Proceeding of the
International Conference on Information and Automation
Lanzhou, China, August 2025

# Region-based Initialization and Spatial-constrained Loop Closure Detection for Efficient Visual-inertial SLAM

Shuyue Lin and Yuxiang Sun

2025 International Conference on Information and Automation (ICIA) | 979-8-3315-2370-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICIA64617.2025.11277738

***Abstract*— This paper presents a robust and efficient visual-inertial simultaneous localization and mapping (SLAM) system, which includes a region-based initialization strategy and a lightweight Mobile-NetVLAD loop closure detection (LCD) module. Specifically, the proposed initialization strategy divides a prior map into structured regions, which facilitates efficient candidate region retrieval and robust initialization across complex environments. For the LCD module, the lightweight Mobile-NetVLAD network is employed to extract keyframe descriptors. To improve the LCD accuracy, historical keyframes with significant position differences are filtered out by spatial constraints. We integrate the above modules into VINS, and evaluate the SLAM system on the public EuRoC dataset and a self-collected dataset. Experimental results demonstrate that the robustness of LCD and localization accuracy is superior to state-of-the-art methods, while maintaining real-time performance.**

## I. Introduction

Simultaneous localization and mapping (SLAM) is a fundamental capability for robots [1]–[5]. In recent years, visual-inertial SLAM (VI-SLAM) and visual-inertial odometry (VIO) have gained widespread attention [6]. The IMU sensor can provide scale and motion information for a pure monocular VIO system, while visual constraints can compensate for the integration drift of IMU when it is stationary. The accuracy of localization can be improved by fusing data from the two sensors. Although VI-SLAM has made remarkable progress in terms of localization performance, they are still faced with two challenges in practical applications, which include the lack of initialization robustness and the generalization ability to detect loop closures.

The use of both visual and inertial sensors increases the dimensionality of system states [7], [8]. Specifically, a typical VIO/VI-SLAM system at least includes 15 degrees-of-freedoms, involving position, orientation, velocity, and bias of the gyroscope and accelerometer, etc. To obtain a stable estimate of the initial state, the system typically requires a number of accumulated observations, which consequently reduces the start-up speed [9]. In addition, most existing initialization methods rely on local observations and do not use global prior maps, making them less generally robust to changing environments.

The existing mainstream VI-SLAM systems [10] currently use the Bag-of-Words (BoW) [11] models for loop closure detection (LCD). Although the method is widely applied, it relies heavily on matching local features among several frames. It is sensitive to illumination, occlusion, and viewpoint changes, which in practice easily lead to false loop detections. In addition, the traditional vocabulary module is built offline and is strongly coupled with training data, making it difficult to adapt to new environments. Thus, it hinders the generalization capability of localization systems. Compared to BoW models, deep learning models, such as NetVLAD [12], are more powerful in describing scene appearances. But deep learning models may still generate wrong matches between different locations with similar visual appearances, leading to failures in LCD.

To provide a solution to the aforementioned issues, we propose a robust and efficient VI-SLAM system featuring a region-based initialization strategy and a lightweight LCD module based on Mobile-NetVLAD [13]. Our code is open-sourced[1]. The main contributions of this work are summarized as follows.

1) We present a region-based initialization method, in which an environment is divided into multiple structured regions. This design allows efficient retrieval of candidate regions and facilitates robust initialization.
2) We propose a LCD strategy, which integrates the Mobile-NetVLAD model and combines spatial consistency constraints. By filtering out keyframes with significant spatial deviations, the system improves both the accuracy and efficiency of LCD.
3) We develop a complete SLAM system by integrating the initialization and LCD components into VINS [14]. The experimental results on both the EuRoc dataset [15] and our self-collected dataset show improvements in localization accuracy.

## II. Related Work

### A. Visual-Inertial Odometry/SLAM

Visual-inertial odometry/SLAM systems [16] are generally divided into tightly coupled [17]–[20] and loosely coupled [21]–[23]. The advantages of loosely coupled systems lie in their simple structure and strong modularity. They model visual and inertial measurements separately and combine their estimation results during the back-end stage. This design limits the interaction between the two types of sensors, making it difficult to fully take advantage of their complementary information. They also encounter issues such as timestamp misalignment, which degrades the overall

This work was supported in part by Hong Kong Research Grants Council under Grant 15222523, and in part by City University of Hong Kong under Grant 9231601. *(Corresponding author: Yuxiang Sun.)*

Shuyue Lin and Yuxiang Sun are with City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong (e-mail: shuyue.lin@cityu.edu.hk; yx.sun@cityu.edu.hk, sun.yuxiang@outlook.com).

