GENG5512 MPE Engineering Research Project Part 2

Final Report

Visual Localization and Navigation for a Humanoid Robot

Lingxiao Ma
24526617
School of Engineering, University of Western Australia

Supervisor: Thomas Braunl
School of Engineering, University of Western Australia

Word count: 6636

School of Engineering
University of Western Australia

Submitted: 18 May 2026

Project Summary
This research focused on realising autonomous localisation and navigation for a humanoid robot in indoor environments. Humanoid robots are widely seen as ideal platforms for real-world applications. Their human-like motion behaviours make them suitable for a broad range of scenarios. However, deploying existing localisation and navigation technologies on humanoid platforms continues to be challenging. Main issues include cumulative drift, sensor noise, and limited computing resources.

A complete navigation pipeline was designed and implemented on the Unitree G1 humanoid robot. The Intel RealSense D435i RGB-D camera provided depth and visual data. RTAB-MAP was used to build a 3D map of the indoor environment from depth images and odometry. An Extended Kalman Filter (EKF) was applied to fuse visual odometry and IMU measurements. This fusion provided stable and accurate pose estimates for the robot. A* was used for global path planning, and the Dynamic Window Approach (DWA) handled local obstacle avoidance. The full system was implemented under the ROS2 framework and tested in the UWA EE 313 Lab.

Functional validation confirmed that the system successfully built a consistent map of the indoor environment. The robot maintained stable localisation throughout the traversal. It navigated to pre-defined waypoints and avoided obstacles in real time. These results show the feasibility of deploying a vision-based localisation and navigation pipeline on a humanoid robot. This paper provides a technical reference for upcoming applications in service robotics and human–robot collaboration.

Acknowledgements
The author wishes to express sincere gratitude to Prof. Thomas Braunl and Dr. Oliver Zhang for their guidance and support throughout this project. His expertise in robotics and continuous feedback were invaluable to the completion of this work.

The author also thanks the School of Engineering at the University of Western Australia for providing access to the Unitree G1 humanoid robot platform.

Nomenclature

Symbol	Description
x	State vector
P	State covariance matrix
F	State transition matrix
Q	Process noise covariance matrix
z	Measurement vector
H	Measurement matrix
R	Measurement noise covariance matrix
K	Kalman gain
I	Identity matrix
v	Linear velocity (m/s)
ω	Angular velocity (rad/s)
Δt	Discrete time step (s)
d	Depth measurement from RGB-D sensor (m)
f	Camera focal length (pixels)
$g(v, \omega)$	Admissible velocity set in DWA
$h(n)$	Heuristic function in A*
$g(n)$	Path cost function in A*

1. Introduction

1.1. Research Problem Definition

In recent years, Humanoid Robots have gradually become a popular trend in robotic research and industrial applications. In contrast to traditional wheeled robots, humanoid robots have more similar body structures and movement patterns to human beings, which allows them to perform operations and interactive tasks in sophisticated and unstructured environments. This type of robot has demonstrated great application potential in scenarios such as services, medical rehabilitation, home assistance, and disaster relief. Thus, enhancing the ability of autonomous navigation in a realistic environment is highly noted in both the academic and industrial sectors.

1.2. Research Context and Gap

Reliable autonomous localization and navigation serve as a fundamental enabler of humanoid robots' practical deployment in complex real-world scenarios. However, existing localization and navigation techniques still face significant limitations in applying to humanoid robot platforms. A method that leverages wheel odometry has long served to estimate displacement based on the rolling motion of wheels. However, the indirect estimation means cannot benefit humanoid robots because the odometry is derived from joint encoders, which are less accurate than wheels. The accumulated error in joint-based odometry becomes amplified due to frequent gait and body posture adjustments. The Inertial Measurement Unit (IMU) is a type of device that can provide gesture and acceleration information, which is frequently used in robot platforms to estimate position and orientation. Whereas the cumulated drift due to long-time estimation will be substantial [19] and cause a loss in accuracy. The Global Positioning System (GPS) has been proven to be an effective solution in outdoor scenes with reliable signal coverage. Yet, with limited signal coverage in an indoor environment and disability to position the z-axis in 3D space, GPS is not robust for indoor localization and navigation. As some state-of-the-art designs, indoor positioning methods based on Bluetooth or ultra-wideband can provide promising, accurate localization for humanoid robots. However, its deployment cost is high and its reliability is insufficient, making it difficult to meet the demands of pervasive applications. Beyond using physical sensing modalities, Simultaneous Localization and Mapping (SLAM) [18] has emerged as a widely adopted computational framework to address challenges of real-time localization. It uses an incremental constructed representation to build maps and leverages visual [8] or LiDAR [3] features for data association to localize real-time position. For humanoid robots, which must operate in dynamic and cluttered indoor environments, SLAM methods have the potential to become a reliable solution.

1.3. Research Hypothesis

Based on the existing solutions and their respective limitations, this research will focus on providing an effective and robust localization and navigation method for humanoid robots, which can enable them to operate autonomously in indoor environments.

It is hypothesized that the integration of visual perception techniques, particularly visual SLAM [8] [9], with navigation and obstacle-avoidance algorithms will serve

as a fundamental enabler for humanoid robots to perform autonomous tasks in known indoor environments.

1.4. Research Significance

The rapid development of artificial intelligence (AI) in recent years shows unlimited potential in perception, reasoning, and decision-making. However, even though AI is powerful, its impact on the physical world remains limited unless it is embodied in a tangible platform. Robots serve as a natural medium to bridge the virtual and physical domains, enabling AI to exert direct influence on real-world tasks. Among various robotic platforms, humanoid robots are particularly well-suited to operate effectively in environments originally designed for humans. To fully realize their potential, fundamental capabilities such as perceiving surroundings, localizing themselves, and navigating to destinations are required. Therefore, research into reliable localization and navigation is important to provide an essential foundation for humanoid robots to act and carry AI in complex real-world environments.

2. Literature Review

2.1. Mapping Approaches

The primary task of Simultaneous Localization and Mapping (SLAM) is to estimate the robot's pose while incrementally building a representation of the surrounding environment. SLAM methods have been categorized in various ways depending on sensing modalities and algorithmic formulations. In this review, we focus on a sensor-based taxonomy and divide existing approaches into LiDAR-based and vision-based methods.

2.1.1. LiDAR-based SLAM

LiDAR-based SLAM also has multiple solutions. For 2D LiDAR SLAM, GMapping [1] employs a Rao-Blackwellized particle filter to construct 2D occupancy grid maps but is limited to relatively small-scale environments due to the particle degeneracy. Hector SLAM [2] leverages scan matching with a fast Gauss-Newton method and does not require odometry input, making it suitable for a lightweight platform. 3D LiDAR SLAM is a technique to build 3D maps using LiDAR. LOAM [3] decouples LiDAR odometry and mapping to achieve real-time 3D mapping with high accuracy and has become a foundation for many subsequent systems. LeGO-LOAM [4] optimizes LOAM for ground vehicles by segmenting point clouds into ground components and non-ground components, which reduces computational cost while maintaining accuracy. Fast-LIO [5] and FAST-LIO2 [6] tightly couple LiDAR and IMU, which deliver state-of-the-art real-time performance (see Figure 2.1). But because of the demands of high-quality sensors and computational cost, they are not suitable for lightweight platforms.

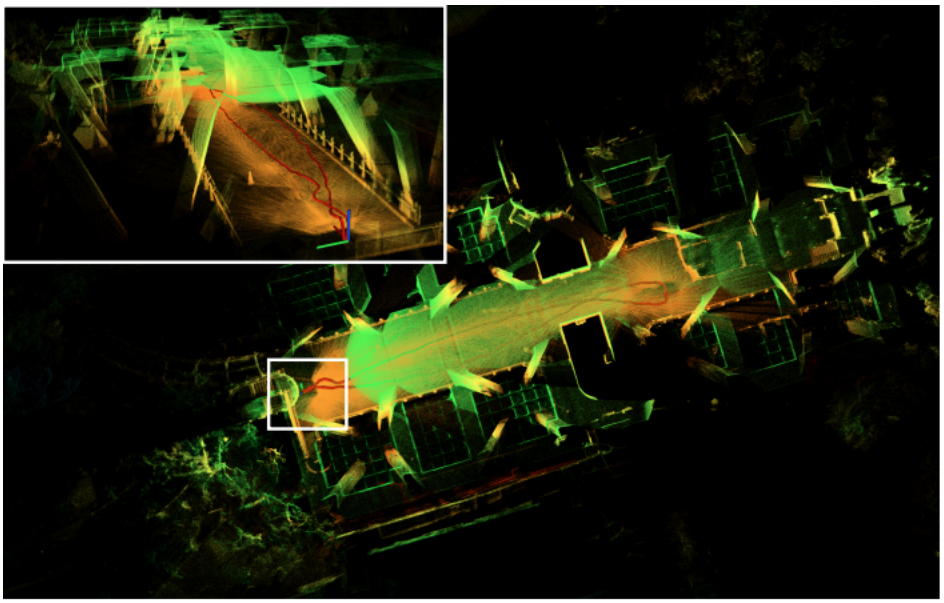

Figure 2.1: 3D point cloud map produced by Fast-LIO2 using tightly-coupled LiDAR–IMU odometry. Image taken from [6].

