2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

A Robust Visual SLAM System for Small-Scale Quadruped Robots in Dynamic Environments

Chengyang Li, Yulai Zhang, Zhiqiang Yu, Xinming Liu, and Qing Shi, *Senior Member, IEEE*

***Abstract*— This paper presents a robust visual SLAM system designed for small-scale quadruped robots (ViQu-SLAM) for accurate localization, especially to mitigate the issue of erroneous data association caused by moving objects in dynamic environments. The proposed approach leverages a self-adaptive framework that integrates semantic segmentation with alterations in the spatial location of categorized map points. Besides, combination of leg odometry derived from forward kinematics with IMU provides scale information for positional transformations between keyframes, thus optimizing the overall localization accuracy of quadruped robots. At last, we performed evaluation across various stages and the results demonstrate competitive performance, with 53.16% reduction in average absolute trajectory error compared to that of ORB-SLAM3 in dynamic benchmark datasets. As a result, ViQu-SLAM, including visual and IMU-fused leg odometry, exhibits promising results on a small quadruped robot, reducing positioning errors in dynamic scenes by an average of 29.36% compared to existing state-of-the-art methods.**

I. INTRODUCTION

For Simultaneous Localization and Mapping (SLAM) tasks of small-scale robots in limited and dynamic environments, vision-based methods usually use features on image sequences to estimate camera poses and reconstruct feature point clouds in 3D virtual space, such as ORB-SLAM [1], RTB-MAP [2] and LSD-SLAM [3], etc. Based on feature extraction, these methods have achieved good results in texture-rich scenes, while they usually have reduced the reliability of state estimation and often suffer from ambiguity in textureless scenes. In comparison, LiDAR-based SLAM [4] is efficient and versatile due to the advanced hardware setup with rich structural information. However, LiDAR-based SLAM systems typically require relatively large equipment. These issues are particularly unacceptable in resource-constrained environments, such as for small robots or during long-duration missions.

To address the practical problem, we propose a robust system for small quadruped robots to achieve precise localization and mapping in dynamic environments. The ViQu-SLAM, whose pipeline is shown in Fig.1, is mainly based on the architecture of the ORB-SLAM3 [5], in which we simultaneously consider object detection and segmentation as well as the changes in the spatial position of classified map points, thereby building a reliable model that can eliminate moving objects in dynamic environments. In addition, we obtained the Leg Odometry (LO) by modeling the forward kinematics and fused it with the IMU data to provide a correction reference for subsequent calculations of position transformations between keyframes. Furthermore, to remove feature points related to moving objects, we obtain a dynamic object mask and extract the object feature points outsides the mask to be processed in the downstream SLAM framework. The mapping results are continuously updated by combining the prior semantic information of the object and the changes in the spatial location of the classified map points in successive frames. In addition, dynamic attribute of the object are obtained by clustering feature points in the same category. This integration improves the localization and mapping accuracy and increases the robustness of the system in dynamic environments. The main contributions are

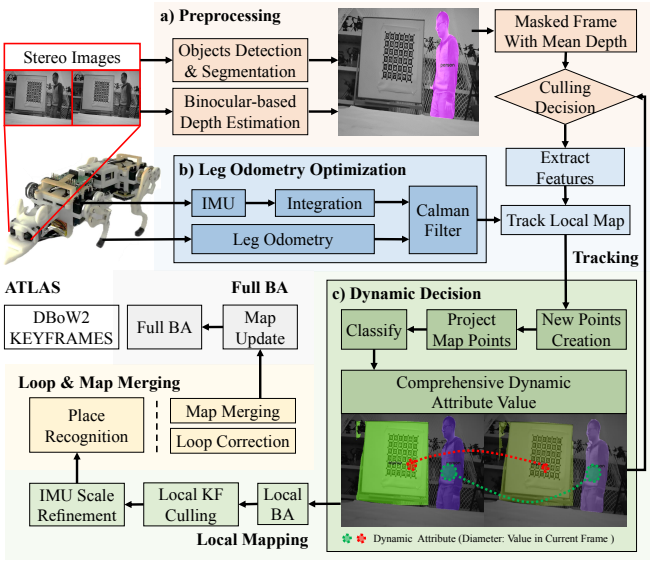


Fig. 1: ViQu-SLAM pipeline. (a) Preprocessing module acquires frames with mask and depth information by applying a detection and segmentation model. (b) Leg odometry module calculates of the leg odometry via forward kinematics, combined with an IMU data for assisted localization. (c) Dynamic object detection and removal module calculates dynamic attribute of categorized map points and culls out the moving objects.