[1]https://github.com/lab-sun/RSC-VINS

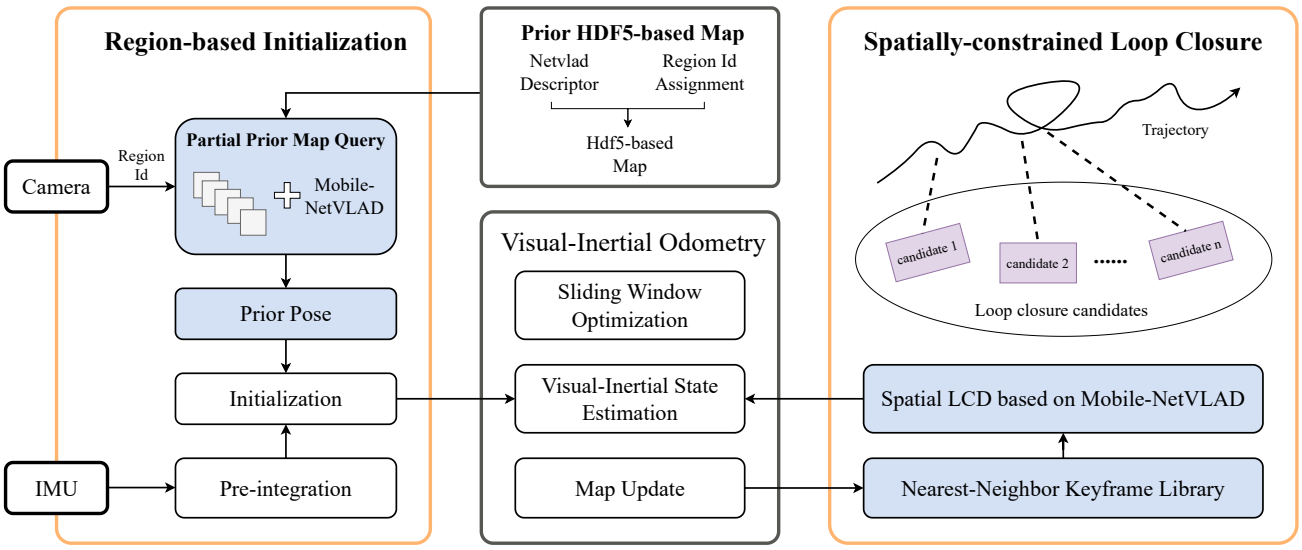



Fig. 1: The overall structure of our proposed system. It takes as input image sequences and IMU data, and outputs state estimation. Our system is built on VINS [14]. The structure of our system mainly consists of the region-based initialization module, the sliding-window optimization-based state estimator, and the Mobile-NetVLAD-based LCD module.

localization performance. In contrast, tightly coupled systems jointly model visual and inertial measurements during back-end optimization. They can obtain more accurate and robust state estimation, and have become the mainstream solution in this filed. However, the increased nonlinearity of the localization system makes the initialization process more difficult. In most cases, some observations are required to achieve a stable initial estimate.

### B. Initialization

To achieve robust and accurate initialization, many studies rely on nonlinear optimization frameworks, where residuals are minimized through a sliding window to estimate the full state. VINS-Mono [14] combines sliding-window optimization with a loosely coupled visual-inertial framework. ORB-SLAM2 [24] utilizes IMU pre-integration and optimization to estimate biases, gravity, and scale. Evangelidis *et al.* [25] proposed closed-form initialization via direct triangulation. However, these methods often require long observations and show poor adaptability to changing environments, resulting in degraded performance in low-texture or occluded scenarios. Recent works [26], [27] introduced prior maps by matching keyframes and traditional descriptors from previously explored environments, enabling faster and more stable initialization. Yet, traditional local descriptors are sensitive to lighting and dynamic elements, leading to incorrect matches. Thus, prior map–based initialization still needs improvement.

