remote sensing

MDPI

Article

SVG-Loop: Semantic–Visual–Geometric Information-Based Loop Closure Detection

Zhian Yuan [1], Ke Xu [1], Xiaoyu Zhou [1], Bin Deng [1] and Yanxin Ma [2,3,*]

[1] College of Electronic Science, National University of Defense Technology, Changsha 410073, China; yuanzhian@nudt.edu.cn (Z.Y.); xuke@nudt.edu.cn (K.X.); zhouxiaoyu512@nudt.edu.cn (X.Z.); dengbin@nudt.edu.cn (B.D.)
[2] College of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073, China
[3] Hunan Key Laboratory for Marine Detection Technology, Changsha 410073, China
* Correspondence: mayanxin@nudt.edu.cn

Abstract: Loop closure detection is an important component of visual simultaneous localization and mapping (SLAM). However, most existing loop closure detection methods are vulnerable to complex environments and use limited information from images. As higher-level image information and multi-information fusion can improve the robustness of place recognition, a semantic–visual–geometric information-based loop closure detection algorithm (SVG-Loop) is proposed in this paper. In detail, to reduce the interference of dynamic features, a semantic bag-of-words model was firstly constructed by connecting visual features with semantic labels. Secondly, in order to improve detection robustness in different scenes, a semantic landmark vector model was designed by encoding the geometric relationship of the semantic graph. Finally, semantic, visual, and geometric information was integrated by fuse calculation of the two modules. Compared with art-of-the-state methods, experiments on the TUM RBG-D dataset, KITTI odometry dataset, and practical environment show that SVG-Loop has advantages in complex environments with varying light, changeable weather, and dynamic interference.

Keywords: loop closure detection; bag of words; panoptic segmentation; visual simultaneous localization and mapping

Citation: Yuan, Z.; Xu, K.; Zhou, X.; Deng, B.; Ma, Y. SVG-Loop: Semantic–Visual–Geometric Information-Based Loop Closure Detection. *Remote Sens.* **2021**, *13*, 3520. https://doi.org/10.3390/rs13173520

Academic Editors: Kai-Wei Chiang, Chenglu Wen, Li-Ta Hsu and Andrea Masiero

Received: 3 August 2021
Accepted: 3 September 2021
Published: 5 September 2021

1. Introduction

Simultaneous localization and mapping (SLAM) [1] is the key technology of automatic navigation and has become a research hotspot in the past decade. SLAM mainly focuses on the problem of robots or vehicles positioning themselves and building maps when they enter an unfamiliar environment [2]. A camera can obtain abundant information as a low-cost, small-scale, and convenient human-computer interaction sensor. Thus, camera-based visual SLAM technology has attracted increasing attention. Loop closure detection is an important module in visual SLAM [3]. Loop closure detection can be employed to judge whether robots return to a previous position and help to eliminate the cumulative error in a front-end odometer [4]. Currently, achieving accurate and efficient loop detection is a challenging problem.

Traditional loop detection algorithms rely on visual appearance. These methods extract various features [5–7] to compare similarities of images. Visual-based algorithms can work effectively in various environments and have become mainstream methods of visual SLAM systems. For example, FAB-MAP [8] utilized Speeded-Up Robust Features (SURF) to build a visual bag-of-words model. Adrien et al. [9] used Scale Invariant Feature Transform (SIFT) to detect the loop. Gálvez-López and Tardós [10] proposed a visual place recognition method using the From Accelerated Segment Test (FAST) keypoint detector and Binary Robust Independent Elementary Features (BRIEF) descriptor. The Oriented FAST

and Rotated BRIEF (ORB) feature were employed to complete the loop closure detection in ORB-SLAM [11–13]. Most of these methods utilize the bag-of-words model [14], which generates words from feature points to build a dictionary structure and then queries the similarity of words in each image to make a loop judgment. These mentioned methods are computationally efficient but only suitable for processing static scenes with limited scene changes. Moreover, these methods make full use of low-level visual features but ignore geometric and structural information. Finally, tiny details can be effectively identified, but macro characteristics are difficult to grasp.

In the task of place recognition in different weather and seasons, sequence-based methods have significant results. Instead of calculating single frames, SeqSLAM [15] acquires candidates with coherent image sequences to deal with perceptual change. Local best matches are found by contrast enhancement and localized template matching. Sayem et al. [16] designed a Fast-SeqSLAM system which narrowed the search scope to the initial match images and used motion continuity to complete an extended search. In DOSeqSLAM [17], fixed-size length of sequences are replaced by dynamic lengths which are related to image similarity. This improvement allows the system to reduce the computational cost of operation and run on-line. Based on the visual features, Konstantinos et al. [18] proposed the concept of Tracked Words, which are generated through feature tracking and matching in the sequence. The distance of Tracked Words is leveraged to set up a voting mechanism for loop closure detection. Renata et al. [19] employed similarity between intervals in image and temporal constraints to find loop closures. The above methods are more robust for resolving the problem of light changes and weather transformation. However, these methods are not sensitive enough to ensure loop closure and are only effective in continuous image sequences.

The development of neural networks in computer vision fields has led to their use in the extraction of high-level features and application to loop detection. Chen et al. [20] leveraged convolutional neural networks (CNN) to extract landmarks as image features. These features were employed to achieve localization and place recognition in different seasons. Wang et al. [21] proposed a robust loop detection algorithm that combined semantic segmentation and a CNN. Semantic segmentation provided semantic topological graphs and landmark regions. The CNN was used to acquire features of landmark regions. Similarity scores of topological graphs and CNN features were calculated to complete loop closure detection. Finman et al. [22] extracted the objects in a dense map and connected them in a sparse map to complete location recognition faster. In [23], graph information formed by the relative positions of landmarks was recorded in an adjacency matrix. Word adjacency matrixes of scenes were leveraged for position matching. Abel et al. [24] proposed an X-View global localization system that employs semantic topology graph descriptor matching to complete place recognition and global localization. The algorithm in [25] relied on visual landmarks extracted by a pre-trained CNN. To solve the issue of viewpoint jitter, an incremental covisibility graph was built to improve the robustness of the system. In [26], a CNN architecture named NetVLAD was developed for place recognition. Furthermore, Patch-NetVLAD, combining both local and global descriptors, was proposed to complete loop closure detection in [27]. Most of these methods have achieved great results. However, the neural-network-based method ignored local details in the process of extracting image features, so scenes with missing semantic and texture information were difficult to process.

Inspired by the above modules, a loop closure detection algorithm based on a semantic bag of words and a semantic landmark vector is proposed in this paper. The proposed method named SVG-Loop integrates semantic–visual–geometric information. Visual information can mine the details of images and increase the computational speed. Geometric and semantic information, being more advanced image content, can improve the robustness of the proposed algorithm in complex environments. SVG-Loop mainly consists of two parts: a semantic bag-of-words model and semantic landmark vector calculation. The semantic bag-of-words model combines visual information and semantic information to perform quick pixel-level matching. The semantic landmark vector is obtained to complete

instance-level matching, which can combine geometric and semantic information. Experiments on public databases TUM RGB-D [28] and KITTI odometry [29] were used to verify the effectiveness of the proposed algorithm. In order to further explore the generalization and robustness of the SVG-Loop algorithm, further extended experiments were conducted on real indoor and outdoor scenes.

In short, the main contributions of this work are as follows:

(1) A semantic bag-of-words model was constructed to reduce the interference caused by dynamic objects and improve the accuracy of image matching.
(2) A semantic landmark vector was designed that can express semantic and geometric information of images and improve the robustness of loop closure detection.
(3) A semantic–visual–geometric information-based loop detection algorithm SVG-Loop is proposed to improve robustness in complex environments.

The remainder of this paper is organized as follows: Section 2 introduces the related work; Section 3 provides the general pipeline and detailed description of SVG-Loop. Section 4 details the process of experiments and comparative analysis of experimental results. Section 5 discusses the results and future research direction. Finally, the conclusion is presented in Section 6.

2. Related Work

Loop closure detection is an important task in the field of monocular SLAM, and it has also been a hot research topic in recent years. References [30–32] give a variety of loop detection and location recognition algorithms. These methods have their advantages in efficiency, accuracy, and generalization. Two types of loop closure detection methods closely related to our work are based on bag-of-words and semantic information approaches.

2.1. Bag-of-Words Model-Based Loop Closure Detection

At present, the vision-based method is a mainstream method of loop closure detection. One of the classic algorithms is FAB-MAP [8], which uses the Chow Liu tree structure to construct a bag of visual words. FAB-MAP achieved excellent results and once became the baseline for loop detection algorithms. Reference [33] provided a fully open-source implementation of FAB-MAP. DBoW2 [10] is another representative algorithm that integrates the BRIEF descriptor and FAST features to speed up the calculation. DBoW2 chooses a k-d tree structure to build a bag-of-words model, which is used for image similarity queries. Reference [34] further improved the efficiency based on DBoW2 by using the ORB descriptor, which retained the scale invariance and rotation invariance. Incremental bag of binary words for appearance-based loop closure detection (IBuILD) was proposed in [35]. An online, incremental formulation of a binary vocabulary was derived for place recognition in IBuILD. In [36], dynamic segmentation of image stream and an online visual words clustering algorithm were designed to define the target place. Then, a nearest neighbor voting scheme was used to select suitable candidates. Reference [37] used the words in the image sequence to generate a sequence visual-word vector containing image-to-image association for matching. Tsintotas [38] proposed a sequence-based loop closure detection approach fusing the SeqSLAM and Bag-of -Words methods. On this basis, an incremental bag-of-tracked-words model was proposed in [39]. Then, nearest-neighbor voting was leveraged to acquire the probabilistic scores to positions in sequence. Bag-of-words model-based loop closure detection is computationally efficient but cannot adequately handle a dynamic environment.

2.2. Semantic-Information-Based Loop Closure Detection

Semantic-information-based loop closure detection leverages semantic segmentation and an instance segmentation network. DS-SLAM [40] and Segnet [41] were employed to segment dynamic objects. DS-SLAM removed the feature points located in the area of the dynamic scene to alleviate dynamic interference in loop detection. In [42], a Mask region-based convolutional neural network (R-CNN) [43] was designed as an independent

thread for instance segmentation. The segmentation results were shared in tracking, local-mapping, and loop closure detection. The authors of [44] utilized a deep CNN to perform semantic segmentation and extract loop feature of images at the same time. In [45], an unsupervised semi-semantic auto-encoder model DeepLab_AE was designed to obtain semantic features of scenes. The authors of [46] proposed a place recognition method via re-identification of salient objects. Oh et al. [47] performed loop detection by matching graphs of detected objects. Foreground information was mainly employed in this algorithm. The X-View [24] global localization system was used to design a semantic topology graph descriptor. The graph structure was stored in node descriptors and applied to complete place recognition and global localization. A semantic loop closure detection (SLCD) approach was designed to reduce the issue of semantic inconsistency in [48]. This method fused instance-level semantic information obtained from object detection and feature information. In [21], a robust loop closure detection algorithm integrated semantic segmentation and a CNN. The CNN was used to extract landmark information divided by a segmentation network. The semantic topological graph and CNN features were jointly used as the basis of loop closure detection judgment. These methods all attempt to combine the semantic and higher-level information to improve the performance of loop closure detection and SLAM systems. However, the extraction and utilization of semantic information are insufficient. This paper proposes the SVG-Loop algorithm, which leverages the latest panoptic segmentation. In SVG-Loop, foreground and background information is effectively utilized to improve the accuracy and robustness of loop closure detection.

