

Department of Human and Engineered Environmental Studies
Graduate School of Frontier Sciences
The University of Tokyo

2024

Master's Thesis

Switching-based Multi-modal SLAM
for Extreme and Degraded Environments
(極限・劣化環境のためのマルチモーダル情報を用いた切替型 SLAM)

Submitted February 01, 2025

Adviser: Professor Atsushi Yamashita (seal)

Student ID Number 47236676

Lee Junwoon (李峻源)

Abstract

LiDAR-visual-inertial simultaneous localization and mapping (SLAM) have demonstrated remarkable performance in many robotic applications. However, their robustness is significantly challenged in environments where either structural or visual degeneration is severe. These challenges came from the reliance on Maximum A Posteriori (MAP) or Maximum Likelihood Estimation (MLE), which suffers with degenerate scenes. While thermal imaging offers unique advantages in low-light or night-time conditions, the performance of SLAM system is limited due to issues such as texturelessness of thermal images and intermittent frame drops.

The objective of this thesis is to develop a LiDAR-thermal-inertial SLAM system that is accurate and robust against both structural and visual degeneration, using a switching strategy between multi-modal sensor fusion grounded in degeneracy detection and infrared thermal features aided by a self-supervised point tracker. The proposed framework consists of three key components: (1) Switch SLAM, a switching-Based LiDAR-inertial-visual SLAM system that dynamically selects optimal odometry sources to avoid long-term degeneration; (2) Self-TIO, a thermal-inertial odometry system using a learning-based feature tracker to solve issues from textureless and intermittent frame drop in thermal imaging; and (3) TC-LTIO, a tightly-coupled LiDAR-thermal-inertial odometry system, which integrates the components from Switch SLAM and Self-TIO to achieve ultimate robustness in both structurally and visually degraded environments.

First, Switch SLAM introduces a switching structure to optimize the initial guess between LiDAR and visual odometry, ensuring that only reliable estimations propagate through the system, thus improving overall multi-modal SLAM performance. Second, Self-TIO introduces self-supervised learning in 16-bit raw image domain to robustly extract and track thermal features, significantly improving the performance of thermal-inertial SLAM. Finally, TC-LTIO introduces a novel LiDAR-thermal-inertial SLAM system that incorporates a tightly-coupled formulation and zero-velocity detection, effectively addressing the limitations of traditional LiDAR-visual-inertial approaches.

In conclusion, the proposed systems contribute to robust SLAM systems for structurally and visually degraded environments by introducing a novel switching strategy and leveraging thermal-inertial odometry with self-supervised learning. Future work could explore continuous formulations, enhance map representations, and adapt the framework to multi-robot SLAM systems.

Contents

Chapter 1	Introduction	1
1.1	Background	2
1.2	Objective	6
1.3	Thesis Structure	7
Chapter 2	Related Work	9
2.1	SLAM	10
2.2	SLAM System Robust To LiDAR SLAM Degeneration	12
2.2.1	LiDAR Odometry Degeneration	13
2.2.2	LiDAR Visual SLAM	14
2.3	Visual SLAM System Using Thermal Images	14
Chapter 3	Switch-SLAM: Switching-Based LiDAR-Inertial-Visual SLAM	17
3.1	Introduction: Switch-SLAM	18
3.2	Proposed Method	20
3.2.1	System Overview	20
3.2.2	LiDAR-Inertial-Visual SLAM	21
3.2.3	Degeneracy Detection of LiDAR Odometry	23
3.2.4	Failure Detection of Visual Odometry	26
3.2.5	Scan-to-Map Matching	26
3.2.6	Frontend Implementation of Switch-SLAM	28
3.2.7	Backend Implementation of Switch-SLAM	29
3.3	Experiments	30
3.3.1	Datasets	30
3.3.2	Accuracy Evaluation	32
3.3.3	Degeneracy Detection Evaluation	36
3.4	Summary	38

Chapter 4	Self-TIO: Thermal-Inertial Odometry via Self-Supervised 16-bit Feature Extractor and Tracker	39
4.1	Introduction: Self-TIO	40
4.2	Proposed Method	42
4.2.1	ThermalLANet	42
4.2.2	Thermal Optical Flow	45
4.2.3	Hybrid Feature Tracker	48
4.2.4	Thermal Inertial Odometry Formulation	48
4.3	Experiments	50
4.3.1	Evaluation on Feature Point Detection	50
4.3.2	Evaluation on Feature Point Tracker	52
4.3.3	Evaluation on State Estimation Performance	54
4.4	Summary	60
Chapter 5	TC-LTIO: Tightly-coupled LiDAR Thermal Inertial Odometry for LiDAR and Visual Odometry Degraded Environments	61
5.1	Introduction: TC-LTIO	62
5.2	Tightly-coupled LiDAR Thermal Inertial Odometry	64
5.2.1	System Overview	64
5.2.2	Feature Extraction	64
5.2.3	Visual Feature Tracking	65
5.2.4	Tightly-Coupled optimization formulation	68
5.2.5	Zero Velocity Detection	69
5.3	Experiments	71
5.4	Summary	75
Chapter 6	Conclusion	77
6.1	Summary	78
6.2	Future Work	80
	Acknowledgements	83
	Reference	85
	Research Publications	97

List of Figures

1.1	Examples of 3D maps generated using LiDAR SLAM.	2
1.2	Examples of MAP-based fusion.	3
1.3	Comparison between visual and thermal cameras.	5
1.4	Structure of this research.	7
3.1	Snapshots and maps from simulated Farm dataset.	19
3.2	System structure of Switch-SLAM.	20
3.3	Representative examples of LiDAR odometry degenerate structures.	23
3.4	Description of the status buffer.	25
3.5	The implementation details of Switch-SLAM.	28
3.6	Simulated environments.	31
3.7	Imaging problem caused by video interruption in ANYmal 3.	31
3.8	Trajectory of proposed and compared LiDAR visual SLAM in Fast Rotate, Plane, Farm, Multi Floor, and Long Corridor dataset.	34
3.9	Resulting maps from the compared methods and Switch-SLAM.	35
3.10	Comparison of degeneracy detection of the state-of-the-art and proposed methods on a part of the Handheld dataset.	37
4.1	System overview of Self-TIO.	42
4.2	Overview of the ThermalLANet network.	43
4.3	Noise augmentation to train proposed network using a transformed 16-bit radiometric image.	43
4.4	Overview of the thermal optical flow network.	46
4.5	Self-supervised training strategies for proposed thermal optical flow network.	47

4.6	Feature point extraction results and inliers when using different methods on the FLIR thermal image train dataset.	51
4.7	Examples of feature point extraction results using different methods.	52
4.8	Comparison results between the hybrid track and compared methods in NUC and stationary scenes.	53
4.9	Trajectories in public dataset.	56
4.10	Experimental platform and environment snapshots.	57
4.11	Trajectories in real-world dataset.	59
5.1	System overview of TC-LTIO.	64
5.2	Proposed optical flow network.	66
5.3	Method to generate synthetic thermal image from RGB image.	66
5.4	Labeling for self-supervised learning from single thermal image.	67
5.5	Experimental result on gate01 and street01.	73
5.6	Resulting map in gate02 and street02.	74

List of Tables

3.1	Dataset details for Switch-SLAM.....	30
3.2	Comparison of ATE [m] on the tested datasets.	33
4.1	Comparison of feature point detection algorithms on the FLIR thermal image test dataset.	50
4.2	Comparison of ATE [m] when using each feature tracking algorithms in grassland dataset.	54
4.3	Comparison of ATE [m] in the VIVID++ dataset.....	55
4.4	Comparison of ATE [%] in the SNU sequence of the STheReO dataset.....	55
4.5	Comparison of ATE [m] in proposed real-world dataset.	58
5.1	Comparison of ATE [m] on M2DGR dataset.	71

Notations and Symbols

Unless specially noted, matrices are capitalized and in bold font, vectors are in bold, italic font.

Acronyms	Meaning
SLAM	Simultaneous localization and mapping.
LiDAR	Light detection and ranging.
IMU	Inertial measurement unit.
GNSS	Global navigation satellite system.
LO	LiDAR odometry.
VIO	Visual inertial odometry.
LVIO	LiDAR visual inertial odometry.
TIO	thermal inertial odometry.
LTIO	LiDAR thermal inertial odometry.
DOF	Degrees of freedom.
MAP	Maximum a posteriori.
ATE	Absolute trajectory error.
RMSE	Root-mean-square error.
NUC	Non-uniformity correction.
FPN	Fixed-pattern noise.
LPN	Low-frequency noise.
CNN	Convolutional neural network.
FPS	Frames per second.

Chapter 1

Introduction

1.1	Background.....	2
1.2	Objective.....	6
1.3	Thesis Structure.....	7

1.1 Background

In recent years, significant progress has been made in autonomous mobile robots. These robots are now applied in various environments, ranging from homes, restaurants, and offices to extreme environments such as factories, farmlands, mountains, and radioactive zones. To construct fully automated mobile robots, localization and mapping are essential. Recently, advances in 3D simultaneous localization and mapping (SLAM) have played an important role in leading to improvements in mobile robot capabilities. These developments allow robots not only to accurately determine their own positions but also to understand their surroundings and generate detailed 3D maps of their environments. The representative 3D maps from SLAM is presented in Fig. 1.1.

SLAM systems can be categorized based on the primary sensor utilized: LiDAR SLAM, which uses light detection and ranging (LiDAR) sensors, and visual SLAM, which relies on visual cameras. Recently, both LiDAR and visual SLAM show effectiveness and generalizability in general and well-controlled environments [5, 6, 7, 8], with LiDAR SLAM demonstrating

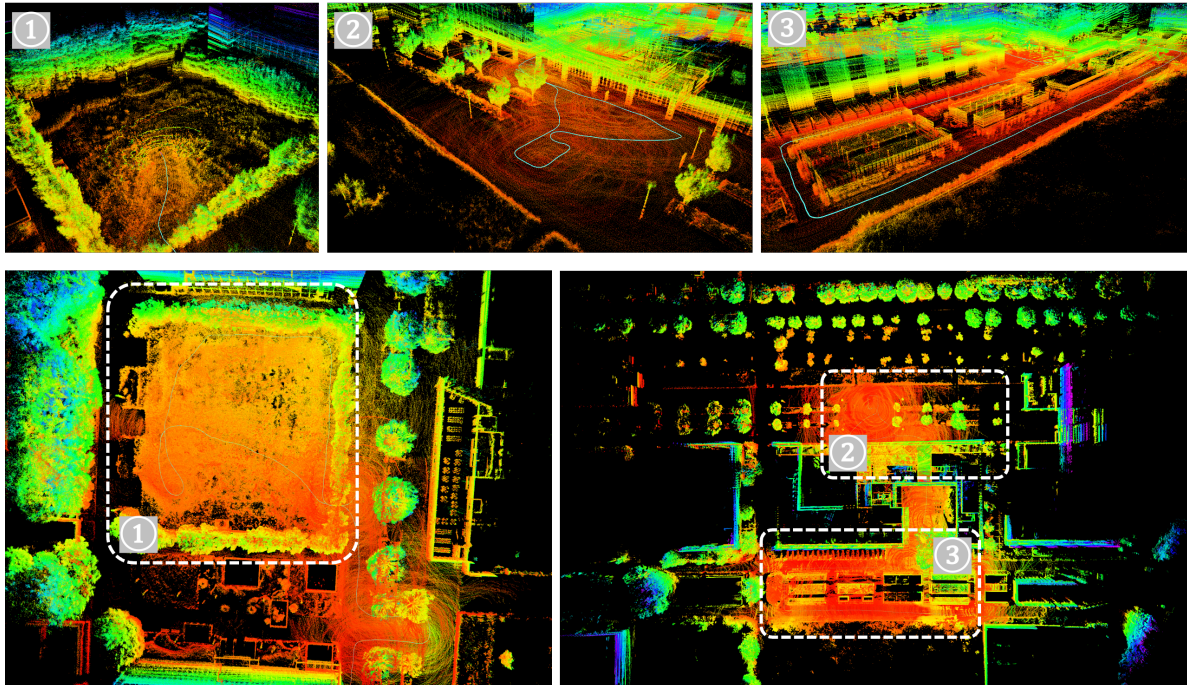


Fig. 1.1: Examples of 3D maps generated using LiDAR SLAM, LIO-SAM [4] on the Kashiwa Campus of the University of Tokyo. The blue line represents the trajectories obtained from SLAM.