### C. Loop Closure Detection

As a key component in SLAM, LCD has received wide attention. It can eliminate accumulated localization drifts during long-term operations. Many existing methods [28] adopt the BoW model to achieve LCD through local feature matching. The method has the advantages of high computational efficiency and easy deployment. However, the model is sensitive to illumination, viewpoint changes, and occlusions, which may easily generate false detections. In recent years, global descriptor LCD methods based on deep learning have been proposed [29], and many localization systems have introduced data-driven LCD models to improve detection performance. For example, D²SLAM [30] combines SuperPoint [31] and NetVLAD models, which improves performance but has a high hardware dependency on CUDA acceleration. The NetVLAD-based LCD used in [32] does not utilize spatial or motion constraints, and the descriptor matching is completely open, leading to an excessive matching range and an increased risk of mismatches. So, how to improve the accuracy and robustness of deep learning-based LCD while ensuring a low computational cost remains an open question.

To overcome the aforementioned challenges, we propose a VI-SLAM system that features the region-based initialization strategy and the spatially constrained loop closure mechanism. The system structurally partitions prior maps to enable efficient candidate retrieval and robust initialization. Meanwhile, it integrates the spatial consistency filter with a lightweight Mobile-NetVLAD descriptor to achieve accurate and efficient loop closure, which ensures reliable localization in complex environments.

## III. The Proposed Method

### A. The Overall Architecture

The structure of the whole SLAM system is shown in Fig. 1. We proposes a new region-based initialization method, and a LCD method based on Mobile-NetVLAD loop detection with the spatial consistency. The input to the SLAM system includes continuous image sequences and high-frequency IMU data. During initialization, the system pre-loads the relevant map region according to a given prior region ID. If no region ID is provided, the whole global

map is loaded. After that, the NetVLAD descriptor of the current image frame is used to search for similar frames in the loaded map. We use their poses as the prior positions for system initialization. Meanwhile, the LCD thread removes false loop candidates that appear visually similar in the images but are distant in physical space. Finally, Mobile-NetVLAD is applied for similarity matching to perform accurate and robust loop correction, which improves the localization accuracy.

### *B. The Region-based Initialization*

To provide robust and accurate initialization for VIO/VI-SLAM systems, this paper proposes a region-based initialization method, which is combined with a feature searching method using multi-threading. This method improves the initialization speed and robustness of the localization system. In the prior map construction stage, the system collects images and their corresponding pose data and assigns each geographic region a unique region ID. Each region map patch stores the NetVLAD features and the poses of all images in this region separately. This information is saved in an HDF5 file structure, where each region ID corresponds to a group in the file, which is convenient for loading data when needed.

During the initialization of the system, if a prior region ID is obtained, the system only loads the corresponding region map patch, which greatly reduces the scope of similar frame retrieval and significantly improves overall matching efficiency. If a prior ID is not available, the system loads the global map to ensure the accuracy and robustness of the initialization process. In order to achieve fast loading and efficient retrieval of complex region maps, we develop a multi-threaded system for NetVLAD descriptor loading and index data construction. This module reads and efficiently normalizes the descriptors and poses of each keyframe in parallel, which reduces the data processing delay. The method balances the scalability of complex maps with the real-time performance of the initialization process, and provides the system with a robust and accurate initialization pose even under challenging conditions. The detailed algorithmic procedure is illustrated in Algorithm 1.

**Algorithm 1:** Region-based Initialization

**Input:** Region ID $\mathcal{A}$, query descriptor $\mathbf{v}_q$, map file $\mathcal{H}$
**Output:** Matched pose $\mathbf{T}^*$ or failure
**1. Parallel Loading and Normalization**;
Initialize $\mathcal{D}, \mathcal{I}, \mathcal{T} \leftarrow \emptyset$;
**foreach** $I_i \in \mathcal{A}$ **do**
  $\mathbf{v}_i \leftarrow \mathcal{H}[\mathcal{A}][I_i][\texttt{netvlad}]$;
  **if** $\|\mathbf{v}_i\|_2 > 0$ **then**
    $\hat{\mathbf{v}}_i \leftarrow \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2}$;
    $\mathbf{T}_i \leftarrow \mathcal{H}[\mathcal{A}][I_i][\texttt{pose}]$;
    Append to $\mathcal{D}, \mathcal{I}, \mathcal{T}$;
**2. Build Index**;
Initialize index $\mathcal{I}_{\text{index}}$;
**foreach** $\hat{\mathbf{v}}_i \in \mathcal{D}$ **do**
  $\mathcal{I}_{\text{index}}.\texttt{add}(\hat{\mathbf{v}}_i)$;
**3. Query**;
$\hat{\mathbf{v}}_q \leftarrow \frac{\mathbf{v}_q}{\|\mathbf{v}_q\|_2}$;
$\{\texttt{idx}_j\}, \{\texttt{dist}_j\} \leftarrow \mathcal{I}_{\text{index}}.\texttt{search}(\hat{\mathbf{v}}_q, k)$;
$j^* \leftarrow \arg\min_j \texttt{dist}_j$;
**return** $\mathbf{T}^* \leftarrow \mathcal{T}[\texttt{idx}_{j^*}]$;

