electronics

MDPI

Article

Stereo SLAM in Dynamic Environments Using Semantic Segmentation

Yongbao Ai, Qianchong Sun, Zhipeng Xi, Na Li, Jianmeng Dong and Xiang Wang *

National Innovation Institute of Defense Technology, Beijing 100071, China; aybjackai@163.com (Y.A.); sunqianchong@hotmail.com (Q.S.); xzp_paper@163.com (Z.X.); spring831004@163.com (N.L.); 15652655581@163.com (J.D.)

* Correspondence: wang_xiang927@163.com

Abstract: As we all know, many dynamic objects appear almost continuously in the real world that are immensely capable of impairing the performance of the majority of vision-based SLAM systems based on the static-world assumption. In order to improve the robustness and accuracy of visual SLAM in high-dynamic environments, a real-time and robust stereo SLAM system for dynamic scenes was proposed. To weaken the influence of dynamic content, the moving-object detection method was put forward in our visual odometry, and then the semantic segmentation network was combined in our stereo SLAM to extract pixel-level contours of dynamic objects. Then, the influences of dynamic objects were significantly weakened and the performance of our system increased markedly in dynamic, complex, and crowed city spaces. Following experiments with both the KITTI Odometry dataset and in a real-life scene, the results showed that our method could dramatically decrease the tracking error or drift, and improve the robustness and stability of our stereo SLAM in high dynamic outdoor scenarios.

Keywords: stereo SLAM; semantic segmentation; moving object detection; dynamic scenarios

Citation: Ai, Y.; Sun, Q.; Xi, Z.; Li, N.; Dong, J.; Wang, X. Stereo SLAM in Dynamic Environments Using Semantic Segmentation. *Electronics* **2023**, *12*, 3112. https://doi.org/10.3390/electronics12143112

Academic Editor: Arturo de la Escalera Hueso

Received: 27 May 2023
Revised: 24 June 2023
Accepted: 10 July 2023
Published: 18 July 2023

1. Introduction

In the 21st century, the increasing maturation of computer vision technology, artificial intelligence, and sensor technology have promoted the gradual transformation of robots from traditional industrial robots to intelligent robots capable of perception, analysis, learning, and decision making. Intelligent robots that have characteristics of interconnectivity, virtuality and reality combination, and man–machine fusion will play a more and more important role in industry, agriculture, transportation, aerospace, national defense, military, public security, and other fields. Nevertheless, realizing efficient and accurate environment perception, autonomous localization, and the navigation of intelligent robots in an unstructured environment composed of static obstacles and high-dynamic objects is a hot topic in the field, and the key technologies include practical Simultaneous Localization and Mapping (SLAM). This refers to simultaneously estimating locations of newly perceived landmarks and the location of the external environment sensing sensor itself, while incrementally mapping an unknown environment. Moreover, SLAM is considered to be the critical technology for visual navigation for several applications, for example, by exploring areas inaccessible to humans or where maneuverability is difficult, in such scenarios as under water, inside mines, or in other narrow spaces.

Cameras, which are common in life, offer the advantages of being fully 3D, cheaper than lasers, convenient, and having a visual perception like humans. More importantly, cameras can provide seemingly endless potential for the extraction of detail, both geometric and photometric, and for semantic and other higher-level scene understanding, as continually elucidated by the prosperous computer vision research community. Hence, visual SLAM systems are a pivotal strategy for the development of future navigation techniques.

The last two decades have seen a significant surge in visual SLAM [1–3] from range-sensor-based systems. And many well-performing SLAM systems have been developed, such as SVO [4], LSD-SLAM [5], and ORB-SLAM1/2/3 [6–8]. Nevertheless, to simplify the problem formulation, the majority of SLAM systems typically assume a static world, where only rigid and non-moving objects are involved. In the real world, it is common to have objects in motion relative to the stationary environment around them, especially in the field of public transportation, where vehicles and pedestrians constantly move back and forth. Although a fraction of moving objects in classical systems can be dealt with by regarding them as noise, the great majority of dynamic objects violate the assumption of a static world. In addition, the accidental appearance of these dynamic objects in front of camera lenses can cause a decrease in the accuracy of the visual SLAM system, and even the phenomenon of image blur caused by fast moving objects, seriously weakening the robustness and stability of the visual SLAM system, which leads to some available visual SLAM systems being limited for real-world applications. Therefore, they require modeling and tracking to ensure that they do not interfere with the location and map generation of the visual SLAM system. Some solutions have been proposed, such as motion target tracking, spatiotemporal consistency, dynamic object segmentation and modeling, and other technologies, to deal with the influence of dynamic objects on the system. Just as Panchpor et al. elucidated in a survey [9], since many problems in SLAM for automatic robots in dynamic environments cannot have a robust solution, there is a large amount of scope for further research and development in this domain.

In this paper, we reveal the advantages of accommodating both the semantic segmentation convolutional neural network and the proposed moving object detection algorithm in a visual SLAM system using just a stereo camera. After conducting experiments on the KITTI Odometry datasets [10] and real-world scenarios, the performance of our system showed a significant improvement in high-dynamic scenes compared to the original ORB-SLAM2 [7] system. It also outperformed the DynaSLAM [11] system, which was specifically designed for dynamic scenes. The novelty of our paper is summed up below:

- A new stereo SLAM system based on the ORB-SLAM2 framework combined with a deep learning method is put forward to decrease the impact of dynamic objects on the camera pose and trajectory estimation. The approach of the semantic segmentation network plays a role in the data preprocessing stage to filter out the feature expression of moving objects.
- A novel motion object detection method is presented to reduce the influence of moving targets on the camera pose and trajectory estimation, which calculates the likelihood of each keyframe point belonging to the dynamic content and distinguishes between dynamic and static goals in scenarios.
- The semantic segmentation neural network ENet [12], which is appropriate for city spaces, is first utilized (to the best of our knowledge) to enhance the performance of the visual SLAM system, which makes our system more robust and practical in high-dynamic and complex city streets, and this has practical engineering applications to a certain extent.

The remainder of the paper is organized as follows: Section 2 presents the relevant work. The architecture of the proposed stereo SLAM system and the details of our method to solve visual SLAM in dynamic scenes are provided at length in Section 3. Thereafter, we show the qualitative and quantitative experimental results in Section 4. Finally, both the conclusion and future research are given in Section 5.