3. Methodology

The pipeline of the proposed SVG-Loop method is shown in Figure 1. SVG-Loop is mainly composed of two modules: semantic bag-of-words model, and semantic landmark vector model. In this section, the construction of the semantic bag-of-words model is introduced firstly. Then, a detailed description of the semantic landmark vector is provided. Finally, a fusion calculation method of the two models is presented.

3.1. Semantic Bag-of-Words Model

Bag of words is a model that transforms an image into a sparse vector, which can be used to calculate the similarity of images by leveraging feature vocabulary. The process of semantic bag-of-words construction includes two steps: semantic–visual word extraction and vocabulary construction.

3.1.1. Semantic–Visual Word Extraction

As shown in Figure 2, each input image is processed by two parts: feature extraction and panoptic segmentation. Corner points with rich visual information of objects are determined by feature extraction. Semantic information in the instance level is acquired by panoptic segmentation. Visual features and corresponding semantic labels are acquired at the same time.

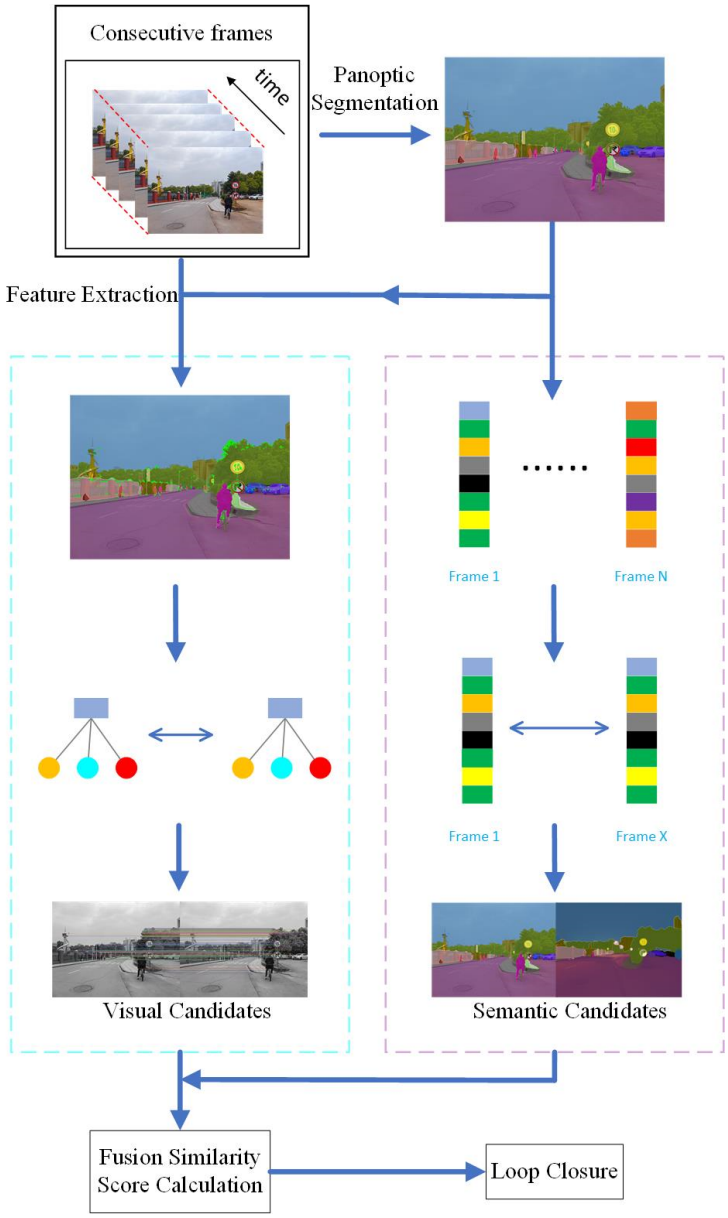


Figure 1. The pipeline of the proposed SVG-Loop. The blue box on the left is the semantic bag-of-words model and the purple box on the right is the semantic landmark vector model.

The extracted feature, which is the visual part of the semantic–visual words, consists of the keypoint and descriptor. To maintain the rotation invariance and scale invariance, the ORB feature [7] is selected in the feature extraction. When viewpoints change, the place recognizer can still work efficiently. The OpenCV library is employed to complete ORB feature extraction and descriptor generation. The quantity of features influences the performance of scene similarity detection: too small a quantity affects the distance matching and too large a quantity reduces the accuracy of keypoints. After the experimental test, 1000 features are chosen from each image for distance matching. Furthermore, although the matching probability is slightly decreased, non-maximum suppression is employed to improve the distribution situation of features.

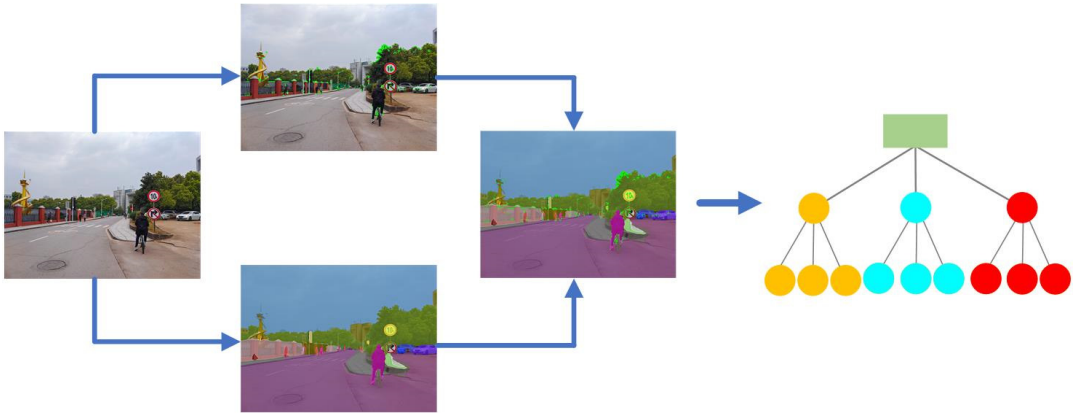


Figure 2. Extraction process of semantic–visual words. Corners are extracted by feature extraction and semantic labels are acquired by panoptic segmentation. Similar descriptors with the same semantic labels are classified as semantic–visual words through clustering.

Panoptic segmentation [49] incorporates semantic segmentation and object detection. Foreground and background information is obtained by panoptic segmentation at the same time. Panoptic feature pyramid networks (Panoptic-FPN) [43] were applied to complete panoptic segmentation in this study. Panoptic-FPN combine a Mask R-CNN and feature pyramid networks (FPN) [50] to achieve pixel-wise semantic segmentation prediction.

After feature extraction and panoptic segmentation, semantic and visual information are combined. Each feature is connected with the corresponding semantic label to generate semantic–visual words. Descriptor vectors and semantic labels of keypoints are stored in words. Because the descriptor consists of a binary vector, the distance between two descriptors can be expressed by Hamming distance, which includes the xor operation of bits.

3.1.2. Vocabulary Construction

The vocabulary structure is constructed offline with a large image dataset that includes a semantic marker. In the semantic bag-of-words model, a k-d tree structure [51] is employed to store the vocabulary space. As shown in Figure 3, the tree structure includes two layers: the semantic layer and the feature layer. In the semantic layer, labels of objects are divided into two categories: dynamic objects and static objects. Static objects in the sensor field of view are excellent landmarks to the loop closure detection algorithms. On the contrary, dynamic objects may become interference factors for place recognition. In order to decrease the undesirable effects of dynamic objects, features of dynamic objects are discarded in the process of semantic verification. To build vocabulary, a series of features with semantic labels were extracted from the image dataset and constitutes the tree structure (Figure 3). Weight W_j^i for semantic–visual words is designed in leaf nodes. According to the term frequency-inverse document frequency (TF-IDF) [14], W_j^i is defined by:

$$W_j^i = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \log\left(\frac{N_{all}}{N_i + 1}\right) \tag{1}$$

where $n_{i,j}$ denotes the number of features i in the image j, N_{all} represents the number of images in training datasets, and N_i indicates the number of images that include the term i. Words with higher scores can provide more effective matching information. In particular, because of the high frequency of human movement, the weights of words belonging to the person node are set to 0.

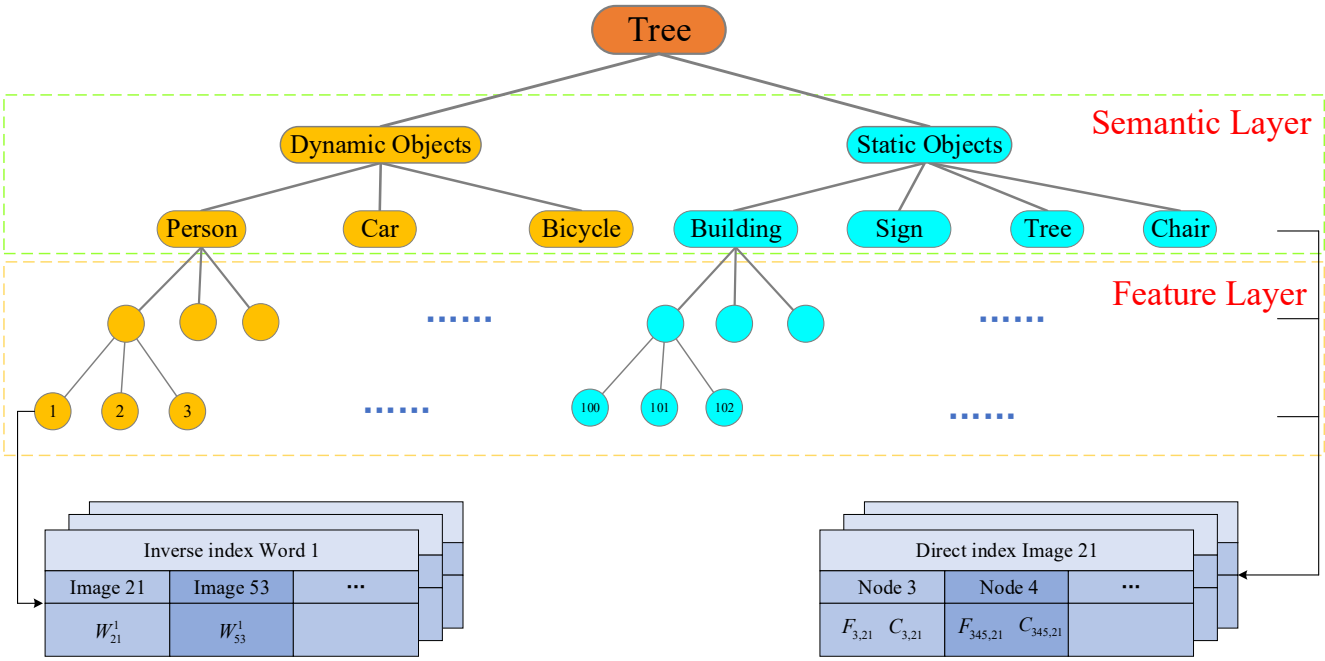


Figure 3. Example of vocabulary structure and index information. Tree structure mainly includes the semantic layer and feature layer. The inverse index records weight of the words in images containing them. The direct index stores features of the images and semantic labels of associated nodes at a different level of the vocabulary tree.

