# Bag-of-Word-Groups (BoWG): A Robust Loop Closure Module for In-pipe Visual-Laser-Inertial SLAM

Xiang Fei[1], Tina Tian[2], Lu Li[2], Howie Choset[2]

***Abstract*— In-pipe simultaneous localization and mapping (SLAM) techniques with photorealistic RGB-D reconstruction capability have the potential to enhance human labor to inspect pipe conditions and localize anomalies, thereby preventing hazardous leaks and explosions. Loop closure detection is vital in the process of SLAM, as it helps reduce the accumulative drift of the robot's estimated odometry and generate a globally consistent map. However, in confined-space environments such as narrow pipes, conventional loop closure methods suffer perceptual aliasing due to feature scarcity and textural repetitiveness. In this research, we aim to develop a robust loop closure module in confined-space environments on top of our prior confined-space dense RGB-D SLAM method, visual-laser-inertial (VLI) SLAM. Specifically, we define the concept of word group based on spatial proximity and positions of features and propose to build and maintain a novel loop closure detection module called Bag-of-Word-Groups (BoWG) online, which provides context-specific feature representation. Besides, we utilize Gaussian pyramids to implement Multi-scale Good Features To Track (MS-GFTT) to detect richer features at various scales for word group analysis. Our method does not require any extra sensor other than a monocular visual camera and can be easily integrated into existing Bag-of-Words (BoW) methods. To validate the proposed method, we conduct real-world experiments in a narrow, feature-sparse pipeline with loops. Experiment results show that our method is robust and can achieve high precision while maintaining acceptable recall when the perceptual aliasing problem is serious. In addition, the proposed method has the potential to be applied to environments other than narrow pipes.**

## I. INTRODUCTION

The proper functioning of pipelines is critical for sustaining the needs of societies and industries, and any disruption or damage to these systems can have severe consequences, both economically and environmentally [8]. This mandates the rigorous inspection and maintenance of pipelines as a matter of paramount importance. To facilitate non-destructive testing (NDT) and condition assessment in narrow pipes, Simultaneous Localization and Mapping (SLAM) techniques can be applied to generate a 3D reconstruction of the pipe interior, which offers valuable insights for detailed pipe inspection and condition tracking.

Existing visual-inertial SLAM methods such as VLI-SLAM [6] previously developed by our group have been adapted to operate in pipes and have achieved promising results in producing dense RGB-D maps with sub-millimeter grade scanning accuracy to facilitate pipe inspection. Although tests indicate that this method can maintain relatively consistent localization in short distances, it experiences unignorable accumulative drift in long-distance situations.

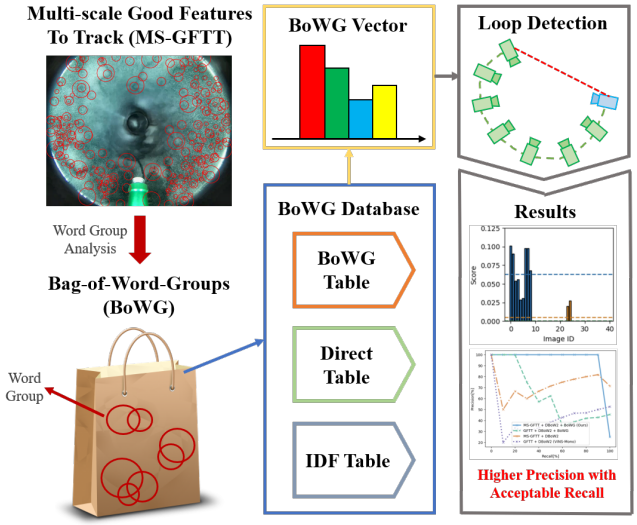



Fig. 1: The proposed BoWG-based Loop Closure Detection Method in Confined Spaces.

To reduce accumulative drift, loop closure is a common technique in SLAM for constraining the odometry estimation by leveraging the loops in the environment [16]. Since pipeline networks often comprise one or multiple loops, loop closure is also applicable and valuable for in-pipe SLAM methods to generate a globally consistent map. However, it is challenging for conventional loop closure methods like [10] to correctly perform loop detection when there is perceptual aliasing, which is particularly severe in confined-space environments like pipes due to feature scarcity and textural repetitiveness [8].

Aiming to address the above problems, we propose a novel visual loop closure detection module termed Bag-of-Word-Groups (BoWG), which helps enhance the precision of loop closure detection in confined spaces where perceptual aliasing is severe. Specifically, the three major contributions of this paper are summarized as follows:

- A feature detector capable of detecting features at different scales based on an image pyramid.
- A definition of word groups exploiting the spatial co-occurrence of detected features, which provides context-

This work was supported by the Department of Energy (DOE)'s Advanced Research Projects Agency-Energy (ARPA-E), REPAIR Program

[1]The author is with the School of Data Science, the Chinese University of Hong Kong, Shenzhen. This work was completed when Xiang Fei served as an intern in the Robotics Institute Summer Scholar at Carnegie Mellon University. xiangfei@link.cuhk.edu.cn

[2]The authors are with the Biorobotics Lab, the Robotics Institute at Carnegie Mellon University, Pittsburgh, PA 15213, USA {yutian, lilu12, choset}@andrew.cmu.edu

specific feature representation.
- A database is established and updated online storing the word group information of each image, which enables efficient loop closure score computation.

Experiment results demonstrate that our method achieves significantly higher precision with acceptable recall compared to existing methods and better distinguishes the loop closure frame compared to previous methods in scenarios where perceptual aliasing is prominent. Besides, our method exhibits robustness to minor data augmentation on visual keyframes. Additionally, tests using datasets from the indoor environment indicate that the applicability of the proposed method extends beyond narrow pipes, making it a promising candidate for loop closure in generic environments as well.