2. Related Work

The Visual SLAM with a stereo camera. Since Davison and Murray described the first traditional stereo-SLAM framework, which was a real-time EKF-based system [13–15], many scholars have made contributions toward this research field. Iocchi et al. combined range mapping and image alignment to reconstruct a planar environment map with just a stereo camera [16]. Se et al. described a stereo-based mobile robot SLAM algorithm

utilizing SIFT features in a small lab scene [17], which was not adapted to large environments or working in challenging outdoor scenes. In [18,19], the authors demonstrated an autonomous low altitude aero-craft system using the EKF SLAM algorithm for terrain mapping with a stereo sensor. Saez et al. presented a full 6-DOF SLAM with a wearable stereo device, which was based on both the ego-motion and the global rectification algorithms [20]. A dense visual SLAM system utilizing Rao-Blackwellized particle filters and SIFT features was demonstrated in [21,22]. The system could also work in stereo mode, with a camera fixed on a robot moving in 2D space. In [23], a six-degrees-of-freedom SLAM with a stereo handheld camera was presented, which was utilized to construct large-scale indoor or outdoor environments on the basis of the conditionally independent divide and conquer algorithm. The many aforementioned SLAM approaches focused on operating in static environments; it is a strong mathematical modeling assumption, so it restricts the system a lot in practical applications. Actually, there are several dynamic objects in real environments, where the moving objects can generate errors in visual SLAM performing in outdoor or indoor dynamic scenes. This is because dynamic features cause bad pose estimation and erroneous data association. For this reason, there are few SLAM systems that specifically address dynamic content, trying to split dynamic and static regions within sequences of images of a dynamic environment. In [24], Lin and Wang presented a stereo-based SLAMMOT method which makes use of the inverse depth parameterization and performs EKF to overcome the observability issue. Kawewong et al. [25] used position-invariant robust features (PIRFs) to simultaneously localize and map in high dynamic environments. Alcatnarilla et al. [26] detected dynamic objects by employing a scene flow representation using stereo cameras for visual odometry [27] in order to improve the robustness of the visual SLAM system [28]. But the algorithm of dense optical flow may detect many pixels in an image as dynamic objects, whereas these pixels remain with static content on account of inevitable measurement noises or optical-flow-inherent problems. In [29], Zou and Tan described a collaborative vision-based SLAM system with multiple cameras, which differs from currently available SLAM systems with just one stereo camera fixed on a single platform. Their system utilized images from each camera to construct a global map and could be run robustly in high dynamic scenes. Their method to deal with dynamic objects is a novel effort, but there is a high hardware cost due to the use of multiple cameras. Fan et al. [30] proposed an image fusion algorithm to get rid of the influence of dynamic content in a stereo-based SLAM. But their system cannot cope with moving objects with a slower speed of movement or large size.

Deep learning models are used in the visual SLAM system, which has proved its superiority among SLAM systems. There are many outstanding studies that have employed deep learning models to replace some non-geometric modules in traditional SLAM systems. B. Bescos et al. [11] presented the DynaSLAM, which uses Mask R-CNN [31] to segment dynamic objects and background inpainting to synthesize a realistic image without dynamic objects. However, since the network speed of Mask R-CNN is slow and its segmentation algorithm architecture is not a multi-threaded operation mode, their system cannot perform in real time. In another paper [32], a weakly supervised semantic segmentation neural network was used in the system, which reduced annotations for training. Kang et al. proposed a DF-SLAM system which used a shallow neural network [33] to extract local descriptors [34]. In literature [35], a two-dimensional detection and classification CNN was used to provide semantic 3D box inferences, then a robust semantic SLAM system in which camera ego-motion and 3D semantic objects in high dynamic environments were tracked was presented with just a stereo camera. In our previous work [36], a robust RGB-D SLAM with a semantic segmentation neural network was put forward. Since the sensor of our system is an RGB-D camera, which is subject to the effects of illumination changes, it is only suitable for indoor dynamic scenes. In this work, we present a robust stereo SLAM that is primarily appropriate for high and complex outdoor dynamic traffic scenes.

Moving object detection (MOD) for the visual SLAM system in dynamic environments. Chiranjeevi and Sengupta used a combination of intensity and texture features to

carry out moving object detection [37]; specifically, they first used background subtraction to estimate the moving object, then utilized spatiotemporal features to eliminate some misclassified background regions, and finally obtained accurate moving object detection results. Their method requires a significant amount of computing resources for background modeling and spatiotemporal feature extraction, so it may face some challenges in practical applications. In another paper [38], an MOD algorithm was designed using the spatial geometric constraint of the stationary landmarks in the environment. Based on the MOD algorithm, the moving objects could be discriminated from the stationary landmarks. Zeng et al. [39] inferred object classes and poses that explained observations, while accounting for contextual relations between objects and the temporal consistency of object poses. They demonstrated an object detection performance superior to Faster R-CNN [40], and accurate 6-DOF object pose estimation compared to 3D registration methods such as ICP, and FPFH [41]. Hu et al. integrated a neural network for MOD and greatly reduced the negative influence of dynamic objects in a visual-inertial SLAM system [42]. The Mask R-CNN [31] was used to segment active moving objects, and motion detection based on LK optical flow [43] was used for passive moving objects in the ORB-SLAM2 (RGB-D) [44]. Due to the significant impact of lighting changes and the limited range of view for RGB-D cameras, this system was only suitable for small-scale indoor applications.

3. System Description

The proposed SLAM system will be described at length in this section. Firstly, the framework of our stereo SLAM is presented. After that, a brief presentation of the real-time semantic segmentation method utilized in our system is given. Then, the moving object detection method that is proposed to calculate the motion corresponding likelihoods among sequential stereo frames is expounded upon in detail. Finally, the means of filtering out outliers in the proposed SLAM system is presented.

3.1. Overview

Nowadays, how to accurately recognize and localize dynamic objects in real scenes, and eliminate their impact on camera pose and trajectory estimation and 3D point cloud map construction, are crucial tasks of visual SLAM. As we all know, ORB-SLAM2 has an outstanding performance in a variety of scenarios, from a handheld camera in the indoor environment to unmanned aerial vehicles flying in outdoor scenes and pilotless automobiles driving around in a road. Accordingly, our method was integrated with the ORB-SLAM2 system so as to achieve the purpose of improving the stability, robustness and reliability of its performance in large-scale scenes of dynamic and complex urban traffic roads.