Along with the vocabulary tree, two kinds of index information are recorded. On the one hand, the direct index stores features of images and semantic labels of connected nodes according to levels. On the other hand, the inverse index stores image IDs associated with the word node and weight of relevance. When new images enter the dataset, both indexes are updated. The direct index of images and inverse index of words accelerate the efficiency of image matching and loop closure verification.

3.1.3. Visual Loop Closure Candidate Detection

After constructing the model, the vocabulary tree is applied to detect the loop closure in image sequences. Image I_j of the input data is transformed into a bag-of-words vector $V_j \in \mathbb{R}^N$. Features within I_j are grouped into word nodes through traversing the vocabulary tree. From the root to the leaves, the feature matches the associated node, which has a minimum Hamming distance at each level.

The similarity between two bag-of-words vectors V_m and V_n is calculated as L_1-score $s(m, n)$:

$$s(m,n) = 1 - \frac{1}{2}\left|\frac{V_m}{|V_m|} - \frac{V_n}{|V_n|}\right| \tag{2}$$

where $s(m, n)$ is normalized to compute the similarity between different images. After illustrating the measurement of similarity, the inverse index is used to accelerate loop closure detection. Image I_j containing word i and the weight W_j^i of word node i is stored in the inverse index structure. When loop closure is detected, only images with part of the same words are computed for similarity. By selecting image pairs according to the inverse index, unnecessary calculations can be reduced.

To prevent adjacent frames from being misjudged as loop closure, a sliding window is set in the process of detection. The current image I_j is the central node of a window whose width is 30. Frames in the window are excluded as loop closure. The maximum similarity score $s_{\max}$ in the window between I_j and the other frames remains. Under the assumption that the camera's movement speed is not too fast, the similarity between adjacent frames in the sliding window is high. Thus, images with similarity scores s_c of I_j more than $s_{\max}$ or without an acquired number of common semantic visual words are discarded. When s_c is

less than the preset threshold α and at the local minimum, the corresponding frame I_c is selected as a visual candidate for I_j and accepted to the verification stage.

In the verification phase, temporal consistency verification [10] and geometric verification [34] are common choices for filtering loop closure candidates. The prerequisite of temporal consistency (the image sequence has a certain loop scene overlap) is strict and does not match the real situation. Thus, the traditional time constraint [52] used to limit loop closure candidates was discarded in this work. However, to ensure the credibility of loop closure, semantic verification to screen loop closure candidates is proposed. In the semantic layer of the vocabulary tree, the similarity of static object nodes of two images must reach 80% before they are passed to the semantic verification. Furthermore, dynamic object nodes are removed to reduce the dynamic interference of different scenes in this step. Geometric verification is completed by computing a fundamental matrix, which is supported by at least 12 correspondences between the matched frames with the RANSAC scheme. In order to speed up feature matching, calculating approximate nearest neighbors in the tree structure replaces the exhaustive comparison method. A direct index is leveraged to compute correspondence points more efficiently. Correspondences are searched from the middle layer of the tree to balance calculation speed and recall rate. At level L, features of the same nodes and semantic labels in the direct index are approximated as nearest neighbors. Correspondences are computed from the nearest neighbor features.

Finally, loop closure candidates that pass semantic and geometric verification are output as visual loop closure candidates I_v. Visual loop closure candidates I_v and their similarity scores S_v enter the process of fusion calculation.

3.2. Semantic Landmark Vector Model

To make full use of semantic and geometric information of images, a semantic landmark vector is designed from the semantic graph of each frame. As shown in Figure 4, firstly, foreground and background information of the image sequence is acquired by the Panoptic-FPN. Next, the semantic landmark vector of each image is generated from the semantic topologic graph. Then, the similarity of different vectors is measured, and similar vectors corresponding to frames are captured. Finally, geometric verification of matched frames is completed, and semantic loop closure candidates are selected.

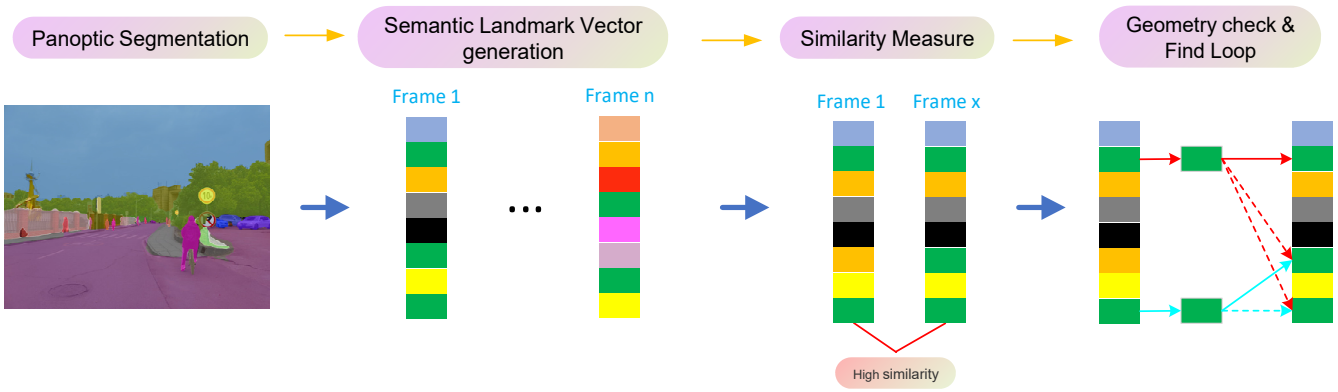
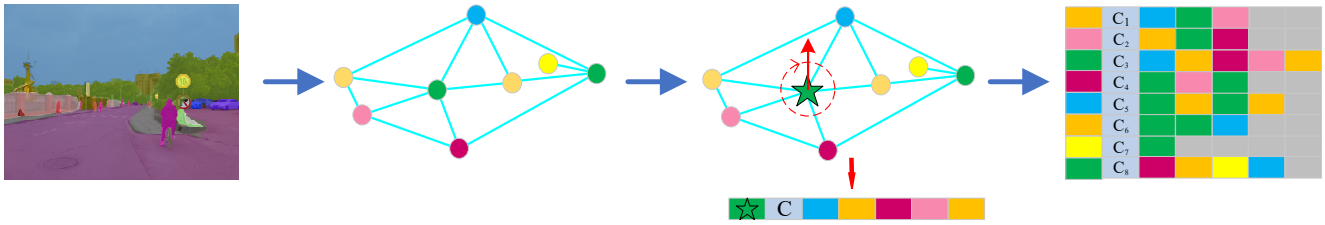


Figure 4. Process of the semantic landmark vector model.

3.2.1. Semantic Descriptor Generation

A semantic descriptor incorporating semantic and geometric information is employed to store the content of the node in the graph. In the process of panoptic segmentation, static and dynamic objects are segmented at the instance level (see Figure 5a). To mitigate the uncertain interference of dynamic objects, dynamic nodes are excluded when constructing the graph (see Figure 5b). As shown in Figure 5c, information of each node is recorded by the semantic descriptor ζ_j^i. The label of the node, the centroid of the node, and labels of connected nodes according to spatial geometric distribution are contained in ζ_j^i. The label

of a node is used to match similar nodes in different graphs. The centroid of the node is leveraged to find the nearest neighbor nodes in the same class. Class labels of connected nodes are stored in a clockwise direction starting from the top, so the spatial relationship between this node and adjacent nodes can be recorded in a consistent order. Finally, all semantic descriptors of nodes in the graph are assembled into semantic landmark vector V_j of image j (see Figure 5d). Length of descriptor ζ_j^i in the image depends on the maximum number of nodes connected, and length of vector V_j is determined by the number of nodes in the semantic topological graph. In descriptors, the missing connected nodes are set to 0.



(**a**) (**b**) (**c**) (**d**)

Figure 5. Generation of semantic descriptor and semantic landmark vector: (**a**) result of panoptic segmentation, (**b**) semantic graph, (**c**) semantic descriptor generation, and (**d**) semantic landmark vector of image.

Compared with existing descriptors expressed by the random walk method [16,19,49], the semantic descriptor proposed in this work eliminates the cumulative error caused by random selection. Moreover, cumulative error increases as the length of random walk increases in the above algorithms. To enrich the node expression, centroids of nodes are added into descriptors. Furthermore, centroid distance can be applied to limit the node drift error caused by segmentation error.

3.2.2. Semantic Loop Closure Candidate Detection

Each image I_j is converted to a semantic topologic graph G_j and acquires the semantic landmark vector V_j. To reduce computation cost and limit the drifting error, a graph distance D_s was designed and calculated as follows:

$$D_s(j,k) = \min \sum_{n=1}^{N} \sum_{m=1}^{M} \left\| C_{j,n}^m - C_{k,n}^m \right\|_2 \tag{3}$$

where $C_{j,n}^m$ and $C_{k,n}^m$ are the centroids of instances in different semantic topologic graphs G_j and G_k, M is the number of objects with the same label, and N is the number of common categories in image pair. The similarity score between semantic landmark vector V_j and V_k only needs to be calculated if $D_s(i,k)$ is less than distance threshold D_{th},. Otherwise, I_k is not regarded as a candidate for I_j.

In the iterative calculation process of $D_s(j,k)$, correspondence between nodes in graphs G_j and G_k is determined. Instead of computing a similarity score S_s of vectors using the sequential method, descriptor similarity is calculated according to the nearest neighbor principle (see Figure 4). Huge errors caused by descriptor alignment drift can be avoided with this option. The similarity measure S_s is computed by Equation (4):

$$S_s(V_j, V_k) = \frac{1}{D_s(j,k)} \cdot \frac{\sum \phi(\zeta_j^i, \zeta_k^i)}{\sum \phi(\zeta_j^i, \zeta_j^i)} \tag{4}$$

where the function $\phi(\zeta_j^i, \zeta_k^i)$ is used to determine the same elements between descriptor ζ_j^i and ζ_k^i. $S_s(V_j, V_k)$ is normalized to [0,1] and can be used to acquire semantic candidates of I_j. As with the semantic bag-of-words model, the sliding window is also set in this

part to reduce the interference of contiguous frames. Firstly, the images in the window are removed from the semantic candidates. The next step is discarding the images with a similarity score less than the threshold S_{th}. Then, because the adjacent frames may have similar graph structures, only the images that have the local minimum of distance D_s with I_j are selected as semantic loop candidates I_s.