## II. RELATED WORKS

### A. Bag-of-Words and Relationships Between Features

The Bag of Words (BoW) technique, originally used in text retrieval [20], has been widely adopted in place recognition, serving as a crucial component of loop closure detection methods. In the BoW approach, a visual vocabulary is created by clustering the extracted features from a set of training images. Each image is represented as a histogram of visual words (cluster centers), where each bin in the histogram corresponds to the frequency of occurrence of a visual word in the image. To implement loop closure detection, histogram matching is performed between the current and the previous keyframes in the pose graph [22]. DBoW2 [7] is a commonly used BoW system, which implements an image database with inverted and direct files to index images and enabling quick queries and feature comparisons. In later research, BoW methods that are updated online and do not require offline training are proposed [14]. In addition, [1] considers the relationship between features when designing the bag of words for topological maps. This research uses feature co-occurrence to provide richer information for images, which utilizes the spatial proximity of different features. However, [1] only considers the relationship between features in pair, and does not take into account the specific positions of features. In this paper, we define various types of word groups based on both spatial proximity and the positions of features. In addition, we design the online BoWG method by considering not only whether a single word occurs or not, but also the relationship between words, which provides context-specific feature representation.

### B. Good Features To Track and Feature Scales

In this paper, we use Good Features To Track (GFTT) [19] to detect features in the keyframes due to the fast speed and the robustness to the low-illumination, texture-less conditions in pipelines as opposed to other feature detectors/descriptors (e.g., SIFT [12], SURF [3], ORB [17], etc.), following the suggestions in [18]. However, GFTT does not have scales to provide the necessary information for analyzing the co-occurrence relationships between features. Inspired by [12], which utilizes Gaussian blur to downsample images, and then obtains scale-invariant features with their sizes according to the Difference of Gaussians (DoG) extreme values at different scales, we obtained the sizes of GFTT features by building Gaussian pyramids. Although scale-invariant feature points are not required for our task and might even be unfavorable for loop closure detection in narrow pipes due to the small number of feature points, the method of obtaining scale information of feature points is meaningful to our research.

## III. METHODOLOGY

### A. Framework of the Proposed BoWG-based Loop Detection Method

The proposed BoWG-based loop closure detection method is developed on top of the loop closure pipeline in VINS-Mono [16], which consists of three main modules: Feature Detection, DBoW2 Processing, and BoWG Processing, as illustrated in Fig. 2. The overall workflow of the proposed online BoWG-based loop closure detection method is as follows:

1. When a new keyframe is detected by the SLAM system, the proposed Multi-scale Good Features To Track (MS-GFTT) method is applied to extract visual features with different sizes and use BRIEF [5] to compute the feature descriptors.
2. The obtained feature descriptors are utilized by DBoW2 to get the words in the keyframe and use the Term Frequency - Inverse Document Frequency (TF-IDF) score [15] from DBoW2 (DBoW2 score) to obtain the loop closure candidates. DBoW2 can be replaced by other Bag-of-Words methods, such as FBoW [13], incremental bag-of-words [2], etc.
3. After getting the size, coordinate (from MS-GFTT) and corresponding word (from DBoW2) of each feature in the keyframe, the word group information of the keyframe is analyzed.
4. The word group information is utilized by the BoWG module to obtain the TF-IDF scores (BoWG scores) of the loop closure candidates. After that, DBoW2 scores and BoWG scores are combined to compute the similarities between the keyframe and the loop closure candidates, then choose the final loop closure frame.
5. After obtaining the final loop closure frame, new word groups in the current keyframe are added into our BoWG database to perform an online update.

### B. Multi-scale Good Features To Track (MS-GFTT)

The proposed Multi-scale Good Features To Track (MS-GFTT) is a feature detector capable of detecting richer features at various scales for word group analysis. Gaussian pyramid is utilized to implement this method.

*1) **Gaussian Pyramid**:* Imagine the pyramid as a set of layers (each layer is an image) in which the higher the layer, the smaller the size, as shown in Fig. 3. To produce higher layer in the Gaussian pyramid from the lower layer, we do the following downsampling steps:

- Convolve the lower layer with a 5x5 Gaussian blur kernel;

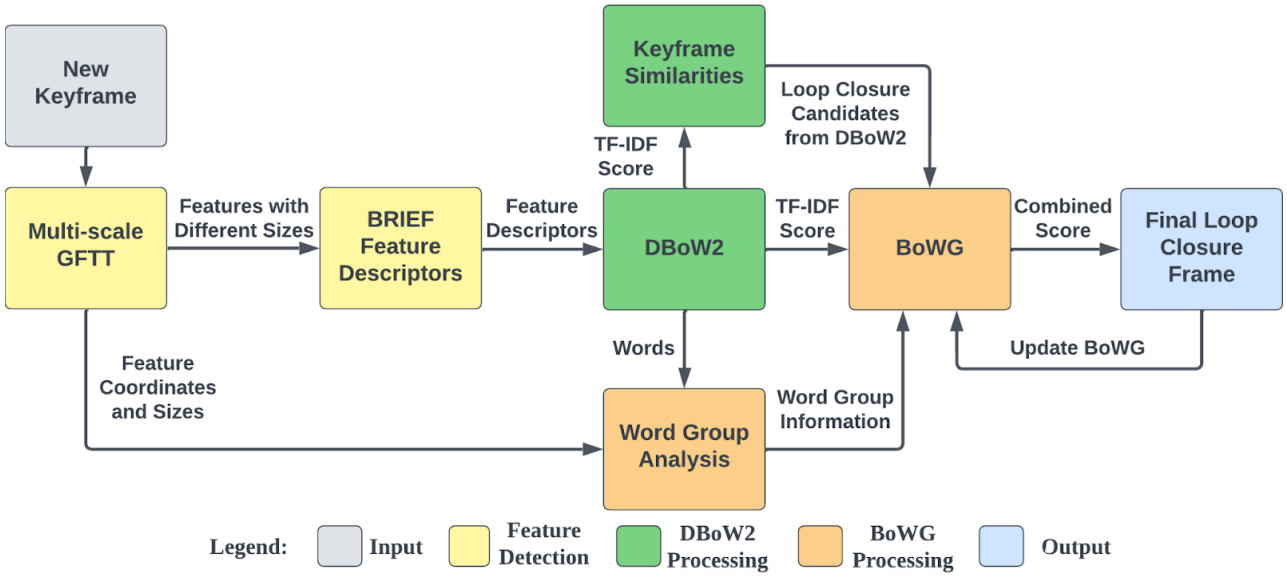



Fig. 2: Framework of the proposed BoWG-based Loop Closure Detection Module.