better accuracy and generalizability than visual SLAM in most cases, as it can capture precise depth information from the sensor. Nonetheless, when it comes to complex and extreme settings, both LiDAR and visual SLAM easily experience drift or failure. These types of drift and failure are caused by degeneration in the optimization process. Here, degeneration denotes limited constraints that do not ensure the solvability of the optimization or filtering process along specific degrees of freedom (DOF) [9]. In other words, the degeneration can be defined as an inefficient solution along specific DOFs caused by insufficient, overly duplicated data distribution, or harsh noisy data during the SLAM problem-solving process. LiDAR SLAM, which mainly relies on structural information, can experience degeneration in environments such as a vast plane, simple-pattern corridor, and tunnel [10]. Conversely, visual SLAM, which primarily utilizes textural information, can encounter degeneration in scenarios such as sensing monotonous white walls, large occlusions, blurring due to aggressive motion, and dark scenes [11].

The fusion of inertial measurement unit (IMU) sensors in the context of SLAM is proposed to address degeneracy problems, as demonstrated in LiDAR-inertial [4, 12, 13] and visual-inertial SLAM [14, 15, 16] systems, which use either filter-based [14] or optimization-based smoothing techniques [17]. Although these LiDAR-inertial and visual-inertial SLAM systems are partially effective in addressing aggressive motion and mitigating short-term degeneration, long-term degeneration still remains a significant problem due to the noisy nature of IMU data. This long-term degeneration situation is obvious in scenarios such as long tunnels, corridors, vast planes, and white walls, as discussed in [11, 18, 19, 20]. LiDAR-visual-(inertial) SLAM systems have been proposed to ensure robustness in such scenarios by using multiple modalities, including both cameras and LiDAR. Whether through loosely-coupled fusion [18, 21, 22, 23] or tightly-coupled fusion [11, 19, 20] between LiDAR and visual sensor, these methods rely

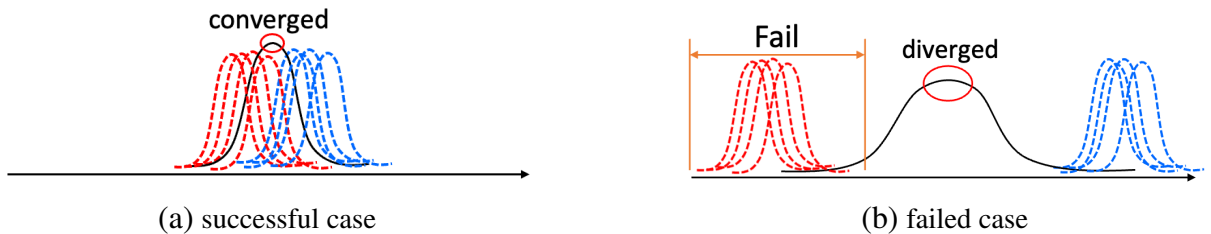


Fig. 1.2: Examples of MAP-based fusion. The red and blue lines represent data distributions, while the black line indicates the result of the MAP estimation.

on maximum a posteriori (MAP) or maximum likelihood (MLE) problems to integrate the data from different sensors. However, as illustrated in Fig. 1.2, statistical or optimization-based fusion methods can lead to divergence in results when dealing with incorrect or overly distributed data. This divergence problem can be critical for sensor-fusion-based SLAM systems, especially when each sub-system experiences degeneration.

On the other hand, to address visual SLAM degeneration caused by night and dark scenes, thermal infrared cameras can be used. Thermal cameras are capable of detecting temperature variations in a scene by capturing infrared radiation emitted from objects. Thermal cameras are classified into refrigerated and non-refrigerated types. In the case of refrigerated types, a secondary refrigerating device is required to cool the camera directly. While this results in lower noise and higher resolution, it also increases the cost and size of the setup. Therefore, refrigerated cameras are typically used for monitoring purposes, where the position of camera remains fixed. Non-refrigerated types, which are more suitable for SLAM applications due to their reliance on natural air cooling, are vulnerable to noise. To address this, manufacturers of non-refrigerated thermal cameras embed a non-uniformity correction (NUC) process to enhance image quality and mitigate noise. However, the NUC process can cause frame drops, and these sudden drops can lead to SLAM failure. Moreover, despite the NUC process, significant noise remains in the images, causing the thermal SLAM system to fail.

Another challenge with thermal cameras is texturelessness, as illustrated in Fig. 1.3. Compared to visual cameras, thermal cameras capture only the infrared region of the electromagnetic spectrum, which limits their ability to fully capture textural information of objects. Moreover, as most thermal cameras capture images in a 16/14-bit format, the raw thermal images need to be converted from 16/14-bit to 8-bit format for further processing and visualization. This conversion is normally required because most image processing libraries, such as OpenCV [24] and Pillow [25], operate on their standard image format of 8-bit, which is predominantly adopted in current visual SLAM systems. Moreover, it is important to note that 16/14-bit raw thermal images utilize only a small portion of the entire 16-bit data range (0–65535), which means, without specific conversion, most of the raw images would not be visible. However, this conversion leads to information loss and exacerbates the lack of texture by compressing the 16-bit data range into 8-bit. Given the textural reliance of visual SLAM, the information loss resulting from the conversion is critical for constructing an accurate SLAM system.

In summary of this section, conventional SLAM systems face two main problems: (1) Statistical or optimization-based fusion between visual and LiDAR features causes

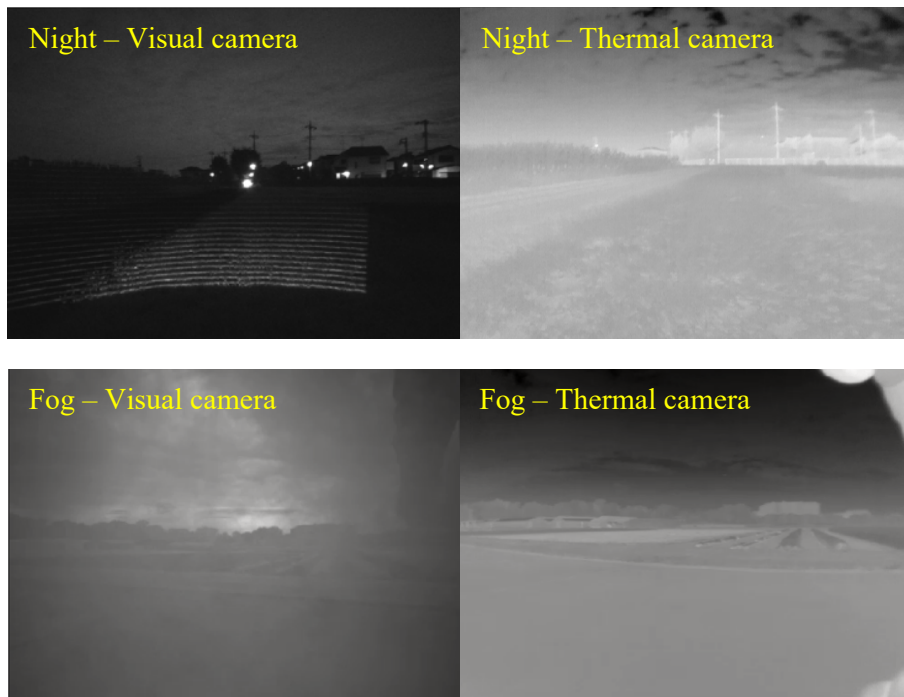


Fig. 1.3: Comparison between visual and thermal cameras. Thermal cameras perform better in night and foggy scenes but have lower texture.

LiDAR-visual SLAM system to degenerate when either feature is excessively outliered. (2) Thermal SLAM, which suffers from issues such as NUC, low-texture, and 8-bit conversion, is not comparable to conventional visual SLAM in terms of accuracy and robustness.

1.2 Objective

As described in Section 1.1, the conventional LiDAR-visual-inertial SLAM systems have limitations when either structural or visual degeneration is severe, due to their reliance on MAP or MLE. Moreover, despite the potential of thermal vision in night and low-light scenes, challenges such as texturelessness and intermittent frame drops remain unresolved. Therefore, the objective of this research is defined as follows: **“Developing a LiDAR-thermal-inertial SLAM system robust to structural and visual degeneration.”**

To achieve this objective, in this research, three methods are proposed: (1) Switch SLAM: Switching-Based LiDAR-Inertial-Visual SLAM, (2) Self-TIO: Thermal-Inertial Odometry via Self-Supervised 16-bit Feature Extractor and Tracker, and (3) TC-LTIO: Tightly-coupled LiDAR Thermal Inertial Odometry for LiDAR and Visual Odometry Degraded Environments.

In Switch SLAM, the main contribution is switching structure, which allows selecting an optimal initial guess between LiDAR and visual odometry. This selection effectively avoids long-term degeneracy and ensures that only reliable estimations propagate through the entire system, therefore improving the overall performance compared to state-of-the-art LiDAR-visual-inertial SLAM systems. In Self-TIO, the key contribution is a feature point tracker for thermal images, built via using a self-supervised optical flow network and 16-bit-based learning strategies. This tracker effectively addresses the issues of texturelessness and frame drops in thermal vision. Moreover, the effective usage of a thermal camera enhances the robustness against long-term visual degeneration. In TC-LTIO, leveraging insights from Switch SLAM and Self-TIO in selective graph optimization and thermal feature tracking, to our knowledge, the world’s first LiDAR-thermal-inertial SLAM system is proposed. This system demonstrates robustness in both structurally and visually degenerate environments, ensuring accurate localization and mapping even under extreme and degraded conditions, thereby directly fulfilling the stated objective.