3.3. Fuse Calculation

After obtaining both visual loop closure candidate I_v and semantic loop closure candidates I_s, I_v, and I_s are calculated jointly to detect real loop closure. Visual similarity score S_v and semantic similarity score S_s of I_v and I_s are leveraged to calculate the loop closure similarity score S according to Equation (5):

$$S = \frac{1}{2}\left(\tan(\frac{\pi}{2}S_v) + \tan(\frac{\pi}{2}S_s)\right) \tag{5}$$

If the similarity score S of the candidate reaches the expected value α, it is judged as a real loop closure. In various environments, semantic bag-of-words and semantic landmark vector models have different sensitivity to the scenes. When one of the models deviates greatly from the ground truth in some scenarios, information of the other model needs to be enhanced by a large margin. The tangent function in Equation (5) is leveraged to non-linearly amplify the dominant model. In the hypothesis, candidates from two models with higher similarity scores provide more loop closure credibility.

4. Experimental Results and Analysis

To carry out a comprehensive and accurate assessment of the system, the SVG-Loop approach was evaluated on two different public datasets and practical environments. The TUM RGB-D dataset [28] was selected as an indoor dataset, and the KITTI odometry dataset [28] was chosen as an outdoor dataset. In addition, the experiments in practical scenes were also conducted in indoor and outdoor environments.

The experiments using the proposed approach were implemented with a desktop computer equipped with an AMD Ryzen7 4800H CPU running at 2.90 GHz and a GTX 1650Ti GPU. A pre-trained model of the Panoptic-FPN [43] based on Pytorch [53] was employed to complete panoptic segmentation.

4.1. Dataset Experiments

4.1.1. Indoor Dataset

The TUM RGB-D dataset [28], which includes 39 sequences of offices, was selected as the indoor dataset to test the SVG-Loop algorithm. The sensor of this dataset is a handheld Kinect RGB-D camera with a resolution of 640 × 480. Only RGB images in sequences were applied to verify different methods. The images contain a slight jitter of viewpoint caused by unsteady hands. In addition, the dynamic objects of the dataset are mainly people. There are two characteristics of the loop closure in the TUM RGB-D dataset caused by motions of the unconstrained 6DOF: sparse and short overlap time. This database was used to test the sensitivity of the SVG-Loop algorithm to loop closure and the ability of the proposed method to capture loop closure quickly.

In order to directly demonstrate the results of loop closure detection methods, SVG-Loop was transplanted to ORB-SLAM3 [13], which is one of the most popular visual SLAM systems. Keyframes of the front-end visual odometer were used as input for the SVG-Loop model. Detected loop closure was merged into the process of global optimization. Then, the trajectory graph of different methods was used to visually display the loop closure detection results. Two monocular visual SLAM systems, LDSO [54] and ORB-SLAM3 [13], that include loop closure detection modules were selected to complete comparative experiments with ORB-SLAM3 + SVG-Loop. Furthermore, trajectory error was computed to evaluate the impact of the SVG-Loop algorithm on SLAM system accuracy. Absolute pose error

(APE) was employed to measure the accuracy of the trajectory of the SLAM systems. The APE of frame i is calculated by Equation (6):

$$E_i = \log\left(T_{gt,i}^{-1} T_{esti,i}\right)^{\vee} \tag{6}$$

where $T_{gt,i}$ is the trajectory of ground truth and $T_{esti,i}$ is the trajectory of estimation. After obtaining the APE, the root mean squared error (RMSE) over all frames is computed by Equation (7):

$$RMSE(E_{1:n}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|E_i\|_2^2} \tag{7}$$

According to Figure 6, there is only one loop closure in the trajectory of the fr2_desk sequence. Among LDSO, ORB-SLAM3, and ORB-SLAM3 + SVG-Loop methods, only ORB-SLAM3 + SVG-Loop detected the loop closure. Similarly, loop closure in the fr3_long_office sequence was only detected successfully by SVG-Loop (see Figure 7).

To quantitatively analyze the effect of loop closure detection, APE is calculated and displayed in a grid graph. According to Figure 8, the RMSEs of LDSO, ORB-SLAM3, and ORB-SLAM3 + SVG-Loop are 1.09064, 0.010287, and 0.008958 m, respectively. The accuracy of the ORB-SLAM3 system is improved by 12.9% due to the superimposed SVG-Loop algorithm. In the fr3_long_office sequence (Figure 9), the RMSEs of LDSO, ORB-SLAM3, and ORB-SLAM3 + SVG-Loop are 0.385381, 0.012877, and 0.010924 m, respectively. The accuracy of the ORB-SLAM3 system with SVG-Loop increased by 15.1%. Thus, experimental results indicate that the SVG-Loop technique improves the accuracy of the ORB-SLAM3 system efficiently.

Experiments using the TUM RGB-D dataset illustrate that SVG-Loop has advantages in the face of loop closure with a short overlap time. Compared with the other two loop closure detection modules, SVG-Loop uses semantic verification to replace the traditional temporal verification. Loop closure can therefore be captured quickly. Furthermore, SVG-Loop can be adapted to the SLAM system and improve its accuracy. However, fast loop closure decision speed may bring the risk of reduced accuracy. Thus, the precision and recall rate of detection methods was verified on the outdoor KITTI odometry dataset with multiple loop closure in the next step.

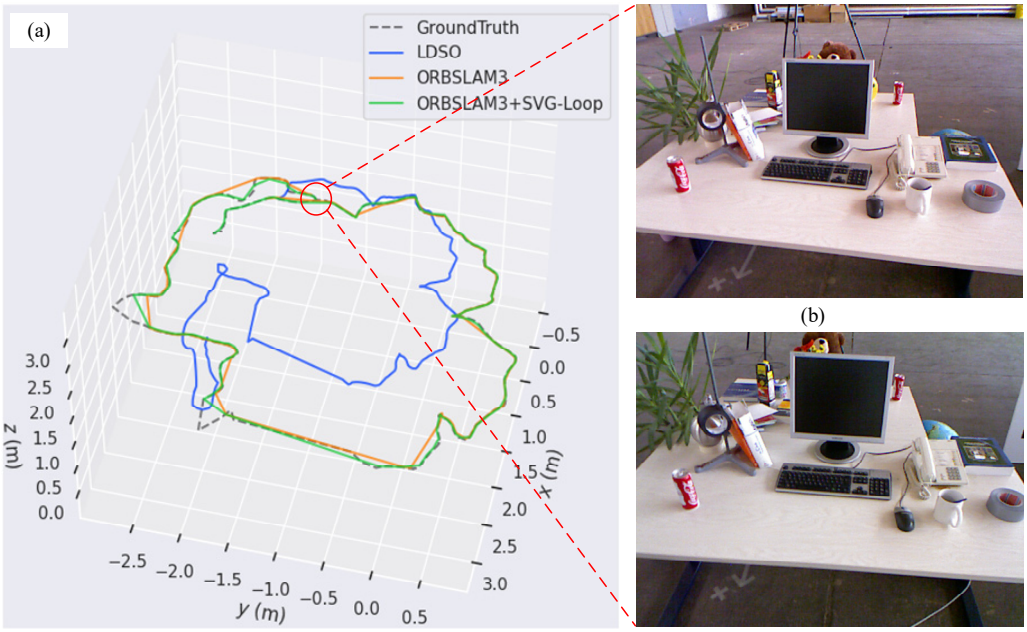


Figure 6. Experimental result for fr2_desk sequence. (**a**) The trajectory graph of LDSO, ORB-SLAM3, and ORB-SLAM3 + SVG-Loop. (**b**) Images of loop closure scene.