This research was supported by Beijing Natural Science Foundation under Grant L233028, and National Natural Science Foundation of China under Grant 62088101.

Chengyang Li is with the Intelligent Robotics Institute, School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081.

Yulai Zhang is with School of Medical Technology, Beijing Institute of Technology, Zhiqiang Yu is with Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing, 314000, China, Xinming Liu and Qing Shi are with the Key Laboratory of Biomimetic Robots and Systems, Beijing Institute of Technology, Ministry of Education, Beijing 100081, China, and also with the Intelligent Robotics Institute, School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China (email: shiqing@bit.edu.cn)

summarized as follows:

- We provide a robust ViQu-SLAM for quadruped robots in dynamic environments that realizes the integration of LO and VIO, whose localization results show higher accuracy and robustness compared to other state-of-the-art methods.
- A framework for mitigating the effect of moving objects is proposed that integrates prior semantic segmentation with spatial transformation detection of classified map points. The correction probability of erroneous data association can exceed 53.16%.
- A real-world evaluation on a customized robot is performed in challenging situations and the experiments show that ViQu-SLAM achieves robust and accurate localization and mapping results.

The rest of the paper is organized as follows: Section II discusses related works, Section III presents our proposed approach in detail, and Section IV compares the experimental results of our system with state-of-the-art systems. Finally, we summarize the whole system in Section V.

II. Related Works

We split analysis of previous works closely related to our main contributions into two parts.

A. SLAM in Dynamic Environments

The development of SLAM systems in recent years has been dedicated to reducing the impact of moving objects in real-world environments, rather than a given static environment. Some methods [6], [7] perform well in reducing localization and mapping errors without modifying existing SLAM frameworks. DS-SLAM [8] utilizes semantic segmentation to identify dynamic objects in the environment, and then culls out dynamic feature points through motion consistency detection. Dynamic-SLAM [9] builds a Single Shot MultiBox Detector (SSD) object detector that incorporates prior knowledge to detect dynamic objects in a new detection thread at the semantic level. Among the stereo camera-based Visual Odometry (VO) systems, DynaSLAM [10] and DynaSLAM II [11] use multi-view geometry and deep learning to detect dynamic objects and build static maps by repairing the background of frames occluded by these dynamics. These systems mainly focus on outlier rejection in the front-end stage of visual SLAM. Most methods [12] incorporate deep learning based feature point detection methods to eliminate moving objects, which is very useful when dealing with known objects. However, there are more general situations where the motion state of the object is inconsistent with the movement state of the trainer, such as a person standing still, or an object being moved by someone, etc. These motion states cannot be defined by simple prior knowledge, or their elimination will reduce the robustness of localization and mapping.

B. Visual Motion Fusion

As an important component of autonomous navigation systems, SLAM has always been regarded as a key perception technology for mobile robots [13], and has recently attracted the attention of research institutions around the world. For legged robots, there are speed changes and morphological changes in each motion mode cycle, which disturb the robot's inertial measurement unit (IMU). Therefore, it is not appropriate to directly use the IMU to perform visual motion fusion odometry on legged robots. To solve this problem, joint angle sensing methods are often combined with IMUs to build ILO [14] which distinguishes the position of the foot to obtain the foot velocity, and further combines it with the robot configuration to obtain the body center speed. In the field of legged robot SLAM systems, sensor fusion usually has good versatility and stability. The Spot robot [15], MIT Cheetah2 [16] and ANYmal [17] use a combination of stereo, RGB-D cameras, or LiDAR to localize themselves and build robot-centric maps. Fusion of multi-sensor data can provide comprehensive information, but most of the above methods use expensive large sensors and cannot solve the SLAM problem in cost-effective lower-legged robots. In addition, the positioning and mapping of micro-robots must take into account real-world conditions such as changing lighting and moving objects.