- Downsample the layer by removing every even-numbered row and column.

*2) **Multi-scale Feature Detection**:* When a new keyframe is collected, we first build its Gaussian pyramid and the pyramid of the mask that is used to block the edge of the keyframe and the laser profiler. After that, Good Features To Track (GFTT) [19] is performed on different layers of images in the pyramid to detect features with different scales, as shown in Fig. 3.

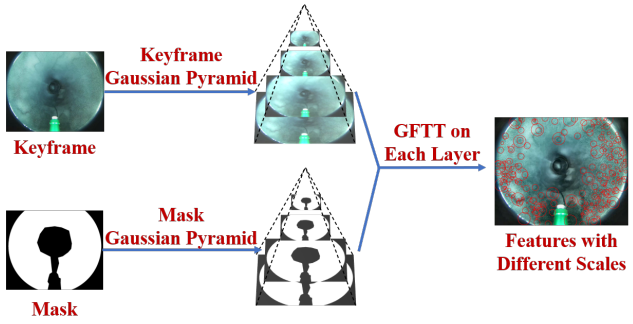



Fig. 3: Schematic diagram of Gaussian Pyramid and MS-GFTT.

### C. Bag-of-Word-Groups (BoWG)

The proposed Bag-of-Word-Groups (BoWG) is a novel loop closure detection module that is updated online and provides context-specific feature representation. This module is used to analyze the relationship between the features of each keyframe, so as to mine more information about the keyframe and to enhance the ability for distinguishing the loop closure frame candidates, thereby improving the loop detection precision. In addition, the proposed BoWG module does not require any extra sensor other than a monocular visual camera. A key data structure in BoWG is the word group (defined below) and four different types of tables: 1) BoWG Vector, 2) BoWG Table, 3) Direct Table, and 4) Inverse Document Frequency (IDF) Table are introduced, which are built and maintained online, as shown in Fig. 5.

***Word Group Definition**:* We define three different types of word groups based on the spatial proximity and positions of features: Multi-Scale Feature, Word Pair, and Word Triplet (Fig. 4).

- Multi-Scale Feature: a feature that can be detected in different layers of the Gaussian Pyramid.
- Word Pair: the distance between two different features is less than the sum of the radius of the two features.
- Word Triplet: the distance between three features from each other is less than the sum of each other's radius.

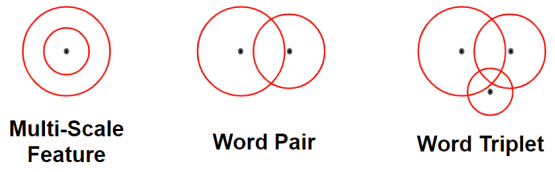



Fig. 4: Three different types of word groups.

To determine whether two word groups are the same, we need to check whether they contain the same words and locate in the same area of the images. If two word groups $A$ and $B$ contain the same words, and the sum of the distance between the pixel coordinates of the corresponding words in $A$ and $B$ is smaller than a threshold $\tau$, then $A$ and $B$ represent the same word group, as the following formula shows:

$$\sum_i \sqrt{(u_A^{(i)} - u_B^{(i)})^2 + (v_A^{(i)} - v_B^{(i)})^2} < \tau \tag{1}$$

where $(u_A^{(i)}, v_A^{(i)})$ and $(u_B^{(i)}, v_B^{(i)})$ represent the pixel coordinates of the word $i$ in $A$ and $B$, respectively.

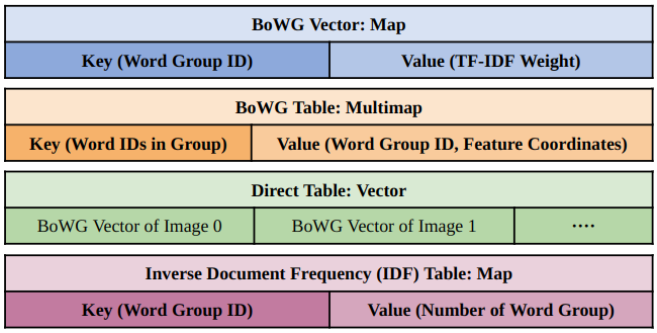



Fig. 5: The data structure of the four types of tables in BoWG method.

*1) **BoWG Vector**:* Each image has one BoWG vector, which is a representation of the image and is used to compute the similarity score between different images. The BoWG vector is implemented using a Map data structure, whose key is the word group IDs that are detected in the image, and the value is the TF-IDF weight of the word group in the image. In addition, the similarity score (TF-IDF score [15]) of two images is:

$$s_b(v_1, v_2) = \begin{cases} 1 & ,\sum_i w_1^{(i)} \cdot w_2^{(i)} > 1 \\ 1 - \sqrt{1 - \sum_i w_1^{(i)} \cdot w_2^{(i)}} & ,\text{else} \end{cases} \quad (2)$$

where $v_1$ and $v_2$ represent the BoWG Vector of the two images, $w_1^{(i)}$ and $w_2^{(i)}$ are the TF-IDF weights of the word group with ID $i$ in these two images.

*2) **BoWG Table**:* The BoWG table stores all the word groups detected in the previous keyframes and their information. The underlying data structure is Multimap [4]. The key is a string consisting of the word IDs in a word group, the value includes the word group ID and feature coordinates, as shown in Fig 5. The reason we use multimap is that word groups with the same keys may be different word groups since they may be located in different areas of the images. Therefore, different word groups may have the same key and word group ID is the only identifier of each word group.

*3) **Direct Table**:* The direct table stores the BoWG Vector of each image, which is used to compute the similarity between the loop closure candidates and the current keyframe. The underlying data structure is Vector and the index of the vector is the image ID.

*4) **Inverse Document Frequency (IDF) Table**:* The IDF table stores the number of each word group in the database and is used to compute the IDF value. The underlying data structure is Map. The key is the word group IDs and the value is the number of each word group in the database.

