Rapid Implementation of an Advanced Visual Localization System for Mobile Robot Navigation

Kunyang Wu, Yang Liu and Guanyu Zhang

Abstract—**The Global Positioning System (GPS) is the most widely used positioning system for outdoor localization and navigation. However, GPS signals are not always available, especially in indoor or urban canyon environments. As such, alternative positioning systems capable of operating in GPS-denied environments are essential. This paper proposes a novel visual positioning system that combines Red-Green-Blue Depth (RGBD) map construction, semantic graph-based image matching, and dynamic localization and tracking. Our system utilizes a multi-modal sensor consisting of LiDAR and camera to acquire data and build a map library of RGBD images with sparse depth information. To initialize localization, we construct semantic graphs from observed and map images and construct image descriptors for matching to obtain approximate positions. To achieve continuous localization, we combine visual odometry with the ASpanFormer image matching method, and correct pose estimates based on the map library to reduce cumulative errors. We also dynamically update the map library in response to environmental changes. The results show that our system achieves superior accuracy and robustness in challenging scenarios, such as lighting variations, dynamic objects, and similar scene distributions.**

Index Terms—**visual localization, GPS-denied environments, multi-modal sensor, map-based correction**

I. Introduction

Localization systems are pivotal for automated machinery, including robots and autonomous vehicles. They provide the capability to determine and orient within an environment, forming the basis for subsequent tasks such as navigation [1] [2], planning [3] [4], and interaction [5]. The Global Positioning System (GPS) [6], which relies on satellite signals to estimate the receiver's position and direction, is the most prevalent positioning system. However, in indoor or urban settings, GPS signals are often unavailable or unreliable, with signals potentially obstructed, reflected, or interfered with by buildings [7], trees, or other obstacles [8]. Consequently, an alternative positioning system capable of operating in a GPS-free environment is essential.

We examine existing methods for positioning in GPS-denied environments and categorize them into two primary groups: SLAM-based methods and Map-based methods [9]. SLAM, or simultaneous localization and mapping, is a technique that allows a device to estimate its own position and orientation while constructing a map of the surrounding environment using sensor data such as camera images [10], LiDAR scans [11], or inertial measurements. V-SLAM uses camera images as the primary sensor data source, applying computer vision techniques to extract features, match them across frames, and estimate the camera pose and the 3D structure of the scene. V-SLAM can provide rich semantic information and low-cost hardware, but it also suffers from illumination changes, occlusions, dynamic objects, and feature sparsity. Some representative works on V-SLAM are ORB-SLAM3 [12] and SOFT2 [13]. LiDAR SLAM uses laser scanners to obtain 3D point clouds of the environment, applying geometric methods to align them and estimate the device pose and the map. LiDAR SLAM can provide high-resolution and accurate measurements, but it also requires expensive and bulky hardware. Some representative works on LiDAR SLAM are LOAM [14] and CT-ICP [15]. Inertial SLAM uses inertial measurement units (IMUs) to measure the linear acceleration and angular velocity of the device, integrating them to obtain the device pose and velocity. Inertial SLAM can provide fast and smooth updates, but it also suffers from sensor errors, drift, and bias. Some representative works on inertial SLAM are OKVIS [16] and VINS-Mono [17]. In general, SLAM-based methods can achieve high accuracy and robustness in GPS-denied environments, but they also face some challenges, such as computational complexity, scalability, loop closure, and drift [18].

Map-based methods form another category of positioning systems in GPS-denied environments. They rely on a pre-built map of the environment and a localization algorithm that matches the sensor data with the map [19]. The map can be constructed offline using SLAM or other methods, or obtained from external sources, such as satellite imagery, aerial photography, or floor plans. The map can be represented in different formats, such as point clouds, grids, graphs, or semantic labels. The localization algorithm can use various techniques, such as particle filters, Kalman filters, or deep learning, to estimate the agent's pose based on the map and the sensor data. Some examples of map-based methods are [20], [21], and [22]. The main advantages of map-based methods are that they can provide more accurate and robust localization than SLAM, and they can reduce the computational cost and complexity of the positioning system. However, map-based methods also have some limitations and challenges. Map-based methods require a prior map of the environment, which may not be available or up-to-date, especially in dynamic or unstructured environments. Furthermore, map-based methods may have difficulty in handling occlusions, illumination changes, or sensor noise,

Manuscript received Feb 1, 2024; revised May 13, 2024. This work was supported in part by the Science and Technology Development Project of Jilin Province under Grant 212551GX010283541.

Kunyang Wu is a doctoral student of the Key Laboratory of Geophysical Exploration Equipment, College of Instrument Science and Electrical Engineering, Jilin University, Changchun 130061, China. (e-mail: wuky23@mails.jlu.edu.cn).

Yang Liu is an associate professor of the Key Laboratory of Geophysical Exploration Equipment, College of Instrument Science and Electrical Engineering, Jilin University, Changchun 130061, China. (e-mail: liu_yang@jlu.edu.cn).