III. Visual-Quadruped SLAM

This study presents a robust and efficient method for dynamic object detection and removal in visual SLAM for cost-efficient quadruped robots. By introducing object detection and segmentation models, we first preprocess the image frames. Then, the dynamic attribute of objects is determined by calculating the depth variation of map points in successive frames. Finally, in the fusion, we utilize the odometry information obtained from the fusion of the LO and IMU to provide a position estimation. The specific details of the design are explained below.

A. Preprocessing

The purpose of preprocessing is to obtain frames with mask and depth information.

Compared to the detection box, eliminating features within the object mask can avoid wasting background information. The preprocessing module helps to reduce the risk of erroneously removing features within the detection box but outside the mask. Specifically, we use the YOLOv8 model to obtain detection results that include information such as image detection bounding boxes and masks, etc.

Additionally, we also consider the depth information that can be computed from the binocular model, compute the depth information at the same location in different frames obtained by triangulation, and use it for dynamic recognition. During preprocessing, the average depth value of the k-th recognized object in the whole mask can be extracted as follows.

$$D_{mean}(k) = \overline{\sum_{(u,v)} D_{\mathrm{k}}(u,v)} \tag{1}$$

where $D_k(u,v)$ represents depth value with coordinates (u,v) of the feature points in the k-th target object mask.

B. Dynamic Object Detection and Removal

This module considers the position and depth of objects to determine their dynamic attribute and the results are used for culling moving objects off.

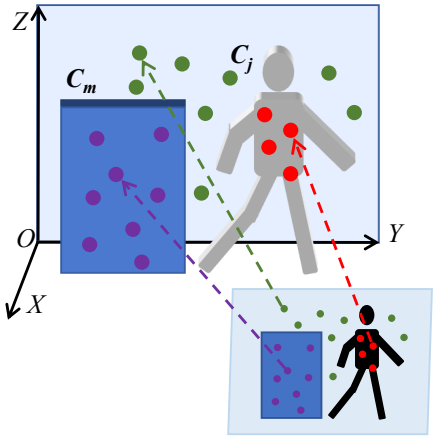


Fig. 2: Map points category decision diagram. C_m and C_j represent different object categories of map points

The projection of map points onto the image plane can be achieved utilizing perspective projection equations, with the known camera intrinsics. As shown in Fig. 2, the dynamic attribute of map points can be determined by checking whether their projection falls within certain masked regions. Specifically, if the pixel p_i is located at the mask of the k-th object, then the map point P_i is assigned a category label. For map points P_k with the same category label k, the mean norm $\overline{w}_k$ and the mean value of mask depth $\overline{d}_k$ of the map point set $\mathcal{P}_k$ can be computed by averaging.

The norm change Δw_k and the depth change Δd_k of the object between consecutive frames of the map point sets $\mathcal{P}_k^{0->(S/2-1)}$ and $\mathcal{P}_k^{(S/2-1)->(S-1)}$ are calculated as follows:

$$\begin{bmatrix}\Delta w_k \\ \Delta d_k\end{bmatrix} = \frac{1}{S/2}\sum_{i=(S/2)-1}^{S-1}\begin{bmatrix}\overline{w}_k^i \\ \overline{d}_k^i\end{bmatrix} - \frac{1}{S/2}\sum_{i=0}^{(S/2)-1}\begin{bmatrix}\overline{w}_k^i \\ \overline{d}_k^i\end{bmatrix} \tag{2}$$

The dynamic attribute A_k for the k-th category can be computed as:

$$A_k = \alpha \cdot \Delta w_k + \beta \cdot \Delta d_k + \gamma \cdot Pr_k \tag{3}$$

where α, β, and γ represent coefficients contributing weights of each component, Pr_k denotes the initial dynamic attribute based on the prior information. Then, the computed dynamic value A_k will be obtained, which will later be used to compare with the thresholds th_d.