2.1.2. Visual-based SLAM

Visual-based SLAM (VSLAM) has become a dominant approach for indoor robotics due to the low cost and lightweight nature of cameras.

Feature-based methods rely on detecting and matching visual features such as corners and descriptors across frames and estimating the motion while building maps. PTAM [7] is the pioneer of tracking and mapping feature points, but it is limited to small-scale environments. ORB-SLAM2 [8] extended to monocular, stereo, and RGB-D inputs with robust loop closure (see Figure 2.2). ORB-SLAM3 [9] optimizes the latest version by integrating inertial data and multi-map support for improved robustness.

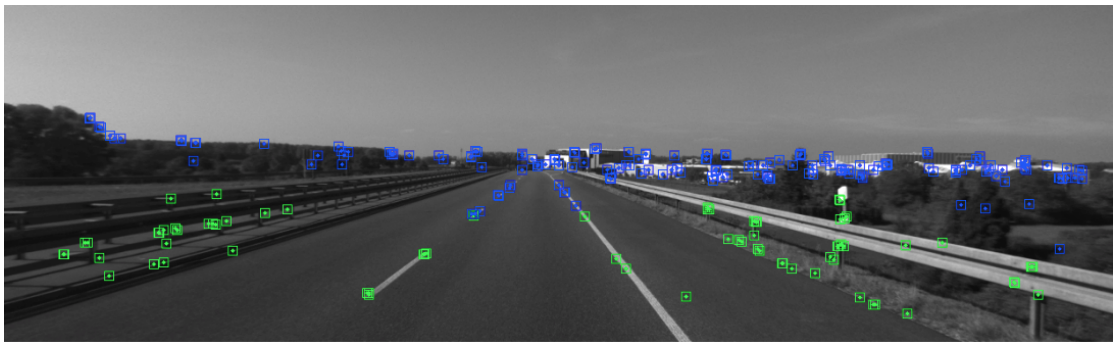

Figure 2.2: Feature tracking by ORB-SLAM2 on the KITTI dataset, where green markers denote near features and blue markers denote far features. Image taken from [8].

Graph-based methods are used to represent robot poses as nodes and constraints as edges, using loop closure and optimization to maintain global consistency. Rtab-Map is the most widely used method in this subcategory. It is a real-time appearance-based SLAM framework that builds large-scale maps using visual loop closure detection and a memory management strategy, making it suitable for long-term indoor mapping and localization tasks.

The visual-inertial fusion method integrates camera data with IMU to reduce drift and improve robustness in long-term navigation. OKVIS [10] implemented tightly coupled optimization of visual and inertial measurements. VINS-Mono [11] and VINS-Fusion demonstrated robust state estimation by fusing camera and IMU data.

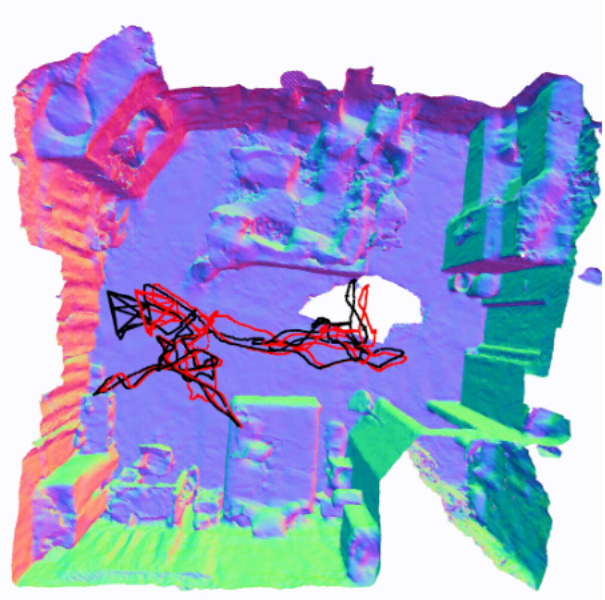

Figure 2.3: Dense scene reconstruction by NICE-SLAM, demonstrating neural implicit representation. Image taken from [13].

Deep learning methods have emerged in recent years. They are employing a deep neural network to learn motion, depth, or scene representations. DeepVO [12] applied CNNs and RNNs for end-to-end visual odometry. NICE-SLAM [13], NICER-SLAM [14], and NeRF-SLAM [15] exploited neural implicit representations to achieve dense, high-quality maps (see Figure 2.3), but their heavy computational demand hinders real-time mapping.

2.2. Localization Techniques

Localization technology aims to estimate the pose of a robot based on a known map. Odometry and methods based on inertial measurement units (IMUs) provide short-term pose estimation by integrating wheel encoders or inertial measurements. Although they are computationally efficient, both methods suffer from cumulative drift and limited long-term accuracy on humanoid robot platforms. Global Positioning System (GPS), Ultra-Wideband (UWB), and Bluetooth can offer absolute positioning, but they are not suitable for humanoid robots operating in indoor environments with poor signal reliability.

The probabilistic method forms the basis of modern positioning technology [16]. The Extended Kalman Filter (EKF) [17] has become one of the most widely used solutions. The EKF models the robot's state as a probability distribution and updates the attitude estimation by combining the predictions with the sensor observation results [18], [19]. It can handle nonlinear systems and is therefore particularly suitable for modeling practical problems.

2.3. Navigation and Obstacle Avoidance Strategies

Navigation strategies provide the decision-making layer that connects localization with actuation. It is responsible for guiding humanoid robots to their destinations. Classical graph-search, such as A* [20] and Dijkstra [21], compute collision-free paths on a grid or graph representation of the environment. Sampling methods such as RRT [22] and its variants extend path planning to high-dimensional

configuration spaces, providing feasible solutions in complex environments. Adaptive methods such as D* [23] and D* Lite [24] further enable incremental replanning in dynamic or partially known maps, which enhances the robustness to environment changes.

For obstacle avoidance, reactive methods are used in real time to handle unforeseen obstacles. Potential Field methods [25] generate attractive and repulsive forces to guide robots. But they suffer from falling into local minima. The Dynamic Window Approach (DWA) samples feasible velocities and optimizes for safety. It achieves reliable obstacle avoidance in cluttered environments.

In many modern frameworks, combining global planning with local obstacle avoidance provides a high-level route planner under dynamic conditions. The layered structure is particularly suitable for humanoid robots.

3. Project Objectives

3.1. Hypothesis

In this research, the Hypothesis is that the humanoid robots can autonomously localize their position and plan routes to navigate to a preset destination by combining Visual SLAM methods and navigation algorithms in a known cluttered indoor environment. Visual SLAM provides promising and stable mapping and localization, and then local path planning methods and obstacle avoidance algorithms offer the robot the ability to make decisions.

3.2. Objective

3.2.1. Mapping

In this research, mapping is to provide a dense and consistent map of the laboratory indoor environment for a humanoid robot. It is expected to exactly extract features in the environment, such as corners, walls, and ground. To evaluate the effectiveness of mapping, the map accuracy is used for measurement. To be specific, the mapping error should be no more than 5 cm. Satisfying this, the mapping result can be thought of as having good robustness and practicality.

3.2.2. Localization

The goal of localization is to enable the robot to achieve continuous, stable, and high-precision autonomous positioning in a known map. The evaluation indicators mainly include the mean error of position and attitude. The stability during long-term operation is also considered. To be specific, the planar position error should be less than 0.3 m and the attitude angle error less than 5°. The TF tree should remain continuously published throughout a 10-minute continuous operation test, and the system should be able to recover quickly from any transient pose estimation loss. Only when these conditions are met can it be considered that the positioning system has good robustness and practicality.