### *D. Combined Score: for finalizing the loop closure frame*

The combined score takes into account both the DBoW2 score and the BoWG score, which reflects the similarities between the current keyframe and loop closure candidates obtained from DBoW2 and is used to determine the final loop closure frame. Since word group matching is stricter than word matching and each image has rich word groups, the BoWG scores are much smaller than DBoW2 score and the score gaps between high and low scoring word groups are greater. Therefore, we use MIN-MAX normalization to process the BoWG score and use it as a factor to multiply with the DBoW2 score to compute the combined score $s_c$:

$$s_c = \frac{s_b - s_{bmin}}{s_b - s_{bmax}} \cdot s_d \quad (3)$$

where $s_d$ is the DBoW2 score, $s_b$ is the BoWG score, $s_{bmin}$ and $s_{bmax}$ are the minimum and maximum BoWG scores of the loop closure candidates obtained from DBoW2.

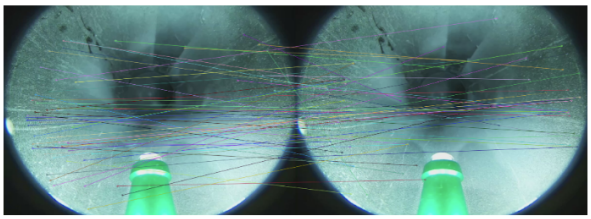
In this way, the combined score has the following characteristics:

- Images with higher DBoW and BoWG scores still have higher combined scores. This characteristic helps to maintain the high score of the true loop closure frame.
- Images with higher DBoW2 scores and lower BoWG scores will have lower scores. This characteristic helps to suppress keyframes that do not belong to the loop closure area but have high DBoW2 scores due to perceptual aliasing.
- The difference between high and low final scores will be greater. This characteristic helps enhance the ability of the loop detection module for distinguishing the true loop closure area and other areas.

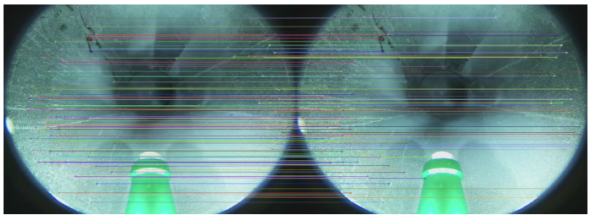
All of the above characteristics help overcome severe perceptual aliasing in narrow pipes.

### *E. Feature Correspondence Outlier Removal*

After detecting loop closure, we create a connection between the current frame and the loop closure frame at the feature level. The feature matching process utilizes BRIEF descriptors matching, where matching pairs are selected based on finding the minimum Hamming distance that is a bitwise XOR operation followed by a bit count.



(a) BRIEF descriptor matching results



(b) Outlier rejection results

Fig. 6: Descriptor matching and outlier removal for feature retrieval after loop detection.

However, directly performing descriptor matching can result in numerous outliers, as demonstrated in Fig. 6a. To

address this issue and validate the loop detection, we use a fundamental matrix test with RANSAC [9] or PnP test with RANSAC [11] to remove outliers in the same way as loop closure in VINS-Mono [16].

### F. Integrated into VLI-SLAM

Since our prior work VLI-SLAM is based on the architecture of VINS-Mono [16], we follow the steps in VINS-Mono to integrate our proposed loop closure detection module into VLI-SLAM system:

1. Sliding-window-based local pose optimization;
2. Loop detection (Our module);
3. Relocalize frames in the current sliding window using the loop constraints;
4. Add the marginalized keyframe from the current window into the pose graph;
5. 4 DOF Pose graph optimization (global pose optimization);
6. Relocalize frames in the sliding window in the optimized pose graph.

## IV. EXPERIMENTS

In this section, the performance of our proposed loop closure detection module is evaluated from the perspectives of precision, recall, and recognition ability for loop closure frames by comparing it with other methods. After that, We test the robustness of the proposed method by performing data augmentation. Finally, we demonstrate the potential of our method to be applied to scenarios other than narrow pipes by testing on the TUM dataset [21].

### A. Experiment Setup

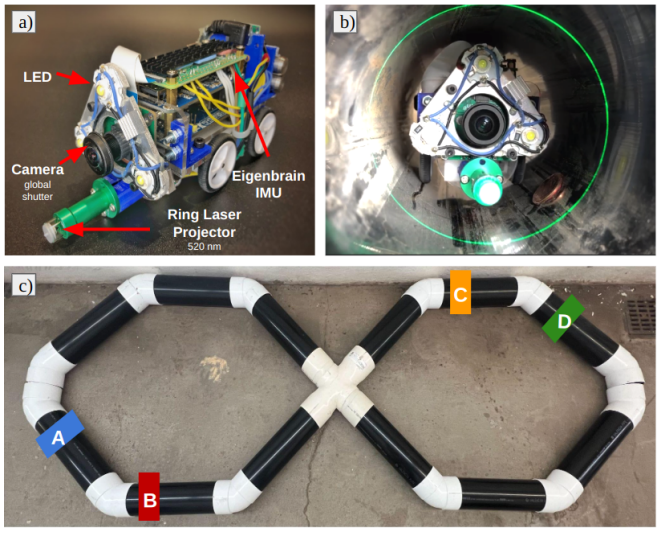



Fig. 7: (a) In-pipe robot (b) the robot moves through the pipe and (c) test site. The pipe diameter is 6".

The in-pipe robot moves through the real-world test pipe with loops whose diameter is 6" and collects image data, as shown in Fig. 7. To test the performance of the proposed BoWG-based loop closure module, we conducted tests as follows:

- Consider 4 small areas (A, B, C, D) in the pipe with loops.
- Each area contains 10 frames (40 frames in total).
- Suppose the robot returns to area A (that is, area A is the true loop closure area) and collects a new image, we calculate the loop closure scores of the 40 previous frames by comparing them with the newly collected image. (Experiments were also carried out when other areas were closed-loop areas and the results can be found in the supplementary materials).

### B. Test Metrics Definition

**Average Score Difference (ASD).** The difference between the average score of the true area and the average score of the remaining area with the highest average score. Higher ASD reflects that the ability of the method for distinguishing the true area and other areas is better.

**Highest Score Difference (HSD).** The difference between the highest score in the true area and the highest score in the remaining areas. Higher HSD reflects that we can adjust the threshold over a larger range to obtain at least one true loop closure frame without errors.