When an object moves, as shown in Fig. 3, the spatial position of the map points observed in successive frames will change. The average spatial position change $\overline{\Delta A_k}$ over several frames can be calculated as follows:

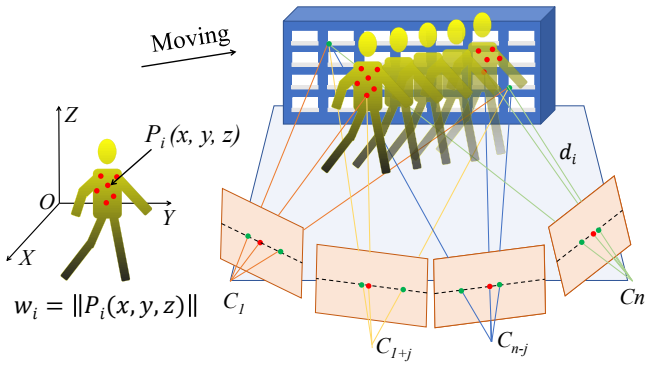


Fig. 3: Dynamic object recognition process. The red points represent map points with the attribute of the object category. When the object moves, the spatial position information of the map points it generates will change.

$$\overline{\Delta A_k} = \frac{1}{N/2}\sum_{i=(N/2)-1}^{N-1}\overline{A}_k^i - \frac{1}{(N/2)}\sum_{i=0}^{(N/2)-1}\overline{A}_k^i \tag{4}$$

where, N represents the number of frames. If $\overline{\Delta A_k}$ exceeds th_d, the target is classified as dynamic.

This approach considers both spatial and temporal changes in the object position and mask depth, as well as any additional dynamic attributes, providing a robust method for dynamic object detection in 3D space.

When a target is identified as dynamic, the feature points within the target mask are no longer extracted in subsequent ORB feature extraction processes. This strategy ensures that feature points located inside the dynamic target are not extracted, thus preventing the generation of related spatial map points and achieving the removal of dynamic targets.

C. Visual-Quadruped Fusion

To provide scale for the visual odometry, we integrate data from the quadruped robot's IMU and LO using a Kalman Filter.

First, with the IMU measurements, we define the state parameters $x = [{}^O\boldsymbol{P}_C, {}^O\boldsymbol{v}_C, {}^O\boldsymbol{P}_i], (i = FL, FR, BL, BR)$, which represent the position and velocity of the body, as well as the position of the four foot-ends in the world coordinate system, respectively. Then, we can establish the state equation of the quadruped robot in matrix form.

$$\begin{bmatrix}{}^O\boldsymbol{P}_C \\ {}^O\boldsymbol{v}_C \\ {}^O\boldsymbol{P}_i\end{bmatrix}_{k+1} = \begin{bmatrix}\boldsymbol{I}_{3\times3} & \Delta t\boldsymbol{I}_{3\times3} & \boldsymbol{0}_{3\times12} \\ \boldsymbol{0}_{3\times3} & \boldsymbol{I}_{3\times3} & \boldsymbol{0}_{3\times12} \\ \boldsymbol{0}_{12\times3} & \boldsymbol{0}_{12\times3} & \boldsymbol{I}_{1\times2}\end{bmatrix}\begin{bmatrix}{}^O\boldsymbol{P}_C \\ {}^O\boldsymbol{v}_C \\ {}^O\boldsymbol{P}_i\end{bmatrix}_k + \begin{bmatrix}\boldsymbol{0}_{3\times3} \\ \Delta t\boldsymbol{I}_{3\times3} \\ \boldsymbol{0}_{12\times3}\end{bmatrix}{}^O\boldsymbol{a} \tag{5}$$

The compact form, $\boldsymbol{x}_{k+1} = A\boldsymbol{x}_k - Bu_k$, defines the a priori prediction equation for the system state that used in the Kalman Filter.

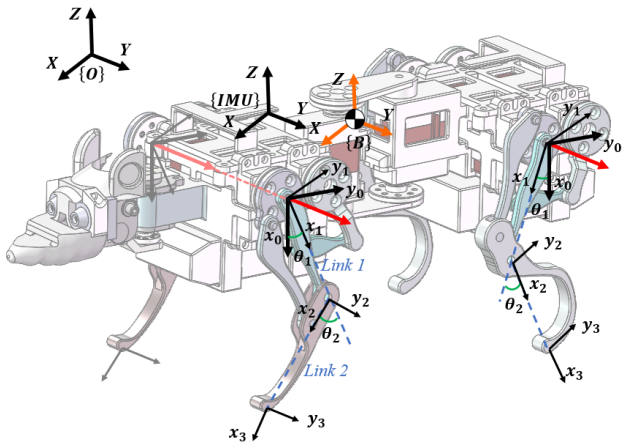


Fig. 4: Coordinate system of kinematic model.