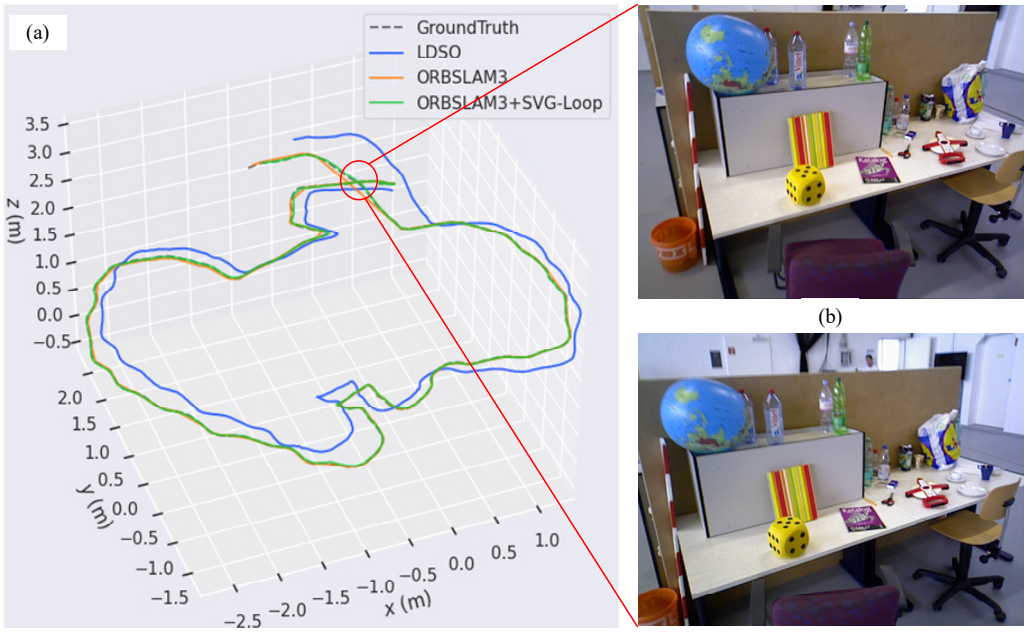
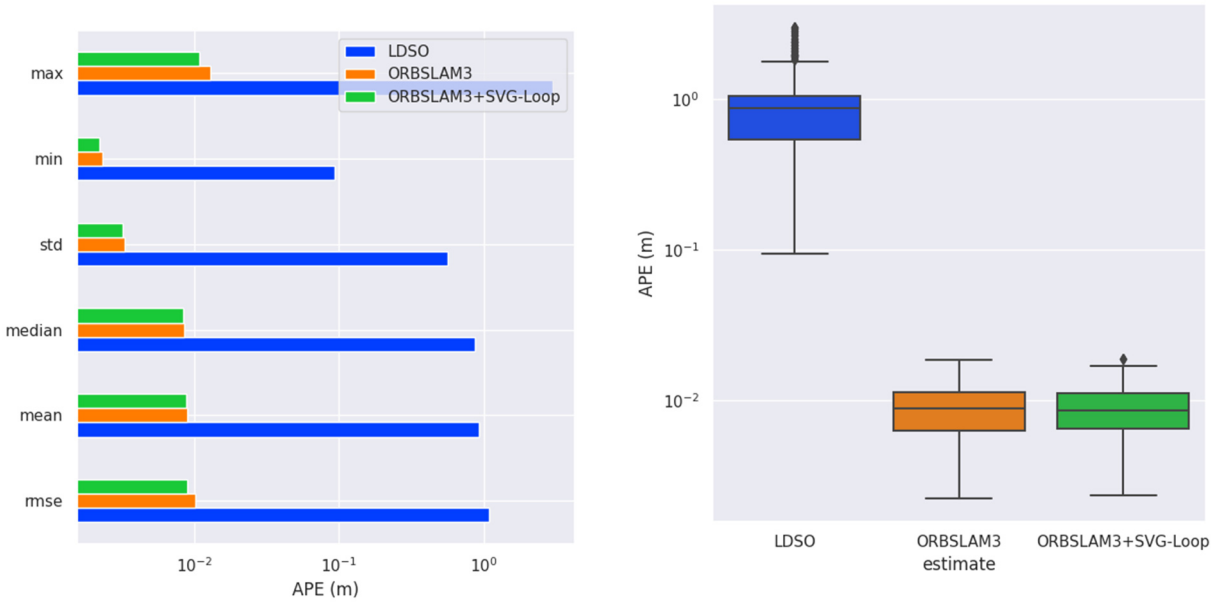
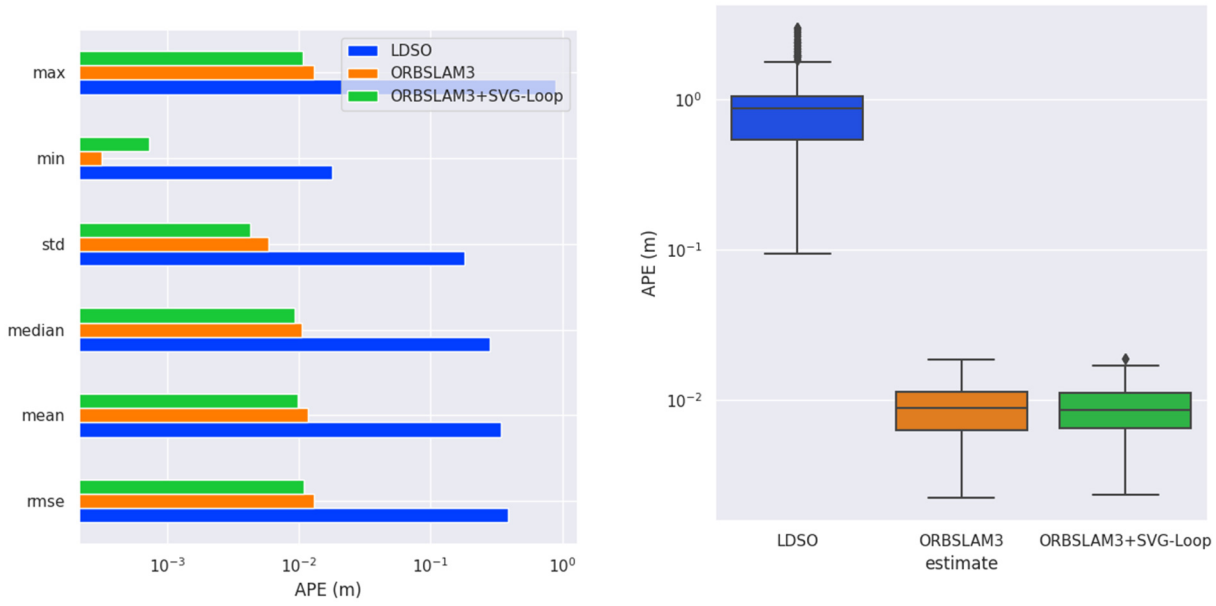


Figure 7. Experimental result for fr3_long_office sequence. (**a**) The trajectory graph of LDSO, ORB-SLAM3, and ORB-SLAM3 + SVG-Loop. (**b**) Images of loop closure scene.



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

Figure 8. Evaluation results of trajectory in fr2_desk sequence. (**a**) Histogram of APE. (**b**) Box diagram of APE.



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

Figure 9. Evaluation results of trajectory in fr3_long_office sequence. (**a**) Histogram of APE. (**b**) Box diagram of APE.

4.1.2. Outdoor Dataset

The KITTI odometry dataset, which was compiled using images of different roads, was selected as the outdoor dataset. Sensors of the KITTI color sequences are PointGray Flea2 color cameras (FL2-14S3C-C). Images in the sequence are 1.4 megapixels, and the viewpoint of the camera is stable. Compared with the TUM dataset, diverse dynamic objects and changes in light that flows over time are contained in scenes of the KITTI dataset. In addition, there is a lengthy overlap time of multiple loop closures in the KITTI odometry dataset. This dataset was leveraged to test the robustness and accuracy of the SVG-Loop method in the outdoor environment.

DBoW2 [34], OpenFABMAP [33], SRLCD [46], and BoWT-LCD [39] were selected to complete comparative experiments with SVG-Loop. DBoW2 and OpenFABMAP are the most popular and practical visual-based methods in loop closure detection. SRLCD is the latest open-source loop closure detection method and is based on salient object information. BoWT-LCD is the art-of-the-state sequence-based loop closure detection model which leveraged the information between images. The precision-recall (POR) curve was chosen as an evaluation metric. The precision-recall metric is calculated as follows:

$$Presicion = \frac{tp}{tp + fp} \tag{8}$$

$$Recall = \frac{tp}{tp + fn} \tag{9}$$

where tp is the number of true positives, fp is the number of false positives, and fn is the number of false negatives. Because error loop closure brings incalculable trajectory deviation in the SLAM system, recall rate at 100% precision has become the most common metric in loop closure detection.

As shown in Figure 10a, the recall rate of SVG-Loop at 100% precision is 73.51%, which outperforms the other methods in sequence 00. According to Figure 10b, the recall rate of SVG-Loop at 100% precision (78.07%), the precision rate of SVG-Loop at 100% recall (30.12%), and the area under curve (AUC) are higher than the other algorithms. In Figure 10c, the recall rate of SVG-Loop at 100% precision is 47.87%, and the precision rate of SVG-Loop at 100% recall is 30.00%, which is the best performing result. SVG-Loop has

the highest precision at any point of the same recall rate in Figure 10d. The above results indicate that SVG-Loop has the best performance out of the five methods in sequences 00, 02, 05, and 06. With more uncontrolled factors, a combination of information in SVG-Loop enhances robustness and precision compared with other loop closure detection methods.

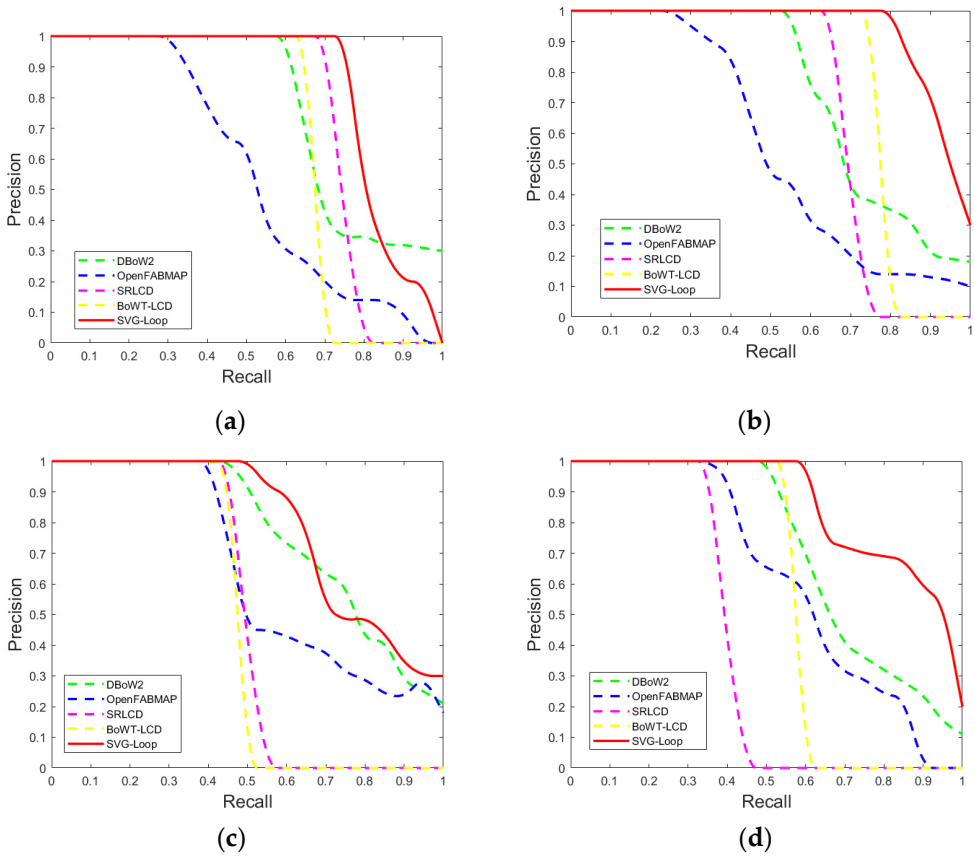


Figure 10. POR curve of DBoW2, OpenFABMAP, SRLCD, BoTW-LCD and SVG-Loop in KITTI dataset. (**a**) Sequence 00. (**b**) Sequence 02. (**c**) Sequence 05. (**d**) Sequence 06.

Table 1 lists the maximum recall rates of the five loop closure detection methods at 100% precision. In sequences 00, 02, 05, and 06, SVG-Loop has the highest recall rate at 100% precision. In addition, the performance of DBoW2 and BoWT-LCD are better than SVG-Loop method in sequences 07 and 09. Performance degradation of SVG-Loop compared with DBoW2 techniques reflects that SVG-Loop loses the advantage of high-level information utilization in monotonous scenes without effective semantic information.

Table 1. The maximum recall rate (%) of five different loop closure detection methods at 100% precision in the KITTI dataset.