Figure 1 shows a flow chart of the stereo SLAM system. It consists of four prime parallel threads: tracking, semantic segmentation, local mapping, and loop closing. The raw stereo RGB images obtained by a ZED camera are dealt with in the tracking and semantic segmentation thread at the same time. At the beginning, ORB feature points of frames are extracted in the tracking thread, and after that dynamic points are filtered out by the proposed moving object detection (MOD) algorithm. Afterwards, there is a wait for the pixel-wise segmentation mask acquired in the semantic segmentation thread. To incorporate the MOD method with semantic segmentation, we were able to discard feature points which indeed belonged to dynamic objects. The reason was that the inherent problem of semantic prior categories it may mistake static content for dynamic, e.g., parked cars or pedestrians waiting for traffic lights. Originally, these two categories of objects were predefined as dynamic objects in the prior category classification of the semantic segmentation algorithm. However, in a real environment, they exist in a static state for a short or long time. If only the semantic segmentation algorithm is used to process such objects, it is easy for it to misclassify. Therefore, the designed stereo SLAM system was also combined with the MOD method to detect the real dynamic areas in the frame and treat the ORB feature points belonging to these areas as outliers. Then, we obtained the transformation

matrix through matching the rest of the stable ORB feature points. Finally, the fifth thread is launched to perform full BA (Bundle Adjustment) after the local mapping and loop closing thread have been implemented, and optimize all camera poses and landmark points of the whole map to eliminate errors accumulated during system operation.

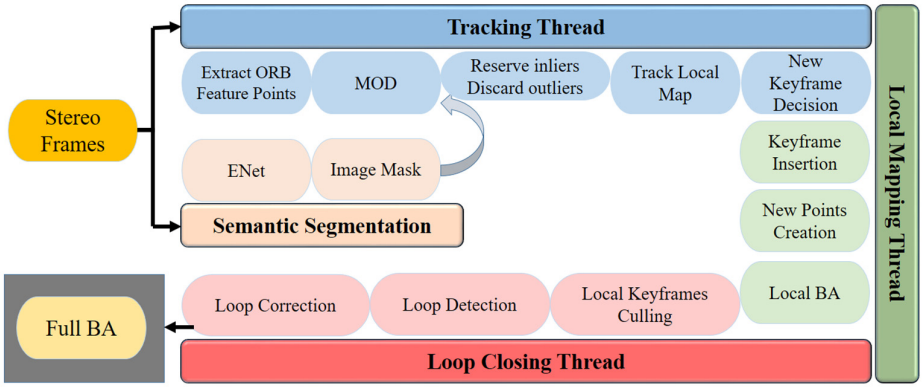


Figure 1. The overview of our VSLAM and main modules: tracking, semantic segmentation, mapping, and global optimization modules.

3.2. Semantic Segmentation

Semantic segmentation plays an important role in understanding the content of images and finding target objects, which is useful and helpful in some practical applications, such as unmanned driving and augmented reality. In our stereo SLAM system, ENet was adopted to provide pixel-wise semantic segmentation masks based on the Caffe implementation using TimoSaemann (https://github.com/TimoSaemann/ENet, 12 March 2018) in real time. And it was also designed to be faster, have fewer parameters, and be more accurate than SegNet, which is utilized in the paper [45]. The ENet is adequate for analyzing applications in the urban street scene [12]. The segmentation network trained on the Cityscapes dataset [46] could segment 19 classes in total.

The ENet neural network takes a color image as input, and outputs the corresponding semantic segmentation mask that labels every pixel in the image with one of the several pre-defined movable categories, for instance, vehicles, persons, and riders. These segmentation masks are easy to use in our SLAM system to accurately separate the dynamic object region and the static background area. The binary masks were input into the tracking thread, and the detail is clarified in Section 3.4.

3.3. Moving Object Detection

The moving object detection (MOD) method sets apart moving objects from the background of video sequence images. In many applications of computer vision, moving object detection is one of the key technologies in image processing [47], and is utilized to identify dynamic objects accurately in the image and keep from bringing on erroneous data association in the SLAM system running in dynamic, crowded real-world environments. Based on the classical LK (Lucas Kanada) optical flow method [43], our MOD method leverages Shi-Tomas corner detection [48] to extract sub-pixel corners in the previous frame, and then the pyramid LK optical flow algorithm is utilized to track the motion and obtain the matched feature points of the current frame. Next, feature points are screened preliminarily according to the following rules: on the condition that (1) the matched pairs are close to the edge of the frame, or (2) the pixel disparity of the 3×3 image block centered on matched pairs is too large, the matched pairs will be ignored. Afterwards, we can find the fundamental matrix through the RANSAC algorithm with a majority of matched feature points, which describes a relationship between any two images of the same scene that is restrained where the projection of points from the scene can appear in both images. The p_0,

p_1 is denoted as the matched points in the previous frame and current frame separately, and P_0, P_1 is their homogeneous coordinate form.

$$\begin{array}{ll} p_0 = [u_0, v_0] & p_1 = [u_1, v_1] \\ P_0 = [u_0, v_0, 1] & P_1 = [u_1, v_1, 1] \end{array} \tag{1}$$

where u, v are the pixel coordinates of the frame. Then, L denotes the epipolar line that maps the feature points from a previous frame to the correspondences search domain of a current frame, which can be solved as follows:

$$L = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = FP_0 = F \begin{bmatrix} u_0 \\ v_0 \\ 1 \end{bmatrix} \tag{2}$$

where X, Y, Z denote line vector and F represents fundamental matrix. The distance from the point P_1 of the current frame to its corresponding epipolar line L is computed by Equation (3):

$$D = \frac{|P_1^T F P_0|}{\sqrt{\|X\|^2 + \|Y\|^2}} \tag{3}$$

where D stands for the distance. If the value of D is higher than the predetermined threshold, matched points will be decided to be dynamic. Figure 2 describes an instance of the four images that can be utilized to compute a sparse optical flow representation of the scene. The original stereo RGB images used in it were collected using a ZED camera in a dynamic real urban road scene. It can be seen that the proposed MOD method can detect moving points in the current frame. However, because the movement of the camera itself interferes with the judgment of the MOD method, some moving points are also detected by mistake on the static buildings in the picture. Therefore, in order to find the real dynamic area in the image, the result obtained by the MOD method is combined with the segmentation mask obtained by real-time semantic segmentation in Section 3.2 to accurately distinguish the dynamic object region and the static background area in frames. The specific implementation process will be discussed in detail in Section 3.4.