Guanyu Zhang is an associate professor of the Key Laboratory of Geophysical Exploration Equipment, College of Instrument Science and Electrical Engineering, Jilin University, Changchun 130061, China. (corresponding author to provide e-mail: zhangguanyu@jlu.edu.cn).

which can affect the map-matching performance [23].

To address the issues present in existing methods, we propose a novel visual positioning system capable of operating in GPS-denied environments, utilizing a multi-modal sensor composed of a LiDAR and a camera. Our system comprises three main modules: RGBD map construction, vision localization, and map update. The RGBD map construction module builds a map of the environment by fusing the LiDAR point clouds and the camera images, generating a map library containing RGBD images and their corresponding poses. The vision localization module consists of two submodules: initialization and dynamic localization and tracking (DLT). The initialization module uses a semantic graph-based image matching method to find the most similar image in the map library and estimate the initial pose of the camera. The DLT module uses a combination of visual odometry and map-based correction to track the camera motion and update the pose estimation. The map-based correction uses a detectorless image matching method based on adaptive spanning transformers to calculate the spatial transformation between the current image and the map image. The map update module evaluates the differences between the current image and the map image, replacing the outdated map image with the new image if necessary.

The main contributions of our paper are as follows:

- We propose a novel visual positioning system capable of operating in GPS-denied environments, utilizing a multi-modal sensor composed of a LiDAR and a camera.
- We propose a novel semantic graph-based image matching method for efficient and robust initialization, capable of handling lighting changes and dynamic objects in the scene.
- The proposed two-threaded dynamic tracking and positioning method is capable of map-based correction to reduce the cumulative error, and the detectorless image matching method relied upon can well handle viewpoint changes and environmental changes in the map.
- We conduct extensive experiments on real road data to demonstrate the effectiveness and superiority of our proposed system over state-of-the-art methods.

The rest of the paper is organized as follows: Section II describes the RGBD map construction module. Section III describes the proposed vision localization method. Section IV presents the experimental results and analysis. Section V concludes the paper and discusses future work.

II. RGBD Map Construction

This section presents a theoretical framework for constructing a comprehensive visual map using a multi-modal sensor, which includes a LiDAR and a camera. The proposed visual map is an enriched dense 3D map that encapsulates the database of camera poses, visual features, and the 3D structures of the scene. This method is particularly suitable for environments deprived of GPS, as it leverages LiDAR-based odometry and designates the LiDAR coordinate system of the initial frame as the global coordinate system.

The acquisition of multi-modal data within the target environment is achieved by employing a lightweight yet precise LiDAR SLAM algorithm, hdl_graph_slam [24], to estimate the sensor poses. Given the disparate frame rates of the LiDAR and the camera, we acquire a set of n_1 LiDAR point cloud frames L_i and the corresponding CT-ICP poses P_i, along with n_2 camera image frames I_j. To synchronize the data, we perform timestamp alignment on the collected data and select a subset of LiDAR point clouds L_k and camera images I_k with timestamp differences below a predefined threshold.

The fusion of each LiDAR frame L_k with its corresponding camera frame I_k is achieved by utilizing the projection equation:

$$I_k = \boldsymbol{K} \cdot \boldsymbol{T}_{\mathrm{L}}^{\mathrm{C}} \cdot L_k \tag{1}$$

where $\boldsymbol{K}$ denotes the camera's intrinsic parameter matrix, and $\boldsymbol{T}_{\mathrm{L}}^{\mathrm{C}}$ signifies the 6-DOF rigid transformation representing the pose of the camera relative to the LiDAR. This fusion results in the generation of RGBD images D_k with sparse depth information.

The poses of these RGBD images in the global coordinate system are derived from the poses of their corresponding LiDAR frames. For a set of images and point clouds in the map, let P_C, P_L, and P_M represent the coordinates of a point in the camera, LiDAR, and map coordinate systems, respectively. We then have the following equations:

$$P_C = R_L^C \cdot P_L + t_L^C \tag{2}$$

$$P_M = R_L \cdot P_L + t_L \tag{3}$$

where, R_L^C and t_L^C denote the rotation matrix and translation vector between the LiDAR and the camera, respectively, as determined by our previously established calibration method [25]. R_L and t_L represent the rotation matrix and translation vector of the current LiDAR coordinate system relative to the map coordinate system, obtained during the mapping process.

Substituting (3) into (2), we obtain:

$$P_M = R_L \cdot R_L^{C^{-1}} \cdot \left(P_C - t_L^C\right) + t_L \tag{4}$$

From (4), the rotation matrix R_C and translation vector t_C of the camera coordinate system relative to the map coordinate system are:

$$R_C = R_L \cdot R_L^{C^{-1}} \tag{5}$$

$$t_C = t_L - R_C \cdot t_L^C \tag{6}$$

In the final step, the system integrates each RGBD image and its associated pose into a unified map node, thereby constructing the map library.