Sequence	DBoW2	OpenFABMAP	SRLCD	BoTW-LCD	SVG-Loop
00	58.83	30.04	68.32	63.23	73.51
02	53.23	24.43	63.38	72.62	78.07
05	44.46	39.23	43.12	42.89	47.87
06	47.71	35.33	33.26	52.85	58.11
07	56.35	30.96	26.20	58.49	50.46
09	57.89	41.87	20.00	74.58	46.12

Figure 11 displays loop closure results of SVG-Loop in sequences 00, 02, 05, and 06. Blue lines are trajectories of sequences, and red lines represent correct loop closure detected by the SVG-Loop method in different time nodes. Compared with the change of horizontal

position, vertical fluctuation is negligible in the KITTI dataset. By converting the vertical axis to frame ID, it can be clearly seen that the spatially overlapping loop closures are detected by SVG-Loop. Under the premise of 100% precision, the area with sparse red lines means that the recall rate of the method is relatively low. On the contrary, the dense red line indicates that the recall rate of the algorithm is high. The performance of the SVG-Loop method on the KITTI dataset indicates that the proposed algorithm could maintain high precision and robustness under the dynamic interference of the outdoor environment.

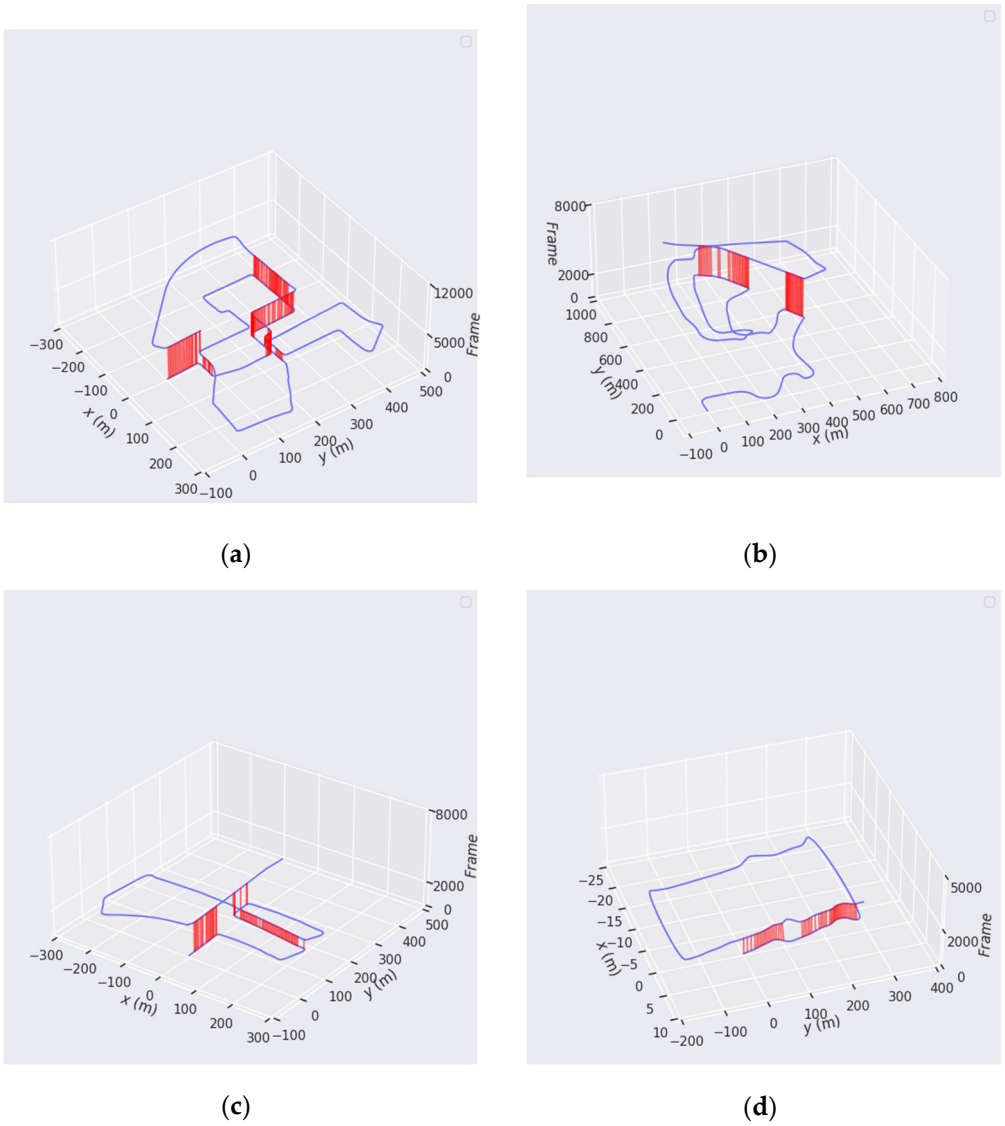


Figure 11. Performance of SVG-Loop method on the KITTI odometry dataset. The horizontal position is maintained, and the vertical axis is replaced with frame ID. (**a**) Results of loop closure detection in sequence 00. (**b**) Results of loop closure detection in sequence 02. (**c**) Results of loop closure detection in sequence 05. (**d**) Results of loop closure detection in sequence 06.

4.2. Practical Environmental Experiments

To further test the accuracy and robustness of the SVG-Loop algorithm with viewpoint jitter, light changes, and dynamic interference, experiments were implemented in different indoor and outdoor environments. As shown in Figure 12a, the camera model used to capture images is a Logitech C922. The resolution of the camera is 960 × 720, and the frame rate is 30 FPS. The field of view (FOV) is 78°, the aperture value is 2.8 and the focus method is an automatic focus (AF). Figure 12b shows some collection scenes in the outdoor test. The experiment was divided into two parts: indoor experiments and outdoor experiments.

(**a**)

(**b**)

Figure 12. Data acquisition sensor and collection scenes. (**a**) Data acquisition sensor in different environments. (**b**) Collection scenes in outdoor test.

4.2.1. Indoor Experiments

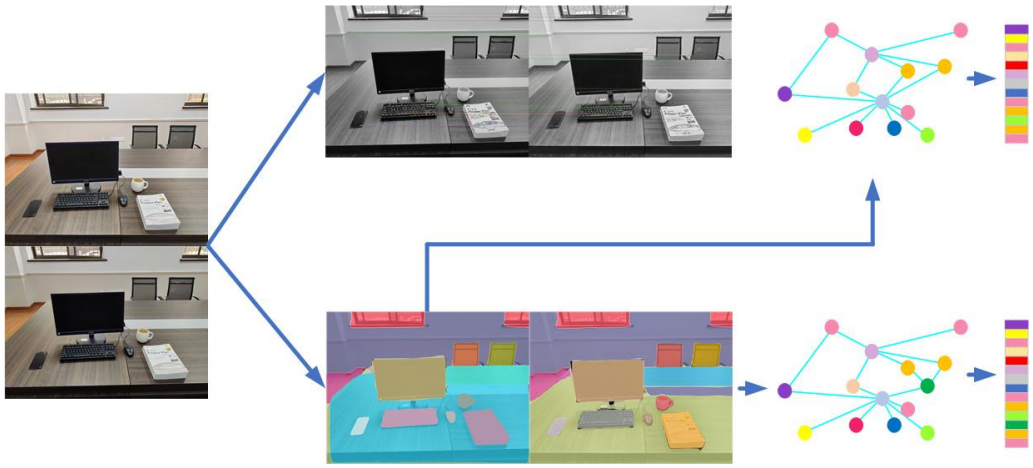
Indoor experiments were designed to test sensitivity to loop closures and robustness to light changes. The camera was held in an office environment for loop closure motion. The data collected in different rooms mainly has two challenges: viewpoint jitter and light change (turning a fluorescent light on and off). The test data was mainly divided into four groups: Room 1, Room 2, Room 3, and Room 4. Room 1 and Room 2 included more reference objects, which can provide landmarks and obvious feature information. The data in Room 3 and Room 4 had the characteristics of strong light changes. Comparative experiments of DBoW2, OpenFABMAP, SRLCD, BoTW-LCD and SVG-Loop methods were completed in different indoor condition.

Table 2 shows the comparative results of the DBoW2, OpenFABMAP, SRLCD, BoTW-LCD, and SVG-Loop methods. According to Table 2, the SVG-Loop algorithm performed better than the other techniques in Room 1, Room 2 and Room 3. However, the recall rate of SVG-Loop at 100% precision is lower 6.53% than BoTW-LCD in Room 4. Light change and absence of reference objects caused the simultaneous reduction of semantic bag-of-word and semantic landmark vector models, which led to the poor performance of the SVG-Loop method.

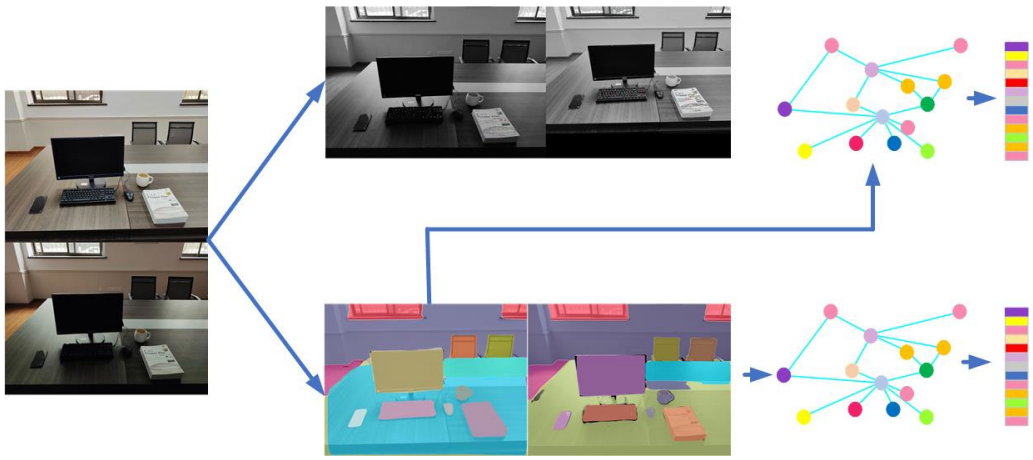
Table 2. The comparative results of five different loop closure detection methods in practical indoor dataset.

Dataset	DBoW2		OpenFABMAP		SRLCD		BoTW-LCD		SVG-Loop	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Room 1	100.00	68.33	100.00	65.00	100.00	58.33	100.00	78.33	100.00	86.67
Room 2	100.00	56.41	100.00	53.84	100.00	21.79	100.00	60.26	100.00	71.79
Room 3	23.56	12.28	15.34	7.02	100.00	38.60	100.00	54.38	100.00	64.91
Room 4	20.00	3.26	13.71	5.43	12.51	6.52	100.00	28.26	100.00	21.73

As shown in Figure 13a, the first loop closure in the meeting room was without light changes. The semantic bag-of-words model can work well and match words in the image pair. At the same time, the semantic graphs and landmark vectors of two-loop scenes are similar. On the contrary, there is an obvious light change in the second loop closure according to Figure 13b. For the semantic bag-of-words model, light change greatly affected the distribution of visual words, which were difficult to match correctly. However, the construction of the semantic graph and landmark vector was robust to the ambient luminosity changes. Finally, the second loop closure was detected successfully. The results of the experiments illustrate that SVG-Loop is robust in the face of light changes and can compensate for the shortcomings of visual-based algorithms.