### C. Test Results

Fig. 8 shows the similarity scores of the 40 frames compared to the new collected frame obtained by different methods when area A is the true loop closure area. For the (GFTT + DBoW2 + BoWG) method, the scale of features is a fixed value. When using the method in VINS-Mono (GFTT + DBoW2), some frames in area C score higher than area A, which may lead to incorrect selection of loop closure frames. Therefore, when area A is the true loop closure area, the perceptual aliasing problem is serious. Precision and recall are computed and shown in Fig. 9.

According to the quantitative results in Table I, the proposed method (MS-GFTT + DBoW2 + BoWG) obtains much higher ASD and HSD values. Besides, our method can achieve significantly higher precision with acceptable recall. When the recall is 50%, the precision of our method is 100% while that of the method in VINS-Mono is 38.5%. When the recall is 90%, the precision of our method is 100% while that of the method in VINS-Mono is 50.0%.

For (GFTT + DBoW2 + BoWG) method, it can also get larger ASD and HSD values compared to the method in VINS-Mono, but in order to achieve acceptable results, it usually requires more stringent hyperparameter tuning (the block size of GFTT, the fixed scale of features, and the distance threshold for identifying word groups) and results in a longer running time. Moreover, it can be seen from the results of other experiments in the supplementary materials that this method of fixing the feature size is not stable.

For the (MS-GFTT + DBoW2) method, the recall has increased, but the ASD and HSD are still low. This indicates that the score difference between the loop closure frame and

TABLE I: Experiment results of different methods when A is the true loop closure area

| Results\Methods | **MS-GFTT + DBoW2 + BoWG (Proposed)** | GFTT + DBoW2 (VINS-Mono) | GFTT + DBoW2 + BoWG | MS-GFTT + DBoW2 |
|---|---|---|---|---|
| ASD | **0.05777** | -0.006797 | 0.01528 | 0.00926 |
| HSD | **0.07399** | -0.00849 | 0.05657 | -0.00175 |
| Precision (50% recall) | **100** | 38.5 | 62.5 | 71.4 |
| Precision (90% recall) | **100** | 50.0 | 42.9 | 81.8 |
| Runtime (s) | 3.53 | **2.48** | 5.30 | 2.94 |

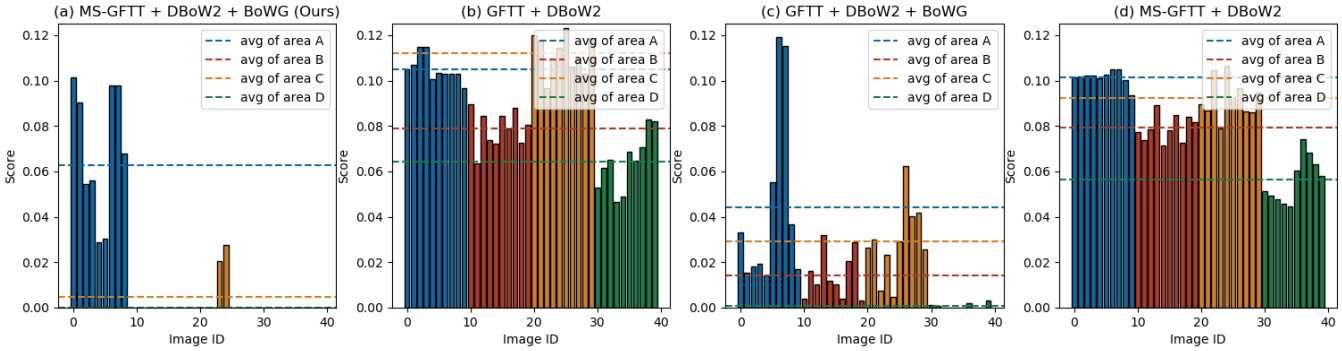



Fig. 8: Similarity scores obtained by different methods when A is the true loop closure area.

frames in other areas is small, which makes it difficult for us to set the threshold in practical use.

Therefore, the proposed (MS-GFTT + DBoW2 + BoWG) method is capable of:

- Achieving much higher precision while maintaining acceptable recall when perceptual aliasing is serious.
- Achieving much higher ASD and HSD, which shows that the method can better distinguish the loop closure area when perceptual aliasing is serious.
- Fast enough to update online.

Taking into account the experiments shown in the supplementary materials in which cases the perceptual aliasing problem is not very serious, the proposed method can still obtain much higher ASD and HSD values and reach high precision with acceptable recall.

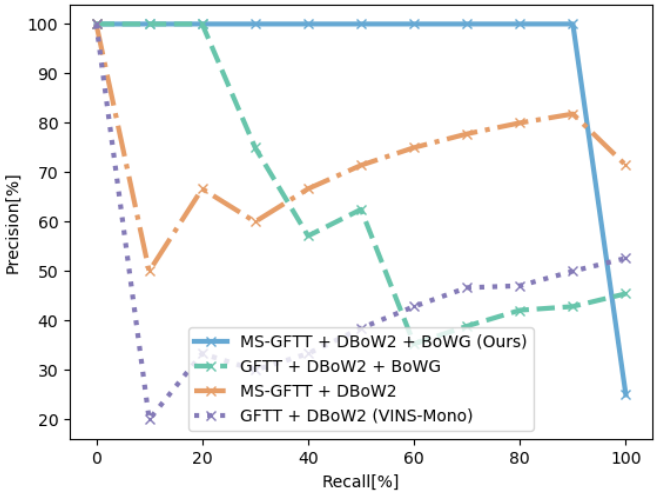



Fig. 9: Precision and recall of different methods when A is the true loop closure area.

### *D. Robustness Test*

To test the robustness of the proposed method, we randomly perform small data augmentation such as rotation (5-10 degrees), translation (10-20 pixels), and scaling (factor is 0.8-0.9) on 5 images from the 40 previous images. Fig. 10 shows the similarity scores of our method with and without data augmentation. Table III illustrates the quantitative results. It can be seen that the results hardly change, which supports that our method is robust and can cope with small changes in the environment.