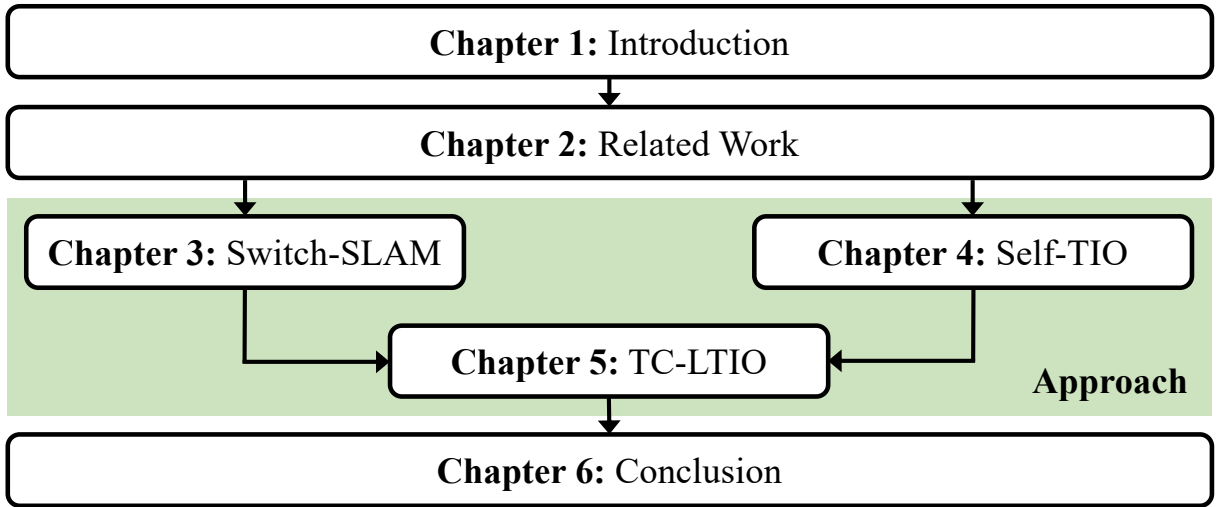


Fig. 1.4: Structure of this research.

1.3 Thesis Structure

This thesis is comprised of six chapters, whose contents illustrated in Fig. 1.4. Research background is introduced in Chapter 1. Related research reviewed in Chapter 2. The proposed methods, Switch-SLAM, Self-TIO, and TC-LTIO, are detailed in Chapters 3, 4, and 5, respectively, to achieve the research objective. The research is summarized and concluded, serving directions for future works, in Chapter 6. Note that this thesis is based on and complies with the copyright rules of published papers by Junwoon Lee *et al.* [1, 2, 3], including all the figures and tables.

Chapter 2

Related Work

2.1	SLAM	10
2.2	SLAM System Robust To LiDAR SLAM Degeneration.....	12
	2.2.1 LiDAR Odometry Degeneration.....	13
	2.2.2 LiDAR Visual SLAM.....	14
2.3	Visual SLAM System Using Thermal Images.....	14

2.1 SLAM

SLAM simultaneously performs the state estimation and mapping by sensors mounted on moving robots, vehicles and drones. The SLAM process is separated into a localization and mapping process. In the localization process, the pose of the SLAM agent is calculated by tracking the specific information of sequential sensor data, such as landmarks, feature points, and the optical flow. In the mapping process, refined sensor data are represented to the map based on the pose calculated in the localization process.

SLAM can be classified into filter-based methods and optimization-based methods. Leonard and Durrant-Whyte first proposed a computational solution to the SLAM problem by extracting geometric landmarks from laser sensor data and matching the geometric landmarks for each scene using the extended Kalman filter (EKF) [26]. However, because filter-based methods, such as EKF, integrate all information of geometric landmarks into the filter, their individual geographic characteristics are lost in the SLAM process. Conversely, Lu and Milios proposed an optimization-based method, which defined a point-to-point least-squares problem, thereby minimizing the distance between the matched points [27]. This optimization-based system maintained better robustness in complex environments compared to filter-based methods.

SLAM can also be classified by the main sensor into visual SLAM, which uses a camera, and LiDAR SLAM, which uses a LiDAR sensor. Due to its low cost, easy configuration, and compact size, the camera is one of the most useful sensors in SLAM systems [28]. Visual SLAM is further classified into monocular visual SLAM, which uses one camera; stereo visual SLAM, which uses two or more cameras; and RGBD SLAM, which uses one RGB camera and one depth camera. Visual SLAM is also categorized into feature-based SLAM, which tracks the sequential scene by matching image features, and direct photogrammetric SLAM, which tracks the scene by calculating the photometric error between two sequential images. Based on the tracking information, two-dimensional pixels of the image are reconstructed into three-dimensional points. Subsequently, the pose is calculated using the movement of tracked features or pixels.

The first proposed visual SLAM was MonoSLAM, which extracts image features using a Shi-Tomasi corner detector [29] and tracks the features encoded with the two-dimension gaussian probability density functions [30, 31]. However, MonoSLAM has low computational efficiency because it processes localization and mapping with a single thread. To address this problem, Klein and Murray proposed parallel tracking and mapping (PTAM) to split

the mapping and localization threads in parallel [32], resulting in higher processing speed than MonoSLAM and achieving real-time performance. To improve accuracy and robustness, PTAM refines the poses and map using bundle adjustment to minimize the reprojection errors of the point features for each scene. Based on PTAM, ORB-SLAM, which is one of the state-of-the-art visual SLAM methods, was proposed [8]. In ORB-SLAM, oriented features from the accelerated segment test and rotated binary robust independent elementary features (ORB) [33] are used, while ensuring time efficiency and incorporating a loop closure for drift correction [8]. Finally, inertial measurement unit (IMU) integration into these visual SLAM systems, specifically visual-inertial SLAM, has been proposed to improve robustness in the face of aggressive motion and tracking failure scenarios [14, 15, 16].

3D LiDAR measures the three-dimensional coordinates of target objects using a time-of-flight of laser beam. One of the main differences between LiDAR SLAM and visual SLAM is that LiDAR directly measures the distance between the sensor and features, whereas visual SLAM estimates depth using photogrammetric methods such as triangulation. Scan matching, which connects two related point cloud pairs from LiDAR, is essential for LiDAR SLAM. One of the representative LiDAR scan matching algorithms is ICP [34]. The ICP algorithm iteratively determines the relative pose between a pair of point clouds by minimizing the sum of the point-to-point distances between paired points. However, since ICP only minimizes the point-to-point distance, mismatches may occur when a complex-shaped point cloud is involved. To address this problem, Segal et al. proposed GICP to [35], which minimizes plane-to-plane distances based on a probabilistic model. The GICP achieves complex structures and incorrect correspondences than ICP. SLAM using the ICP and GICP algorithms has also been proposed [36, 37, 38], including IMU-integrated LiDAR-inertial SLAM [13]. However, a major disadvantage of ICP and GICP is their computational time when matching large point cloud sets. This issue is particularly obvious for online SLAM, where the computational time of matching directly impacts performance and accuracy.

To address this, feature-based matching methods have been proposed to improve the computational efficiency of LiDAR SLAM. Zhang and Singh proposed low-drift and real-time LiDAR odometry and mapping (LOAM) [5]. LOAM extracts edge and planar features based on the smoothness values calculated from nearby points on the same scan line and matches these features according to their geometric properties. Moreover, by dividing the SLAM system into LiDAR odometry (10 Hz) and mapping (1 Hz), LOAM improves both computational efficiency and accuracy. To further enhance computational efficiency, Shan and Englott proposed lightweight and ground-optimized LiDAR odometry and mapping for variable

terrain (LeGO-LOAM) [6], specifically designed for ground vehicles. In LeGO-LOAM, point cloud of the ground is segmented and planar features from the segmented ground and edge features from the nonground point cloud are extracted, as most planar features are found on the ground, unlike edge features. After extracting edge and planar points, LeGO-LOAM employs a two-step optimization process for motion estimation, reducing the computational cost. As a result, LeGO-LOAM outperforms LOAM in computational efficiency while maintaining the accuracy. As the IMU-integrated version of LeGO-LOAM, LIO-SAM [4] is proposed, using the IMU preintegration technique [17] to improve robustness in aggressive motion and short-term degeneracy.

A disadvantage of using a camera in mapping is that the depth information corresponding to the image pixel is not immediately known. Instead, the depth is indirectly obtained by solving a calculation problem, such as triangulation. Additionally, cameras are easily affected by lighting conditions. Meanwhile, the disadvantage of LiDAR sensor is its lower point cloud density compared to the pixel density of the visual image. This makes it more difficult to extract valid features in structureless environments, such as vast planes and corridors, when using LiDAR rather than visual sensors.

To address these limitations, SLAM systems combining LiDAR and visual sensors have been proposed. Sun et al. proposed an EKF-based SLAM that performs data association of vertical lines from a LiDAR and a camera by minimizing the Mahalanobis distance between the angles of each vertical line [39]. Using the vertical lines from the image and line coordinates from LiDAR, this SLAM system exhibited higher accuracy than laser-only EKF SLAM. Another fusion method using 3D LiDAR and a monocular camera is the real-time depth-enhanced monocular odometry (DEMO), which extracts image feature points and associates their depth using parsed depth from LiDAR [21]. Visual-LiDAR odometry and mapping (VLOAM) [22] is another conventional fusion-based SLAM method. VLOAM estimates motion based on DEMO at a high frequency (60 Hz), and subsequently, the motion and point cloud map are refined at a low frequency (1 Hz) using LOAM with high accuracy.

2.2 SLAM System Robust To LiDAR SLAM Degeneration

Note that single sensor-based SLAM is subject to several limitations that arise from inherent constraints imposed by the sensors. For example, LiDAR SLAM methods, including [4, 5, 13], which rely on structural information for associating each point cloud pair, tend to degrade in environments lacking distinct structures, such as long corridors and vast open fields.

Conversely, visual SLAM methods, including [7, 8, 14, 40], which rely on textural information for associating each image pair, face challenges in scenarios involving aggressive motion, rapidly changing light conditions, and texture-less environments. To deal with these problems, in this research, LiDAR odometry degeneration and LiDAR-visual SLAM system are focused.

2.2.1 LiDAR Odometry Degeneration

Zhang *et al.* analyzed the eigenvalues of the Jacobian matrix of scan-matching cost to detect the degeneracy of LiDAR odometry [9]. They then separated non-localizable and localizable degrees of freedom (DOF) and only optimized nonlinear equations along with the localizable DOF. This method effectively detects and addresses the degeneration. However, this approach relies on a heuristic threshold for the eigenvalues, which may not generalize well across different environments.

Ren *et al.* introduced a degeneracy indicator, defined along with the fluctuation of the optimization vector [41]. This indicator, when incorporated into a factor graph, achieved better detection accuracy compared to the eigenvalues-based method. However, this method requires calculation of the entire optimization processes in scan matching, which is time-consuming. Additionally, the heuristic factors are still necessary to appropriately scale the degeneracy indicator.