(**a**)



(**b**)

Figure 13. Example of experimental results with the light change in an indoor environment. (**a**) Process of loop closure detection without light change. (**b**) Process of loop closure detection with light change.

4.2.2. Outdoor Experiments

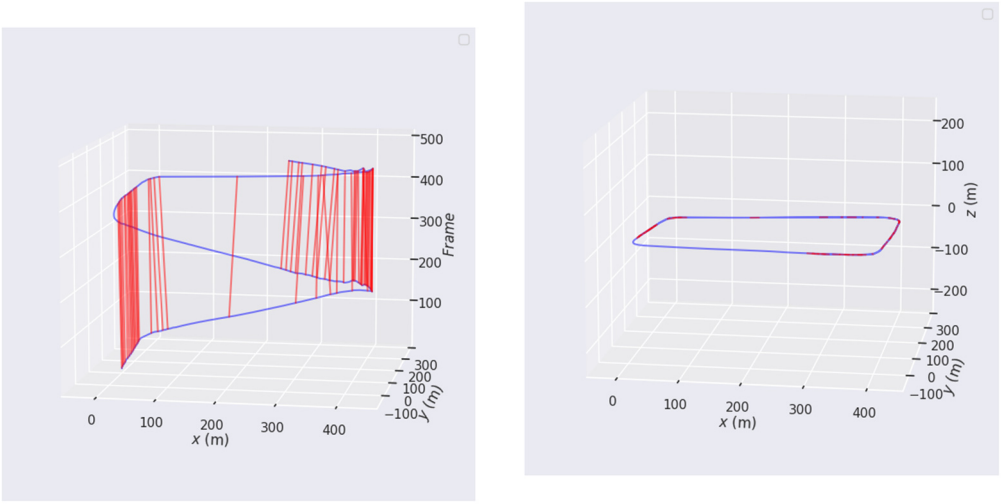
Outdoor experiments were leveraged to further test the accuracy and robustness of SVG-Loop in more complex environments. The outdoor dataset included light changes, weather variations, and dynamic objects. As shown in Figure 12a, the camera was fixed on the car to capture images on different streets. Three loops, which were selected to complete verification experiments, are shown in Figure 14. Loop 1 was located on the main road, with strong light and more dynamic interference. Loop 2 passed through the building complex, with more light and shadow changes. Loop 3 data were collected at 8 a.m., 12 a.m., 6 p.m., and 10 p.m. with different weather. These loops included three conditions: weather change, light change, and dynamic object interference.

Figure 14. Trajectory of the data collection vehicle.

After obtaining the loop closure video data in an outdoor environment, the Global Navigation Satellite System (GNSS) was leveraged to obtain the longitude and latitude of the location in real-time. Then, the longitudes and latitudes of the positions were matched with the loop data and converted into 3D space coordinates (28°23′34.33″ N, 113°00′72.25″ E is the coordinate origin; altitude is set to 0). Figure 15 shows the detection results of SVG-Loop in loop 1 with different dimensions. In outdoor experiments, DBoW2, OpenFABMAP, SRLCD, BoTW-LCD and SVG-Loop algorithms were verified in different loop data. The comparative results of above five methods in practical outdoor dataset are presented in Table 3. As shown in Table 3, the precision and recall rates of SVG-Loop method are higher than the other algorithms. Compared to the state-of-the-art, the recall rates at 100% precision increased by 10.43%, 14.17%, and 4.88%, respectively.

Table 3. The comparative results of five different loop closure detection methods in practical outdoor dataset.

Dataset	DBoW2		OpenFABMAP		SRLCD		BoTW-LCD		SVG-Loop	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Loop 1	100.00	30.36	100.00	23.00	100.00	30.06	100.00	24.23	100.00	40.79
Loop 2	100.00	42.20	100.00	38.04	100.00	45.95	100.00	49.06	100.00	63.20
Loop 3	5.12	1.03	3.11	0.86	74.96	16.23	100.00	19.17	100.00	21.41



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

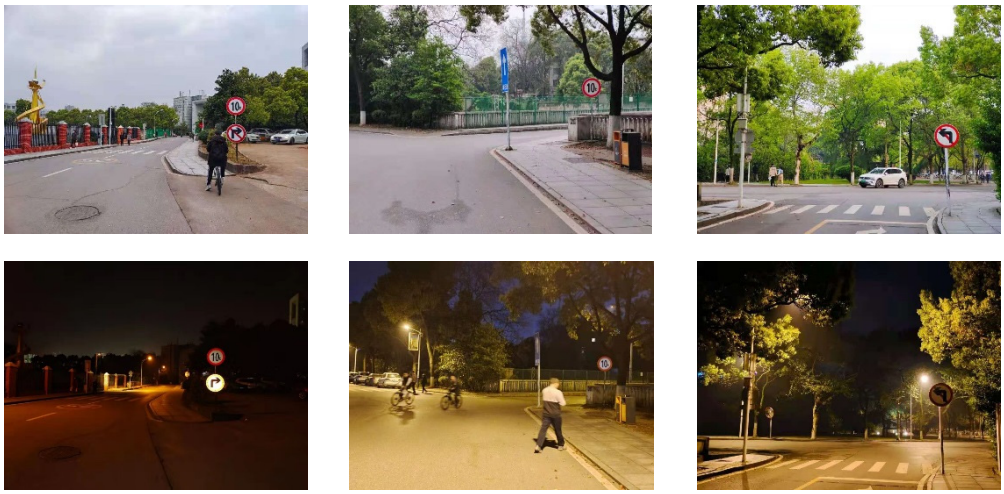
Figure 15. Performance of the SVG-Loop method on the practical outdoor dataset. (**a**) Two-dimensional position trajectory over time and detection result of SVG-Loop in loop 1. (**b**) Three-dimensional position trajectory and the detection result of SVG-Loop in loop 1.

Figure 15 shows the detection results of SVG-Loop in different dimensions. As shown in Figure 15a, loop closures detected by SVG-Loop are correct, and the precision of the proposed method remains 100%. According to Figure 15b, the red line area shows that the recall rate of the SVG-Loop algorithm is 40.79% at 100% precision. In contrast, the maximum recall rates at 100% precision of DBoW2, OpenFABMAP, SRLCD and BoWT-LCD are 30.36%, 23.00%, 30.06% and 24.23%, respectively, in loop 1.

Compared with the indoor dataset, the viewpoint of outdoor data is stable, but uncontrollable factors increase. Figure 16 shows some correct loop scenes detected by SVG-Loop in different environments. Loop scenes in Figure 16a,b contain diverse dynamic objects such as pedestrians, electric vehicles, and cars. Furthermore, Figure 16c shows light and weather changes. Experimental results shown in Figure 16 indicate that the SVG-Loop algorithm is robust for loop closure detection in complex environments that include light changes, weather changes, and dynamic object interference.

(**a**)

(**b**)



(**c**)

Figure 16. Examples of correct loop scenes detected by SVG-Loop. Each column represents a pair of loop scenes. (**a**) Results of SVG-Loop in loop 1. (**b**) Results of SVG-Loop in loop 2. (**c**) Results of SVG-Loop in loop 3.

5. Discussion

5.1. Experiments Analysis

To verify the SVG-Loop method in different environments, extensive experiments were implemented to quantitatively and qualitatively analyze the results of loop closure detection. Experiments were divided into the following four parts:

- The indoor dataset (TUM RGB-D dataset) consists of images taken in stable circumstances. There are no light changes and only a few dynamic objects in sequences, which are selected to complete experiments. Results in Figures 6 and 7 show that the SVG-Loop model is sensitive to loop closure. Figures 8 and 9 indicate that SVG-Loop can combine the SLAM system to achieve higher localization accuracy in an environment where loop closures exist.
- The outdoor dataset (KITTI odometry dataset) contains various dynamic objects but no dramatic light changes. Experiments in this part were leveraged to test the robustness of the SVG-Loop method in an outdoor environment with dynamic interference. According to Figure 11, the SVG-Loop algorithm can overcome some of the dynamic interference and complete loop closure detection.
- The practical indoor experiments included light changes but no dynamic objects. The SVG-Loop method is robust to light changes of different levels and angles, such as in Figure 13. Compared with other visual-based methods, Table 2 shows that SVG-Loop is sensitive to loops and can capture loops quickly and effectively. However, the simultaneous appearance of light changes and lack of semantic landmarks will cause a serious decline in the recall rate of the SVG-Loop algorithm.
- The practical outdoor dataset constructs the most complex situation of the four parts. Drastic light changes, different weather changes, and high-frequency dynamic objects are included in the dataset. According to Table 3, SVG-Loop is robust to outdoor light alters, weather changes, and the movement of dynamic objects. Figures 15 and 16 illustrate that the SVG-Loop model has the potential to detect loop closure for a SLAM system in complex environments.

5.2. Experiment Implementation and Optimization Possibilities

Considering different application scenarios, experimental implementation can be divided into two types: off-line and on-line. In off-line mode, images in datasets are processed sequentially. Table 4 shows the processing time of the SVG-Loop algorithm in different datasets. Panoptic segmentation is the most time-consuming module. If the algorithm is running on-line, the sliding step must be adjusted to match the input speed. When the SVG-Loop is loaded into the SLAM system, the key frames output by the front-end odometer can be used as the detection input to avoid the setting of sliding steps.

Table 4. Processing time per image of SVG-Loop method in KITTI, TUM, and Practical datasets.

		Average Time (ms)		
		KITTI	**TUM**	**Practical Datasets**
Panoptic segmentation		231.5	187.3	251.8
Semantic Bag of Words	Feature extraction	13.2	11.4	14.3
	Vocabulary generation	3.6	3.5	3.4
Semantic landmark Vector	Graph construction	16.9	15.6	18.0
	Vector generation	1.9	1.7	1.8
Loop closure detection		36.5	33.1	38.6
Total		303.6	252.6	327.9

In addition to the above experiment implementation, two optimization possibilities have to be emphasized. Firstly, extended experiments in practical indoor dataset show that

the proposed algorithm still cannot solve the problem of viewpoint jitter and tilt very well. In the face of viewpoint jitter and tilt, the recall rate can be improved by decreasing D_{th}, but at the same time, there is a risk of reducing precision. In future work, two directions will be considered as solutions. Relative spatial distance information of semantic nodes will be added to the descriptors, and the rotation invariant of the semantic descriptor will be explored. Secondly, the results of panoptic segmentation impact this approach substantially. To achieve more robust and accurate results, different targeted pre-training models can be employed according to different environments.

6. Conclusions

In this paper, a loop closure detection method named SVG-Loop is proposed. The SVG-Loop algorithm combines semantic–visual–geometric information to complete loop closure detection in complex environments. SVG-Loop mainly consists of two parts: a semantic bag-of-words model and a semantic landmark vector model. The former determines the visual loop closure candidates by combining visual words and semantic information. The latter provides semantic loop closure candidates through comparing semantic descriptors. Experiments using the TUM RGB-D dataset, KITTI dataset, and practical environments indicate that the SVG-Loop algorithm can effectively complete loop detection and has an advantage in complex environments.