III. Theoretical Framework for Vision Localization

The proposed vision localization system, as depicted in Fig. 1, operates on a sequential workflow. The process begins with an initialization phase that establishes the initial position within the map. Following successful initialization, the system maintains continuous positioning through a synergistic integration of odometry and map-based corrections. This robust framework effectively mitigates odometry drift and dynamically updates the map, ensuring adaptability to evolving environments.

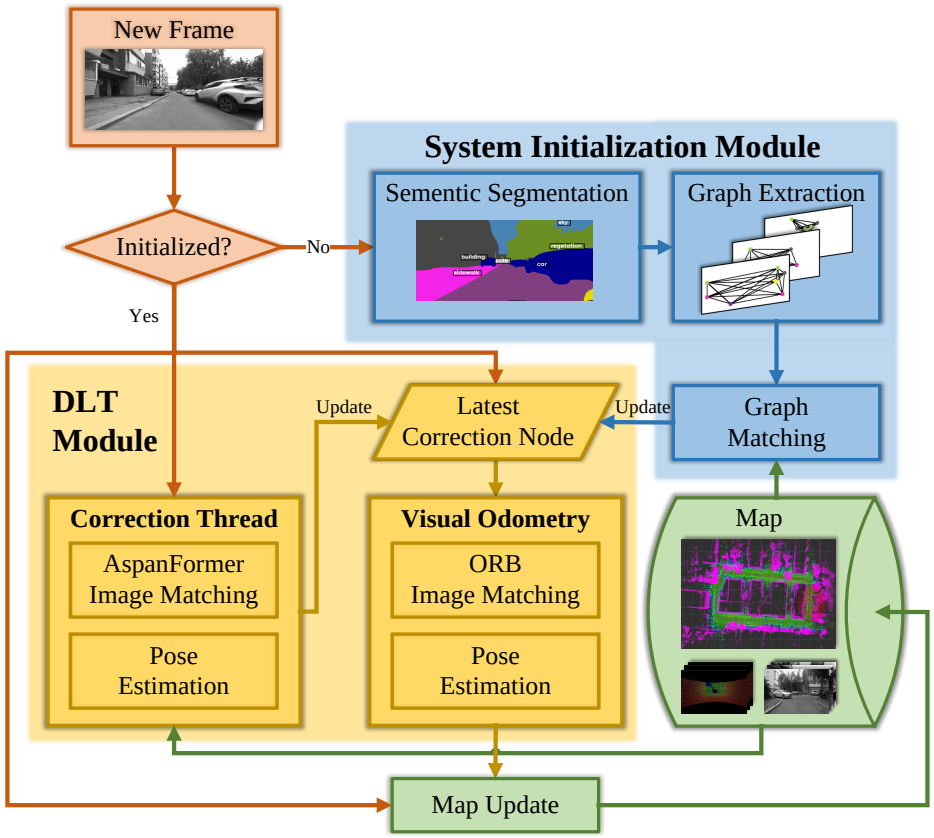

Fig. 1. Overview of the proposed positioning system.

A. Localization Initialization Module

The initialization module employs an innovative semantic graph-based image matching approach to achieve efficient localization. The Mask2Former model [26], pre-trained on Cityscapes, is utilized to segment images into 18 semantic classes. We concentrate on 11 static classes for graph construction, thereby ensuring robustness against dynamic elements in the scene.

Semantic graphs are constructed at multiple scales to capture variations in object density across different distance ranges. At each scale, recognized objects are represented as graph nodes using their 3D geometric centers, and nodes of closely located objects with identical semantic labels are merged. Each node encapsulates its 3D coordinates and semantic label, while each edge signifies the spatial distance between two nodes, representing their undirected spatial relationship. As roads invariably form semantic blocks, we construct the descriptor G for the graph by designating the road node as the central node. Other nodes are inserted into the descriptor in a clockwise order, starting from the node directly above the road node. The distance between each node and the central node, and the pixel range of each node, are also incorporated into the descriptor. Fig. 2 illustrates the process of descriptor extraction.

Descriptors are extracted for the collected images and images in the map library, denoted as G_c and G_m, and concatenated into one-dimensional vectors x and y. The Pearson correlation coefficient is then used to calculate the similarity between the vectors, defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - y)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{7}$$

where $\bar{x}$ and $\bar{y}$ are the mean values of x and y, respectively. The Pearson correlation coefficient is a statistical measure that quantifies the strength of the linear relationship between two variables, ranging from -1 to 1. A value near 1 implies a strong positive correlation, indicating that the two vectors are similar. A value near -1 implies a strong negative correlation, indicating that the two vectors are dissimilar. A value near 0 implies no linear correlation, indicating that the two vectors are independent. The advantage of using the Pearson correlation coefficient for similarity calculation is that it is invariant to scaling and shifting of the vectors, which can be caused by illumination changes or noise in the images.