3.2.3. Navigation

The goal of navigation is to plan an efficient and feasible path for the robot from the starting point to the target point. The evaluation metric is planning efficiency. Every route planning should take less than 1 second. Satisfying this, the navigation system can be considered to have practical real-time capability.

4. Implementation and Methodology

4.1. Research Equipment and Environment

4.1.1. Research Equipment

Figure 4.1: Unitree G1 Humanoid Robot Parameter

In the research, the Unitree G1 Humanoid Robot is used as the experiment platform. This robot has motion characteristics similar to the human body. Its sensing system is equipped with an Intel RealSense D435i RGB-D camera, which can simultaneously obtain color images and depth information. The D435i RGB-D camera is also integrated with an IMU that can provide acceleration and angular velocity data. Although the Unitree G1 platform also integrates an onboard 3D LiDAR, this research deliberately adopts a pure vision pipeline based on the D435i RGB-D camera in order to align with the project's core research scope on visual SLAM for humanoid platforms. These real-time data can be used for SLAM and orientation estimation. In addition, the joint encoder of G1 can provide real-time feedback on changes in joint angles. They offer additional input for motion estimation and gait control. Figure 4.1 shows the parameters of Unitree G1.

The computing part will rely on a laptop as the core processing platform. This notebook runs the ROS2 framework and is equipped with an NVIDIA RTX 5070Ti GPU.

4.1.2. Research Environment

The experiment will be carried out in an indoor environment in the UWA EE 313 LAB. The static obstacles, such as tables, chairs, and cabinets. In some experiments, dynamic obstacles are also artificially set up. It is used to verify the ability to avoid dynamic obstacles.

4.2. Implementation Details

When determining the technical solution in this study, three core engineering principles were mainly followed. The first is the robustness to the gait vibration of humanoid robots. Secondly, there is the issue of computing efficiency. Because the system is to deploy on embedding platform, the real-time response speed of the system is maintained. he last aspect is environmental adaptability, ensuring that the algorithm can adapt to the messy and unstructured indoor environment of a laboratory.

4.2.1. Mapping Stage

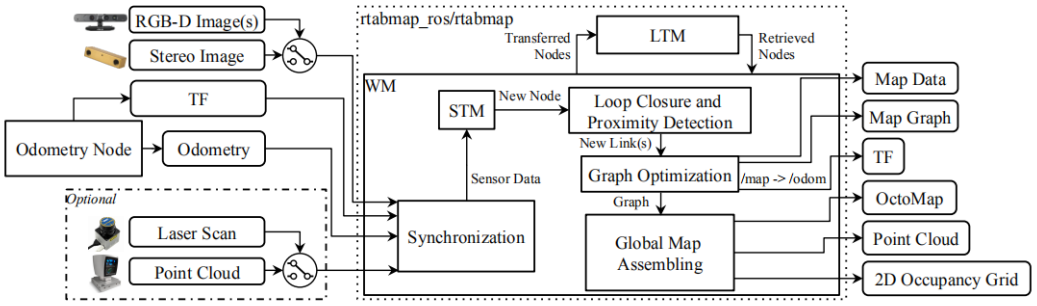

Figure 4.2: Illustration of RTAB-MAP. Image taken from RTAB-MAP official documentation (http://introlab.github.io/rtabmap).

RTAB-MAP (Real-Time Appearance-Based Mapping) is the framework (Figure 4.2) used in the mapping stage. The input includes depth data and color information provided by an RGB-D camera. The system leverages the Odometry Node to complete the primary pose estimation and then synchronizes the results with information such as images and point clouds. New nodes will be stored in Short Term Memory (STM) and used to perform loop closure and proximity detection to recognize the similarity between the current frames and the previous frames. Whenever the closure is detected, the system will leverage graph optimization to adjust the overall topology diagram. The optimized nodes will be transferred to Long Term Memory (LTM) to ensure consistency. In a global view, RTAB-MAP finally delivers multiple map formats.

In this research, the G1 acquires color and depth images from the Intel RealSense D435i RGB-D camera. And at the same time, it uses the IMU sensor to provide attitude and acceleration information. Data are used as input for RTAB-MAP, and then the map and the odometry with respect to the map are generated. During the experiment, the robot was controlled to walk around a designated path in a preset indoor environment to ensure coverage of the main spatial areas.

Compared with the pure feature point methods, such as ORB-SLAM3 mentioned in the literature review, RTAB-MAP has a unique long-term and short-term memory mechanism and appearance-based closure detection. This makes it much more robust when dealing with sharp turns or violent shaking. The LiDAR methods are excellent solutions, and The Unitree G1 platform itself integrates 3D LiDAR. But this research deliberately chose the pure visual pipeline based on RGB-D cameras. This not only closely aligns with the core research scope of this project,

but also because visual sensors can capture the rich color and texture features in the real world.

4.2.2. Localization Stage

Localization-only mode of RTAB-MAP is used to provide vision-based positioning service. In this mode, The global map within the system has been frozen. RTAB-MAP stop adding nodes to LTM to reduce computational overhead. During the movement of the robot, the system will extract the RGB-D visual features sent back by the current Intel RealSense D435i in real time and continuously compare them with the saved static global map. Once the visual features of the current frame successfully match the historical nodes in the map, RTAB-MAP can precisely calculate the absolute pose of the robot in the current global map coordinate system through the perspective N-point algorithm (PnP) or feature alignment.

However, due to the intense trunk swaying of humanoid robots when they walk, pure visual feature matching may experience a brief loss. Besides, Its update frequency is limited by the camera frame rate. Therefore, this visual pose estimation output by RTAB-MAP is not directly used for control but serves as a high-precision low-frequency "observation".

Extended Kalman Filter (EKF) is to fuse multiple sources of data. The Kalman Filter leverages the state space model of the system, and recursively estimates the optimal state. EKF is the non-linear extension suitable for a non-linear motion model. It assumes the system state as (4.1):

$$x_k = f(x_{k-1}, u_k) + w_k \tag{4.1}$$

Where x_k is the system state vector at time stamp k, $u_k$is the input control, and $f(\cdot)$ is the nonlinear state transition function. $w_k \sim \mathcal{N}(0, Q)$ is Wiener Process. Observation model is (4.2):

$$z_k = h(x_k) + v_k \tag{4.2}$$

Here, z_k represents the observable quantity, $h(\cdot)$ is the nonlinear observation function, and $v_k \sim \mathcal{N}(0, R)$represents the observation noise.

EKF can be generally regarded as two parts in a single step. The prediction is to estimate the k by $k-1$. The formulas (4.3) and (4.4) are the a priori estimations of state and variance.

$$\hat{x}_{k|k-1} = f\left(\hat{x}_{k-1\,|\,k-1}, u_{k-1}\right) \tag{4.3}$$

$$P_{k\,|\,k-1} = F_k P_{k-1\,|\,k-1} F_k^T + Q_{k-1} \tag{4.4}$$

Then update the state estimation, the Kalman Gain is calculated by (4.5);

$$K_k = P_{k\,|\,k-1} H_k^T \left(H_k P_{k\,|\,k-1} H_k^T + R\right)^{-1} \tag{4.5}$$

The Kalman Gain is used as a weight to balance prediction and observation. If the observation noise is relatively large, the prediction result is dominant. When the reliability of the observation is relatively high, the weight of the observation results is greater. Then the updating process is performed by (4.6) and (4.7):

$$\hat{x}_{k\mid k} = \hat{x}_{k\mid k-1} + K_k\left(z_k - h\left(\hat{x}_{k|k-1}\right)\right) \quad (4.6)$$

$$P_{k|k} = (I - K_k H_k)P_{k|k-1} \quad (4.7)$$

Through this step, the system fuses the predicted values with the observed values, making the state estimation closer to the true value.

In application to Unitree G1, EKF is the core data fusion module used to integrate multiple sensor data. It fuses the Unitree SDK odometry information with the visual localization information generated by RTAB-MAP. Through this cycle of prediction and update, EKF achieves dynamic trade-offs among different data sources. The final output of the odom2map pose provides precise location information for subsequent path planning and obstacle avoidance.

In the selection of positioning schemes, the Extended Kalman filter is considered superior to the traditional linear filter. Because it can model the complex nonlinear motion characteristics of humanoid robots more accurately. Although pure visual odometers are susceptible to "image jumps" caused by gait vibrations, EKF greatly smooths the pose output by weighting and fusing high-frequency IMU predictions with stable visual observations. It provides reliable continuous positioning data for the subsequent navigation module.