However, there are two limitations to the proposed method. On the one hand, SVG-Loop cannot easily adapt to dramatic viewpoint jitter. On the other hand, the proposed algorithm is highly dependent on the results of panoptic segmentation. In future work, relative spatial distance information and the rotation invariant of the semantic descriptor will be explored. In addition, extensive experiments for each component of SVG-Loop will be designed for a more detailed study.

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

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 61871386) and the Natural Science Fund for Distinguished Young Scholars of Hunan Province, China (Grant No. 2019JJ20022).

Acknowledgments: We would like to express our sincere thanks to Jianwei Wan and Ling Wang who provided hardware support and useful advice for our paper. We are also very grateful for Xingwei Cao who helped us with our data collection.

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

References

1. Smith, R.C.; Cheeseman, P. On the Representation and Estimation of Spatial Uncertainty. *Int. J. Robot. Res.* **1986**, *5*, 56–68. [CrossRef]
2. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, D.; Neira, J.; Reid, I.; Leonard, J.J. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robot.* **2016**, *32*, 1309–1332. [CrossRef]
3. Ho, K.L.; Newman, P. Detecting loop closure with scene sequences. *Int. J. Comput. Vis.* **2007**, *74*, 261–286. [CrossRef]
4. Williams, B.; Klein, G.; Reid, I. Automatic relocalization and loop closing for real-time monocular SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1699–1712. [CrossRef]
5. Geavlete, B.; Stanescu, F.; Moldoveanu, C.; Jecu, M.; Adou, L.; Ene, C.; Bulai, C.; Geavlete, P. 227 The test of time for new advances in BPH endoscopic treatment—Prospective, randomized comparisons of bipolar plasma enucleation versus open prostatectomy and continuous versus standard plasma vaporization and monopolar TURP. *Eur. Urol. Suppl.* **2014**, *13*, e227. [CrossRef]
6. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
7. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571. [CrossRef]

8. Cummins, M.; Newman, P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.* **2008**, *27*, 647–665. [CrossRef]
9. Angeli, A.; Filliat, D.; Doncieux, S.; Meyer, J.A. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Trans. Robot.* **2008**, *24*, 1027–1037. [CrossRef]
10. Gálvez-López, D.; Tardós, J.D. Bags of binary words for fast place recognition in image sequences. *IEEE Trans. Robot.* **2012**, *28*, 1188–1197. [CrossRef]
11. 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]
12. 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]
13. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM. *IEEE Trans. Robot.* **2020**, 1–15. [CrossRef]
14. Sivic, J.; Zisserman, A. Video google: A text retrieval approach to object matching in videos. *Proc. IEEE Int. Conf. Comput. Vis.* **2003**, *2*, 1470–1477. [CrossRef]
15. Milford, M.J.; Wyeth, G.F. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In Proceedings of the IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 1643–1649. [CrossRef]
16. Siam, S.M.; Zhang, H. Fast-SeqSLAM: A fast appearance based place recognition algorithm. In Proceedings of the IEEE International Conference on Robotics and Automation, Marina Bay Sands, Singapore, 29 May–3 June 2017; pp. 5702–5708. [CrossRef]
17. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. DOSeqSLAM: Dynamic on-line sequence based loop closure detection algorithm for SLAM. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques, Kraków, Poland, 16–18 October 2018; pp. 1–6. [CrossRef]
18. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Probabilistic appearance-based place recognition through bag of tracked words. *IEEE Robot. Autom. Lett.* **2019**, *4*, 1737–1744. [CrossRef]
19. Neuland, R.; Rodrigues, F.; Pittol, D.; Jaulin, L.; Maffei, R.; Kolberg, M.; Prestes, E. Interval Inspired Approach Based on Temporal Sequence Constraints to Place Recognition. *J. Intell. Robot. Syst. Theory Appl.* **2021**, *102*, 1–24. [CrossRef]
20. Chen, Z.; Maffra, F.; Sa, I.; Chli, M. Only look once, mining distinctive landmarks from ConvNet for visual place recognition. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Vancouver, BC, Canada, 24–28 September 2017; pp. 9–16. [CrossRef]
21. Wang, Y.; Qiu, Y.; Cheng, P.; Duan, X. Robust loop closure detection integrating visual–spatial–semantic information via topological graphs and CNN features. *Remote Sens.* **2020**, *12*, 3890. [CrossRef]
22. Finman, R.; Paull, L.; Leonard, J.J. Toward Object-based Place Recognition in Dense RGB-D Maps. In Proceedings of the IEEE International Conference on Robotics and Automation workshop, Seattle, WA, USA, 26–30 May 2015.
23. Stumm, E.; Mei, C.; Lacroix, S.; Chli, M. Location graphs for visual place recognition. In Proceedings of the IEEE International Conference on Robotics and Automation, Seattle, WA, USA, 26–30 May 2015; pp. 5475–5480. [CrossRef]
24. Gawel, A.; Del Don, C.; Siegwart, R.; Nieto, J.; Cadena, C. X-View: Graph-Based Semantic Multi-view Localization. *IEEE Robot. Autom. Lett.* **2018**, *3*, 1687–1694. [CrossRef]
25. Cascianelli, S.; Costante, G.; Bellocchio, E.; Valigi, P.; Fravolini, M.L.; Ciarfuglia, T.A. Robust visual semi-semantic loop closure detection by a covisibility graph and CNN features. *Robot. Auton. Syst.* **2017**, *92*, 53–65. [CrossRef]
26. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1437–1451. [CrossRef]
27. Hausler, S.; Garg, S.; Xu, M.; Milford, M.; Fischer, T. Patch-NetVLAD: Multi-Scale Fusion of Locally-Global Descriptors for Place Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.
28. Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, 7–12 October 2012; pp. 573–580. [CrossRef]
29. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
30. Lowry, S.; Sunderhauf, N.; Newman, P.; Leonard, J.J.; Cox, D.; Corke, P.; Milford, M.J. Visual Place Recognition: A Survey. *IEEE Trans. Robot.* **2016**, *32*, 1–19. [CrossRef]
31. Masone, C.; Caputo, B. A Survey on Deep Visual Place Recognition. *IEEE Access* **2021**, *9*, 19516–19547. [CrossRef]
32. Chen, Y.; Gan, W.; Zhang, L.; Liu, C.; Wang, X. A survey on visual place recognition for mobile robots localization. In Proceedings of the 2017 14th Web Information Systems and Applications Conference, Liuzhou, China, 11–12 November 2017; pp. 187–192. [CrossRef]
33. Glover, A.; Maddern, W.; Warren, M.; Reid, S.; Milford, M.; Wyeth, G. OpenFABMAP: An open source toolbox for appearance-based loop closure detection. In Proceedings of the IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 4730–4735. [CrossRef]

34. Mur-Artal, R.; Tardós, J.D. Fast relocalisation and loop closing in keyframe-based SLAM. In Proceedings of the IEEE International Conference on Robotics and Automation, Hong Kong, China, 31 May–7 June 2014; pp. 846–853. [CrossRef]
35. Khan, S.; Wollherr, D. IBuILD: Incremental bag of Binary words for appearance based loop closure detection. In Proceedings of the IEEE International Conference on Robotics and Automation, Seattle, WA, USA, 26–30 May 2015; pp. 5441–5447. [CrossRef]
36. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Assigning visual words to places for loop closure detection. In Proceedings of the IEEE International Conference on Robotics and Automation, Brisbane, Australia, 21–25 May 2018; pp. 5979–5985. [CrossRef]
37. Bampis, L.; Amanatiadis, A.; Gasteratos, A. Fast loop-closure detection using visual-word-vectors from image sequences. *Int. J. Robot. Res.* **2018**, *37*, 62–82. [CrossRef]
38. Tsintotas, K.A.; Bampis, L.; Rallis, S.; Gasteratos, A. SeqSLAM with bag of visual words for appearance based loop closure detection. *Mech. Mach. Sci.* **2019**, *67*, 580–587. [CrossRef]
39. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Modest-vocabulary loop-closure detection with incremental bag of tracked words. *Robot. Auton. Syst.* **2021**, *141*, 103782. [CrossRef]
40. Yu, C.; Liu, Z.; Liu, X.-J.; Xie, F.; Yang, Y.; Wei, Q.; Fei, Q. 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; pp. 1168–1174.
41. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation}. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
42. Zhang, Z.; Zhang, J.; Tang, Q. Mask R-CNN Based Semantic RGB-D SLAM for Dynamic Scenes. In Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM, Hong Kong, China, 8–12 July 2019; pp. 1151–1156. [CrossRef]
43. Kirillov, A.; Girshick, R.; He, K.; Dollár, P. Panoptic FPN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 6392–6401. [CrossRef]
44. Merrill, N.; Huang, G. CALC2.0: Combining Appearance, Semantic and Geometric Information for Robust and Efficient Visual Loop Closure. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Macao, China, 4–8 November 2019; pp. 4554–4561. [CrossRef]
45. Wang, Z.; Peng, Z.; Guan, Y.; Wu, L. Two-Stage vSLAM Loop Closure Detection Based on Sequence Node Matching and Semi-Semantic Autoencoder. *J. Intell. Robot. Syst. Theory Appl.* **2021**, *101*, 29. [CrossRef]
46. Wang, H.; Wang, C.; Xie, L. Online visual place recognition via saliency re-identification. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Macao, China, 4–8 November 2019; pp. 5030–5036. [CrossRef]
47. Oh, J.H.; Jeon, J.D.; Lee, B.H. Place recognition for visual loop-closures using similarities of object graphs. *Electron. Lett.* **2015**, *51*, 44–46. [CrossRef]
48. Chen, H.; Zhang, G.; Ye, Y. Semantic Loop Closure Detection with Instance-Level Inconsistency Removal in Dynamic Industrial Scenes. *IEEE Trans. Ind. Inform.* **2021**, *17*, 2030–2040. [CrossRef]
49. Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollar, P. Panoptic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020; pp. 9396–9405. [CrossRef]
50. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
51. Muja, M.; Lowe, D.G. Fast approximate nearest neighbors with automatic algorithm configuration. In Proceedings of the VISAPP 4th International Conference on Computer Vision Theory and Applications, Lisboa, Portugal, 5–8 February 2009; Volume 1, pp. 331–340. [CrossRef]
52. Cadena, C.; Gálvez-López, D.; Tardós, J.D.; Neira, J. Robust place recognition with stereo sequences. *IEEE Trans. Robot.* **2012**, *28*, 871–885. [CrossRef]
53. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 5, pp. 1–4.
54. Gao, X.; Wang, R.; Demmel, N.; Cremers, D. LDSO: Direct Sparse Odometry with Loop Closure. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Madrid, Spain, 1–5 October 2018; pp. 2198–2204. [CrossRef]