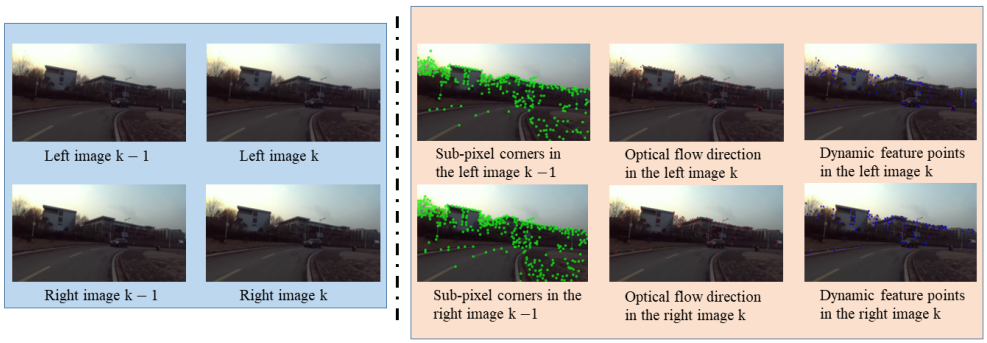


Figure 2. The schematic diagram of moving object detection based on Sparse Optical Flow method.

3.4. Outliers Removal

One of core tasks for visual SLAM in dynamic environments is the rejection of landmarks which in fact belong to dynamic objects. However, if the semantic segmentation method is only used to differentiate static and dynamic regions in frames, the SLAM system will fail to estimate lifelong models when predetermined moving objects, for instance, seated people or parked cars, keep static for a short time. What is worse, in an extremely challenging environment where dynamic objects may occupy almost the whole image view, there may be many residual correspondences declared as inliers, but they may actually belong to moving objects. This leads to large errors in the trajectory estimation and map-

ping of the SLAM system. Hence, we propose to integrate semantic segmentation with the MOD method to filter out outliers successfully.

The urban street scenarios comprise stationary and moving objects. The main error obtained in the measures is the presence of dynamic objects, since the total features fraction is localized on the moving objects. Therefore, it is important to avoid them during the system process. Then, we can illustrate how to identify regions of frames that belong to dynamic content through combining semantic segmentation with the MOD method. With the MOD algorithm explained in Section 3.3, we can derive motion likelihoods which can be utilized to differentiate dynamic objects, aiding the stereo SLAM performance in crowded and high dynamic environments. Once we have computed the dynamic feature points in the current frame, it is necessary to take into account the image mask of semantic segmentation using the ENet network in order to filter out moving objects. According to the experience from multiple tests, we should set a fixed threshold N = 5 so as to detect moving objects in a reliable way. If the dynamic feature points falling into the region of predefined semantic masks are more than N, the segmentation area belongs to moving objects, or conversely to static content. Afterwards, in this way, we can sweep away ORB feature points located on the dynamic areas of frames from the SLAM process, yielding more robust and accurate camera pose and trajectory results. The overall process of outlier removal is depicted in Figure 3.

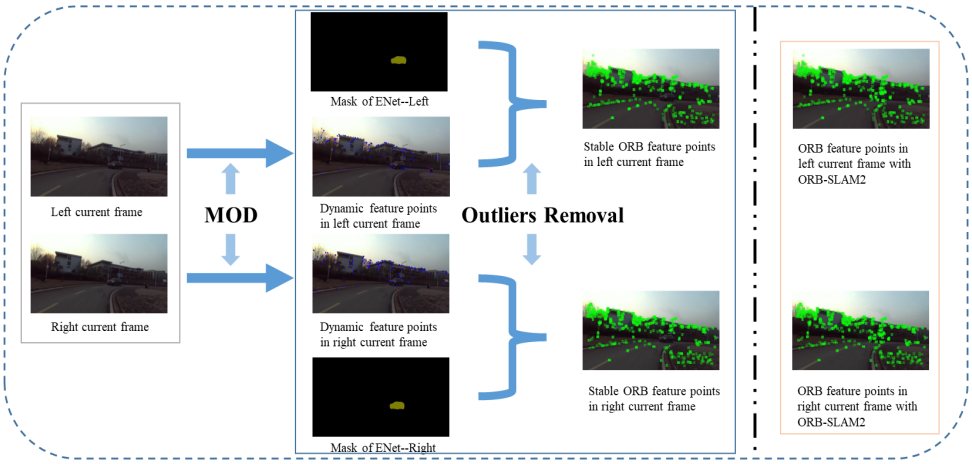


Figure 3. A couple of images captured by the ZED Stereo Camera. On the left side of the demarcation line the process of outlier removal is depicted. The right side is the ORB feature points result of frames without our approach. It can be seen that the ORB feature points of a dynamic car are not discarded in the ORB-SLAM2 system, so trajectory errors come out.

4. Experiments

To analyze the effectiveness of our method quantitatively, the performance of the proposed stereo SLAM system in dynamic scenarios was assessed with the KITTI Odometry dataset, at first. Then, the time required for tracking in our SLAM was counted to test its real-time performance. Furthermore, the stereo SLAM was integrated with the ROS system and we qualitatively tested it on a physical robot in the dynamic urban traffic scene to evaluate its accuracy, robustness, and practicability. Most experiments were executed on a computer with Intel i7 CPU, NVIDIA GeForce GTX TITAN X GPU, and 12 GB memory. The physical robot was a TurtleBot3, and stereo image sequences were captured with the ZED camera, which provided accurate camera calibration parameters and stereo rectification.

4.1. Evaluation Using KITTI Benchmark Dataset

The KITTI Odometry dataset [10] is one of the largest computer vision algorithm evaluation datasets suitable for autonomous driving scenarios, and it is captured by vehicle equipment driving around a middle-sized city, in the countryside, and on the freeway,

which provides many sequences in dynamic scenarios with accurate ground truth trajectories directly attained from the output of the GPS/IMU localization unit projected into the coordinate system of the left camera after rectification. The dataset contains 1240 × 376 stereo color and grayscale images captured at 10 Hz. The 01 sequence is collected on the freeway, and the 04 sequence is from a city road. Both of them were primarily applied in our experiments because they had high dynamic scenes.