### *C. The Spatially Consistent LCD*

The original VINS framework adopts the BoW model for image description and matching. Our proposed system replaces BoW with Mobile-NetVLAD. Although NetVLAD can effectively generate global descriptors of images to determine whether an image has been re-observed, it may still make mistakes when different images have similar appearances. Thus, LCD based on NetVLAD is prone to mismatches in regions with different locations but similar visual appearances. To alleviate this problem, we present a LCD strategy with spatial consistency constraints, which considers the visual similarity of images in relation to their spatial locations.

During system operation, the LCD thread continuously maintains the NetVLAD descriptors of keyframes and their corresponding position information. The system takes the position of the keyframe to be detected as the center and retrieves possible loop candidate frames only within a certain spatial range around it. During candidate selection, the system considers the similarity between the candidate frame and the current frame, as well as the distance between their positions. Based on this constraint, the loop thread can determine whether a valid loop closure is formed, thus significantly reducing the false detection rate in complex scenarios such as those with repetitive structures.

Specifically, when the current keyframe $F_i$ enters the LCD stage, the system first extracts its corresponding NetVLAD descriptor $\mathbf{v}_i \in \mathbb{R}^{4096}$, and stores the descriptor along with its associated 3D position $\mathbf{p}_t$ into a cache. To reduce false positives caused by visual similarity, the system adopts the spatial consistency constraint. Only those historical keyframes whose euclidean distance to the current keyframe is less than a predefined threshold $\tau$ are considered as candidates, while also excluding temporally adjacent frames to effectively suppress false matches between consecutive frames. The candidate set $\mathcal{C}_t$ is defined as:

$$\mathcal{C}_t = \left\{ (\mathbf{v}_t, \mathbf{v}_i) \mid \|\mathbf{p}_i - \mathbf{p}_t\| < \tau,\ t < i - \delta \right\}, \tag{1}$$

where $\delta$ denotes the minimum frame interval used to prevent repeated matching between nearby frames. For each candidate frame in the set, its descriptor $\mathbf{v}_i$ is compared with the descriptor of the current frame $\mathbf{v}_t$ to compute a similarity score based on their inner product:

$$s_i = \langle \mathbf{v}_t, \mathbf{v}_i \rangle = \sum_{j=1}^{4096} v_{t,j} \cdot v_{i,j}, \tag{2}$$

where $s_i$ represents the similarity score between the current frame $F_t$ and candidate frame $F_i$. If the highest similarity

TABLE I: Time cost (ms) for loading and retrieval of prior maps with 100 images.

| Method | Map Loading | Single Image Retrieval |
|---|---|---|
| time | 10 | < 2 (4 threads) |

TABLE II: Comparison of loop closure time (ms) between Mobile-NetVLAD and the traditional BoW method.

| Method | Descriptor Computation | Loop Detection | Total |
|---|---|---|---|
| DBoW2 | 7.767 | 7.368 | 15.041 |
| NetVLAD | 22.047 | 1.364 | 23.411 |

score $s_{i^*}$ exceeds a predefined threshold $\theta$, the system considers the corresponding pair as a valid LCD result. Otherwise, it is treated as a non-match.

## IV. Experimental Results and Discussions

We conduct experiments on the public EuRoC dataset and a self-collected dataset. We first perform a statistical analysis of the time cost for the prior map loading and LCD based on Mobile-NetVLAD. Then, we compare our system with the current mainstream SLAM frameworks on the EuRoC dataset. Finally, we test our system on a self-collected handheld dataset to validate its effectiveness in real-world applications. All the algorithms run on a computer with an Intel Core i7-14700F CPU and without GPU parallelization.

### A. Efficiency Evaluation

To validate the real-time performance, we conduct detailed time cost statistics of the prior map loading and retrieval during the initialization stage, as well as the LCD method. Specifically, the map loading and retrieval test is based on a prior map containing 100 images. As shown in Tab. I, the time cost to load 100 images is about 10 ms. With four threads running, the retrieval time of a single image is maintained within 2 ms. Note that the map loading is completed before the localization system starts, so it does not affect the online performance of the whole system. Thus, the retrieval time meets the requirements of a real-time system.