The similarity score between the two images is then given by:

$$s = \frac{r+1}{2} \tag{8}$$

which normalizes the Pearson correlation coefficient to the range of 0 to 1. The highest similarity score corresponds to the image in the map image library that is most similar to the currently collected image.

B. Dynamic Localization and Tracking Module

The Dynamic Localization and Tracking (DLT) module employs a dual-threaded architecture, comprising visual odometry and correction threads. The visual odometry thread serves as the primary constraint, while the correction thread

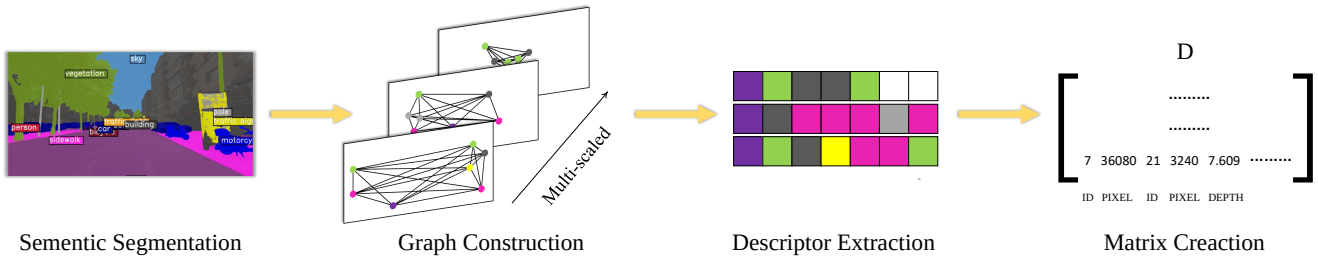

Fig. 2. The process of descriptor matrix creation.

addresses the cumulative error intrinsic to odometry. The visual odometry thread persistently updates the pose between successive calibration nodes, providing initial estimates for the ensuing correction processes. The output of the positioning system hinges on the attitude estimation conducted by the visual odometry thread, which uses the latest correction node as a reference point.

The visual odometry thread estimates the camera motion using a sequence of images captured by the onboard stereo camera. The ORB algorithm is utilized to extract and match feature points across consecutive frames, while the RANSAC algorithm is employed to discard outliers and compute the essential matrix. This matrix can be decomposed to obtain the relative rotation and translation between camera poses. The scale is recovered using the depth information from the stereo camera. The visual odometry thread outputs the camera pose as a 4x4 transformation matrix, composed of rotation and translation components.

The correction thread computes the spatial transformation between the observed and map images via image feature matching, and finalizes the current positioning correction based on the map image's position information. To ensure robustness against challenges such as outdated map data or significant lighting and viewpoint variations, we leverage ASpanFormer [27], a cutting-edge detectorless image matching method. Its adaptive spanning transformer architecture facilitates robust feature alignment amidst such complexities. A semantic segmentation network is integrated within the matching process to identify and mask dynamic elements within the visual scene, mitigating potential estimation biases.

Consider a robot located at point A, with its initial attitude determined by the localization system's odometry thread before correction. Firstly, j images in closest proximity to the robot's current location are extracted from the map library and subjected to image matching with the observed images. The map images with depth information are used to transform 2D feature point correspondences into 2D-3D correspondences, thereby formulating a PnP (Perspective-n-Point) problem. By solving the PnP problem, the rotation matrix R_{map}^{obs} and translation vector t_{map}^{obs} between the camera coordinate systems of the observation and map images can be determined. Combined with the position information associated with the map image, i.e., the rotation matrix R_{map} and translation vector t_{map} relative to the map coordinate system, the positioning information can be corrected to obtain the rotation matrix R_{corr} and translation vector t_{corr} relative to the map coordinate system after the observation image is corrected :

$$R_{corr} = R_{map} \cdot {R_{map}^{obs}}^{-1} \tag{9}$$

$$t_{corr} = t_{map} - R_{corr} \cdot t_{map}^{obs} \tag{10}$$

where RC denotes the corrected pose at point A. Until the next correction node is reached, the odometry thread performs relative attitude estimations between frames based on RC as the output of the positioning system.

C. Map Update Module

The Map Update Module is designed to ensure the consistency and precision of the map library by identifying and updating environmental changes. Two types of changes are considered that could impact the quality of the map: semantic and illumination changes. Semantic changes encompass the addition, removal, or alteration of objects or structures in the scene, such as buildings, trees, or vehicles. Illumination changes pertain to variations in lighting conditions due to factors like weather, time of day, or seasons. Both types of changes can influence the performance of the image matching and localization modules, necessitating appropriate handling.

Semantic changes are detected using the semantic similarity score s, computed in the initialization module. If s falls below a predefined threshold τ_s, it indicates significant semantic differences between the current and map images, necessitating an update of the map image. Illumination changes are detected using image brightness. RGB images are converted to grayscale, and the average pixel intensity b is computed for each image. If the absolute difference between the current image brightness and the map image brightness exceeds a predefined threshold τ_b, it signifies a significant change in illumination conditions, prompting an update of the map image.