4.2.3. Navigation Stage

Navigation is the decision-making module for the system. It is used to receive localization information and pose information and plan the best path to the destination. Then, the path would be transferred to a particular motion control.

Global Planning is responsible for finding an optimal path from the current position to the destination. A* combines the actual cost $g(n)$ with the heuristic estimation $h(n)$. $f(n) = g(n) + h(n)$ is the evaluation function to guide the searching process. This way, it can not only ensure the feasibility and proximity to the optimum of the path but also improve the search efficiency. A* is applied to the two-dimensional occupation raster map generated by RTAB-Map for calculating the global path of the robot from the current position to the target point. Due to the fact that the movement of humanoid robots in complex indoor environments requires consideration of environmental layout and obstacle distribution.

Local Planning uses the Dynamic Window Approach (DWA). It enumerates possible combinations of velocities under the given constraints of velocity and acceleration, and predicts short-term trajectories using kinematic models. Then, the trajectory's proximity to the target, path alignment, and safety with respect to obstacles are comprehensively evaluated through the scoring function. Based on the condition, the optimal velocity will be chosen. In this research, DWA is used to implement a reaction to sudden obstacles.

Nav2 (Navigation2) is a powerful navigation stack in ROS2 environment. It leverage Behavior Trees to arrange complex navigation tasks. Global planning is handled by the Planner Server of Nav2, which is configured to calculate the shortest path A* on the global cost map. Meanwhile, the Controller Server utilizes the DWB local planner (that is, the advanced implementation of DWA in Nav2) to evaluate feasible trajectories. It generates secure speed instructions based on

real-time local cost maps. Nav2 is capable of dynamically updating its cost map and safely driving Unitree G1 to its destination by issuing instructions to the /cmd_vel topic.

The navigation framework chose the mature Nav2 because it utilizes the "behavior tree" to achieve a high degree of modularization in task scheduling. Considering stability, Nav2 can reach an industrial level. In terms of the specific algorithm combination, A* is chosen for global planning because it can efficiently find the theoretical optimal path on the 2D occupied map. The choice of DWA for local planning is due to its extremely high response speed. It enable robot balance the optimality of the path and the safety of real-time obstacle avoidance.

4.3. Overall Pipeline

The pipeline integrates multiple sensing and computing modules to enable humanoid robots to navigate autonomously in indoor environments. RGB-D and IMU data are first input into RTAB-Map to generate dense maps of the environment. Meanwhile, the Pose estimated based on visual information is generated by RTAB-MAP. In the localization stage, a sparse map is gcUnitree SDK and visual localization information will be fused by EKF, and generate odometry based on the map. Finally, in the navigation stage, the A* and DWA algorithms are used to implement path planning and obstacle avoidance. The velocity data will be sent to the /cmd_vel topic to control the robot's move to the destination. The pipeline overview shows as Figure 4.3.

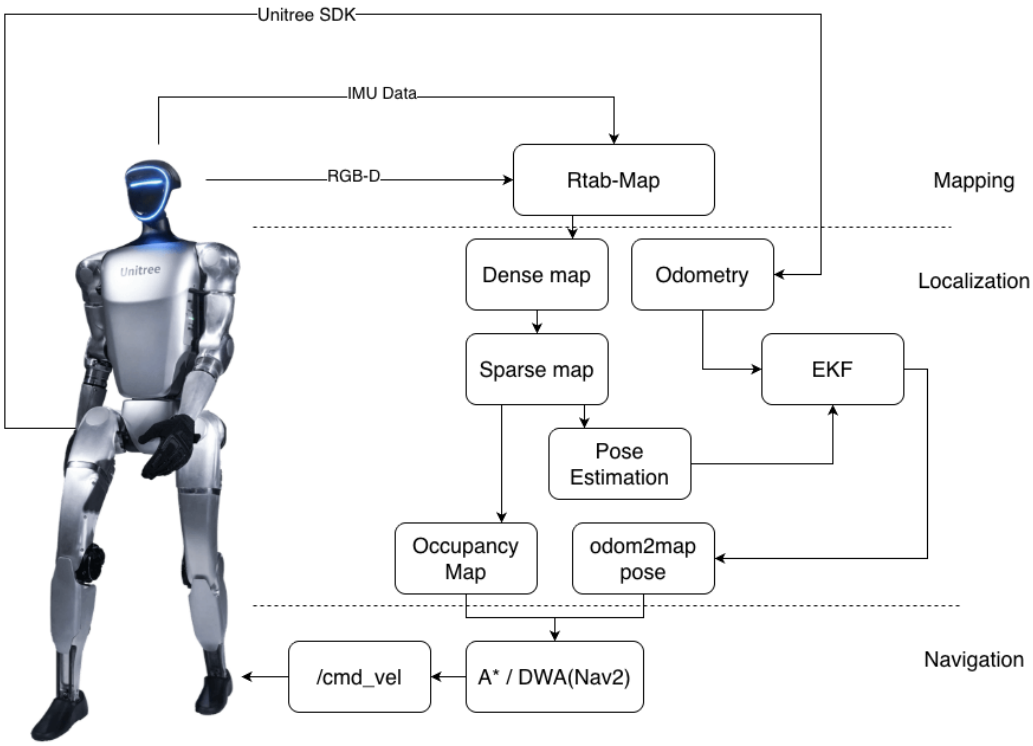

Figure 4.3: Pipeline illustration of the implementation

This specific hierarchical architecture is specifically designed to address the inherent challenges of humanoid robot platforms. Due to Severe trunk swaying and nonlinear dynamic changes, relying solely on Visual Odometry is highly prone to tracking loss and severe cumulative drift. This research introduce EKF to fuse vision-based localization information with physical odometry inside robot. By dynamically update positioning information up to multi-sourced data enhance the robustness of localization function.

On the other hand, decoupling of global planning and local controller ensured the robot is able to process long-term path planning and real-time unseen obstacle avoidance. The low-frequency global optimal path planning is carried out by using the downsampled 2D Occupancy Map, which greatly saves on-board computing resources.

4.4. Experimental Procedures and Measurement Methodology

4.4.1. System Initialization

The experiment was conducted in the UWA EE 313 laboratory. Before starting the test, place the Unitree G1 at the preset starting zero point in the laboratory. Then make sure there is no obstruction within the field of view of the Intel RealSense D435i camera. After the system starts up, first confirm that the TF coordinate transformation tree is loaded correctly and record the initial pose as the reference benchmark for subsequent navigation.

4.4.2. Operational Procedures

During the mapping stage, the robot is controlled to walk in the laboratory at a constant speed of approximately 0.2m/s through a remote controller. To ensure the consistency of the map, the path planning covers the main areas of the laboratory and deliberately makes slow turns in feature-rich areas (such as corners).

During the navigation stage, set the target waypoints on the constructed static map. The navigation task is carried out in two groups: one group is an open path without obstacles, and the other group places static obstacles artificially in the path. These are to test the coordination ability between A* global planning and DWA local obstacle avoidance under the Nav2 framework.

4.4.3. Ground Truth and Accuracy Verification

Due to the lack of an external dynamic capture system in the laboratory, this study adopted the physical marker-assisted method to obtain the ground truth. In the autonomous navigation test, whenever the robot reaches the preset point, the actual physical deviation between the robot center and the ground marking is manually measured. And synchronously record the coordinate values output by the ROS system to calculate the Euclidean distance between them. On the other hand, the mapping accuracy is verified by comparing the obstacle spacing measured in the software with the physical distance measured by the tape measure.

4.5. Health and Safety Protocols

This research strictly adheres to the UWA EE 313 laboratory safety guidelines. To ensure operational safety, a "deadman switch" mechanism is integrated into the robot's motion control logic. The Unitree G1 executes the motion command only when the operator continuously presses the button**.** Once the operator releases

the button, the robot will immediately stop. During non-testing or software debugging periods, the robot is always safely suspended on the bracket**.**

5. System Evaluation and Results

This chapter will present and clarify the experimental results and comprehensive discussion of the autonomous localization and navigation system implemented on the Unitree G1 humanoid robot. The research systematically evaluated 3 key components: The consistency of mapping, the accuracy of positioning, and the efficiency of navigation. The following sections will analyze these findings in combination with the initial goals set for the project. In the end, this chapter will discuss the inherent limitations of the proposed pure visual pipeline.