To remove heuristic factors, Nubert *et al.* proposed a learning-based approach to directly detect degeneracy from a LiDAR scan using a 3D convolutional neural network [42]. This method achieved a competitive performance compared to the threshold-based method [9]. However, the inference time of the neural network is generally more time-consuming than the eigenvalues-based method. This limitation can pose challenges for real-time operation when integrating the learning-based approach into LiDAR-visual SLAM frameworks.

In this research, a non-heuristic degeneracy detection that eliminates the need for heuristic tuning of the threshold is proposed. While the proposed approach shares similarities with eigenvalue-based methods [5, 9] in terms of leveraging the eigenvalues of the Jacobian matrix of the scan-matching cost for computational efficiency, the research primarily focuses on the 3-DOF vulnerable to degeneration and defines the threshold using a Chi-squared test to enhance generalizability.

2.2.2 LiDAR Visual SLAM

Shan *et al.* proposed tightly-coupled LiDAR-visual-inertial odometry via smoothing and mapping (LVI-SAM) [18] that combined LIO-SAM [4] for LiDAR and VINS-mono [14] for visual odometry. This method effectively addresses LiDAR degenerate environments and outperforms the accuracy of other visual and LiDAR SLAM. However, LVI-SAM is susceptible to failures in the visual SLAM subsystem, as it relies on visual odometry for an initial guess of the scan-matching.

Lin *et al.* proposed a robust, real-time, LiDAR-Inertial-Visual tightly-coupled state estimator and mapping (R2LIVE) [11] that combined measurements of IMU, Fast-LIO [12], and VINS-mono [14] via iterated Kalman filter. This approach achieves high accuracy and high robustness in various environments featuring degeneration of LiDAR or visual odometry.

Zhao *et al.* proposed IMU centric multi-modal fusion of IMU odometry, visual odometry, and LiDAR odometry (Super Odometry) [23]. This approach utilizes the IMU odometry, optimized using the poses from prior visual, LiDAR odometry, as the initial guess for present visual and LiDAR odometry. This capability enables SLAM system to operate effectively in environments that are subject to geometric, visual, or combined degradation, ensuring robustness.

However, since these systems rely on MAP-based multi-modal fusion such as Kalman filter or factor graph optimization, the final pose can diverge when individual sensors experience long-term failures. To address this issue, this research proposes the switching structure that directly avoids the information related to failure or degeneration from the optimization process of motion estimation.

2.3 Visual SLAM System Using Thermal Images

To effectively address issues from visually degraded scenes, visual-inertial odometry system using a thermal camera is focused on. To construct robust and accurate thermal-inertial odometry, the following issues should be resolved: low contrast, unique noise patterns, and frame drops caused by NUC, which make feature extraction and tracking challenging. Therefore, in this section, recent thermal-inertial odometry methods that focus on addressing these challenges are discussed.

Wang *et al.* emphasized the use of edges, which are relatively robust and stable even in noisy and low-contrast thermal images [43]. In this method, feature points are extracted from edge images derived from difference of Gaussians (DoG) filters, and then tracked using

the KLT tracker in the original edge or distance-transformed edge images. Although this method outperforms state-of-the-art visual-inertial methods [14], its robustness can be limited in environments lacking prominent edges. Moreover, this method has limitation in handling rapid scene changes caused by NUC.

Zhao *et al.* proposed TP-TIO [44], which makes SuperPoint [45] more fit in to thermal images by using thermal noise augmentation. Additionally, this method employs a 16-bit KLT tracker, initialized with IMU preintegration measurements, as the feature tracker. However, this approach can be challenged when sudden frame drops occur due to NUC, as the KLT tracker has limitations when large differences appear between sequential scenes. Moreover, this approach relies on 8-bit thermal images when extracting features, which can lead to information loss.

Jiang *et al.* proposed a thermal-inertial SLAM system that relies on a feature tracker using learning-based optical flow to handle the frame drops caused by NUC and provides robust tracking in noisy images [46]. In this method, singular value decomposition for image is used to remove severe FPN. Furthermore, feature points extracted based on gradient magnitude are tracked using a lightened RAFT-based optical flow [47]. To train this optical flow network, both supervised learning with a visible image dataset and self-supervised learning with a thermal dataset are employed. However, these training strategies are limited to the 8-bit scale domain. Moreover, this method relies on a gradient-based feature extractor, which concentrates the feature distributions in a specific region within the entire image. It can lead to tracking loss in scenarios involving aggressive motion and significant occlusion of these concentrated features.

Saputra *et al.* proposed an end-to-end thermal-inertial SLAM network [48]. This method directly trains on 16-bit radiometric images, incorporating not only regression loss to ground truth 6-DOF poses but also hallucination loss [49] between the thermal and corresponding visible images. Although DeepTIO effectively balances information from IMU and thermal images, and demonstrates improved robustness even with poorly calibrated extrinsic matrices [50], its reliance on visible images during training and limited generalizability of end-to-end learning can be perceived as weaknesses.

In this thesis, a thermal-inertial odometry system is proposed based on a 16-bit self-supervised feature point tracker. While my approach shares similarities with the work of Zhang *et al.* [46] in leveraging a learning-based lightweight optical flow network, this research primarily focuses on self-supervised learning strategies to address the scarcity of thermal datasets and 16-bit domain learning, which does not require 8-bit conversion.

Chapter 3

Switch-SLAM: Switching-Based LiDAR-Inertial-Visual SLAM

3.1	Introduction: Switch-SLAM.....	18
3.2	Proposed Method	20
3.2.1	System Overview	20
3.2.2	LiDAR-Inertial-Visual SLAM.....	21
3.2.3	Degeneracy Detection of LiDAR Odometry	23
3.2.4	Failure Detection of Visual Odometry	26
3.2.5	Scan-to-Map Matching.....	26
3.2.6	Frontend Implementation of Switch-SLAM	28
3.2.7	Backend Implementation of Switch-SLAM.....	29
3.3	Experiments	30
3.3.1	Datasets	30
3.3.2	Accuracy Evaluation.....	32
3.3.3	Degeneracy Detection Evaluation	36
3.4	Summary	38

3.1 Introduction: Switch-SLAM

In recent years, significant progress has been made in 3D simultaneous localization and mapping (SLAM), leading to notable advancements in the capabilities of mobile robots. These developments have enhanced the capabilities of mobile robots in terms of understanding their surroundings, precisely determining their positions, and creating detailed maps of their environments. However, SLAM is subject to several limitations that arise from inherent constraints imposed by sensors. For example, LiDAR SLAM [4, 5, 13] tend to degenerate in environments lacking distinct structures such as long corridors and vast open fields. Otherwise, visual SLAM [7, 8, 14] face challenges in scenarios involving aggressive motions, rapidly changing light conditions, and texture-less environments.

To handle these issues, various LiDAR-visual SLAM methods have been developed, including those in [11, 18, 19, 20, 22, 23, 51], which integrate information from a LiDAR and camera. However, these methods have limitations when handling persistent degeneracy that exceeds the system capabilities. These limitations primarily arise from their reliance on fusion methods using maximum a posteriori (MAP) estimation, such as iterated Kalman filters [52] and factor graph optimization [53]. Consequently, long-term failure information can detrimentally impact the overall system performance.

To address these limitations, switching-based LiDAR-inertial-visual SLAM (Switch-SLAM) is proposed. Switch-SLAM parallelly processes LiDAR and visual odometry and selects the appropriate sensor odometry by non-heuristic degeneracy detection, as shown in Fig. 3.1. Switch-SLAM incorporates a switching structure that effectively avoids failure information from propagating throughout the system, thereby mitigating the negative impact on performance. The main contributions of proposed work are as follows:

- **Switching structure:** The switching structure allows selection of an optimal initial estimation between LiDAR and visual odometry, both of which are propagated with IMU measurements. This selection efficiently avoids long-term degeneracy and ensures that only reliable estimations propagate through the entire system, improving the overall performance. Note that the system uniquely uses visual and LiDAR odometry simultaneously only during the switching process, employing MAP fusion to refine the discrepancies between the odometry results via the proposed status buffer.
- **Non-heuristic degeneracy detection:** Non-heuristic degeneracy detection checks the convergence of the optimization process by employing a predefined threshold, grounded

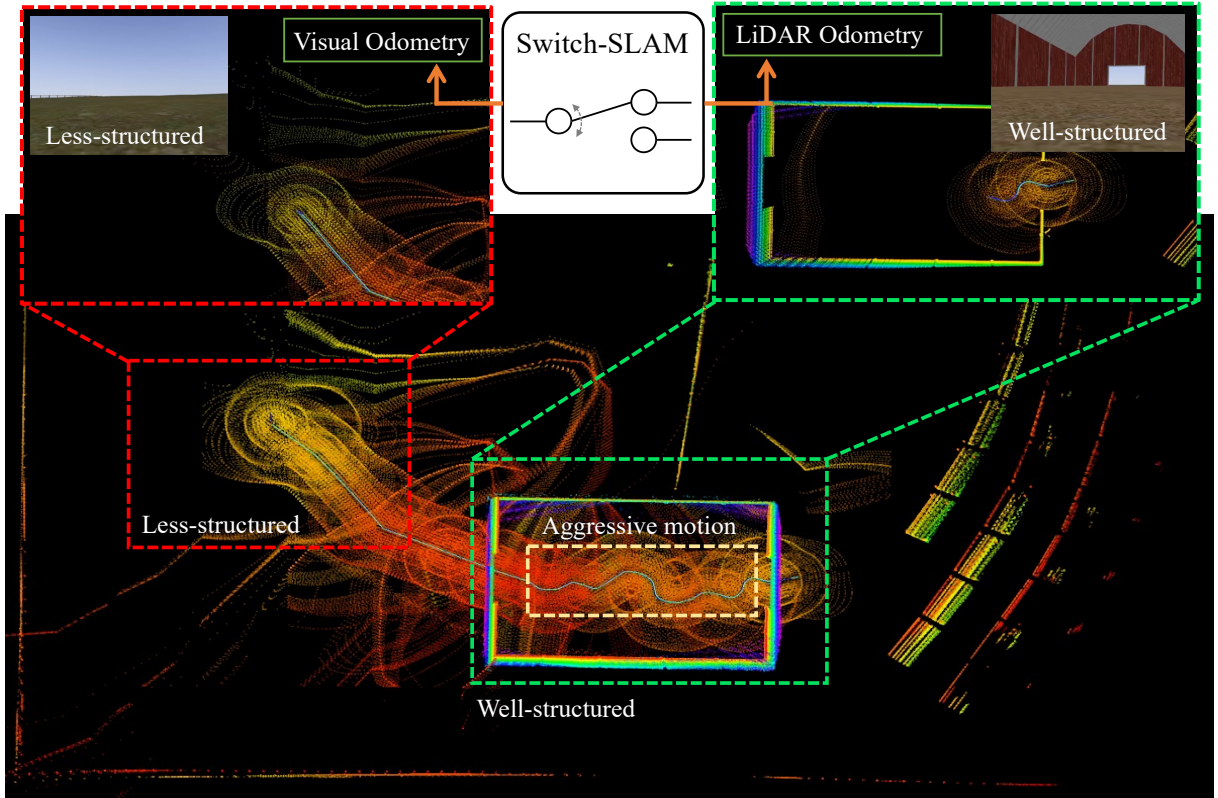


Fig. 3.1: Snapshots and maps from simulated Farm dataset that exhibit both aggressive motions and less-structured environments.

in physical assumptions and statistical significance. This detection enhances the ability to identify degenerate scenarios effectively without the heuristic tuning of the threshold, making it adaptable to various environmental conditions.

