph.D. Thesis, ETH Zurich, Zurich, Switzerland, 2012.
68. Palézieux, N.; Nägeli, T.; Hilliges, O. Duo-VIO: Fast, light-weight, stereo inertial odometry. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016.
69. Mourikis, A.I.; Roumeliotis, S.I. A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Rome, Italy, 10–14 April 2007.
70. Bloesch, M.; Omari, S.; Hutter, M.; Siegwart, R. Robust visual inertial odometry using a direct EKF-based approach. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015.
71. Bloesch, M.; Burri, M.; Omari, S.; Hutter, M.; Siegwart, R. Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback. *Int. J. Robot. Res.* **2017**, *36*, 1053–1072. [CrossRef]
72. Leutenegger, S.; Furgale, P.; Rabaud, V.; Chli, M.; Konolige, K.; Siegwart, R. Keyframe-based visual-inertial SLAM using nonlinear optimization. In Proceedings of the Robotis Science and Systems (RSS), Berlin, Germany, 24–28 June 2013.
73. Leutenegger, S.; Lynen, S.; Bosse, M.; Siegwart, R.; Furgale, P. Keyframe-based visual–inertial odometry using nonlinear optimization. *Int. J. Robot. Res.* **2015**, *34*, 314–334. [CrossRef]
74. Mur-Artal, R.; Tardós, J.D. Visual-inertial monocular SLAM with map reuse. *IEEE Robot. Autom. Lett.* **2017**, *2*, 796–803. [CrossRef]
75. Qin, T.; Li, P.; Shen, S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [CrossRef]
76. Qin, T.; Pan, J.; Cao, S.; Shen, S. A general optimization-based framework for local odometry estimation with multiple sensors. *arXiv* **2019**, arXiv:1901.03638.
77. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and mul-timap SLAM. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [CrossRef]
78. Li, X.; Li, Y.; Örnek, E.P.; Lin, J.; Tombari, F. Co-planar parametrization for stereo-SLAM and visual-inertial odometry. *IEEE Robot. Autom. Lett.* **2020**, *5*, 6972–6979. [CrossRef]
79. Xie, H.; Chen, W.; Wang, J.; Wang, H. Hierarchical quadtree feature optical flow tracking based sparse pose-graph visual-inertial SLAM. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Online, 31 May–15 June 2020.
80. Seiskari, O.; Rantalankila, P.; Kannala, J.; Ylilammi, J.; Rahtu, E.; Solin, A. HybVIO: Pushing the Limits of Real-time Visual-inertial Odometry. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022.
81. Rahman, S.; Li, A.Q.; Rekleitis, I. Sonar visual inertial SLAM of underwater structures. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018.
82. Rahman, S.; Li, A.Q.; Rekleitis, I. Svin2: An underwater SLAM system using sonar, visual, inertial, and depth sensor. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 4–8 November 2019.
83. Rahman, S.; Li, A.Q.; Rekleitis, I. Contour based reconstruction of underwater structures using sonar, visual, inertial, and depth sensor. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 4–8 November 2019.
84. Cebollada, S.; Payá, L.; Flores, M.; Peidro, A.; Reinoso, O. A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data. *Expert Syst. Appl.* **2021**, *167*, 114195. [CrossRef]
85. Sartipi, K.; Do, T.; Ke, T.; Vuong, K.; Roumeliotis, S.I. Deep depth estimation from visual-inertial SLAM. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 4 October 2020–24 January 2021.
86. Duan, C.; Junginger, S.; Huang, J.; Jin, K.; Thurow, K. Deep learning for visual SLAM in transportation robotics: A review. *Transp. Saf. Environ.* **2019**, *1*, 177–184. [CrossRef]
87. Zhao, C.; Sun, Q.; Zhang, C.; Tang, Y.; Qian, F. Monocular depth estimation based on deep learning: An overview. *Sci. China Technol. Sci.* **2020**, *63*, 1612–1627. [CrossRef]
88. Ming, Y.; Meng, X.; Fan, C.; Yu, H. Deep learning for monocular depth estimation: A review. *Neurocomputing* **2021**, *438*, 14–33. [CrossRef]

89. Arshad, S.; Kim, G.W. Role of deep learning in loop closure detection for visual and lidar SLAM: A survey. *Sensors* **2021**, *21*, 1243. [CrossRef]
90. Sualeh, M.; Kim, G.W. Simultaneous localization and mapping in the epoch of semantics: A survey. *Int. J. Control. Autom. Syst.* **2019**, *17*, 729–742. [CrossRef]
91. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 25–34. [CrossRef]
92. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
93. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
94. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. Lift: Learned invariant feature transform. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
95. Li, D.; Shi, X.; Long, Q.; Liu, S.; Yang, W.; Wang, F.; Wei, Q.; Qiao, F. DXSLAM: A robust and efficient visual SLAM system with deep features. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 4 October 2020–24 January 2021.
96. Tateno, K.; Tombari, F.; Laina, I.; Navab, N. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
97. Li, R.; Wang, S.; Long, Z.; Gu, D. UnDeepVo: Monocular visual odometry through unsupervised deep learning. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018.