Tab. II shows the time cost between Mobile-NetVLAD and the traditional vocabulary-based method in LCD. Although the descriptor computation time of Mobile-NetVLAD is higher than that of the traditional method and accounts for the majority of the LCD time, the retrieval stage requires significantly less time cost, demonstrating an effective global matching capability and efficient candidate filtering. On the whole, although Mobile-NetVLAD has a higher time cost in the feature extraction process, the overall LCD can still maintain near real-time performance due to its strong descriptive ability and high detection efficiency.

### B. Comparative Experiments

The EuRoC dataset includes 11 sequences, covering challenging scenarios such as nighttime and low textureness. In the experiments, we name and compare the systems under different module combinations. Specifically, *Ours-prior* denotes a system that uses region-based initialization but does not use the Mobile-NetVLAD LCD. *Ours-loop* represents a system that uses the Mobile-NetVLAD LCD but does not introduce the proposed initialization. *Ours* integrates both of these strategies. We compare these three configurations with VINS-Mono [14] on the EuRoC to evaluate the impact of each strategy on system performance.

TABLE III: The RMSE (m) of localization accuracy for VINS-Mono and *Ours-prior* in the first 5 seconds of each sequence. ↓ means reduced error, and ↑ means increased error. Best in bold, second-best underlined.

| Dataset | VINS-Mono | Ours-prior | Improved |
|---|---|---|---|
| MH_01_easy | 0.0093 | **0.0072** | 23.02% ↓ |
| MH_02_easy | 0.0095 | **0.0083** | 12.23% ↓ |
| MH_03_medium | **0.0099** | 0.0107 | 8.64% ↑ |
| MH_04_medium | 0.0161 | **0.0128** | 20.03% ↓ |
| MH_05_difficult | **0.0146** | 0.0192 | 23.91% ↑ |
| V1_01_easy | 0.0097 | **0.0067** | 31.37% ↓ |
| V1_02_medium | 0.0152 | **0.0150** | 0.74% ↓ |
| V1_03_difficult | 0.0145 | **0.0132** | 8.74% ↓ |
| V2_01_easy | 0.0068 | **0.0060** | 11.94% ↓ |
| V2_02_medium | 0.0207 | **0.0133** | 35.57% ↓ |
| V3_03_difficult | **0.0143** | 0.0154 | 8.18% ↑ |

TABLE IV: The RMSE (m) of localization accuracy for VINS-Mono, *Ours-loop*, and *Ours* on the EuRoC. Best in bold, second-best underlined.

| Dataset | VINS-Mono | Ours-loop | Ours |
|---|---|---|---|
| MH_01_easy | 0.0690 | **0.0582** | 0.0597 |
| MH_02_easy | 0.0696 | 0.0536 | **0.0472** |
| MH_03_medium | **0.0749** | 0.0917 | 0.0835 |
| MH_04_medium | 0.1330 | 0.1170 | **0.1022** |
| MH_05_difficult | 0.1462 | 0.1351 | **0.1315** |
| V1_01_easy | 0.0693 | 0.0505 | **0.0427** |
| V1_02_medium | 0.0695 | 0.0574 | **0.0545** |
| V1_03_difficult | 0.1380 | **0.1293** | 0.1333 |
| V2_01_easy | **0.0470** | 0.0649 | 0.0675 |
| V2_02_medium | 0.0982 | **0.0865** | 0.0881 |
| V3_03_difficult | 0.2120 | 0.1566 | **0.1538** |

For the evaluation of the initialization strategy, only the first 5 seconds of each sequence are chosen for the calculation. This is because the system completes its initialization during this period. The pre-run period reflects the short-term stability and accuracy of the initialization, so as to avoid the cumulative errors from interfering with the evaluation of the initialization results. The results of the specific accuracy comparison are shown in Tab. III. In most of the sequences, the accuracy of the region-based initialization is higher than that of VINS-Mono. For example, in sequences V1_01_easy and V2_02_medium, the accuracy of the first 5-second trajectory is improved by 31.37% and 35.57% compared with that of VINS, respectively. Meanwhile, The overall accuracy on EuRoC is improved by 8.68%. The box plot of the first 5-second trajectories for all sequences is shown in Fig. 2, and *Ours-prior* performs better in the most of the sequences.