- **Experiments on various scenarios:** Switch-SLAM is evaluated by conducting extensive experiments in diverse environments. These scenarios involve degeneracy in both LiDAR and visual odometry, providing a comprehensive evaluation of the system performance. Consequently, Switch-SLAM shows its advantages and effectiveness in challenging scenarios when compared against other state-of-the-art SLAM.

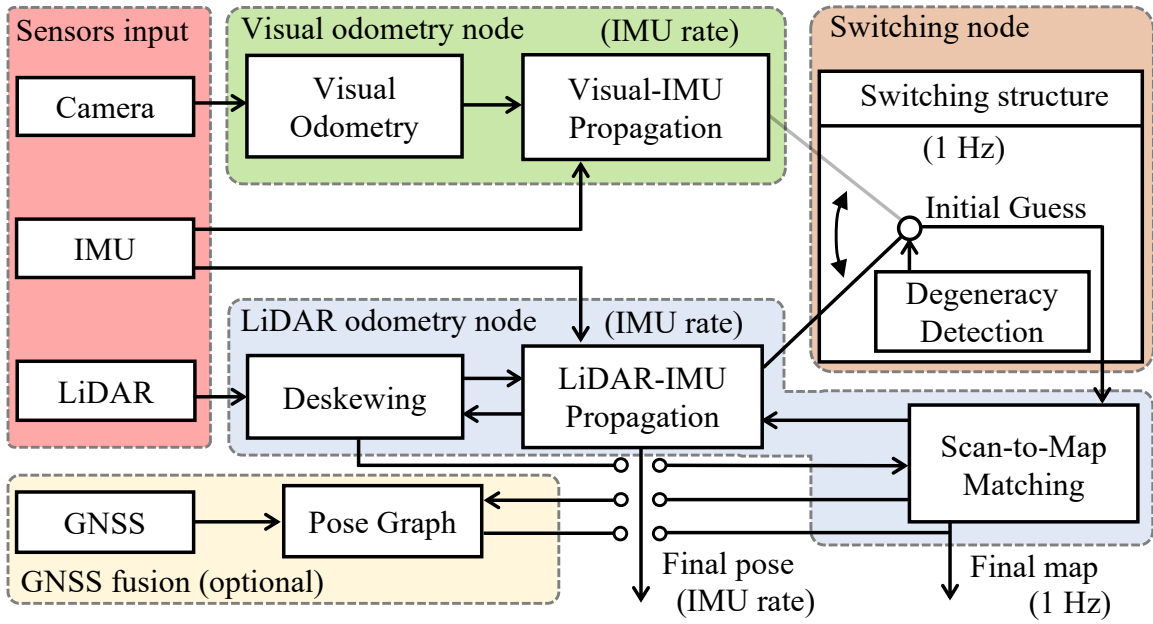


Fig. 3.2: System structure of Switch-SLAM.

3.2 Proposed Method

3.2.1 System Overview

The overview of the proposed method is shown in Fig. 3.2. The proposed approach consists of three main components: visual odometry, LiDAR odometry, and a switching node.

In the visual odometry node, the pose is estimated with sliding window optimization of tracked features, employing the method proposed in [14]. The estimated pose from visual odometry is then propagated at the frequency of the IMU measurements.

In the LiDAR odometry node, the LiDAR distortion resulting from ego-motion is corrected using the poses obtained from the switching structure. Subsequently, scan-to-map matching is conducted utilizing the geometric features proposed in [5], with an initial guess provided by the switching node. The estimated pose from the scan-to-map matching is also propagated at the IMU frequency.

In the switching node, the initial guess for the scan-to-map matching is selected between the poses derived from LiDAR-IMU and visual-IMU propagation, based on the results of degeneracy detection. The proposed work also includes a global navigation satellite system

(GNSS) option [54], which is fused with the final pose from the scan-to-map matching using pose graph optimization.

3.2.2 LiDAR-Inertial-Visual SLAM

I. IMU Preintegration

As proposed in [17], IMU preintegration is utilized to integrate a high-frequency IMU with individual sensor odometry. The IMU preintegration factors ($\Delta \mathbf{p}_{ij}$, $\Delta \mathbf{v}_{ij}$, and $\Delta \mathbf{R}_{ij}$) from time i to j can be as follows:

$$\Delta \mathbf{p}_{ij} = \mathbf{R}_i^\top (\mathbf{p}_j - \mathbf{p}_i - \mathbf{v}_i \Delta t_{ij} - \frac{1}{2} \mathbf{g} \Delta t_{ij}^2) + \delta \mathbf{p}_{ij}, \quad (3.1)$$

$$\Delta \mathbf{v}_{ij} = \mathbf{R}_i^\top (\mathbf{v}_j - \mathbf{v}_i - \mathbf{g} \Delta t_{ij}) + \delta \mathbf{v}_{ij}, \quad (3.2)$$

$$\Delta \mathbf{R}_{ij} = \mathbf{R}_i^\top \mathbf{R}_j \text{Exp}(\delta \phi_{ij}), \quad (3.3)$$

where \mathbf{p} , \mathbf{v} , \mathbf{g} , and \mathbf{R} denote the translation, linear velocity, gravity vector, and rotation matrix of the IMU state, respectively. $\delta \phi_{ij}$, $\delta \mathbf{v}_{ij}$, and $\delta \mathbf{p}_{ij}$ are process noises with Gaussian distribution. Furthermore, \mathbf{p}_j , \mathbf{v}_j , and \mathbf{R}_j can be solved from \mathbf{p}_i , \mathbf{v}_i , and \mathbf{R}_i as follows:

$$\mathbf{p}_j = \mathbf{p}_i + \frac{1}{2} \mathbf{g} \Delta t^2 + \frac{1}{2} \sum_{k=i}^{j-1} (2\mathbf{v}_k \Delta t + \mathbf{R}_k (\mathbf{a}_k - \mathbf{b}_k^a - \boldsymbol{\eta}_k^a) \Delta t^2), \quad (3.4)$$

$$\mathbf{v}_j = \mathbf{v}_i + \mathbf{g} \Delta t + \sum_{k=i}^{j-1} \mathbf{R}_k (\mathbf{a}_k - \mathbf{b}_k^a - \boldsymbol{\eta}_k^a) \Delta t, \quad (3.5)$$

$$\mathbf{R}_j = \mathbf{R}_i \prod_{k=i}^{j-1} \text{Exp}((\boldsymbol{\omega}_k - \mathbf{b}_k^g - \boldsymbol{\eta}_k^g) \Delta t^2), \quad (3.6)$$

where \mathbf{a} and $\boldsymbol{\omega}$ denote linear acceleration and rotation rate of IMU measurements, $\boldsymbol{\eta}$ and \mathbf{b} represent sensor noise and bias. Moreover, Δt denotes the sensor frequency of the IMU. After integrating the IMU preintegration factor with each sensor odometry factor, the last estimated pose is directly propagated using high-frequency IMU measurements, enabling the system to utilize high-frequency poses.

II. LiDAR odometry

LiDAR odometry is performed by scan-to-map matching, as proposed in [4, 5]. In this process, planar and edge features are extracted from each LiDAR scan by evaluating the smoothness of the local surface along the same scan line. Moreover, features in j -th scan and those in i -th map are associated using a nearest neighbor search. With this association established, the distances

between the extracted features in the scan and corresponding points in the map can be calculated as follows:

$$d^e = \frac{\|(\mathbf{p}_j^e - \mathbf{p}_{i,1}^e) \times (\mathbf{p}_j^e - \mathbf{p}_{i,2}^e)\|}{\|\mathbf{p}_{i,1}^e - \mathbf{p}_{i,2}^e\|}, \quad (3.7)$$

$$d^p = \frac{\|(\mathbf{p}_j^p - \mathbf{p}_{i,1}^p)((\mathbf{p}_{i,1}^p - \mathbf{p}_{i,2}^p) \times (\mathbf{p}_{i,1}^p - \mathbf{p}_{i,3}^p))\|}{\|(\mathbf{p}_{i,1}^p - \mathbf{p}_{i,2}^p) \times (\mathbf{p}_{i,1}^p - \mathbf{p}_{i,3}^p)\|}, \quad (3.8)$$

where d^e denotes the distance between \mathbf{p}_j^e and corresponding prior edge features $\mathbf{p}_{i,1}^e$ and $\mathbf{p}_{i,2}^e$. d^p denotes the distance between \mathbf{p}_j^p and corresponding prior planar features $\mathbf{p}_{i,1}^p$, $\mathbf{p}_{i,2}^p$ and $\mathbf{p}_{i,3}^p$. Additional details can be found in [5].

To solve 6-DOF pose \mathbf{x} , scan-to-map matching optimization is defined using distances \mathbf{d}_l (subscript l denotes LiDAR) stacked with all d^e and d^p as follows:

$$\mathbf{f}_l(\mathbf{x}) = \mathbf{d}_l, \quad (3.9)$$

where \mathbf{f}_l is the nonlinear matching cost derived from eqs. (3.7) and (3.8). Moreover, eq. (3.9) can be solved iteratively using the Levenberg-Marquardt method [55] as follows:

$$\mathbf{x} \leftarrow \mathbf{x} - (\mathbf{J}_l^\top \mathbf{J}_l + \lambda \text{diag}(\mathbf{J}_l^\top \mathbf{J}_l))^{-1} \mathbf{J}_l^\top \mathbf{f}_l(\mathbf{x}), \quad (3.10)$$

where $\mathbf{J}_l = \frac{\partial \mathbf{f}_l}{\partial \mathbf{x}}$ is the Jacobian matrix of \mathbf{f}_l and λ is the damping factor. Consequently, eq. (3.10) can be simplified as follows:

$$\delta \mathbf{x} = -\mathbf{H}_l^{-1} \mathbf{J}_l^\top \mathbf{d}_l, \quad (3.11)$$

where $\delta \mathbf{x}$ denotes the transformation increment. The pose from eq. (3.11) is also propagated with IMU measurements.

III Visual odometry

The method in [14] is adapted for the proposed visual odometry submodule. This method effectively addresses the scale problem of monocular vision by initialization with the alignment of visual and IMU motion. After initialization, the sliding window optimization is performed for bundle adjustment, and the pose derived from the optimization is propagated with IMU measurements.