Next, we establish the observation equation of the system by modeling the forward kinematics of the quadruped robot, as shown in the Fig. 4. Taking the left forelimb as an example, the coordinates of the foot-end can be obtained by the following equation:

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} c_1 & c_{12} \\ s_1 & s_{12} \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} \tag{6}$$

Where x_3 and y_3 represent the location of the foot-end with respect to the body frame $\{B\}$, x_0 and y_0 denote the hip or shoulder joint location with respect to $\{B\}$. c_1 denotes $\cos\theta_1$, and c_{12} denotes $\cos(\theta_1 + \theta_2)$. l_1 and l_2 stand for the length of the upper limb and lower limb, respectively.

As a result, we obtained a spatial description of the foot-end of the left forelimb in the body coordinate system. Assuming that the foot-end of the robotic rat does not slip and is not subject to lateral force at the same time. In this case, when one leg changes to the support phase, the position of the center of mass relative to the stationary foot-end can be obtained as follows.

$$^{3}\boldsymbol{P}_c = -\begin{bmatrix} \frac{1}{2}bl - s_1 l_1 - s_{12} l_2 \\ \frac{1}{2}bw \\ -c_1 l_1 - c_{12} l_2 \end{bmatrix} \tag{7}$$

Finally, the position of the robot center of mass in the world coordinate system $^{O}\boldsymbol{P}_C$ can be calculated by position accumulation as $^{O}\boldsymbol{P}_{C,k+1} = {}^{O}\boldsymbol{P}_{C,k} + {}^{O}\boldsymbol{P}_C$.

With the above analysis, we obtained the observation equation by the linear mathematical model and the state equation from IMU integration. This makes it easy to associate the Kalman Filter with obtaining more comprehensive information about the system, so that it gives a complete result and makes it more accurate with the predictions gathered by the robot's internal sensors.

IV. Experimental Validation

We conducted a comprehensive evaluation across multiple stages and the results demonstrate the excellent performance of ViQu-SLAM compared with state-of-the-art works.

The accuracy is evaluated by the Root-Mean-Square-Error (RMSE), which is derived from the Absolute Trajectory Error (ATE). To align all trajectories with respect to the ground truth, we used the SE(3) Umeyama alignment method.

A. Dynamic Feature Detection and Removal

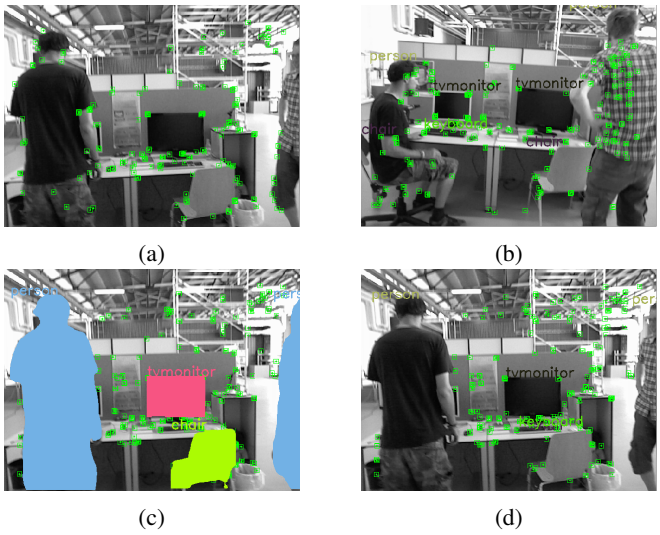
To verify the performance of the proposed algorithm in scenes with dynamic objects, the experiments mainly use four highly dynamic sequences $friburg3\_walking*$ from the TUM RGB-D dataset [18]. To evaluate the performance improvement over the original system, we compared it to ORB-SLAM3, the results of DS-SLAM [8] and Dynamic-VINS [19] from their original published papers are also included as baselines in Table I.

TABLE I
RESULTS OF RMSE OF ATE [m] ON TUM RGB-D *fr3_walking* DATASETS. **BEST RESULT** IS HIGHLIGHTED IN BOLD, WHILE SECOND-BEST RESULT IS UNDERLINED.