At each correction node, the current image is compared with the map image using the semantic similarity score and image brightness. If either exceeds the corresponding threshold, the map image is marked as outdated and replaced with the current image. The position and orientation of the map image are also updated using the output of the correction thread. This approach ensures that the map library consistently reflects the most recent state of the environment.

IV. Experimental Results

Fig. 3 illustrates our mobile robotic experimental platform, equipped with a suite of sensors including a binocular camera, LiDAR, and a GPS receiver. The binocular camera is a ZED2 model, and the LiDAR is a HESAI Pandar XT-32

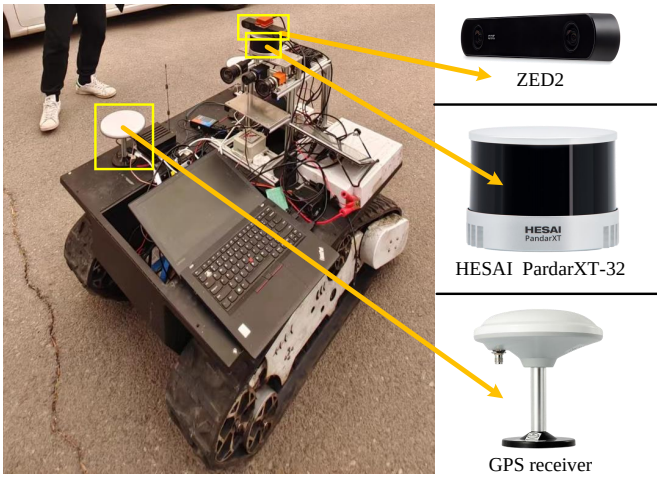

Fig. 3. The mobile robotic experimental platform, outfitted with sensors such as a camera, LiDAR, GPS receiver, among others.

model. The GPS receiver, utilizing RTK technology in conjunction with a mobile base station, achieves centimeter-level positioning and provides ground truth for the experimental data. The onboard computing platform of the robot comprises an industrial computer with 8GB RAM, powered by an Intel i5-8257U processor, and operates under Ubuntu 18.04.

For the experiments, real-world road data was collected from the community of Changchun, China, as indicated in red in Fig. 4. The selected experimental routes presented challenges, including random dynamic objects, lighting variations, and similar scene distributions. Subsequently, the mobile robot was deployed to construct the map of the experimental area. Fig. 4 presents the final map structure, which includes the point cloud map generated by the LiDAR SLAM system, the RGB image of each key point, and the corresponding depth image obtained by point cloud projection.

A. Image Search Experiment during Initialization

To assess the efficacy of our proposed method in comparison to other state-of-the-art techniques, we randomly gather images at the map scale and match them against images in the map library. We employ three widely-used image matching methods for comparison: ORB [28], SURF [29], and AspanFormer [27]. The efficiency is gauged by computing the time complexity of matching each image with its corresponding map image. Additionally, robustness is evaluated by examining the ability to handle variations in lighting and environmental changes induced by dynamic objects.

We curate three sets of images to evaluate the adaptability of the methods under diverse conditions:

- Set 1: Images captured in the afternoon, exhibiting different lighting conditions compared to the morning build.
- Set 2: Images gathered at deviations from the build trajectory, offering different perspectives and orientations compared to the original images.
- Set 3: Images collected with environmental alterations, such as vehicles parked on the roadside.

Our proposed method demonstrates superior performance across all three experimental sets, achieving 100% accuracy

TABLE I
IMAGE SEARCH EXPERIMENT RESULTS

Method	Group1		Group2		Group3	
	accuracy	time	accuracy	time	accuracy	time
ORB	76.7%	**1.68s**	63.3%	**1.71s**	70.0%	**1.68s**
SURF	96.7%	9.73s	86.6%	9.78s	93.3%	9.69s
Aspan	**100%**	14.61s	**100%**	14.32s	**100%**	14.43s
Proposed	**100%**	4.01s	**100%**	3.94s	**100%**	4.03s

while maintaining significantly lower time complexity. In Set 1, where images were captured under varying lighting conditions, ORB achieved 76.7% accuracy in 1.68 seconds. SURF improved accuracy to 96.7%, but required a longer duration of 9.73 seconds. AspanFormer achieved perfect accuracy but incurred a higher time complexity of 14.61 seconds. Set 2, involving images collected while deviating from the build trajectory, saw our proposed method outperforming the other methods with perfect accuracy and requiring only 3.94 seconds for image matching - thereby demonstrating its robustness to changes in viewing angle and orientation. In Set 3, where environmental changes such as vehicles parked on the roadside were introduced into the image, both AspanFormer and our proposed method maintained perfect accuracy; however, our method was more than three times faster. The results indicate that while traditional methods like ORB are faster, they compromise on accuracy. Conversely, learning-based methods like AspanFormer are accurate but computationally intensive. The proposed semantic graph-based image matching method ensures high accuracy while optimizing computational efficiency, rendering it ideal for real-time applications where both speed and accuracy are paramount.