To verify the effectiveness of our method in this section, besides comparing it with the ORB-SLAM2 system which our stereo SLAM proposed to improve its performance in high dynamic environments, the OpenVSLAM was chosen to measure the performance of our system in the static scene, and it was based on an indirect SLAM algorithm with sparse features, such as ProSLAM and UcoSLAM [49]. The OpenVSLAM framework could apply sequence images or videos collected by different types of cameras (monocular, stereo, and so on) to locate the current position of the camera in real time and reconstruct the surrounding environment in three-dimensional space. It has advantages over the ORB-SLAM2 system in the speed and performance of the algorithm, and can quickly locate the newly captured image based on the pre-built map. However, the OpenVSLAM system was also constructed on the basis of the static world assumption. Theoretically, the performance of our stereo SLAM system is more robust and stable than it in high dynamic scenarios. In addition, the DynaSLAM (stereo) is a representative visual SLAM applicable to dynamic scenes, which is used to compare the performance in high dynamic scenarios.

The metric of absolute pose error (APE) is used to measure the performance of a visual SLAM system. And the metric of relative pose error (RPE) is suitable for measuring the drift of a visual odometry. Therefore, the metrics APE and RPE were computed for the quantitative evaluation analysis. And the values of root-mean-square error (RMSE) and Standard Deviation (STD) can give evidence of the robustness and stability of the system. Therefore, they were chosen as evaluation indexes. Figures 4 and 5, respectively, show APEs and RPEs on the 11 sequences of KITTI dataset. Furthermore, compared with the original ORB-SLAM2 system, the RMSE of APE improved by up to 80%, and the STD metric improved by up to 78%. The RMSE of RPE improved by up to 50%, and the STD metric improved by up to 60%. It had a comparable performance to OpenVSLAM with respect to the tracking accuracy in most static sequences (00, 02, 03, 05–10). These fully prove that the proposed stereo SLAM system had much greater stability, accuracy and robustness no matter whether it was in a high dynamic environment or a low dynamic or static environment.

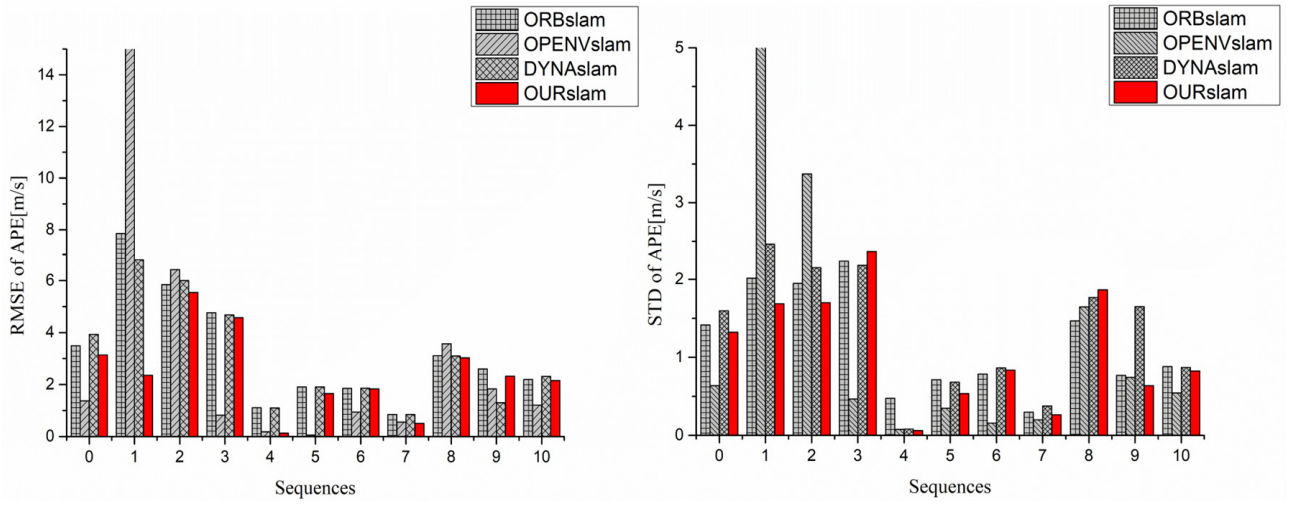


Figure 4. Absolute pose errors on the 11 sequences in the KITTI odometry dataset. Lower is better.

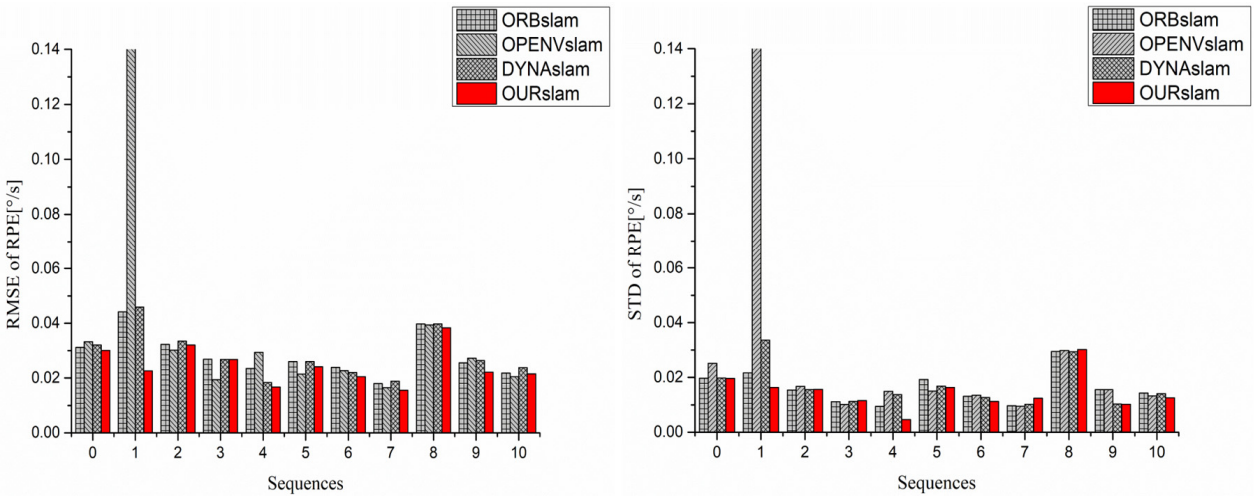
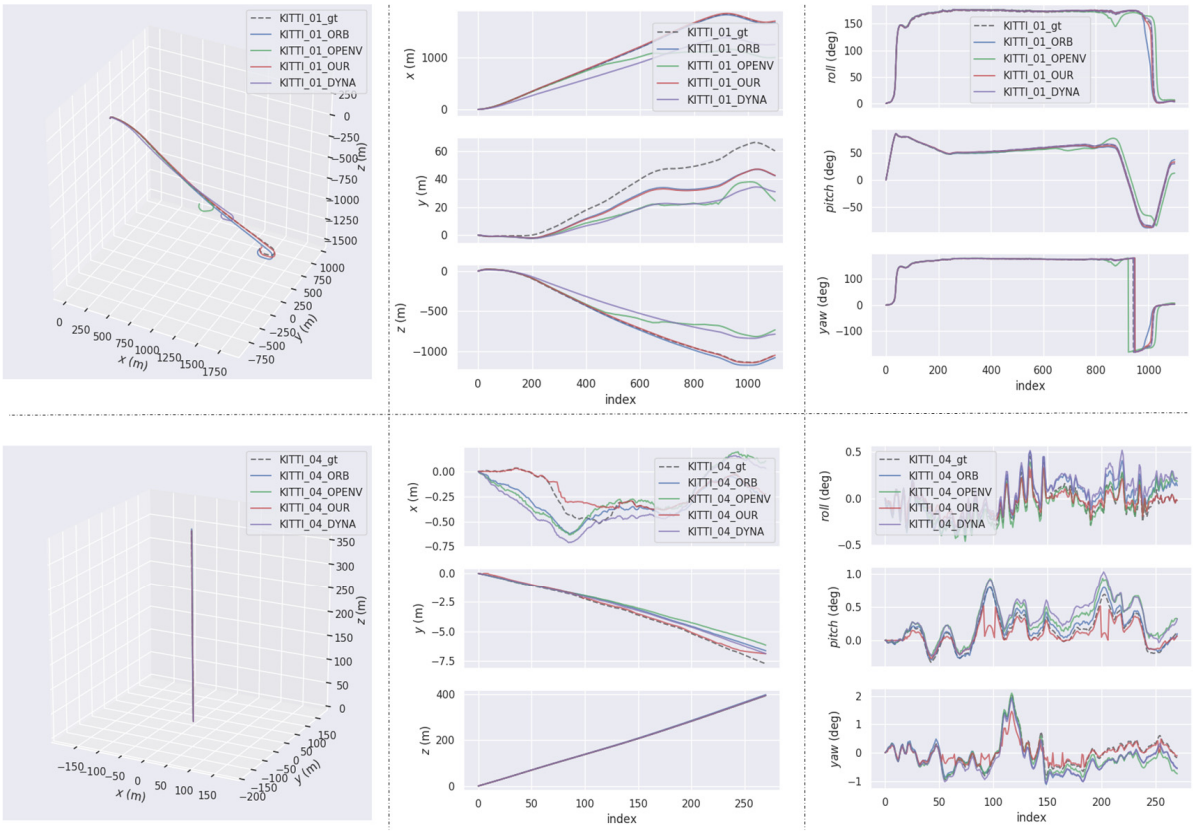