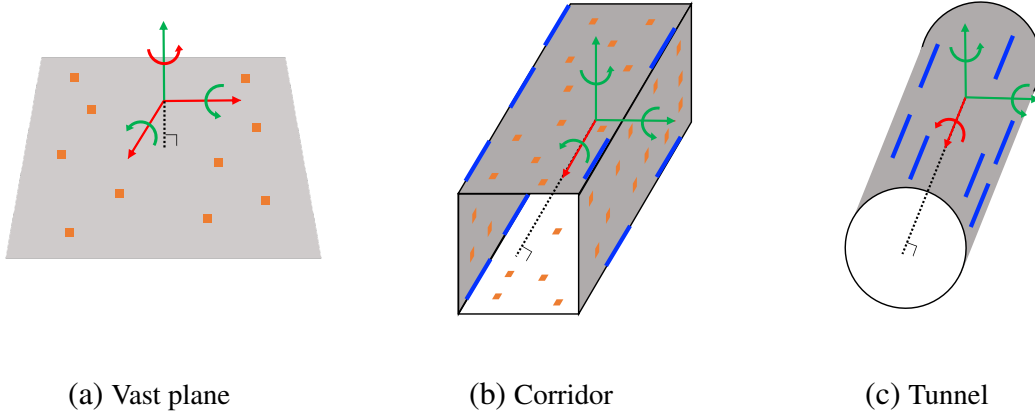


Fig. 3.3: Representative examples of LiDAR odometry degenerate structures and their corresponding well-conditioned/degenerate DOFs. Green arrows denote well-conditioned DOFs. Red arrows denote degenerate DOFs. Orange patches denote planar features. Blue lines denote edge features.

Following [14], nonlinear optimization of visual-inertial odometry is calculated using the Gauss-Newton method as follows:

$$\mathbf{x} \leftarrow \mathbf{x} - (\mathbf{J}_v^\top \mathbf{J}_v)^{-1} \mathbf{J}_v^\top \mathbf{f}_v(\mathbf{x}), \quad (3.12)$$

where \mathbf{f}_v (subscript v denotes visual) denotes the cost function of visual-inertial odometry and $\mathbf{J}_v = \frac{\partial \mathbf{f}_v}{\partial \mathbf{x}}$ denotes the corresponding Jacobian matrix. eq. (3.12) can be simplified as follows:

$$\delta \mathbf{x} = -\mathbf{H}_v^{-1} \mathbf{J}_v^\top \mathbf{d}_v. \quad (3.13)$$

3.2.3 Degeneracy Detection of LiDAR Odometry

Most degenerate cases in LiDAR originate from structure-less environments, such as a vast open field, long corridor, and tunnel-like structure, as depicted in Fig. 3.3. In the case of a vast plane, one translational DOF in the direction perpendicular to the plane and two rotational DOFs, excluding the axis perpendicular to the plane, are constrained. Conversely, in the case of a corridor or tunnel, two translational DOFs, excluding the horizontal DOF in the penetrated direction of each structure, and more than two rotational DOFs are constrained. Similarly, the previous studies about structural degeneracy presented multiple structure-less shapes in which LiDAR odometry degenerates; however, none of them also exceeds 3-DOFs [56]. Consequently, either plane or edge features still exist within the sensing range of LiDAR,

making LiDAR odometry rarely degenerate beyond 3-DOFs. Therefore, this research can make physical assumptions that degeneracy rarely occurs in more than three DOFs out of the six DOFs when a plane or edge feature is present. Consequently, proposed method primarily focuses on the degeneracy of the other 3-DOF directions.

Eigenvalues of \mathbf{H}_1 in eq. (3.11) are utilized to detect degeneracy, where d_1 , d_2 , and d_3 denote the most degenerate DOFs. The corresponding three eigenvalues, denoted as $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3]$, are extracted as the three smallest values from the eigenvalues of \mathbf{H}_1 in ascending order. Then, $\boldsymbol{\lambda}$ is normalized to $\bar{\boldsymbol{\lambda}} = [\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3]$. This research defines a non-heuristic threshold of normalized eigenvalues using the Chi-squared test [57]. The Chi-squared test is a statistical test assessing whether two categorical variables are critically associated. Therefore, by using the Chi-squared test, the boundary line at which statistical significance between the expected and observed values is lost can be defined as the non-heuristic threshold. In the proposed case, the DOF of the Chi-squared test can be set as 2 because $\bar{\boldsymbol{\lambda}}$ is normalized to 1 and one value can be determined when the other two are observed. The null hypothesis posits that each normalized eigenvalue follows its respective expectation within 95% interval, allowing outliers of the observed values at a 5% significance level around the expected value. Therefore, the Chi-squared test formulation to reject the null hypothesis, thereby accepting the alternative hypothesis, is folmulated as:

$$(\bar{\boldsymbol{\lambda}} - \mathbf{e}_m)^2 / \mathbf{e}_m > 0.103, \quad (3.14)$$

where 0.103 denotes the Chi-squared value for 2-DOFs at a 95% confidence level, and \mathbf{e}_m denotes the expectation values of each eigenvalue. Note that although the Chi-squared value is the only user-defined parameter, it remains constant across all the experiments.

According to [58], the eigenvalues distribution of a randomly constructed real symmetric matrix with finite dimensions can be approximated with a semicircle distribution [59]. Therefore, to determine \mathbf{e}_m , this research assumes that each distribution of $\bar{\boldsymbol{\lambda}}$ follows a symmetric probability distribution within the range defined by its maximum and minimum values. Moreover, note that the norm of $\bar{\boldsymbol{\lambda}}$ is 1, and $\bar{\lambda}_1 \leq \bar{\lambda}_2 \leq \bar{\lambda}_3$. Therefore, as $\bar{\lambda}_1$ can take a maximum value of $1/\sqrt{3}$ (when $\bar{\lambda}_1 = \bar{\lambda}_2 = \bar{\lambda}_3$) and a minimum value of 0, the expectation of $\bar{\lambda}_1$, denoted as e_1 , is $1/2\sqrt{3} \approx 0.289$. For λ_2 , the maximum value is $1/\sqrt{2}$ (when $\bar{\lambda}_1 = 0$ and $\bar{\lambda}_2 = \bar{\lambda}_3$) and the average minimum value is e_1 (when $\bar{\lambda}_1 = \bar{\lambda}_2$). Thus, e_2 is 0.498. Similarly, for $\bar{\lambda}_3$, the maximum value is 1, (when $\bar{\lambda}_1 = \bar{\lambda}_2 = 0$), and the average minimum value is e_2 .

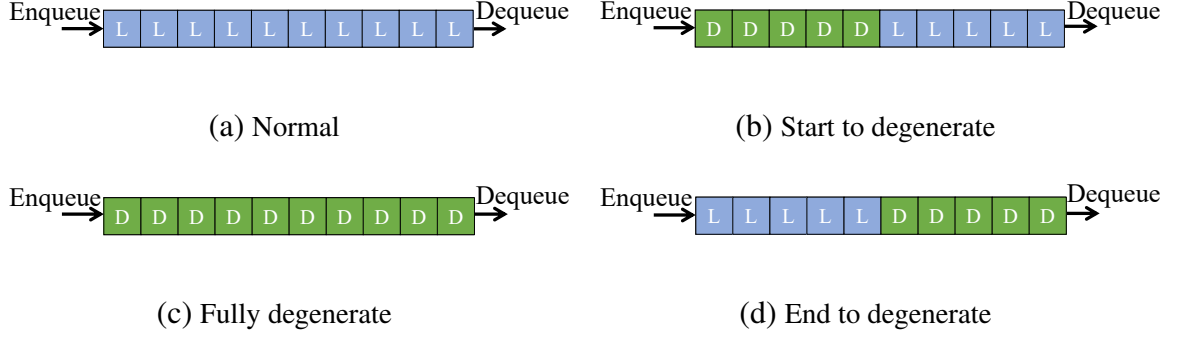


Fig. 3.4: Description of the status buffer. “D” denotes the sequences with degenerate LiDAR odometry in at least one DOF. “L” denotes the sequences with well-conditioned LiDAR odometry.

(when $\bar{\lambda}_2 = \bar{\lambda}_3$). Thus, e_3 is 0.749. Following eq. (3.14), the degeneracy threshold of $\bar{\lambda}$, $\bar{\lambda}_t$ is:

$$\bar{\lambda}_t = \mathbf{e}_m - \sqrt{0.103 * \mathbf{e}_m}. \quad (3.15)$$

Thus, the non-heuristic threshold $\bar{\lambda}_t$ is determined as $[0.12, 0.27, 0.48]$ from eq. (3.15), as the corresponding values of \mathbf{e}_m are 0.289, 0.498, and 0.749, respectively. If any $\bar{\lambda}$ value is lower than the corresponding value of $\bar{\lambda}_t$, the initial guess is “switched” from the value of LiDAR odometry to visual odometry to stabilize the entire scan-to-map optimization process and aid in estimating each eigenvalue stably. Inversely, if all $\bar{\lambda}$ values are greater than the corresponding value of $\bar{\lambda}_t$, the system sets the initial guess based on pure LiDAR odometry.

However, problems still remain when dealing with states that are close to the defined threshold. This situation can rapidly change the value of the initial guess and result in non-smooth outcomes during the scan-to-map optimization process. To address this problem, this research employs the status buffer method, which prevents the status from changing radically. In this method, the present status is classified as “Normal,” “Start/End to degenerate,” and “Fully degenerate” from a queue Q_s with a pre-defined size, which continuously stores past status information, as shown in Fig. 3.4.

Therefore, this research employs linear interpolation to bridge the piercing gap in the estimated state between LiDAR and visual odometry during the start or end of LiDAR odometry degeneracy, as shown in Figs. 3.4(b) and 3.4(d). The initial guess of the 6-DOF state, $\mathbf{T}_k \in \mathcal{M}$, is interpolated during the start or end of the degenerate status using prior state \mathbf{T}_{k-1} , differential state of LiDAR odometry $\delta \mathbf{T}_{k-1,k}^l = \mathbf{T}_k^l \boxminus \mathbf{T}_{k-1}^l \in \mathbb{R}^n$, and differential state of visual

odometry $\delta \mathbf{T}_{k-1,k}^v = \mathbf{T}_k^v \boxminus \mathbf{T}_{k-1}^v$ as follows:

$$\mathbf{T}_k = \mathbf{T}_{k-1} \boxplus \sqrt{3} \bar{\lambda}_1 \delta \mathbf{T}_{k-1,k}^l \boxplus (1 - \sqrt{3} \bar{\lambda}_1) \delta \mathbf{T}_{k-1,k}^v, \quad (3.16)$$

Here, the maximum value of $\sqrt{3} \bar{\lambda}_1$ is 1. “ \boxplus ” and “ \boxminus ” denote the operations to map the elements to and from a given manifold \mathcal{M} and its tangent space \mathbb{R}^n [60].

3.2.4 Failure Detection of Visual Odometry

The minimum eigenvalue of the Hessian matrix of visual odometry is unstable and remains large after failure, as shown in [61]. Therefore, this research adapts failure detection of visual odometry as proposed in [14] instead of degeneracy detection in proposed system. The number of tracked features, bias changes, and positional/rotational changes between consecutive keyframes are used for failure detection. If any of these values exceed the predefined threshold, the system treats the current state as a failure. Moreover, when the failure is detected, the state of visual odometry, denoted as S_{vio} , is set to “fail,” and the system attempts re-initialization. Until successful re-initialization is achieved, the entire system relies on pure LiDAR odometry.