B. Quantitative Assessment of Localization Precision

We employ ground truth poses, derived from a GPS receiver equipped with RTK technology, as the benchmark for evaluating the localization precision. The performance of our proposed DLT module is compared with two alternative methods:

- ZED2 odometry: an inherent feature of the ZED2 stereo camera.
- V-LOAM: a cutting-edge Visual-LiDAR SLAM system [30]

We conduct a quantitative examination of the positioning errors associated with the three methods, where the error is quantified as the Euclidean distance between the estimated pose and the ground truth pose. A boxplot representing the estimated positioning error for each method is illustrated in Fig. 5. The boxplot depicts the distribution of error values, encompassing the median, lower and upper quartiles, and minimum and maximum values. The box plot substantiates that our DLT module exhibits the lowest median and quartile values along with the narrowest range of error values, signifying its superior accuracy and consistency. The V-LOAM method displays higher median and quartile values and a broader range of error values, suggesting its moderate accuracy and consistency. The ZED2 odometry method presents the highest median and quartile values and the broadest range of error values, indicating its low accuracy and consistency,

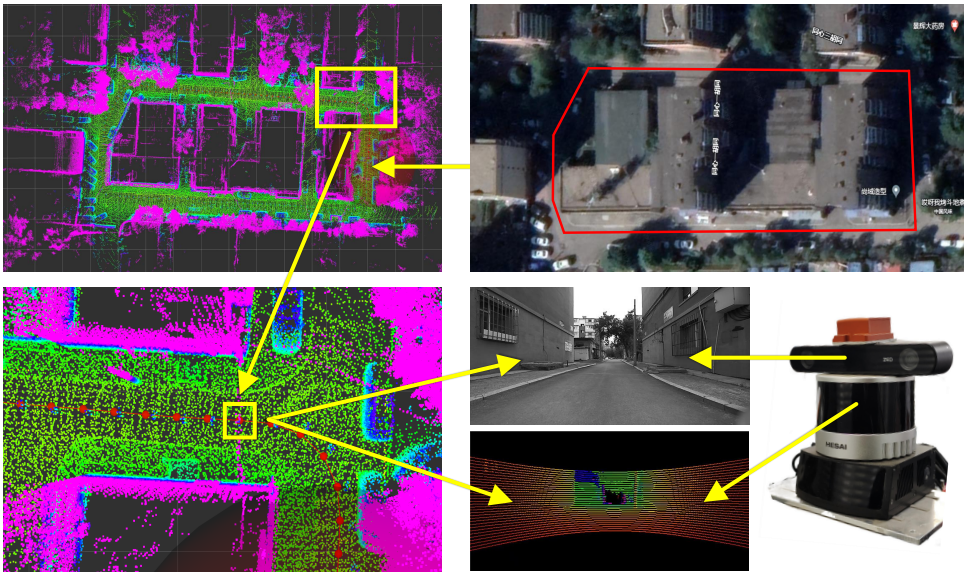

Fig. 4. The process of descriptor extraction.

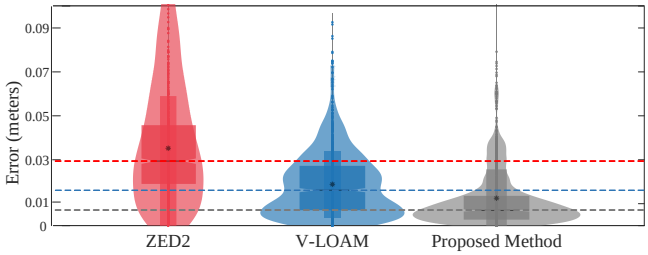

Fig. 5. Comparison of box plots for three methods.

TABLE II
COMPARATIVE RESULTS OF THE PROPOSED METHOD WITH ZED2 AND V-LOAM METHODS

	ZED2	V-LOAM	Proposed Method
Test Length(m)	1184	1184	1184
Max errors(m)	0.274	0.093	**0.079**
Mean errors(m)	0.029	0.015	**0.007**

which is attributed to the inherent cumulative error of visual odometry. TABLE II enumerates the maximum error and average error of the three methods. The proposed method eradicates the cumulative error engendered by the visual front-end through the correction thread, and the average positioning precision can be sustained within 0.01m, achieving centimeter-level positioning.

To delve deeper into the localization error of our DLT module, we also plot the error along the x, y, and z axes, as depicted in Fig. 6. The error is computed as the discrepancy between the estimated pose and the ground truth pose in each axis. The figure reveals that the error in the x and y axes is relatively minuscule and stable, indicating that our DLT module can precisely estimate the horizontal position. The error in the z axis is marginally larger and more fluctuating, suggesting that the vertical position estimation poses more challenges. However, the error in the z axis remains within an acceptable range, and does not impede the overall performance of our DLT module. The error analysis along the xyz axes corroborates the high accuracy and consistency of our DLT module in 3D localization.