Sequence	ORB-SLAM3	DS-SLAM	Dynamic-VINS	**ViQu-SLAM**	Improvement
fr3_w_half	0.3424	0.0303	0.0608	**0.0261**	92.37%
fr3_w_xyz	0.2747	**0.0247**	0.0486	0.1258	54.20%
fr3_w_rpy	0.1551	0.4442	**0.0629**	0.0720	53.62%
fr3_w_static	0.0176	0.0081	**0.0077**	0.0124	29.67%

The experimental results in Table I provide the performance of our system compared to existing state-of-the-art methods. We observe a consistent improvement in RMSE over ORB-SLAM3 for all sequences. For instance, in the $fr3\_w\_half$ sequence, our system achieves a remarkable RMSE of 0.0261 meters, outperforming ORB-SLAM3 by 92.37%. Similar trends are observed in other sequences, with an average enhancement exceeding 57% on the RMSE metric compared to ORB-SLAM3. Furthermore, our system also demonstrates competitive performance against DS-SLAM and Dynamic-VINS, demonstrating its superiority in localization accuracy. Our algorithm has a smaller RMSE because we introduce the adaptive dynamic attributes. By assigning initial dynamic attributes to the detected objects and updating their dynamic attributes according to the depth changes of the objects in successive frames, the approach makes the experimental results show the best results. Fig. 5 illustrates the effectiveness of dynamic feature detection and removal. Initially, the dynamic attributes of the two people are unknown, and ViQu-SLAM extracts feature points from their respective areas in the frame. Subsequently, as the people move, our algorithm accurately identifies and removes their dynamic features.

These results highlight the effectiveness of our approach in improving localization accuracy in dynamic environments. Moreover, the reduction in RMSE across different sequences demonstrates the robustness and generalizability of our system across different scenarios, further validating its efficacy for real-world applications.



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

Fig. 5: Dynamic feature detection and removal. (a) shows the feature extraction result of ORB-SLAM3. (b) shows the original result when people are still at the beginning of the sequence. (c) shows the semantic segmentation result, and (d) shows the result after the system has detected the people's movement.

B. Evaluation of ViQu-SLAM on OpenLORIS Benchmark

Our evaluation, conducted on twelve selected sequences from four scenes in the OpenLORIS-Scene[20], encompasses a comprehensive examination of various indoor sequences, each presenting distinct challenges in terms of dynamic object presence, scene complexity, and illumination variations. The results of ORB-SLAM3 in Stereo-Inertial (S-I) mode, VINS-Fusion in Stereo-Wheel (S-W) mode, and RVWO are from [21].

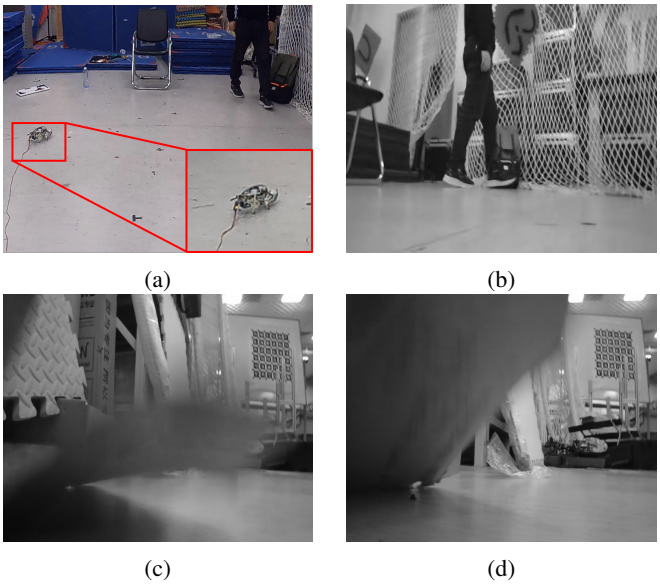
TABLE II
RESULTS OF RMSE OF ATE [m] ON OpenLORIS DATASETS. **BEST RESULT** IS HIGHLIGHTED IN BOLD, '-' INDICATES FAILURE.