5.1. Final System Design and Implementation

This section shows the final integration of the visual positioning and navigation system on the Unitree G1 humanoid robot platform. The system has achieved full-function operation under the ROS2 framework. The system fully takes advantage of the outstanding computational power provided by the laptop to support High-resolution depth processing and real-time path planning. Figure 5.1 shows the physical system that was field-tested in the laboratory environment of UWA EE 313.

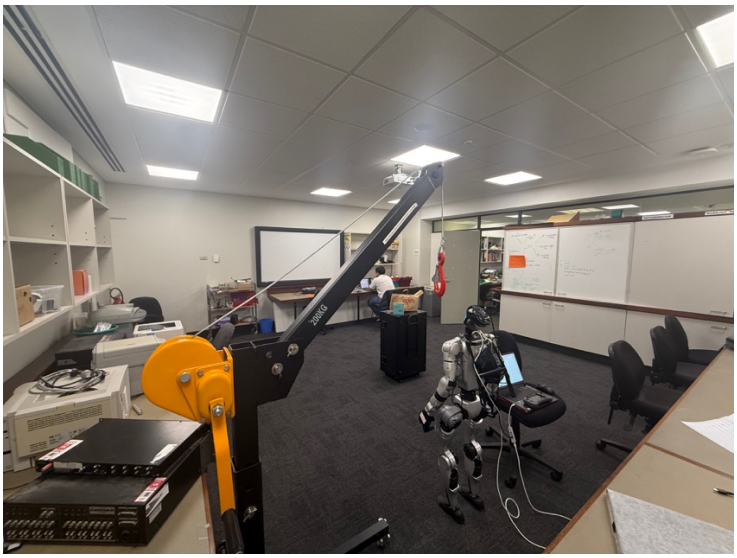

Figure 5.1: Unitree G1 deployed in the UWA EE 313 laboratory

At the software level, Figure 5.2 shows the real-time system monitoring interface implemented through RViz. The figure clearly shows that the RGB-D raw data collected from the Intel RealSense D435i has been successfully transformed into a high-precision spatial point cloud.

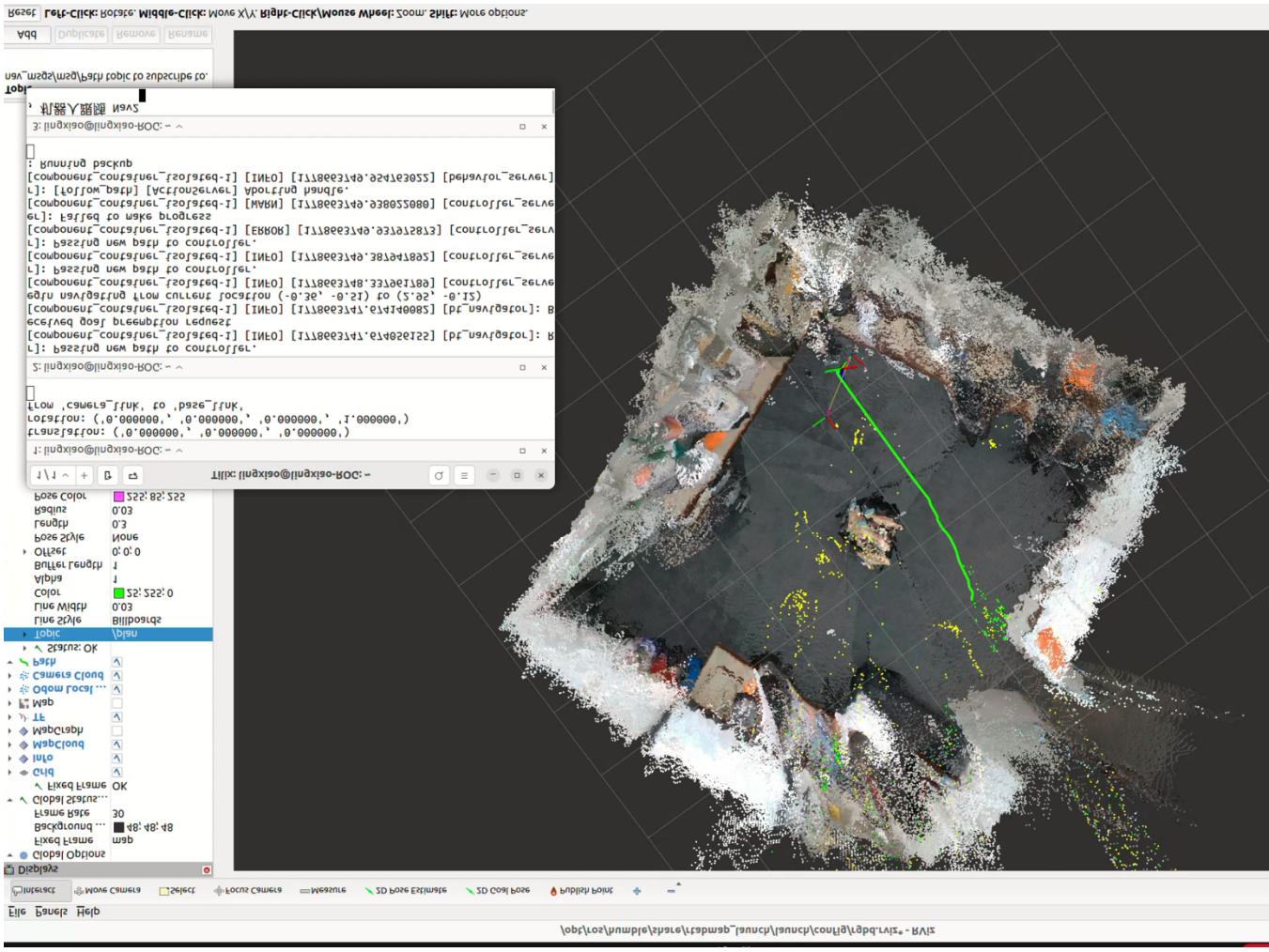

Figure 5.2: RViz real-time monitoring interface

Figure 5.3 shows the complete TF tree generated by the system. The tree is rooted at the global map frame (map), which connects to the odometry frame (odom) at approximately 20 Hz, then to camera_link, from which base_link and all sensor-specific frames branch out. This confirms that the full transformation chain from map to the robot base and all sensor endpoints is correctly loaded and continuously updated.