V. CONCLUSION

This paper introduced a novel visual positioning system designed for mobile robot navigation within the environments devoid of GPS. The system is composed of three primary modules: system initialization, vision localization, and map update. It capitalizes on multi-modal sensors, semantic graph-based image matching, and an adaptive spanning transformer to accomplish robust and precise localization. The system's performance was evaluated using real road data gathered from a demanding urban scenario. Experimental outcomes demonstrated that our system surpassed state-of-the-art methods in terms of localization precision, efficiency, and adaptability. The system is versatile and can be deployed in a variety of applications necessitating reliable and exact positioning, including autonomous driving, indoor navigation, and augmented reality.

REFERENCES

[1] Zhangfang Hu, Wenhao Wang, Kuilin Zhu, Hongyao Zhou, and Jiangtao Chen, "Loop Closure Detection Algorithm Based on Attention Mechanism," *IAENG International Journal of Computer Science*, vol. 50, no.2, pp592-598, 2023

[2] Anna Gorbenko, and Vladimir Popov, "Visual Landmark Selection for Mobile Robot Navigation," *IAENG International Journal of Computer Science*, vol. 40, no. 3, pp134-142, 2013

[3] Khairunnisa' Ahmad Shahrim, Abdul Hadi Abd Rahman, and Shidrokh Goudarzi, "Hazardous Human Activity Recognition in Hospital Environment Using Deep Learning," *IAENG International Journal of Applied Mathematics*, vol. 52, no.3, pp748-753, 2022

[4] Lie Yu, Lei Ding, and Yukang Tian, "Tracking Control for Intelligent Tracing Car based on Novel Path Tracking Strategy," *IAENG International Journal of Applied Mathematics*, vol. 53, no.2, pp664-669, 2023

[5] Spachos, Petros, and Konstantinos N. Plataniotis, "BLE Beacons for Indoor Positioning at an Interactive IoT-Based Smart Museum," *IEEE Systems Journal*, vol. 14, no. 3, pp3483-3493, 2020

[6] Zhang, Ethan, and Neda Masoud, "Increasing GPS Localization Accuracy With Reinforcement Learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp2615-2626, 2020

[7] Yanhui Lv, Lei Chen, Deyu Zhang, and Jielin Liu, "Research on Indoor Mobile Node Localization Based on Wireless Sensor Networks," *Engineering Letters*, vol. 29, no.2, pp562-574, 2021

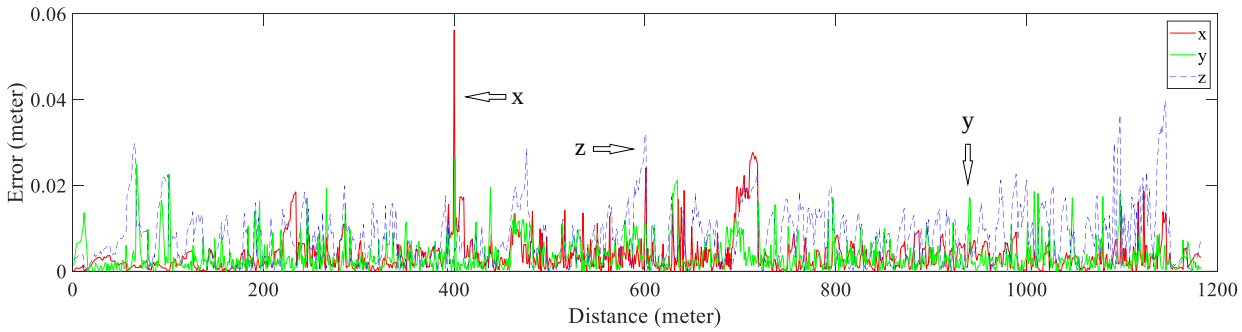

Fig. 6. The localization error of the proposed method.

[8] Xiuwei Xia, and Qian Sun, "An On-line Calibration Method of Star Sensor/Inertial Navigation System for Marine," *IAENG International Journal of Computer Science*, vol. 47, no.1, pp19-24, 2020

[9] Landgraf, Zoe, Fabian Falck, Michael Bloesch, Stefan Leutenegger, and Andrew J. Davison, "Comparing View-Based and Map-Based Semantic Labelling in Real-Time SLAM," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp6884-6890, 2020

[10] Yuan Luo, YongChao Zeng, RunZhe Lv, and WenHao Wang, "Dual-stream VO: Visual Odometry Based on LSTM Dual-Stream Convolutional Neural Network," *Engineering Letters*, vol. 30, no.3, pp926-934, 2022

[11] Yuqiong Wang, Tianqi Gu, Binbin Sun, Mengxue Xie, Song Gao, and Ke Sun, "Research on Motion Distortion Correction Method of Intelligent Vehicle Point Cloud Based on High Frequency Inertial Measurement Unit," *IAENG International Journal of Applied Mathematics*, vol. 53, no.1, pp1-8, 2023