Figure 5. Relative pose errors on the 11 sequences in the KITTI odometry dataset. Lower is better.

The trajectories of four VSLAM systems on KITTI high dynamic sequences are described in Figure 6. It was found that our trajectories were the closest to the ground-truth trajectories in dynamic scenes, which clearly showed that our stereo SLAM system was the most accurate and robust among four vision-based SLAM systems in the high dynamic sequences of 01 and 04. In addition, compared with the trajectory of the 04 sequence, the 01 sequence was more intricate. It can be inferred from this that our SLAM system will be more stable and robust than others in more challenging dynamic environments.



(a) Trajectories in a XYZ Coordinate System (b) Trajectories in the x/y/z plane (c) Trajectories in the r/p/y orientation

Figure 6. The trajectories of four VSLAM systems on KITTI high dynamic sequences. The top row is the result of the 01 sequence and the bottom row is the result of the 04 sequence.

4.2. Time Analysis

In most practical applications of visual SLAM, such as intelligent robots and virtual and augmented reality on mobile devices, the real-time performance is crucial. Then, we tried out the tracking times of our SLAM with the 05 sequence of the KITTI benchmark dataset. The configurations of the test platform were as those introduced at the beginning. The DynaSLAM (stereo) is not compared here because it is not a real-time system. Table 1 shows the test results. It is obvious that the run time of our SLAM was numerically close to ORB-SLAM2, which was sufficiently short for practical applications, although it was a little slower than OpenVSLAM.

Table 1. The mean and median tracking times of three frameworks.

	ORB-SLAM2	**OpenVSLAM**	**OurSLAM**
Mean [ms/frame]	65.35	55.46	68.56
Median [ms/frame]	66.43	56.25	69.21

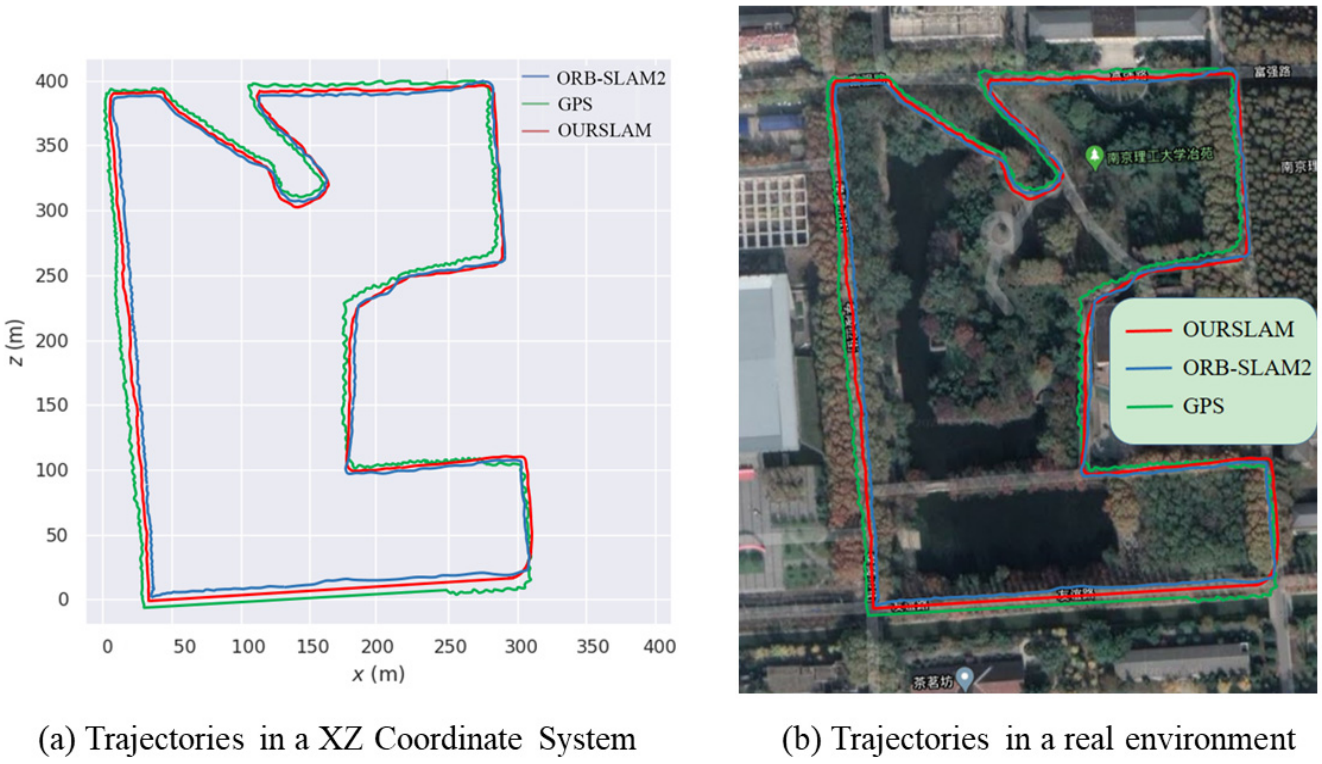
4.3. Evaluation Test in Real Environment