Sequence	ORB-SLAM3	VINS-Fusion	RVWO	**ViQu-SLAM**	Improvment
corr. 1-3	0.990	0.525	**0.159**	0.1929	80.51%
corr. 1-5	1.131	0.504	0.420	**0.1467**	87.03%
office1-1	0.109	0.104	**0.020**	0.0239	78.08%
office1-3	-	0.037	0.022	**0.0126**	-
office1-5	0.235	0.323	**0.081**	0.0339	85.59%
home1-1	0.406	0.570	0.221	**0.1419**	65.05%
home1-5	0.318	0.374	0.069	0.1666	47.62%
cafe1-1	0.116	0.324	0.213	**0.0887**	23.52%

The evaluation results, quantified in Table II, illustrate the RMSE values for different sequences across the evaluated methods. Notably, ORB-SLAM3 and VINS-Fusion suffer more drifts, especially, ORB-SLAM3 fails to track in the *office1-3* sequence. The RVWO system can adapt to dynamic environments and performs well in corridor and office scenarios. In addition, our ViQu-SLAM has the best performance in four scenarios, and in some sequences, the accuracy is improved by more than 80% compared to ORB-SLAM3. These findings underscore the adaptability and efficacy of our ViQu-SLAM algorithm in addressing the complexities of dynamic indoor environments, emphasizing its potential for real-world applications requiring accurate and reliable localization and mapping capabilities.

C. Evaluation on a Quadruped Robot Rat

In this section, we performed a real-world validation on a quadruped robot, as shown in Fig. 6(a), which is equipped with a stereo RGB camera and an IMU. The experimental setup mainly consisted of a prototype of a quadruped robot, and a standard optical tracking system (Motive: Tracker) with a sampling frequency of 180 Hz.



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

Fig. 6: Experimental setup and visual challenging conditions. (a) Quadruped robotic rat used in the evaluation, (b) Moving objects, (c) Blurred Image, (d) Closing to the camera.

We collected a total of four test sequences. In the first sequence, named *explore_1*, the robot walked in a straight line with no dynamic scenes. The second and third sequences recorded the robot exploring challenging environments, including moving people. The fourth and most challenging sequence included multiple moving objects, blurry images, and closing to the camera, as shown in Fig. 6. All sequences experienced jitter and slippage during the motion of the robotic rat.

Through experiments, we observed that our algorithm has a relatively low RMSE compared to ILO, ORB-SLAM3, and DynaVINS in all sequences. As shown in Table III, ViQu-SLAM achieved RMSE of 0.0095, 0.0153, and 0.2199 in *explore_2*, *explore_3*, and *explore_4*, respectively, which is lower than the other three algorithms. Among them, the localization accuracy of ILO is second only to ours. While in *sequence_1*, since there is no dynamic target, the result of ORB-SLAM3 is optimal.

Besides, we compared trajectory performance of ILO, ORB-SLAM3, and ViQu-SLAM in sequence *explore_4*. The

TABLE III
RESULTS OF RMSE OF ATE [m] ON A DATASET RECORDED ON OUR SETUP. **BEST RESULT** IS HIGHLIGHTED IN BOLD, WHILE SECOND-BEST RESULT IS UNDERLINED.

Sequence	ILO	ORB-SLAM3	DynaVINS	**ViQu-SLAM**
explore_1	0.0647	**0.0589**	0.1809	0.1276
explore_2	0.1800	0.0318	0.1152	**0.0095**
explore_3	0.0143	0.0188	0.0183	**0.0153**
explore_4	0.3253	0.5448	0.6422	**0.2199**

trajectories in both x and y directions are shown in the Fig. 7, throughout the entire 250s period, the trajectory of ILO and ViQu SLAM closely followed the ground truth, but ORB-SLAM3 showed significant deviation from it, especially during the 100s-250s period in x direction. Because of the contribution of leg odometry, our trajectories are more competitive in the x-direction. but deviation still happens in the y-direction. Although ViQu-SLAM has relatively large errors compared to ORB-SLAM3 in the y direction, overall, the localization result of ViQu-SLAM are relatively desired.