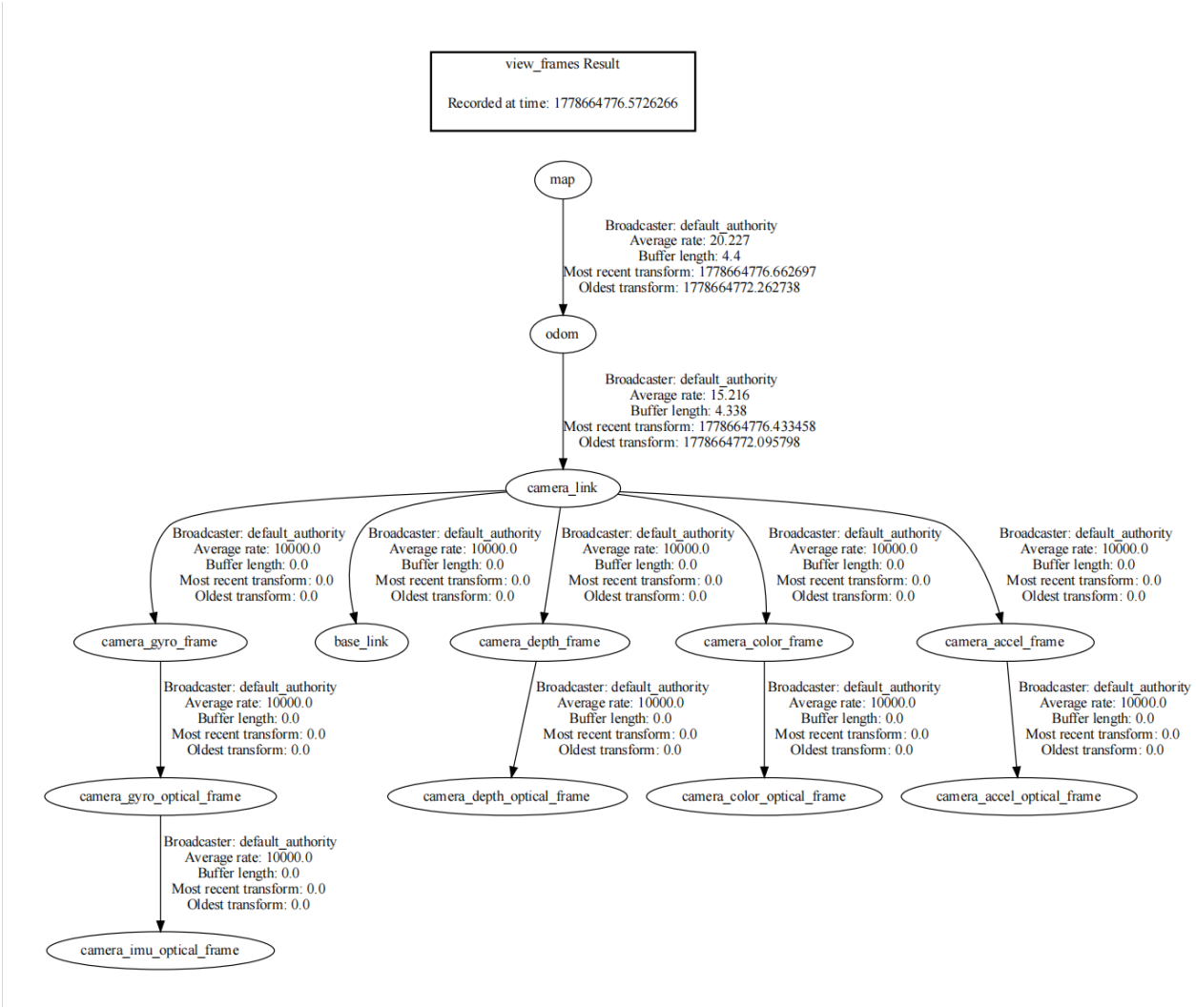

Figure 5.3: Complete TF transformation tree

The initial functional tests verified the communication logic among the various modules of the system. RTAB-MAP can stably receive and fuse odometer feedback. Moreover, there was no coordinate fracture or obvious conversion delay in the coordinate transform chain. The entire pipeline of the robot positioning and navigation system has been normally activated and is ready for subsequent evaluation and testing.

5.2. Mapping Performance

The mapping experiment was conducted in the UWA EE 313 laboratory. During the mapping process, the system controls the robot to perfor remote traversal at a constant speed of approximately 0.2 m/s. To ensure the full extraction of map features, the robot's trajectory was designed to cover the entire circumference of the laboratory and perform slow in-place rotations at the feature-dense corners.

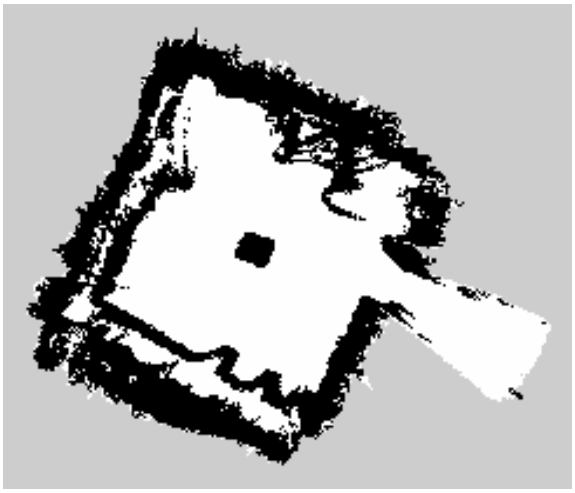

Figure 5.4: 2D occupancy grid map output by RTAB-MAP

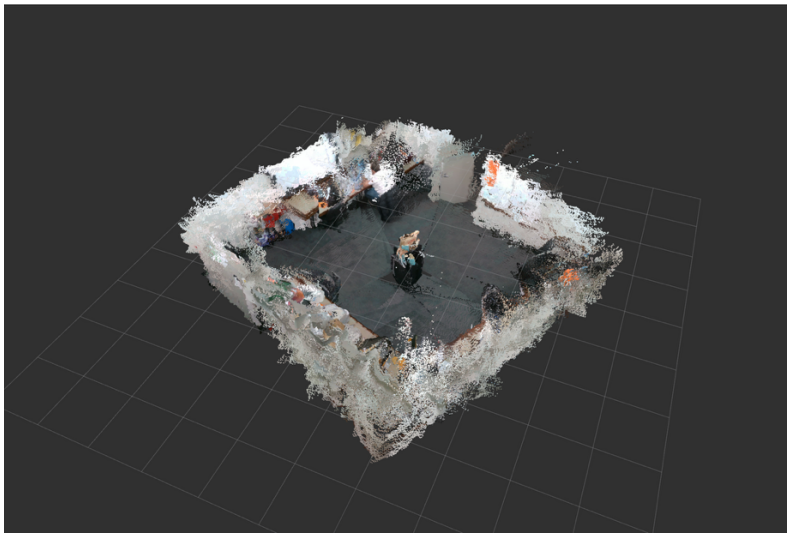

Figure 5.5: RGB-coloured 3D point cloud map generated by RTAB-MAP

Figures 5.4 and 5.5 respectively show the 2D Occupancy Grid Map and 3D dense Point Cloud Map that were finally output in the mapping stage. From a qualitative perspective, the system successfully captured and reconstructed the geometric structure of the laboratory. In Figure 5.4, the occupied raster map presents clear boundaries, and the black areas accurately depict the projections of walls and static obstacles such as tables and cabinets. The white area clearly demarcates the free space for passage. The raster Resolution was set at 0.05 meters per pixel. And there were no obvious blurred boundaries on the map as a whole.

The 3D point cloud map in Figure 5.5 was rendered with high fidelity using the RGB color information provided by the Intel RealSense D435i. All objects and obstacles in environment can be distinguished clearly. Thanks to the excellent short-term/long-term memory (STM/LTM) management mechanism of RTAB-MAP, the system did not experience any obvious memory overflow or lag when processing high-resolution point clouds.

In the mapping performance, what is worth mentioning is the overall acceptable performance of coping with cumulative drift. Due to the inevitable trunk swaying

of humanoid robots during walking, the pure visual odometer experienced a slight trajectory deviation at the initial stage of mapping. However, when the robot completed one round of traversal and approached the starting area again, RTAB-MAP successfully triggered the global loop closure detection through visual feature matching based on appearance. The system immediately and automatically executed the pose map optimization, which helps correct the accumulated error previously. So that the final output global map could be good enough for further downstream tasks.

To quantitatively verify the measurement accuracy of the mapping and evaluate whether it meets the core objective of "mapping error no more than 5 cm" set in Chapter 3. This research measured three static features: the width of a desk, the width of a door, and the length of the room. The actual physical distances of these three locations measured in reality using a tape measure are 0.80 m, 1.20 m, and 6.00 m, respectively. Subsequently, the ranging tool in RViz is used to measure the same feature points in the generated 3D map. The computational values in reconstructed map are 0.82 m, 1.15 m, and 5.96 m, respectively.

After comparative calculation, the absolute errors of these three groups of measurements are respectively 2 cm, 5 cm and 4 cm, and the maximum absolute error is 5 cm. This result meets the preset 5 cm error threshold. It strongly proves that the proposed pure visual mapping pipeline has a high degree of measurement accuracy in complex indoor environments.

5.3. Localization Accuracy

Continuous and stable positioning is a prerequisite for humanoid robots to perform navigation tasks. In order to evaluate the accuracy of the system in the localization-only mode, the research evaluates error by computing the error between real-world relative coordinates and localization system outputs. A total of 10 independent positioning tests were conducted in the experiment. The robot is manually controlled to move to a group of specific positions with known coordinates. Observed values are recorded, and the error is computed with the ground truth.