Tab. IV displays the evaluation results on each whole trajectory of EuRoC. *Ours-loop* and *Ours* are generally better than VINS-Mono in terms of localization accuracy. The MH_05_difficult sequence is recorded in a black night scene. *Ours* shows a significant localization accuracy improvement (roughly 10.08%) in this scene. In V1_01_easy sequence,

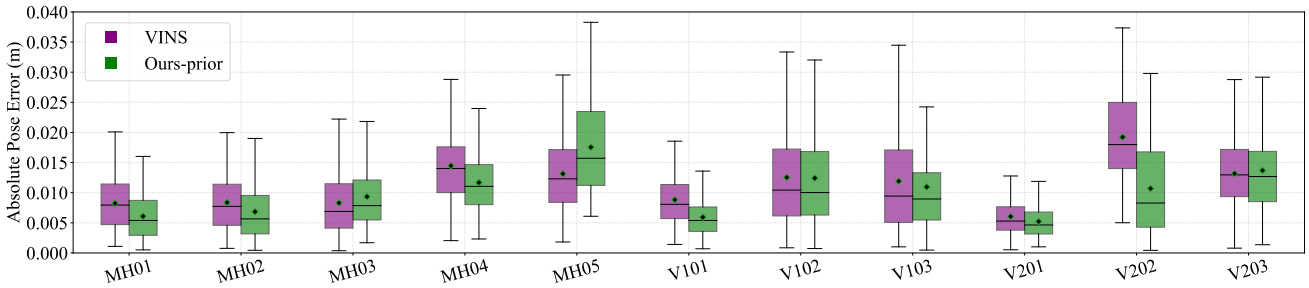



Fig. 2: Boxplot of region-based initialization. Most sequences show reduced errors, indicating that the proposed initialization method performs better.

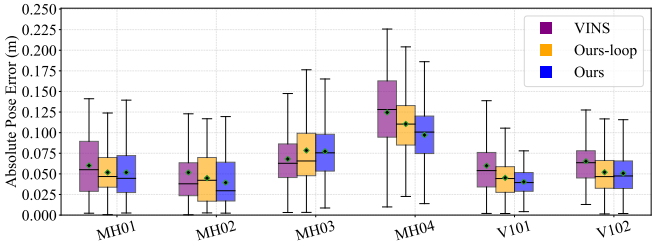



Fig. 3: Boxplot of VINS-Mono, *Ours-loop* and *Ours* in whole trajectory on EuRoC. Overall, *Ours-loop* yields slightly lower errors than VINS-Mono, and *Ours* performs marginally better than *Ours-loop*.

where the sequence is recorded in a room containing white walls, both *Ours-loop* and *Ours* localization accuracy have been improved by 27.08% and 38.26%. Fig. 3 displays the box plots of some of the sequences. Different initializations lead to varying error propagation, causing differences in loop closure pose estimates and thus resulting in varying trajectory optimization outcomes. Therefore, the performance of *Ours-loop* is better than *Ours* for some sequences in Tab. IV.

### C. Practical Test

Three trajectories with approximate lengths of 979.30 m, 184.74 m, and 247.74 m are captured in real environments using an Intel RealSense D435i camera. As shown in Fig. 4, each trajectory starts from a specific location and eventually returns to its corresponding starting point, forming a complete closed loop. This setup provides a practical scenario to evaluate the performance of LCD in the absence of real ground truth for trajectories as a reference. We evaluate the overall robustness and accuracy of the proposed system according to whether it can successfully complete LCD as well as the localization accuracy when returning to the starting point. The experimental results show that the proposed method not only detects loop closure more quickly but also corrects the accumulated error more accurately.

Tab. V compares the Loop Closure Median Drift (LCMD) between our method and VINS across three scenarios. We extract the first and last N frames from the trajectory. For each segment, we compute its geometric center and then calculate the coordinate-wise median of these points to obtain robust start and end positions. LCMD is defined as the Euclidean distance between these two robust positions. In particular, in

TABLE V: The LCMD (m) for 5, 10, 20 frames, and the trajectory lengths in the three scenes.

| Scene | Length (m) | Method | LCMD (5) | LCMD (10) | LCMD (20) |
|---|---|---|---|---|---|
| Scene 1 | 979.30 | VINS | 25.633 | 25.650 | 25.670 |
| | | Ours | 0.414 | 0.456 | 0.477 |
| Scene 2 | 184.74 | VINS | 1.788 | 1.044 | 1.236 |
| | | Ours | 0.193 | 0.206 | 0.201 |
| Scene 3 | 247.74 | VINS | 2.471 | 2.429 | 2.377 |
| | | Ours | 0.425 | 0.350 | 0.362 |

Scene 1 with long-distance loop closure, the drift of VINS-Mono is more than 25 meters, while our method maintains the error within 0.5 meters, which fully demonstrates the high accuracy and strong convergence ability of the proposed loop closure strategy. In shorter trajectories, such as Scenarios 2 and 3, although the error of VINS-Mono is already relatively small, our method still shows clear advantages. This shows that the system has excellent robustness and stability in both extreme cases and regular medium-distance trajectories. In summary, the experimental results verify that the proposed loop closure strategy can achieve reliable and accurate LCD and drift correction in various environments.