TABLE II: experiment results of robustness test

| Aug\Results | ASD | HSD | Precision (50% recall) | Precision (90% recall) | Runtime (s) |
|---|---|---|---|---|---|
| Without Aug | 0.05777 | 0.07399 | 100 | 100 | 3.53 |
| With Aug | 0.05598 | 0.07399 | 100 | 100 | 3.65 |

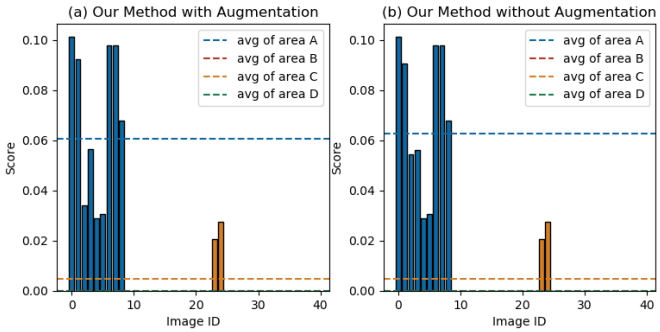



Fig. 10: Similarity scores in the robustness test when A is the true loop closure area.

### *E. Test on Other Environments*

To test whether our method can be applied in other environments, we conduct experiments on the Sequence

'freiburg1_room' of TUM Dataset [21], which is an indoor office environment and it is well suited for evaluating loop closure. Using a design similar to the previous experiments, select four small areas A, B, C, and D on the TUM data sequence, as shown in Fig. 11. The similarity scores of our method and the method in VINS-Mono are shown in Fig. 12. Table I illustrates the quantitative results. It can be seen that the our method can reach high precision with acceptable recall. In addtion, the HSD obtained by our method is significantly higher, which means that it is much easier for us to set the threshold in practical use. Therefore, the proposed method has the potential to be applied on scenarios other than narrow pipes.

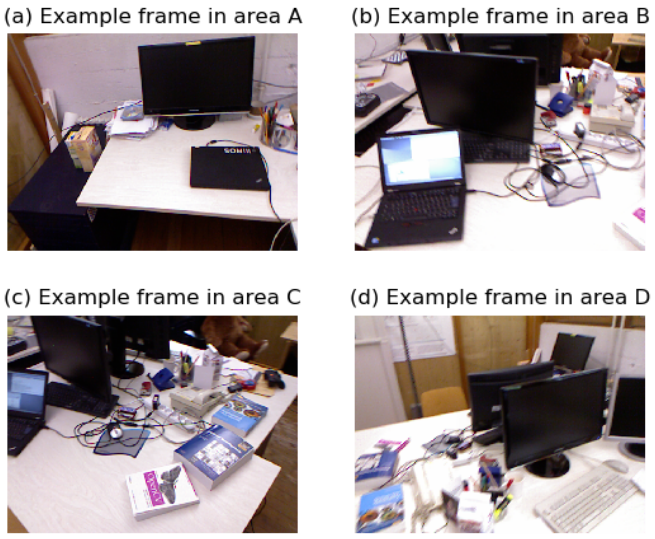



Fig. 11: Frames in different areas of TUM dataset.

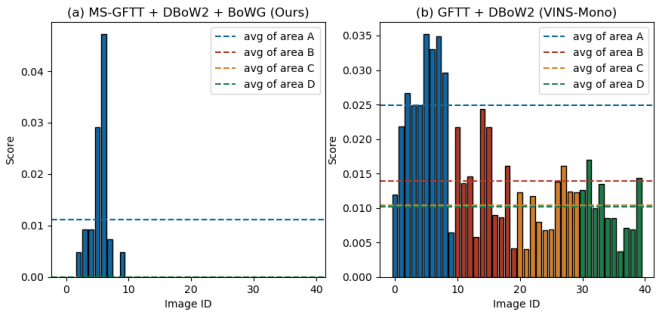



Fig. 12: Similarity scores of the TUM dataset when A is the true loop closure area.

TABLE III: experiment results on TUM dataset

| Methods\Results | ASD | HSD | Precision (70% recall) | Runtime (s) |
|---|---|---|---|---|
| **MS-GFTT + DBoW2 + BoWG (Ours)** | **0.01119** | **0.04714** | **100** | 3.37 |
| GFTT + DBoW2 (VINS-Mono) | 0.01098 | 0.01081 | **100** | **0.87** |

## V. CONCLUSION AND DISCUSSION

This paper presents a novel module called Bag-of-Word-Groups (BoWG) to improve the precision of loop closure detection. The concept of word group is defined based on the spatial proximity and positions of features. We first utilize Gaussian pyramid to implement Multi-scale Good Features To Track (MS-GFTT) to detect richer features at various scales for word group analysis. After that, BoWG is built and updated online for loop detection, which does not require any extra sensor other than a monocular visual camera and can be easily integrated into existing bag-of-words methods based on the framework proposed in this paper. Experiments show that our method can achieve much higher precision with acceptable recall and better distinguish the loop closure area compared to previous methods when perceptual aliasing is serious. Besides, robustness tests illustrate that the proposed method is robust and can cope with small changes in the environment. In addition, our method has the potential to be applied to scenarios other than narrow pipes.

While our method offers many advantages, there are several limitations. First, the proposed BoWG module may fail if the viewpoint changes greatly when the robot returns to the previous position. This is because excessive viewpoint changes lead to large displacements of word groups in the images, making it difficult to accurately identify word groups based on a fixed distance threshold. Therefore, adapting to significant viewpoint changes is a crucial issue for future research. Additionally, the proposed method is affected by multiple hyperparameters (the block size of GFTT, the number of layers in the image pyramid, and the distance threshold for identifying word groups). Improving the stability of our method with respect to the hyperparameters is also an interesting future topic. For the next phase of our work, we plan to incorporate the proposed loop closure detection module into VLI-SLAM to increase the localization accuracy and mapping consistency.

## ACKNOWLEDGMENT

This work was supported by the Department of Energy (DOE)'s Advanced Research Projects Agency-Energy (ARPA-E), REPAIR Program; Carnegie Mellon University Robotics Institute Summer Scholar (RISS) program; the Biorobotics Lab; Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS); and the Chinese University of Hong Kong, Shenzhen (CUHK, Shenzhen). The authors would like to thank Rachel Burcin and Prof. John Dolan for their support and organization of the RISS program.

## REFERENCES

[1] High performance loop closure detection using bag of word pairs. *Robotics and Autonomous Systems*, 77:55–65, 2016.