3.2.5 Scan-to-Map Matching

Scan-to-map matching can fail because estimations of directions to degenerate DOFs can be unstable in structure-less environments. To prevent the effect of a degenerate DOF on the optimization process, eq. (3.11) considering the degenerate DOF is remapped. Given \mathbf{H}_1 and its eigendecomposition as $\mathbf{U} \mathbf{\Lambda} \mathbf{U}^{-1}$, the optimization process, when the state of LiDAR odometry is well-conditioned or visual odometry fails, is as follows:

$$\delta \mathbf{x} = -(\mathbf{U} \mathbf{\Lambda} \mathbf{U}^{-1})^{-1} \mathbf{J}_1^\top \mathbf{d}_1. \quad (3.17)$$

When the state of LiDAR odometry is degenerate in at least one DOF and visual odometry does not fail, the optimization process is remapped by fusing visual and LiDAR odometry in a tightly coupled way as follows:

$$\begin{aligned} \delta \mathbf{x} = \underset{\delta \mathbf{x}}{\operatorname{argmin}} & \left(\underbrace{\|\delta \mathbf{x} + (\mathbf{H}_v^{-1} \mathbf{J}_v^\top \mathbf{d}_v)\|}_{\mathbf{e}_v(\delta \mathbf{x})}^2 \right. \\ & \left. + \underbrace{\|\delta \mathbf{x} + (\mathbf{U} \mathbf{\Lambda}_p \mathbf{U}^{-1})^{-1} \mathbf{J}_1^\top \mathbf{d}_1\|}_{\mathbf{e}_l(\delta \mathbf{x})}^2 \right), \end{aligned} \quad (3.18)$$

Algorithm 3.1: Switching node with degeneracy detection

Input: Prior status \mathbf{T}_{k-1} , \mathbf{H}_1 in eq. (3.11),
status buffer queue Q_s with size n , status of VO S_{vio} ,
differential state of LO $\delta\mathbf{T}_{k-1,k}^l$, and VO $\delta\mathbf{T}_{k-1,k}^v$,
Output: Final status \mathbf{T}_k

- 1: 3-DOF normalized eigenvalues $\bar{\lambda} = \text{eigen}_{d_1, d_2, d_3}(\mathbf{H}_1)$
- 2: **if** $S_{\text{vio}} == \text{fail} \vee \forall i, \bar{\lambda}(i) \geq \bar{\lambda}_t(i)$ **then**
- 3: //Use LO propagation as the initial guess
 $\mathbf{T}_k^{\text{init}} = \mathbf{T}_{k-1} \boxplus \delta\mathbf{T}_{k-1,k}^l$
- 4: **else if** $\text{check}(Q_s) == \text{"Start/End to degenerate"}$ **then**
- 5: //Use an interpolation of VO and LO as the initial guess
 $\mathbf{T}_k^{\text{init}} = \mathbf{T}_{k-1} \boxplus \sqrt{3} \bar{\lambda}_1 \delta\mathbf{T}_{k-1,k}^l \boxplus (1 - \sqrt{3} \bar{\lambda}_1) \delta\mathbf{T}_{k-1,k}^v$
- 6: **else**
- 7: //Use VO propagation as the initial guess
 $\mathbf{T}_k^{\text{init}} = \mathbf{T}_{k-1} \boxplus \delta\mathbf{T}_{k-1,k}^v$
- 8: **end if**
- 9: //Scan to map matching
Update \mathbf{T}_k with $\mathbf{T}_k^{\text{init}}$ following eq. (3.19)
- 10: //Update status buffer queue
Dequeue Q_s and Enqueue current status to Q_s .
- 11: **return** \mathbf{T}_k

where Λ_p denotes the matrix with eigenvalues removed corresponding to degenerate DOFs from Λ .

When both LiDAR odometry degeneracy and visual odometry failure occur, the optimization process is executed only along the well-conditioned DOFs. In this case, the IMU preintegration significantly impacts the undetermined directions. Consequently, the entire process of scan-to-map matching is

$$\delta\mathbf{x} = \begin{cases} -\mathbf{H}_1^{-1} \mathbf{J}_1^\top \mathbf{d}_1, & \text{if } \forall i, \bar{\lambda}(i) \geq \bar{\lambda}_t(i) \\ -(\mathbf{U} \Lambda_p \mathbf{U}^{-1})^{-1} \mathbf{J}_1^\top \mathbf{d}_1, & \text{else if } S_{\text{vo}} = \text{fail} \\ \underset{\delta\mathbf{x}}{\text{argmin}} (||\mathbf{e}_v(\delta\mathbf{x})||^2 + ||\mathbf{e}_l(\delta\mathbf{x})||^2), & \text{otherwise} \end{cases} \quad (3.19)$$

Note that although eq. (3.18) relies on MAP fusion, proposed switching structure ensures robustness of the multimodal system, preventing failure or degeneration of one element from affecting the overall fusion process. The entire processes in the switching node are described in Algorithm 3.1.

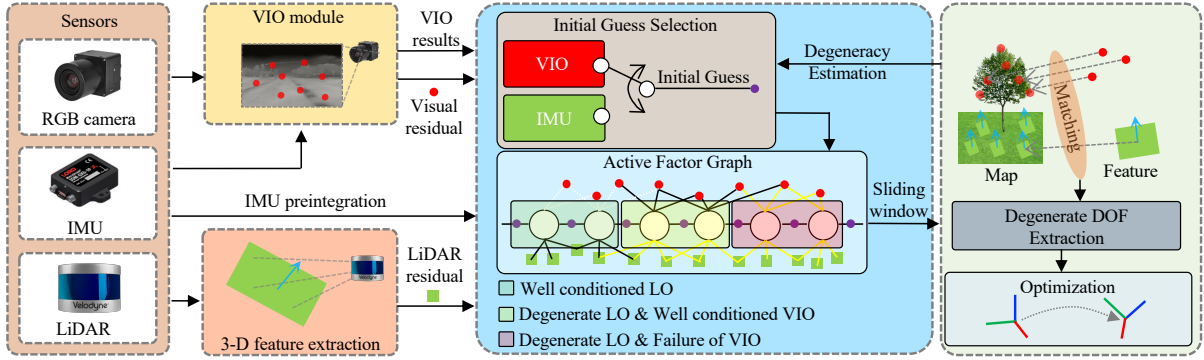


Fig. 3.5: The implementation details of Switch-SLAM

3.2.6 Frontend Implementation of Switch-SLAM

I. Implementation Overview

To implement frontend of Switch-SLAM as described in Section 3.2.2, active factor graph is newly proposed to fuse MAP-based and hard-switching-based method. Each factor is selectively connected or disconnected along with switching conditions.

The overview of the proposed implementation is described in Fig. 3.5. The proposed approach consists of a visual inertial odometry (VIO) module, 3-D feature extraction, active factor graph, and degeneracy-aware optimization. Firstly, VINS-MONO [14] is used for the VIO module to derive VIO results and visual residual that contains visual-inertial bundle adjustment formulation. Then, LOAM [5] is used for feature extraction to extract plane features and LiDAR residual calculated as point-to-plane distance. After the initial guess is selected between VIO and IMU preintegration [17] along with degeneracy detection results, an active factor graph is constructed for sliding window optimization. Finally, degeneracy-aware optimization is performed to selectively optimize only directions of well-conditioned DOF.

II. Active Factor Graph

Based on the factor graph theory [53], the active factor graph utilizes three types of factors: the initial guess factor, the LiDAR residual factor, and the visual residual factor. The factor graph is actively constructed along with S_l and S_{vio} . Here, S_{vio} and S_l denote the states of VIO and LO, respectively.

When LO is well-conditioned, the initial guess, propagated with IMU preintegration and LiDAR odometry, is used for the initial guess factor. Moreover, only the LiDAR residual in eq. (3.19) and the initial guess factor are used to construct the factor graph, excluding the

visual residual. Note that visual odometry is relatively less accurate compared to LiDAR odometry in well-structured environments, which makes the multimodal fusion less accurate in such environments.

When LO is degenerate and VIO is well-conditioned, the VIO result, propagated at the IMU rate, is used for the initial guess factor. Furthermore, both the LiDAR and visual residuals are inserted into the factor graph to help escape from LiDAR degeneration. Note that visual residual is denoted in [14].

When LO is degenerate and VIO experiences failure simultaneously, the IMU preintegration value is utilized for the initial guess factor. The LiDAR residual and initial guess factor are then utilized to construct the factor graph. In this case, the initial guess factor has a significant impact on the directions of the degenerate DOF in optimization process, reducing drift even in structurally and visually degenerate situations.

3.2.7 Backend Implementation of Switch-SLAM

Switch-SLAM contains pose graph optimization to efficiently fuse the GNSS or loop closure with the system. To optimize the pose graph, iSAM2 [62], which is time-efficient and robust in large environments, is used. compared to the general nonlinear optimization such as Gauss-Newton or Levenberg-Marquardt method. Moreover, Scan context [63] is employed for the loop-closing submodule, which shows high accuracy and robustness.

3.3 Experiments

In this section, the evaluation of the accuracy and robustness of the proposed method with various datasets containing sensor degeneracy is presented. Furthermore, the effectiveness of the proposed degeneracy detection is discussed.

3.3.1 Datasets

This research prepares various datasets with different environments, as summarized in Table 3.1 and shown in Fig. 3.6. First, proposed method is evaluated on simulated datasets: Plane, Fast Rotate, and Farm datasets. The Plane dataset demonstrates degenerate environments for LiDAR SLAM in the entire area owing to the presence of predominantly planar structures. The Farm dataset contains both degeneration of visual SLAM, in the blue line region, caused by fast rotations (also shown in the Fast Rotate dataset) and degeneration of LiDAR SLAM, in the red line region, caused by the predominance of plane-only structure. Note that the Farm dataset is specifically designed to evaluate scenarios with visual and structural degeneration. In real-world farmland settings, such challenges are common, as farming vehicles equipped with continuous tracks often experience rapid rotations and navigate expansive, open fields. All the simulations are conducted with ROS [64] and in the Gazebo simulator [65] using open-sourced environments. The sensor suite of the simulated robot contains Velodyne VLP-16 operating at 10 Hz, 640×480 RGB camera operating at 60 Hz, and Gazebo basic plugin 9-axis IMU operating at 200 Hz.