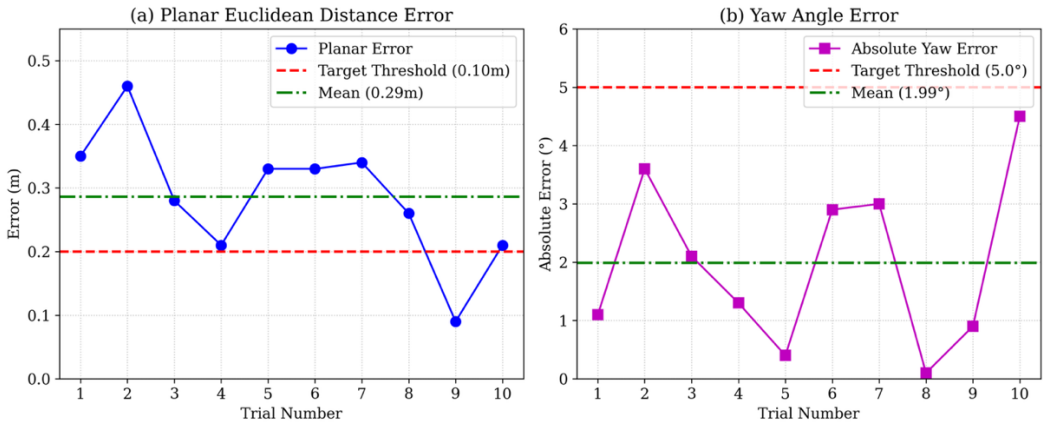

Figure 5.6: Planar position error across 10 localization trials

The X-axis deviation, Y-axis deviation, planar Euclidean distance error, and yaw angle error in each of these 10 tests are recorded in Appendix A. Figure 5.6 presents the planar error fluctuations of these 10 tests.

Through the comprehensive calculation of 10 sets of test data, the mean of the system's planar position error is 0.29 m, and the average yaw angle error is 1.99 degrees. Although the planar error failed to reach the extremely strict theoretical threshold of 0.1m preset in Chapter Three, in practical engineering applications, this performance is already very close to the ideal state. It can fully guarantee the safety requirements of regular indoor navigation. This deviation mainly stems from the unique mechanical constraints of humanoid robots. Without the assistance of any external motion capture system, achieving an absolute positioning accuracy of approximately 0.3m solely relying on the built-in pure vision and IMU fusion pipeline has fully demonstrated the system's high reliability and practical value in complex and dynamic indoor environments.

In addition to the accuracy of a single positioning, the stability of the system during operation is equally crucial. This study conducted a 10-minute continuous cruise test in a laboratory environment. It should be particularly noted that setting the test duration to 10 minutes is not limited by the software algorithm. Due to the inherent communication bottleneck between the underlying hardware of Unitree G1 and laptops, the connection would unpredictable disconnected. Therefore, this test aims to verify the robustness of the algorithm pipeline within the longest reliable time window allowed by the hardware. Throughout the entire 10-minute testing period, the TF tree maintained continuous release. Even sometimes the pose estimation loss is still unavoidable, but the system can quickly recognize new poses while moving in the environment.

In the design of this system, EKF efficiently weighted and fused the high-frequency Unitree SDK odometry information with the low-frequency but globally consistent RTAB-MAP visual localization information. This multi-sensor fusion strategy greatly smooths the positioning output curve and effectively filters out the noise disturbance caused by mechanical gait.

5.4. Path Planning Performance

After verifying the consistency of mapping and the accuracy of positioning, this section focuses on evaluating the global path planning and trajectory execution capabilities of the Nav2 navigation stack in a statically known environment.

The test was carried out in the area with more complex terrain in the UWA EE 313 laboratory. The testers, through the RViz visualization front end, assigned the robot long-range Goal

poses spanning multiple obstacle areas on a 2D occupied raster map. To ensure the robot's physical safety, the system has preset an obstacle expansion radius of 0.5 meters in the global costmap.

Figure 5.7: Nav2 global path planned by A*

As shown in Figure 5.7, the Behavior Tree of Nav2 responds promptly after receiving the target. The A* algorithm conducts heuristic search within the security boundary. An optimal collision-free trajectory that avoids static obstacles, such as laboratory benches and load-bearing columns, has been successfully generated. The system background log shows that the average calculation time from receiving the target coordinates to generating this long-distance global path is around 0.7 seconds. This result not only falls far short of the preset efficiency indicator of 1.0 second but also demonstrates the system's outstanding computing power scheduling capability when dealing with high-resolution maps.

During the subsequent path execution phase, the upper-level Nav2 navigation stack continuously generates smooth linear velocity and angular velocity instructions and publishes them to the /cmd_vel topic. However, since G1 is a complex bipedal walking platform, the standard kinematic instructions generated under the ROS2 framework cannot directly drive its hardware joints. For this purpose, the system subscribes to the /cmd_vel topic through a specially deployed motion transformation node and invokes the underlying Unitree SDK in real time. High-level speed instructions are precisely mapped and forwarded to the core Gait Controller of the robot. Due to the personnel safety regulations in the laboratory and the steady-state walking limitations of bipedal robots, the maximum linear speed issued through the SDK is strictly restricted to 0.2 m/s. During the actual

movement, thanks to the low-delay pose feedback provided by the EKF filtering mechanism described in the previous section, the robot demonstrated excellent yaw locking ability when strictly following the global trajectory. When passing through the right-angle bend of the corridor, there was no obvious Oscillation or deviation and disorientation.

5.5. Discussion and System Limitations

Although this research successfully achieved vision-based autonomous positioning and navigation on Unitree G1, the experimental process also fully exposed the inherent limitations of the current mainstream technology pipeline when facing humanoid robot platforms. To promote further development in this field, it is necessary to conduct critical reflection on the following core limitations.

5.5.1. Sensory Limitations of Pure Vision

The current system is highly dependent on pure vision SLAM based on traditional feature point matching. It was observed in the experiment that when the D435i camera of the robot was confronted with areas lacking texture (such as large white walls or reflective glass in corridors), the number of effective feature points extracted by RTAB-MAP would drop sharply. It is highly likely to cause the temporary failure of the visual odometer and a sudden drop in positioning confidence. In addition, due to the limitations of binocular baseline length and infrared projection power, vision sensors generate significant depth noise at long distances. This directly leads to severe "Ghosting" and distortion phenomena in the constructed 3D point cloud at the edges of obstacles, which restricts the absolute ranging accuracy of the map from the physical hardware level.

5.5.2. Feature-SLAM vs. NeRF-SLAM

RTAB-MAP is a SLAM framework based on traditional feature engineering and appearance matching. Although it has demonstrated good real-time performance on platforms with limited computing power. However, the maps it constructs are essentially sparse or semi-dense discrete geometric point sets. It lacks a continuous understanding of the three-dimensional structure of the environment and high-fidelity rendering capabilities. Although modern Neural Implicit Representations techniques such as NICE-SLAM and NeRF-SLAM can provide extremely dense 3D reconstruction, they require massive parallel tensor computations and extremely high video memory bandwidth. Under the premise of simultaneously undertaking high-frequency gait control and path planning, it also cannot meet the real-time rendering requirements of NeRF-SLAM. Therefore, the adoption of traditional feature point SLAM is a forced compromise limited by the current edge computing power.

5.5.3. Dimensionality Downgrade of Humanoid Locomotion

In system integration, there is deep-seated architectural contradictions that an attempt was made to drive a 3D humanoid robot with extremely high degrees of freedom using a 2D navigation stack designed for a wheeled chassis. Nav2's projection mechanism is essentially a "dimensional reduction and compression" of the environment. It regards Unitree G1 as a two-dimensional particle that can only slide on a plane, completely eliminating the topological advantages of bipedal robots in stepping over low obstacles, taking side-steps, or adapting to complex terrains.

5.5.4. Heuristic Search Navigating Method to VLA/VLN

The traditional A* global planning and cost mapping mechanism lacks any semantic understanding ability. The algorithm cannot distinguish whether the obstacle in front of it is an impenetrable wall or an empty cardboard box that can be pushed open by a mechanical arm. With the development of Embodied intelligence, decoupling the navigation system into a highly modular traditional pipeline of "mapping, positioning, planning, and control" has become increasingly rigid. Exploring end-to-end vision-language-action large models (VLA) and vision-language navigation (VLN) enables robots to directly map high-dimensional multimodal perception inputs to underlying joint torque outputs.

6. Conclusions and Future Work

6.1. Conclusions

This research successfully constructed a high-fidelity indoor 2D/3D map using Intel RealSense D435i and RTAB-MAP, and the mapping error was successfully controlled within the preset threshold of 5cm. This provides an extremely clear and reliable physical boundary for the subsequent generation of the global cost map.