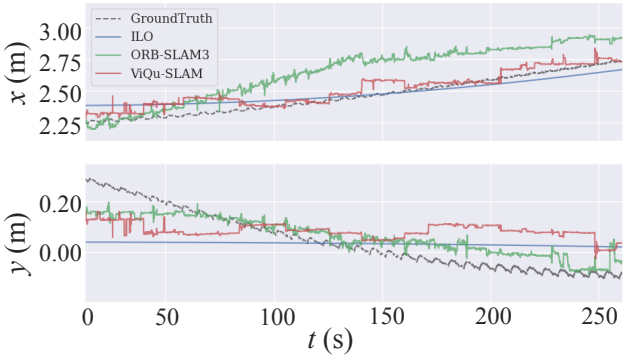


Fig. 7: Trajectories of ViQu-SLAM, and other systems, generated on $explore\_4$ sequence compared to ground truth.

V. CONCLUSIONS

This study introduces a robust and efficient method for dynamic object detection and removal in visual SLAM for quadruped robots. By integrating advanced techniques including object detection, semantic segmentation, and dynamic map point determination. Furthermore, leg odemetry is incorporated to achieve significant improvements in robustness and efficiency. Experimental validation across various datasets and scenarios confirms the superior performance of the algorithm, highlighting its potential for real-world applications in mapping and localization tasks in dynamic environments.

REFERENCES

[1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[2] M. Labbe and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 734–745, 2013.

[3] D. Caruso, J. Engel, and D. Cremers, "Large-scale direct slam for omnidirectional cameras," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 141–148, IEEE, 2015.

[4] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time.," in *Robotics: Science and systems*, vol. 2, pp. 1–9, Berkeley, CA, 2014.

[5] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[6] B. Bescos, J. Neira, R. Siegwart, and C. Cadena, "Empty cities: Image inpainting for a dynamic-object-invariant space," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 5460–5466, IEEE, 2019.

[7] R. Uittenbogaard, "Moving object detection and image inpainting in street-view imagery," 2018.

[8] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "Ds-slam: A semantic visual slam towards dynamic environments," in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 1168–1174, IEEE, 2018.

[9] L. Xiao, J. Wang, X. Qiu, Z. Rong, and X. Zou, "Dynamic-slam: Semantic monocular visual localization and mapping based on deep learning in dynamic environment," *Robotics and Autonomous Systems*, vol. 117, pp. 1–16, 2019.

[10] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "Dynaslam: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4076–4083, 2018.

[11] B. Bescos, C. Campos, J. D. Tardós, and J. Neira, "Dynaslam ii: Tightly-coupled multi-object tracking and slam," *IEEE robotics and automation letters*, vol. 6, no. 3, pp. 5191–5198, 2021.

[12] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8918–8927, 2021.

[13] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*, pp. 834–849, Springer, 2014.

[14] D. Belter and M. R. Nowicki, "Optimization-based legged odometry and sensor fusion for legged robot continuous localization," *Robotics and Autonomous Systems*, vol. 111, pp. 110–124, 2019.

[15] M. Lozano, S. Mead, M. Cavazza, and F. Charles, "Search-based planning: A method for character behaviour," in *Proceedings of the 3rd on Intelligent Games & Simulation. Congreso GameOn-2002, Gran Bretaña*, 2002.

[16] D. J. Hyun, S. Seok, J. Lee, and S. Kim, "High speed trot-running: Implementation of a hierarchical controller using proprioceptive impedance control on the mit cheetah," *The International Journal of Robotics Research*, vol. 33, no. 11, pp. 1417–1445, 2014.

[17] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch, *et al.*, "Anymal-a highly mobile and dynamic quadrupedal robot," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 38–44, IEEE, 2016.

[18] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580, IEEE, 2012.

[19] J. Liu, X. Li, Y. Liu, and H. Chen, "Rgb-d inertial odometry for a resource-restricted robot in dynamic environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9573–9580, 2022.

[20] X. Shi, D. Li, P. Zhao, Q. Tian, Y. Tian, Q. Long, C. Zhu, J. Song, F. Qiao, L. Song, Y. Guo, Z. Wang, Y. Zhang, B. Qin, W. Yang, F. Wang, R. H. M. Chan, and Q. She, "Are we ready for service robots? the openloris-scene datasets for lifelong slam," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3139–3145, 2020.

[21] J. Mahmoud, A. Penkovskiy, H. T. L. Vuong, A. Burkov, and S. Kolyubin, "Rvwo: A robust visual-wheel slam system for mobile robots in dynamic environments," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3468–3474, IEEE, 2023.