[2] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 24(5):1027–1037, 2008.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[4] J. Bergin. *Sets, Maps, Multisets, and MultiMaps*, pages 239–266. Springer New York, New York, NY, 1998.

[5] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 778–792, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[6] D. Cheng, H. Shi, A. Xu, M. Schwerin, M. Crivella, L. Li, and H. Choset. Visual-laser-inertial slam using a compact 3d scanner for confined space. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5699–5705, 2021.

[7] D. Gálvez-López and J. D. Tardós. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, October 2012.

[8] A. Gunatilake, S. Kodagoda, and K. Thiyagarajan. Battery-free uhf-rfid sensors-based slam for in-pipe robot perception. *IEEE Sensors Journal*, 22(20):20019–20026, 2022.

[9] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003.

[10] M. Labbé and F. Michaud. Online global loop closure detection for large-scale multi-session graph-based slam. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2661–2666, 2014.

[11] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 81(2):155–166, Feb 2009.

[12] T. Lindeberg. Scale invariant feature transform, 2012. QC 20120524.

[13] R. Muñoz-Salinas and R. Medina-Carnicer. Ucoslam: Simultaneous localization and mapping by fusion of keypoints and squared planar markers. *Pattern Recognition*, 101:107193, 2020.

[14] T. Nicosevici and R. Garcia. Automatic visual bag-of-words for online robot navigation and mapping. *IEEE Transactions on Robotics*, 28(4):886–898, 2012.

[15] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2161–2168, 2006.

[16] T. Qin, P. Li, and S. Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[17] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011.

[18] Z. Shang and Z. Shen. Dual-function depth camera array for inline 3d reconstruction of complex pipelines. *Automation in Construction*, 152:104893, 2023.

[19] J. Shi and Tomasi. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[20] Sivic and Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, 2003.

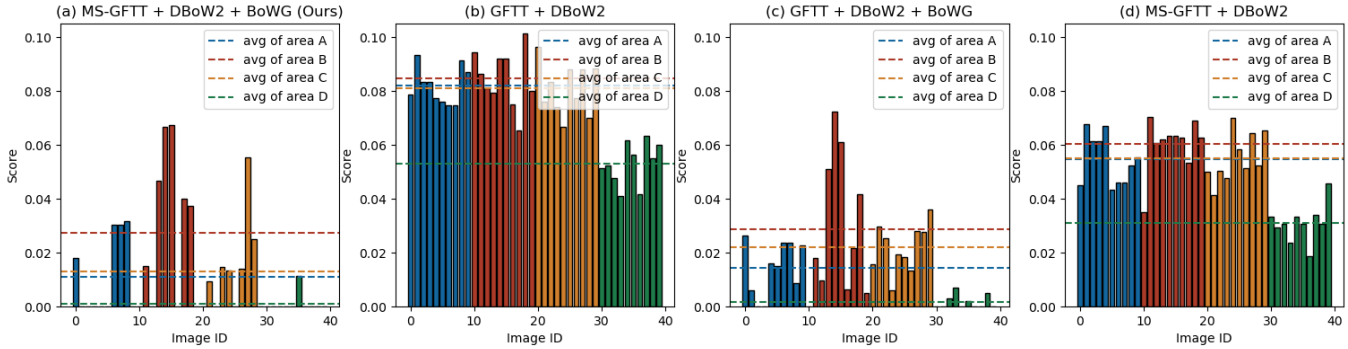
[21] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[22] H. Zhang, B. Li, and D. Yang. Keyframe detection for appearance-based visual slam. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2071–2076. IEEE, 2010.

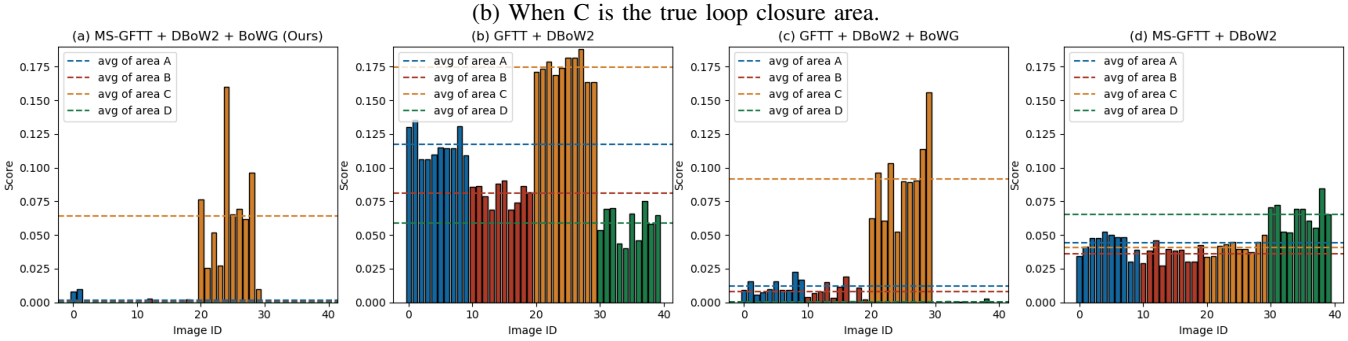
## SUPPLEMENTARY MATERIALS

### *A. Results when B, C, D is the true loop closure area*

(a) When B is the true loop closure area.




(b) When C is the true loop closure area.




(c) When D is the true loop closure area.