## V. Conclusions and Future Work

This paper presents a robust and efficient VI-SLAM system featuring a region-based initialization and a lightweight Mobile-NetVLAD loop closure module. Experimental results demonstrate our superior localization accuracy and real-time performance compared to existing methods. Future work will focus on network optimization and system evaluation to more complex environments.

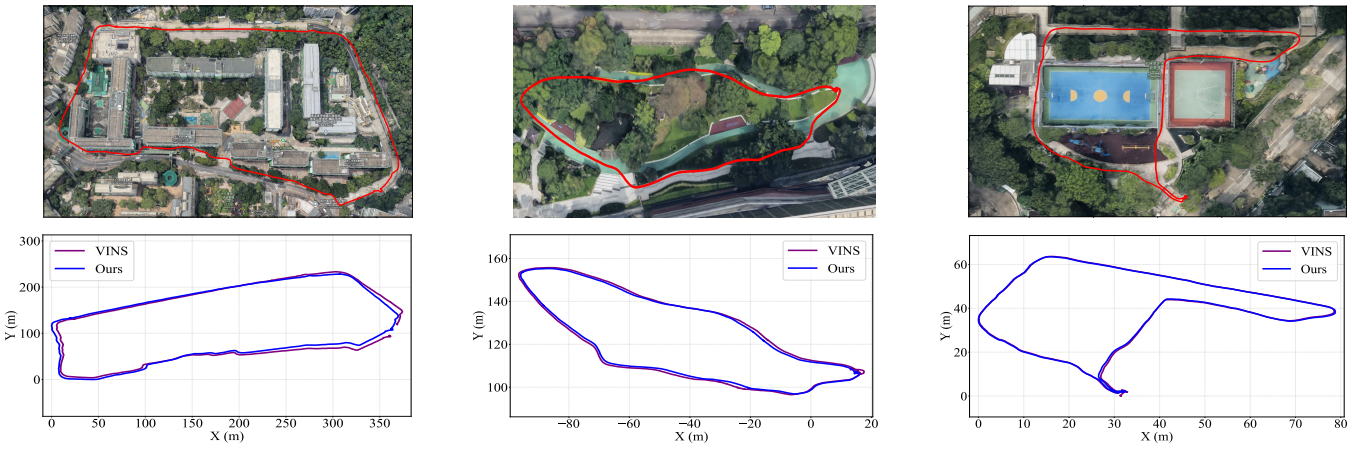
## References

[1] J. Ruan, B. Li, Y. Wang, and Y. Sun, "Slamesh: Real-time lidar simultaneous localization and meshing," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 3546–3552.

[2] Y. Sun, M. Liu, and M. Q.-H. Meng, "Improving rgb-d slam in dynamic environments: A motion removal approach," *Robotics and Autonomous Systems*, vol. 89, pp. 110–122, 2017.

[3] H. Huang, H. Ye, Y. Sun, L. Wang, and M. Liu, "Incorporating learnt local and global embeddings into monocular visual slam," *Autonomous Robots*, vol. 45, no. 6, pp. 789–803, 2021.

[4] Y. Sun, M. Liu, and M. Q.-H. Meng, "Motion removal for reliable rgb-d slam in dynamic environments," *Robotics and Autonomous Systems*, vol. 108, pp. 115–128, 2018.




Scene 1 Scene 2 Scene 3

Fig. 4: Trajectories overlaid on Google Maps (top) and estimated trajectories (bottom). The three scenarios are collected in open areas near City University of Hong Kong, Kowloon, Hong Kong. The red curves are manually-depicted reference trajectories (not based on GPS or any other ground-truth data). Therefore, the LCMD to evaluate the quality of LCD.


[5] H. Huang, Y. Sun, H. Ye, and M. Liu, "Metric monocular localization using signed distance fields," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1195–1201.