98. Kim, J.J.Y.; Urschler, M.; Riddle, P.J.; Wicker, J.S. SymbioLCD: Ensemble-Based Loop Closure Detection using CNN-Extracted Objects and Visual Bag-of-Words. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021.
99. Bescos, B.; Fácil, J.M.; Civera, J.; Neira, J. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robot. Autom. Lett.* **2018**, *3*, 4076–4083. [CrossRef]
100. Liu, Y.; Miura, J. RDS-SLAM: Real-time dynamic SLAM using semantic segmentation methods. *IEEE Access* **2021**, *9*, 23772–23785. [CrossRef]
101. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
102. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.; Woo, W. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 28–37.
103. Wang, S.; Clark, R.; Wen, H.; Trigoni, N. DeepVo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Marina Bay Sands, Singapore, 29 May–3 June 2017.
104. Walch, F.; Hazirbas, C.; Leal-Taixe, L.; Sattler, T.; Hilsenbeck, S.; Cremers, D. Image-based localization using LSTMs for structured feature correlation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
105. Xue, F.; Wang, Q.; Wang, X.; Dong, W.; Wang, J.; Zha, H. Guided feature selection for deep visual odometry. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018.
106. Chancán, M.; Milford, M. DeepSeqSLAM: A trainable CNN+ RNN for joint global description and sequence-based place recognition. *arXiv* **2020**, arXiv:2011.08518.
107. Clark, R.; Wang, S.; Wen, H.; Markham, A.; Trigoni, N. VINet: Visual-inertial odometry as a sequence-to-sequence learning problem. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
108. Chen, C.; Rosa, S.; Miao, Y.; Lu, C.X.; Wu, W.; Markham, A.; Trigoni, N. Selective sensor fusion for neural visual-inertial odometry. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Los Angeles, CA, USA, 16–20 June 2019.
109. Han, L.; Lin, Y.; Du, G.; Lian, S. DeepVIO: Self-supervised deep learning of monocular visual inertial odometry using 3d geometric constraints. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 4–8 November 2019.
110. Chen, D.; Wang, N.; Xu, R.; Xie, W.; Bao, H.; Zhang, G. RNIN-VIO: Robust Neural Inertial Navigation Aided Visual-Inertial Odometry in Challenging Scenes. In Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Bari, Italy, 4–8 October 2021.
111. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 27–36.
112. Almalioglu, Y.; Saputra, M.R.U.; deGusmao, P.P.B.; Markham, A.; Trigoni, N. GanVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In Proceedings of the International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019.
113. Feng, T.; Gu, D. SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. *IEEE Robot. Autom. Lett.* **2019**, *4*, 4431–4437. [CrossRef]
114. Almalioglu, Y.; Turan, M.; Saputra, M.R.U.; Gusmao, P.P.B.; Markham, A.; Trigoni, N. SelfVIO: Self-supervised deep monocular visual-inertial odometry and depth estimation. *Neural Netw. arXiv* **2019**, arXiv:1911.09968. [CrossRef] [PubMed]

115. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81. [CrossRef]
116. Xue, F.; Wu, X.; Cai, S.; Wang, J. Learning multi-view camera relocalization with graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020.
117. Sarlin, P.E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning feature matching with graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020.
118. Taherkhani, A.; Belatreche, A.; Li, Y.; Cosma, G.; Maguire, L.P.; McGinnity, T.M. A review of learning in biologically plausible spiking neural networks. *Neural Netw.* **2020**, *122*, 253–272. [CrossRef] [PubMed]
119. Milford, M.J.; Wyeth, G.F.; Prasser, D. RatSLAM: A hippocampal model for simultaneous localization and mapping. In Proceedings of the International Conference on Robotics and Automation (ICRA), New Orleans, LA, USA, 26 April–1 May 2004.
120. Ball, D.; Heath, S.; Wiles, J.; Wyeth, G.; Corke, P.; Milford, M. OpenRatSLAM: An open source brain-based SLAM system. *Auton. Robot.* **2013**, *34*, 149–176. [CrossRef]
121. Silveira, L.; Guth, F.; Drews-Jr, P.; Ballester, P.; Machado, M.; Codevilla, F.; Duarte-Filho, N.; Botelho, S. An open-source bio-inspired solution to underwater SLAM. *IFAC-PapersOnLine* **2015**, *48*, 212–217. [CrossRef]
122. Yu, F.; Shang, J.; Hu, Y.; Milford, M. NeuroSLAM: A brain-inspired SLAM system for 3D environments. *Biol. Cybern.* **2019**, *113*, 515–545. [CrossRef]
123. Qin, J. Visual-Navigation-and-Positioning. Available online: https://github.com/qinjiangying/visual-navigation-and-positioning (accessed on 30 May 2022).
124. Ferrera, M.; Creuze, V.; Moras, J.; Trouve-Peloux, P. AQUALOC: An underwater dataset for visual–inertial–pressure localization. *Int. J. Robot. Res.* **2019**, *38*, 1549–1559. [CrossRef]