In order to illustrate the stability and availability of our system, we carried out a large-scale visual SLAM experiment in cluttered urban outdoor scenarios, with some independently moving objects; for example, pedestrians, riders, and cars. We took advantage of a TurtleBot3 robot to perform our SLAM in outdoor scenes. It had to come back to the same place of the route so as to close the loop and correct the accumulated drift during the SLAM process. Images with a resolution 320 $\times$ 240 captured with a ZED camera at 30 Hz were processed on an NVIDIA Jetson TX2 platform. The outdoor sequence was composed of 12,034 stereo pairs gathered in a public square of our city, and the experiment lasted about 15 min. The trajectory was about 978 m long from the initial position. The quantitative comparison results are shown in Table 2, and it turns out that the proposed stereo SLAM is more accurate and robust than ORB-SLAM2 in dynamic real city spaces.

Table 2. The ATE and RPE results of two SLAM systems.

Item	**ORB-SLAM2**		**OurSLAM**		**Improvements**	
	RMSE	**STD**	**RMSE**	**STD**	**RMSE**	**STD**
ATE [m/s]	20.387	10.021	4.632	2.987	77.28%	70.19%
RPE [°/s]	0.203	0.324	0.085	0.039	58.13%	87.96%

Figure 7a shows the trajectory performed by the robot platform in the city space scenario, considering our stereo SLAM algorithm (in red) and the ORB-SLAM2 system (in blue), as well as the estimated trajectory using a commercial GPS (in green). As can be observed, the ORB-SLAM2 could not acquire an exact camera trajectory on the major road of the map, since there were many dynamic objects (e.g., cars or riders) impairing its stability. In contrast, our stereo SLAM system can cope with moving objects befittingly, and the calculated trajectory is in correspondence with the real camera trajectory. And the trajectory estimated by GPS is close to the one acquired with our SLAM in most cases; however, there are some places in the sequence where the measurement errors are large. This is because these situations are the areas where there are many tall buildings or dense woods. In these areas, GPS is liable to fail because of low satellite visibility conditions. Figure 7b describes the same comparison, but displays results on an aerial image view of the sequence.



(a) Trajectories in a XZ Coordinate System (b) Trajectories in a real environment

Figure 7. Comparison of VSLAM and GPS estimated camera trajectories in urban dynamic environments. (**a**) Our SLAM, ORB-SLAM2, and GPS. (**b**) Aerial image view of the sequence.

5. Conclusions

A real-time stereo visual odometry method that can cope with dynamic scenes with some independent moving objects has been proposed in this paper. The semantic segmentation and moving object detection method was incorporated into the ORB-SLAM2 (stereo) system, which made some significant performance improvements in urban outdoor high dynamic scenarios. We were able to deal with dynamic objects and improve the performance of the visual odometry considerably, and consequently more robust and accurate localization and mapping results were obtained in crowded and high dynamic environments. In the end, we carried out experimental results to illustrate the improved accuracy of our proposed models and the efficiency and availability of our implementation.

There is still room for amendment. In our stereo SLAM system, the deep neural network utilized in the semantic segmentation thread was the supervised method. Therefore, the model may only just predict the right results when big differences turn up between training environments and actual scenarios. In the future, we could employ self-supervised or unsupervised deep learning means so as to deal with this.

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

Funding: This research was funded by the Military Scientific Research Project, grant number JK20211A020057.

Data Availability Statement: The data presented in this study are available in the article.