[6] A. Samadzadeh and A. Nickabadi, "Srvio: Super robust visual inertial odometry for dynamic environments and challenging loop-closure conditions," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 2878–2891, 2023.

[7] Y. He, B. Xu, Z. Ouyang, and H. Li, "A rotation-translation-decoupled solution for robust and efficient visual-inertial initialization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 739–748.

[8] J. He, M. Li, Y. Wang, and H. Wang, "Ple-slam: A visual-inertial slam based on point-line features and efficient imu initialization," *IEEE Sensors Journal*, 2025.

[9] H. Huang, Z. Liu, L. Zhang, D. Wang, and J. Wang, "Via-slam: An underwater visual-inertial-acoustic slam with integrated dvl," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[10] H. Zhang, J. Huo, Y. Huang, and Q. Liu, "Real-time dynamic visual-inertial slam and object tracking based on lightweight deep feature extraction matching," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[11] C.-F. Tsai, "Bag-of-words representation in image annotation: a review," *International Scholarly Research Notices*, vol. 2012, no. 1, p. 376804, 2012.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[13] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, "Leveraging deep visual descriptors for hierarchical efficient localization," in *Conference on Robot Learning*. PMLR, 2018, pp. 456–465.

[14] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE transactions on robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[15] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[16] K. Wang, J. Guo, K. Chen, and J. Lu, "An in-depth examination of slam methods: Challenges, advancements, and applications in complex scenes for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[17] Z. Zheng, S. Lin, and C. Yang, "Rld-slam: A robust lightweight vi-slam for dynamic environments leveraging semantics and motion information," *IEEE Transactions on Industrial Electronics*, 2024.

[18] S. Xu, K. Zhang, and S. Wang, "Aqua-slam: Tightly-coupled underwater acoustic-visual-inertial slam with sensor calibration," *IEEE Transactions on Robotics*, 2025.

[19] X. Peng, P. Tong, X. Yang, C. Wang, and A.-M. Zou, "Idmf-vins: Improving visual-inertial slam for complex dynamic environments with motion consistency and feature filtering," *IEEE Sensors Journal*, 2025.

[20] Y. Ge, L. Zhang, Y. Wu, and D. Hu, "Pipo-slam: Lightweight visual-inertial slam with preintegration merging theory and pose-only descriptions of multiple view geometry," *IEEE Transactions on Robotics*, vol. 40, pp. 2046–2059, 2024.

[21] L. Zhang, W. Ye, J. Yan, H. Zhang, J. Betz, and H. Yin, "Loosely coupled stereo vins based on point-line features tracking with feedback loops," *IEEE Transactions on Vehicular Technology*, 2024.

[22] S. Rahman, R. DiPietro, D. Kedarisetti, and V. Kulathumani, "Large-scale indoor mapping with failure detection and recovery in slam," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 12 294–12 301.

[23] Y. Li, F. Liu, J. Zhang, J. Wang, H. Han, and C. Hao, "A binocular vision/imu msckf localization method considering dynamic initialization and loopback detection," *IEEE Sensors Journal*, 2024.

[24] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[25] G. Evangelidis and B. Micusik, "Revisiting visual-inertial structure-from-motion for odometry and slam initialization," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1415–1422, 2021.

[26] Z. Zhang, Y. Jiao, S. Huang, R. Xiong, and Y. Wang, "Map-based visual-inertial localization: Consistency and complexity," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1407–1414, 2023.

[27] Z. Zhang, Y. Song, S. Huang, R. Xiong, and Y. Wang, "Toward consistent and efficient map-based visual-inertial localization: Theory framework and filter design," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 2892–2911, 2023.

[28] C. Liu, H. Yu, P. Cheng, W. Sun, J. Civera, and X. Chen, "Pe-vins: Accurate monocular visual-inertial slam with point-edge features," *IEEE Transactions on Intelligent Vehicles*, 2024.

[29] B. Liu, F. Tang, Y. Fu, Y. Yang, and Y. Wu, "A flexible and efficient loop closure detection based on motion knowledge," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 241–11 247.

[30] H. Xu, P. Liu, X. Chen, and S. Shen, "$d^2$slam: Decentralized and distributed collaborative visual-inertial slam system for aerial swarm," *IEEE Transactions on Robotics*, 2024.

[31] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[32] T. Cieslewski, S. Choudhary, and D. Scaramuzza, "Data-efficient decentralized visual slam," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 2466–2473.