In terms of positioning, the multi-sensor fusion of the high-frequency IMU and the visual odometer was carried out through the EKF filter, achieving an extremely low mean yaw angle error of 1.99°. This effectively smoothed out the visual high-frequency jitter and observation noise caused by mechanical gait. Although limited by the intense nonlinear gait impact of the bipedal robot and the instantaneous motion blur of the camera, the mean planar positioning error (0.29 m) failed to reach the original theoretical target of 0.1 m set in Chapter 3. However, considering the highly complex kinematics of the humanoid robot chassis, such sub-meter moderate drift is completely acceptable in practical engineering and falls within the revised 0.3 m threshold.

This accuracy has fully demonstrated the high practical value of the pure vision solution in indoor navigation without the assistance of an external motion capture system. And it is fully capable of ensuring that the robot can safely and accurately complete room-level point-to-point autonomous movement tasks in non-extremely compact indoor environments such as corridors or laboratories.

6.2. Future Work

Based on the research findings of this project and the limitations of the existing system, future research can be further expanded from the following three dimensions:

Multimodal Sensor Integration: To overcome the vulnerability of pure vision SLAM in texture-free areas or complex lighting conditions, future work should exploit the 3D LiDAR already integrated on the Unitree G1 platform. Building a multimodal fusion architecture of vision – LiDAR–inertial tight coupling can eliminate depth perception noise and environmental dependence at the hardware source.

3D Stepping-aware Navigation: Abandon the traditional two-dimensional occupying grid map and develop a full-dimensional motion planner specifically designed for bipedal robots. This planner should be capable of integrating discrete footholds calculations with the 3D topological structure of the environment, fully unlocking the physical potential of humanoid robots to traverse complex terrains.

Vision-Language-Action Paradigm: Future systems should gradually evolve from "Heuristic Search" lacking semantic understanding to end-to-end, embodied, intelligent large models. By directly mapping multimodal vision and natural language instructions to the underlying joint control, humanoid robots are able to achieve advanced cognitive and interactive navigation capabilities.

References

[1] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with Rao-Blackwellized particle filters," IEEE Trans. Robot., vol. 23, no. 1, pp. 34–46, Feb. 2007, doi: 10.1109/TRO.2006.889486.

[2] S. Saat, W. Abd Rashid, M. Tumari, and M. Saealal, "HECTORSLAM 2D mapping for simultaneous localization and mapping (SLAM)," J. Phys.: Conf. Ser., vol. 1529, no. 4, p. 042032, Apr. 2020, doi: 10.1088/1742-6596/1529/4/042032.

[3] J. Zhang and S. Singh, "Low-drift and real-time lidar odometry and mapping," Auton. Robots, vol. 41, no. 2, pp. 401–416, Feb. 2017, doi: 10.1007/s10514-016-9548-2.

[4] T. Shan and B. Englot, "LeGO-LOAM: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Madrid, Spain, Oct. 2018, pp. 4758–4765, doi: 10.1109/IROS.2018.8594299.

[5] W. Xu and F. Zhang, "FAST-LIO: A fast, robust LiDAR-inertial odometry package by tightly-coupled iterated Kalman filter," IEEE Robot. Autom. Lett., vol. 6, no. 2, pp. 3317–3324, Apr. 2021, doi: 10.1109/LRA.2021.3064227.

[6] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "FAST-LIO2: Fast direct LiDAR-inertial odometry," IEEE Trans. Robot., vol. 38, no. 4, pp. 2053–2073, Aug. 2022, doi: 10.1109/TRO.2022.3141876.

[7] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in Proc. 6th IEEE/ACM Int. Symp. Mixed Augmented Reality, Nara, Japan, Nov. 2007, pp. 1–10, doi: 10.1109/ISMAR.2007.4538852.

[8] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," IEEE Trans. Robot., vol. 33, no. 5, pp. 1255–1262, Oct. 2017, doi: 10.1109/TRO.2017.2705103.

[9] C. Campos, R. Elvira, J. J. Gómez Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," IEEE Trans. Robot., vol. 37, no. 6, pp. 1874–1890, Dec. 2021, doi: 10.1109/TRO.2021.3075644.

[10] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," Int. J. Robot. Res., vol. 34, no. 3, pp. 314–334, Mar. 2015, doi: 10.1177/0278364914554813.

[11] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," IEEE Trans. Robot., vol. 34, no. 4, pp. 1004–1020, Aug. 2018, doi: 10.1109/TRO.2018.2853729.

[12] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in Proc. IEEE Int. Conf. Robotics and Automation (ICRA), Singapore, May 2017, pp. 2043–2050, doi: 10.1109/ICRA.2017.7989236.

[13] Z. Zhu et al., "NICE-SLAM: Neural implicit scalable encoding for SLAM," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 12776–12786, doi: 10.1109/CVPR52688.2022.01245.

[14] Z. Zhu et al., "NICER-SLAM: Neural implicit scene encoding for RGB SLAM," in Proc. Int. Conf. 3D Vision (3DV), Davos, Switzerland, Mar. 2024, pp. 42–52, doi: 10.1109/3DV62453.2024.00096.

[15] A. Rosinol, J. J. Leonard, and L. Carlone, "NeRF-SLAM: Real-time dense monocular SLAM with neural radiance fields," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Detroit, MI, USA, Oct. 2023, pp. 3437–3444, doi: 10.1109/IROS55552.2023.10341922.

[16] R. C. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," Int. J. Robot. Res., vol. 5, no. 4, pp. 56–68, Dec. 1986, doi: 10.1177/027836498600500404.

[17] S. J. Julier and J. K. Uhlmann, "New extension of the Kalman filter to nonlinear systems," in Proc. SPIE Signal Process., Sensor Fusion, Target Recognit. VI, vol. 3068, Orlando, FL, USA, Jul. 1997, pp. 182–193, doi: 10.1117/12.280797.

[18] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (SLAM) problem," IEEE Trans. Robot. Autom., vol. 17, no. 3, pp. 229–241, Jun. 2001, doi: 10.1109/70.938381.

[19] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in Proc. IEEE Int. Conf. Robotics and Automation (ICRA), Rome, Italy, Apr. 2007, pp. 3565–3572, doi: 10.1109/ROBOT.2007.364024.

[20] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," IEEE Trans. Syst. Sci. Cybern., vol. 4, no. 2, pp. 100–107, Jul. 1968, doi: 10.1109/TSSC.1968.300136.

[21] E. W. Dijkstra, "A note on two problems in connexion with graphs," Numerische Mathematik, vol. 1, no. 1, pp. 269–271, Dec. 1959, doi: 10.1007/BF01386390.

[22] V. Vonásek, J. Faigl, T. Krajník, and L. Přeučil, "RRT-path – A guided rapidly exploring random tree," in Robot Motion and Control 2009 (Lect. Notes Control Inf. Sci., vol. 396), K. R. Kozłowski, Ed. London, U.K.: Springer, 2009, pp. 307–316, doi: 10.1007/978-1-84882-985-5_28.

[23] A. Stentz, "Optimal and efficient path planning for partially-known environments," in Proc. IEEE Int. Conf. Robotics and Automation (ICRA), San Diego, CA, USA, May 1994, vol. 4, pp. 3310–3317, doi: 10.1109/ROBOT.1994.351061.

[24] Z. Ren, S. Rathinam, M. Likhachev, and H. Choset, "Multi-objective path-based D* Lite," IEEE Robot. Autom. Lett., vol. 7, no. 2, pp. 3318–3325, Apr. 2022, doi: 10.1109/LRA.2022.3146918.

[25] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," Int. J. Robot. Res., vol. 5, no. 1, pp. 90–98, Mar. 1986, doi: 10.1177/027836498600500106.

[26] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," IEEE Robot. Autom. Mag., vol. 4, no. 1, pp. 23–33, Mar. 1997, doi: 10.1109/100.580977.

Appendices

Appendix A: Localization Accuracy Test Data

Table A.1: Localization accuracy measurement results for 10 navigation trials.

Trial	*X Error (m)*	*Y Error (m)*	*Planar Distance (m)*	*Yaw Error (°)*
1	-0.09	-0.34	0.35	+1.1°
2	+0.32	+0.33	0.46	-3.6°
3	+0.16	+0.23	0.28	-2.1°
4	+0.07	-0.20	0.21	-1.3°
5	-0.24	-0.22	0.33	-0.4°
6	-0.24	-0.22	0.33	+2.9°
7	-0.31	-0.14	0.34	-3.0°
8	+0.26	+0.02	0.26	+0.1°
9	+0.07	-0.05	0.09	+0.9°
10	-0.15	-0.15	0.21	-4.5°
Mean	**0.19**	**0.19**	**0.29**	**1.99°**