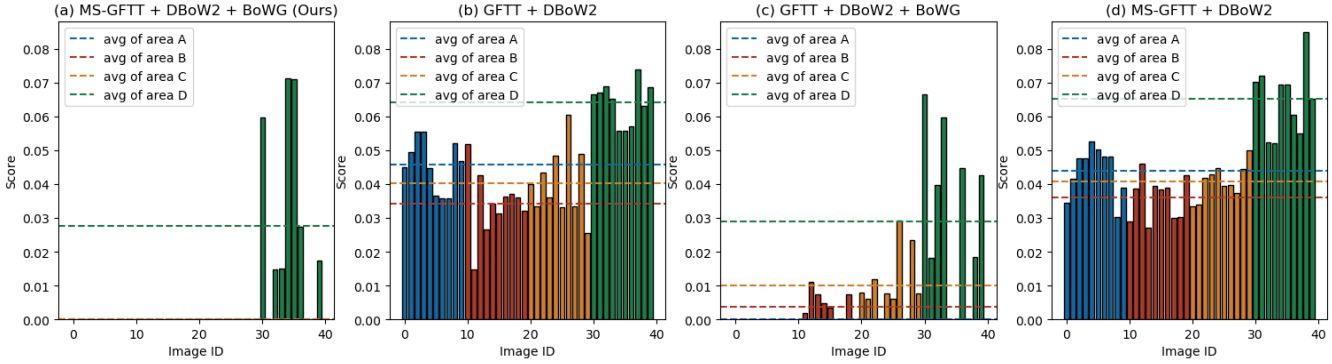
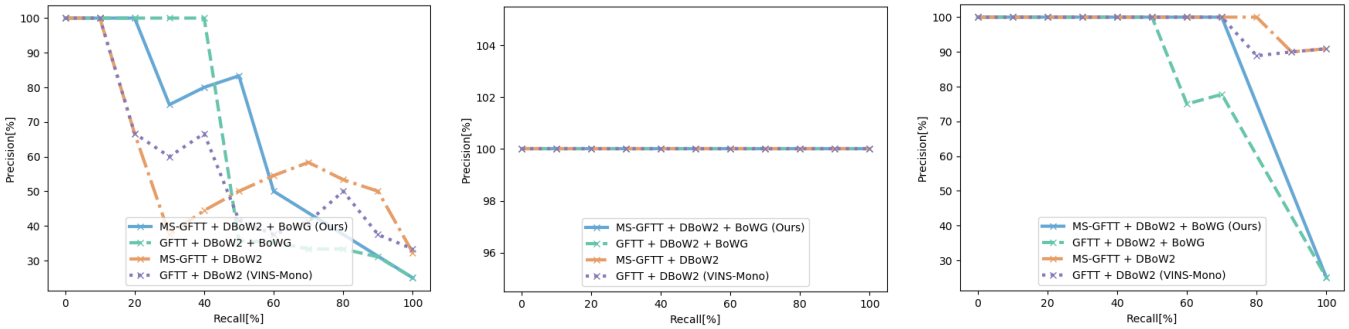
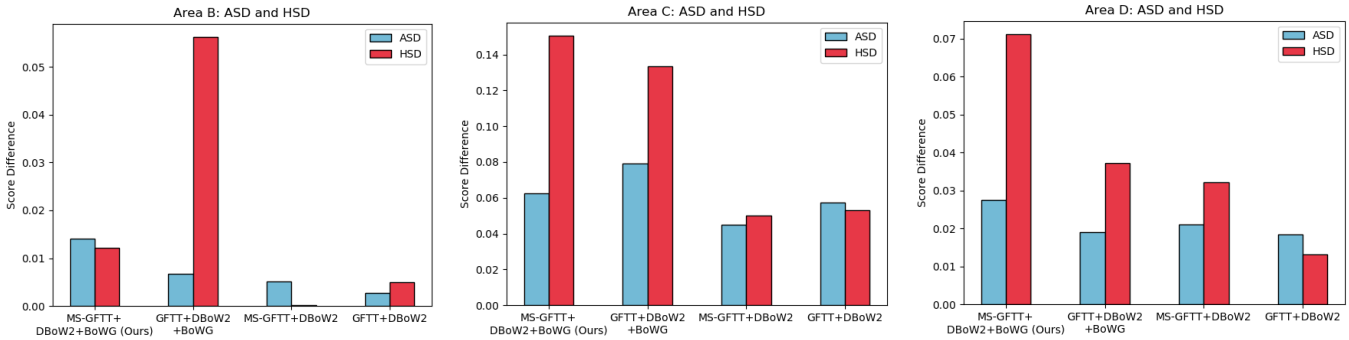



Fig. 13: Similarity scores of different methods when B, C, D is the true loop closure area.




(a) When B is the true loop closure area. (b) When C is the true loop closure area. (c) When D is the true loop closure area.

Fig. 14: Precision and recall of different methods when B, C, D is the true loop closure area.




(a) When B is the true loop closure area. (b) When C is the true loop closure area. (c) When D is the true loop closure area.

Fig. 15: ASD and HSD of different methods when B, C, D is the true loop closure area.

TABLE IV: Experiment results of different methods when B, C, D is the true loop closure area

| Results \Algorithm | **MS-GFTT + DBoW2 + BoWG (Proposed)** | GFTT + DBoW2 (VINS-Mono) | GFTT + DBoW2 + BoWG | MS-GFTT + DBoW2 |
|---|---|---|---|---|
| ASD | **0.01413** | 0.00274 | 0.00671 | 0.00507 |
| HSD | 0.01211 | 0.00491 | **0.05623** | 0.00013 |
| Precision (50% recall) | **83.3** | 41.7 | 35.7 | 50.0 |
| Runtime (s) | 3.58 | **2.59** | 5.20 | 2.79 |

(a) When B is the true loop closure area

| Results \Algorithm | **MS-GFTT + DBoW2 + BoWG (Proposed)** | GFTT + DBoW2 (VINS-Mono) | GFTT + DBoW2 + BoWG | MS-GFTT + DBoW2 |
|---|---|---|---|---|
| ASD | 0.06252 | 0.05726 | **0.07924** | 0.04482 |
| HSD | **0.15040** | 0.05286 | 0.13350 | 0.05016 |
| Precision (50% recall) | **100** | **100** | **100** | **100** |
| Precision (90% recall) | **100** | **100** | **100** | **100** |
| Runtime (s) | 3.65 | **2.58** | 5.26 | 2.87 |

(b) When C is the true loop closure area

| Results \Algorithm | **MS-GFTT + DBoW2 + BoWG (Proposed)** | GFTT + DBoW2 (VINS-Mono) | GFTT + DBoW2 + BoWG | MS-GFTT + DBoW2 |
|---|---|---|---|---|
| ASD | **0.02762** | 0.01851 | 0.01895 | 0.02115 |
| HSD | **0.07117** | 0.01328 | 0.03721 | 0.03218 |
| Precision (70% recall) | **100** | **100** | 77.8 | **100** |
| Runtime (s) | 3.59 | **2.43** | 5.32 | 2.91 |

(c) When D is the true loop closure area