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

References

1. Davison, A.; Reid, I.; Molton, N.; Stasse, O. MonoSLAM: Realtime single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef] [PubMed]
2. Neira, J.; Davison, A.; Leonard, J. Guest editorial, special issue in visual slam. *IEEE Trans. Robot.* **2008**, *24*, 929–931. [CrossRef]
3. Klein, G.; Murray, D. Parallel tracking and mapping for small AR workspaces. In Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, Japan, 13–16 November 2007.
4. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. In Proceedings of the IEEE International Conference on Robotics and Automation, Hong Kong, China, 31 May–7 June 2014.
5. Engel, J.; Schops, 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*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 834–849.
6. 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]
7. Mur-Artal, R.; Tardos, J.D. ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
8. Campos, C.; Elvira, R.; Rodriguez, J.; Montiel, J.; Tardós, J. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [CrossRef]
9. Panchpor, A.A.; Shue, S.; Conrad, J.M. A survey of methods for mobile robot localization and mapping in dynamic indoor environments. In Proceedings of the 2018 Conference on Signal Processing and Communication Engineering Systems, Vijayawada, India, 4–5 January 2018.
10. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets Robotics: The KITTI Dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
11. Bescos, B.; Facil, J.M.; Civera, J.; Neira, J. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robot. Autom. Lett.* **2018**, *3*, 4076–4083. [CrossRef]
12. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147.
13. Davison, A. Mobile Robot Navigation Using Active Vision. Ph.D. Dissertation, University Oxford, Oxford, UK, 1998.
14. Davison, A.J.; Murray, D.W. Simultaneous localization and map building using active vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 865–880. [CrossRef]
15. Davison, A.; Kita, N. 3-D simultaneous localisation and map building using active vision for a robot moving on undulating terrain. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. 384–391.
16. Iocchi, L.; Konolige, K.; Bajracharya, M. Visually realistic mapping of a planar environment with stereo. In *International Symposium on Experimental Robotics*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 521–532.
17. Se, S.; Lowe, D.; Little, J. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *Int. J. Robot. Res.* **2002**, *21*, 735–758. [CrossRef]
18. Jung, I.; Lacroix, S. High resolution terrain mapping using low altitude aerial stereo imagery. In Proceedings of the 9th International Conference on Computer Vision, Nice, France, 14–17 October 2003; Volume 2, pp. 946–951.
19. Hygounenc, E.; Jung, I.; Soueres, P.; Lacroix, S. The autonomous blimp project of LAAS-CNRS: Achievements in flight control and terrain mapping. *Int. J. Robot. Res.* **2004**, *23*, 473–511. [CrossRef]
20. Saez, J.; Escolano, F.; Penalver, A. First Steps towards Stereobased 6DOF SLAM for the visually impaired. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition-Workshops, San Diego, CA, USA, 21–23 September 2005; Volume 3, p. 23.
21. Sim, R.; Elinas, P.; Griffin, M.; Little, J. Vision-based SLAM using the Rao–Blackwellised particle filter. In Proceedings of the International Joint Conference on Artificial Intelligence-Workshop Reason, Uncertainty Robot, Edinburgh, UK, 30 July–5 August 2005; pp. 9–16.
22. Sim, R.; Elinas, P.; Little, J. A study of the Rao–Blackwellised particle filter for efficient and accurate vision-based SLAM. *Int. J. Comput. Vis.* **2007**, *74*, 303–318. [CrossRef]
23. Paz, L.M.; PiniÉs, P.; TardÓs, J.D.; Neira, J. Large-Scale 6-DOF SLAM With Stereo-in-Hand. *IEEE Trans. Robot.* **2008**, *24*, 946–957. [CrossRef]
24. Lin, K.H.; Wang, C.C. Stereo-based simultaneous localization, mapping and moving object tracking. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010.
25. Kawewong, A.; Tongprasit, N.; Tangruamsub, S.; Hasegawa, O. Online and Incremental Appearance-based SLAM in Highly Dynamic Environments. *Int. J. Robot. Res.* **2011**, *30*, 33–55. [CrossRef]
26. Alcantarilla, P.F.; Yebes, J.J.; Almazán, J.; Bergasa, L.M. On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments. In Proceedings of the IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 1290–1297.
27. Kaess, M.; Ni, K.; Dellaert, F. Flow separation for fast and robust stereo odometry. In Proceedings of the IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 3539–3544.

28. Karlsson, N.; Di Bernardo, E.; Ostrowski, J.; Goncalves, L.; Pirjanian, P.; Munich, M.E. The vSLAM algorithm for robust localization and mapping. In Proceedings of the IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18–22 April 2005; pp. 24–29.
29. Zou, D.; Tan, P. CoSLAM: Collaborative Visual SLAM in Dynamic Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 354–366. [CrossRef]
30. Fan, Y.; Han, H.; Tang, Y.; Zhi, T. Dynamic objects elimination in SLAM based on image fusion. *Pattern Recognit. Lett.* **2019**, *127*, 191–201. [CrossRef]
31. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
32. Sun, T.; Sun, Y.; Liu, M.; Yeung, D.Y. Movable-Object-Aware Visual SLAM via Weakly Supervised Semantic Segmentation. *arXiv* **2019**, arXiv:1906.03629.
33. Balntas, V.; Riba, E.; Ponsa, D.; Mikolajczyk, K. Learning local feature descriptors with triplets and shallow convolutional neural networks. *Br. Mach. Vis. Conf.* **2016**, *1*, 3.
34. Kang, R.; Shi, J.; Li, X.; Liu, Y. DF-SLAM: A Deep-Learning Enhanced Visual SLAM System based on Deep Local Features. *arXiv* **2019**, arXiv:1901.07223.
35. Li, P.; Qin, T.; Shen, S. Stereo Vision-based Semantic 3D Object and Ego-motion Tracking for Autonomous Driving. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
36. Ai, Y.; Rui, T.; Lu, M.; Fu, L.; Liu, S.; Wang, S. DDL-SLAM: A robust RGB-D SLAM in dynamic environments combined with Deep Learning. *IEEE Access* **2020**, *8*, 162335–162342. [CrossRef]
37. Chiranjeevi, P.; Sengupta, S. Moving object detection in the presence of dynamic backgrounds using intensity and textural features. *J. Electron. Imaging* **2011**, *20*, 3009. [CrossRef]
38. Wang, Y.T.; Chi, C.T.; Hung, S.K. Robot Visual Simultaneous Localization and Mapping in Dynamic Environments. *J. Comput. Theor. Nanosci.* **2012**, *8*, 229–234. [CrossRef]
39. Zeng, Z.; Zhou, Y.; Odest, C.J.; Karthik, D. Semantic Mapping with Simultaneous Object Detection and Localization. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Madrid, Spain, 1–5 October 2018; pp. 911–918.
40. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
41. Rusu, R.B.; Blodow, N.; Beetz, M. Fast point feature histograms (fpfh) for 3D registration. In Proceedings of the Robotics and Automation, IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 3212–3217.
42. Hu, J.; Fang, H.; Yang, Q.; Zha, W. MOD-SLAM: Visual SLAM with Moving Object Detection in Dynamic Environments. In Proceedings of the 40th Chinese Control Conference, Shanghai, China, 26–28 July 2021; pp. 4302–4307.
43. Lucas, B.D.; Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, Aichi, Japan, 23–29 August 1997.
44. Xie, W.; Liu, P.X.; Zheng, M. Moving Object Segmentation and Detection for Robust RGBD-SLAM in Dynamic Environments. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5001008. [CrossRef]
45. Yu, C.; Liu, Z.X.; Liu, X.J.; Xie, F.G.; Yang, Y.; Wei, Q.; Qiao, F. DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, Madrid, Spain, 1–5 October 2018.
46. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
47. Zhao, Y.; Shi, H.; Chen, X.; Li, X.; Wang, C. An overview of object detection and tracking. In Proceedings of the 2015 IEEE International Conference on Information and Automation, Lijiang, China, 8–10 August 2015.
48. Shi, J.; Tomasi, C. Good Features to Track. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hilton Head, SC, USA, 13–15 June 2000; p. 600.
49. Sumikura, S.; Shibuya, M.; Sakurada, K. OpenVSLAM: A Versatile Visual SLAM Framework. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2292–2295.