[12] Campos, Carlos, Richard Elvira, Juan J. Gómez Rodríguez, José MM Montiel, and Juan D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp1874-1890, 2021

[13] Cvišić, Igor, Ivan Marković, and Ivan Petrović, "SOFT2: Stereo Visual Odometry for Road Vehicles Based on a Point-to-Epipolar-Line Metric," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp273-288, 2022

[14] Zhang, Ji, and Sanjiv Singh, "LOAM: Lidar Odometry and Mapping in Real-time," in *Robotics: Science and systems*, vol. 2, no. 9. Berkeley, CA, pp1-9, 2014

[15] Dellenbach, Pierre, Jean-Emmanuel Deschaud, Bastien Jacquet, and François Goulette, "CT-ICP: Real-time elastic LiDAR odometry with loop closure," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, pp5580-5586, 2022

[16] Leutenegger, Stefan, Paul Furgale, Vincent Rabaud, Margarita Chli, Kurt Konolige, and Roland Siegwart, "Keyframe-based Visual Inertial SLAM using Nonlinear Optimization," *Proceedings of Robotis Science and Systems (RSS) 2013*, 2013

[17] Qin, Tong, Peiliang Li, and Shaojie Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp1004-1020, 2018

[18] Khairuddin, Alif Ridzuan, Mohamad Shukor Talib, and Habibollah Haron, "Review on simultaneous localization and mapping (SLAM)," in *2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*. IEEE, pp85-90, 2015

[19] Chalvatzaras, Athanasios, Ioannis Pratikakis, and Angelos A. Amanatiadis, "A Survey on Map-Based Localization Techniques for Autonomous Vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp1574-1596, 2022

[20] Li, Yicheng, Zhaozheng Hu, Yingfeng Cai, Huawei Wu, Zhixiong Li, and Miguel Angel Sotelo, "Visual Map-Based Localization for Intelligent Vehicles From Multi-View Site Matching," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp1068-1079, 2020

[21] Li, Yicheng, Yingfeng Cai, Zhixiong Li, Shizhe Feng, Hai Wang, and Miguel Angel Sotelo, "Map-Based Localization for Intelligent Vehicles from Bi-Sensor Data Fusion," *Expert Systems with Applications*, vol. 203, p117586, 2022

[22] Sadli, Rahmad, Mohamed Afkir, Abdenour Hadid, Atika Rivenq, and Abdelmalik Taleb-Ahmed, "Map-Matching-Based Localization Using Camera and Low-Cost GPS for Lane-Level Accuracy," *Sensors*, vol. 22, no. 7, p2434, 2022

[23] Li, Liang, Ming Yang, Bing Wang, and Chunxiang Wang, "An overview on sensor map based localization for automated driving," *2017 Joint Urban Remote Sensing Event (JURSE)*, pp1-4, 2017

[24] Koide, Kenji, Jun Miura, and Emanuele Menegatti, "A portable three-dimensional LIDAR-based system for long-term and wide-area people behavior measurement," *International Journal of Advanced Robotic Systems*, vol. 16, no. 2, p1729881419841532, 2019

[25] Zhang, Guanyu, Kunyang Wu, Jun Lin, Tianhao Wang, and Yang Liu, "Automatic extrinsic parameter calibration for camera-lidar fusion using spherical target," *IEEE Robotics and Automation Letters*, 2024

[26] Cheng, Bowen, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-Attention Mask Transformer for Universal Image Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp1290-1299, 2022

[27] Chen, Hongkai, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan, "ASpanFormer: Detector-Free Image Matching with Adaptive Span Transformer," in *European Conference on Computer Vision*. Springer, pp20-36, 2022

[28] Rublee, Ethan, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*. Ieee, pp2564-2571, 2011

[29] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer, pp404-417, 2006

[30] Zhang, Ji, and Sanjiv Singh, "Visual-lidar Odometry and Mapping: Low-drift, Robust, and Fast," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp2174-2181, 2015

Kunyang Wu received the B.S. degree in measuring and control technology and instrumentation, in 2021 from Jilin University, Changchun, China, where he is currently working toward the Ph.D. degree in detection technique and automatic device. His current research focus is on the field of detection technique and automatic device, with a specific interest in multi-modal sensor calibration, sensor fusion, and the utilization of perception systems for autonomous navigation and vision applications.

Yang Liu received the Ph.D. degree from Jilin University, Changchun, China, in 2020. He is currently an Associate Professor at Jilin University. His research interests include stereo vision, 3-D reconstruction, texture filtering, and SLAM.

Guanyu Zhang received the M.S. and Ph.D. degrees in mechanical engineering from Jilin University, Changchun, China, in 2015. He is currently an Associate Professor with the Department of Instrument Science and Electrical Engineering, Jilin University. His research interests include mechatronics system design and artificial intelligence control technology research.