Table 3.1: Dataset details for Switch-SLAM

Type	Dataset	LiDAR SLAM	Visual SLAM	Total Distance (m)
Simulation	Fast Rotate	Well-constraints	Degenerate	115
	Plane	Degenerate	Well-constraints	109
	Farm	Degenerate	Degenerate	536
Real-world	Handheld	Degenerate	Well-constraints	2850
	Multi Floor	Degenerate	Degenerate	270
	Long Corridor	Degenerate	Degenerate	616
	ANYmal 1	Degenerate	Well-constraints	240
	ANYmal 2	Degenerate	Well-constraints	687
	ANYmal 3	Degenerate	Degenerate	311
	ANYmal 4	Well-constraints	Well-constraints	500

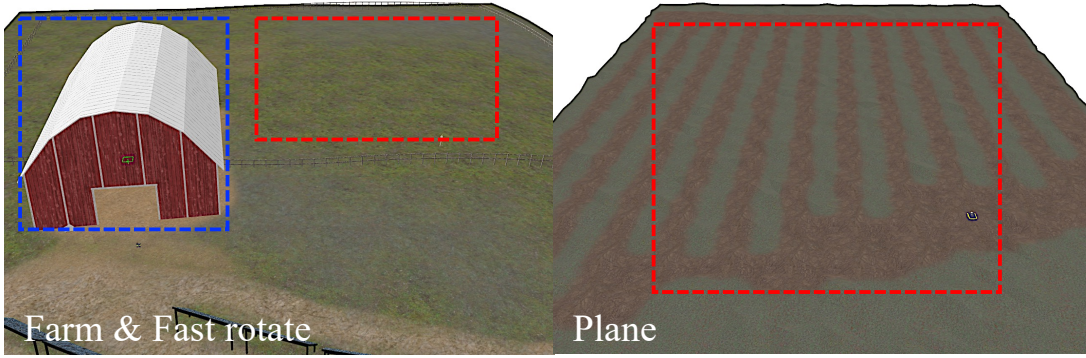


Fig. 3.6: Simulated environments. The red region indicates the region of LiDAR SLAM degeneration. The blue region indicates the region of visual SLAM degeneration due to fast rotation.

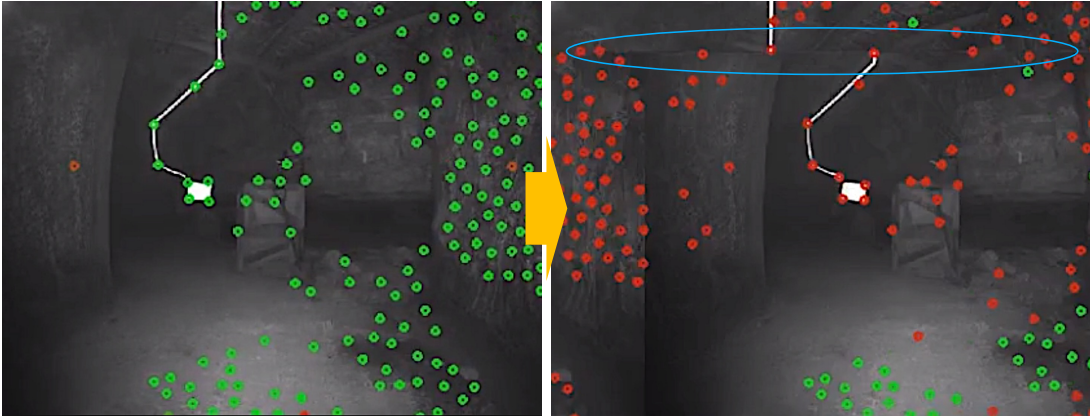


Fig. 3.7: Imaging problem caused by video interruption in ANYmal 3. Green dots represent tracked features and red dots indicate untracked features. The blue region highlights the abnormal parts induced by the imaging problem.

Second, proposed method is evaluated on real-world and open-sourced datasets: Handheld [18], CERBERUS DARPA subterranean challenge [66] and SubT-MRS [67] datasets. The Handheld dataset contains degeneration of LiDAR odometry caused by vast open fields. Because the CERBERUS dataset (ANYmal 1, 2, 3, and 4) lacks the degeneration of LiDAR, the horizontal field-of-view of LiDAR is limited at 180° to create more structure-less situations for each scan. This setup induces LiDAR degradation in ANYmal 1, 2, and 3. Additionally, in this experiment, the front right camera (cam1) is only used in the CERBERUS dataset. This camera experiences a single video interruption momentarily at approximately 10 frames in ANYmal 3, as shown in Fig. 3.7, leading to the failure of visual odometry. The Multi Floor and Long Corridor dataset (SubT-MRS dataset) contain structure-less and visually challenging scenes simultaneously.

The proposed method, Switch-SLAM is compared with the state-of-the-art of LiDAR [5, 4], visual [14], and LiDAR-visual odometry [11, 19, 51, 18]. All the methods are executed with Ubuntu OS and on an Intel i7-1165G7 CPU in real-time operation. The GNSS integration and loop closure are not used. All the experimental results are presented as averages obtained from each set of three repeated tests.

3.3.2 Accuracy Evaluation

The entire results of the evaluation of accuracy and trajectories are shown in Table 3.2 and Fig. 3.8. On the Fast Rotate dataset, LIO-SAM shows the best performance among the compared methods, whereas VINS-MONO fails in their localization because of aggressive rotation. Compared LiDAR visual inertial odometry (LVIO) methods demonstrate a larger drift than pure LiDAR-based methods. Proposed method is competitive with LIO-SAM because Switch-SLAM works as pure LiDAR SLAM in well-structured environments using the switching structure. On the Plane dataset, which mainly contains less-structured ground-only environments, the proposed method and VIN-MONO exhibit the best performance among the compared methods, whereas the LiDAR-based methods fail in their localization. Proposed method also outperforms state-of-the-art of LiDAR-visual SLAM because Switch-SLAM mainly employs visual odometry for its initial guess of scan matching in less-structured environments.

On the Farm dataset, which contains both aggressive motion and less-structured environments, LiDAR odometry fails in the phase of mapping less-structured environments, whereas visual odometry fails in the phase of aggressive motion. Conversely, the proposed method outperforms not only compared LiDAR and visual SLAM but also the state-of-the-art LVIO methods. This result is attributed to the switching structure, which allows for appropriate status transitions based on the given environmental conditions.

In the Handheld dataset, the proposed method is competitive with LVI-SAM, whereas it outperforms the other compared methods, as shown in Table 3.2. Note that the Handheld dataset contains a few LiDAR SLAM degeneracy phases, which makes no significant difference between LVI-SAM and the proposed method. When visual SLAM degeneracy is prolonged such as in the Fast Rotate and Farm datasets, LVI-SAM can drift significantly compared to the proposed method.

On the Multi Floor and Long Corridor dataset, the proposed method shows the best performance among the compared methods. Most of the compared methods suffer with scenes

Table 3.2: Comparison of ATE [m] (Maximum, RMSE) on the tested datasets.

Dataset	Fast Rotate		Plane		Farn		Handheld		Multi Floor		Long Corridor		ANYmal 1		ANYmal 2		ANYmal 3		ANYmal 4	
	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE
LOAM	1.41	0.44	-	-	-	-	-	-	17.9	10.6	25.6	12.6	10.26	6.63	9.67	5.05	7.81	2.39	5.79	3.90
LIO-SAM	0.72	0.21	-	-	-	-	-	-	-	-	17.6	7.64	8.38	3.77	-	-	7.10	3.52	2.47	1.02
VINS-MONO	-	-	1.17	0.41	-	-	21.4	10.3	12.8	6.30	23.8	11.8	24.9	9.08	36.9	15.1	-	-	8.55	3.52
LYI-SAM	8.82	1.82	1.82	0.69	28.8	8.52	5.75	3.27	1.23	-	8.62	4.37	5.83	2.41	9.53	3.28	-	-	6.42	3.75
R2LIVE	1.67	0.64	19.5	8.53	8.52	4.21	-	-	35.2	18.5	-	-	-	-	14.5	7.29	8.60	3.73	3.90	1.18
R3LIVE	10.1	6.43	9.01	5.84	58.6	34.7	-	-	32.4	19.0	14.5	7.63	-	-	-	-	6.77	2.06	27.1	14.0
FAST-LIVO	11.3	7.12	-	-	51.2	26.5	-	-	-	-	-	-	-	-	4.87	1.48	-	-	-	-
Switch-SLAM	1.50	0.23	1.27	0.35	1.10	0.38	3.07	1.25	3.63	1.61	5.09	2.42	2.96	1.29	3.41	1.37	3.68	1.61	2.42	1.05

“-” denotes the failure of localization. The units are in meters.

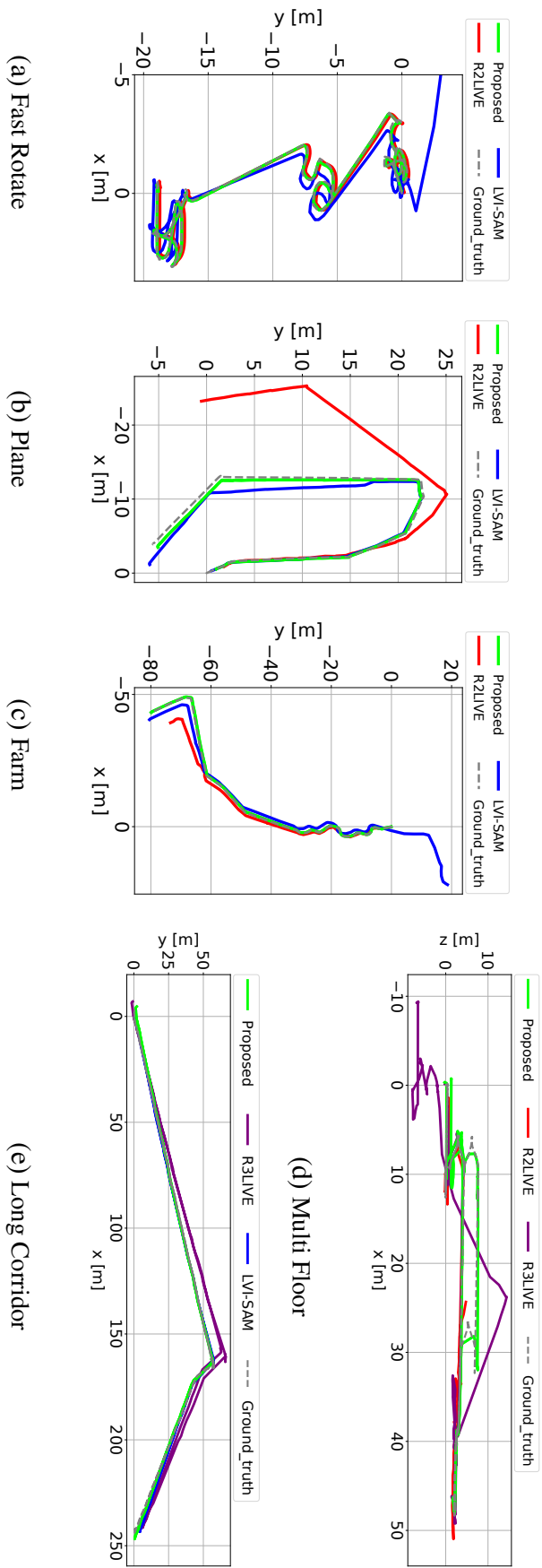


Fig. 3.8: Trajectory of proposed and compared LiDAR visual SLAM in Fast Rotate, Plane, Farm, Multi Floor, and Long Corridor dataset.

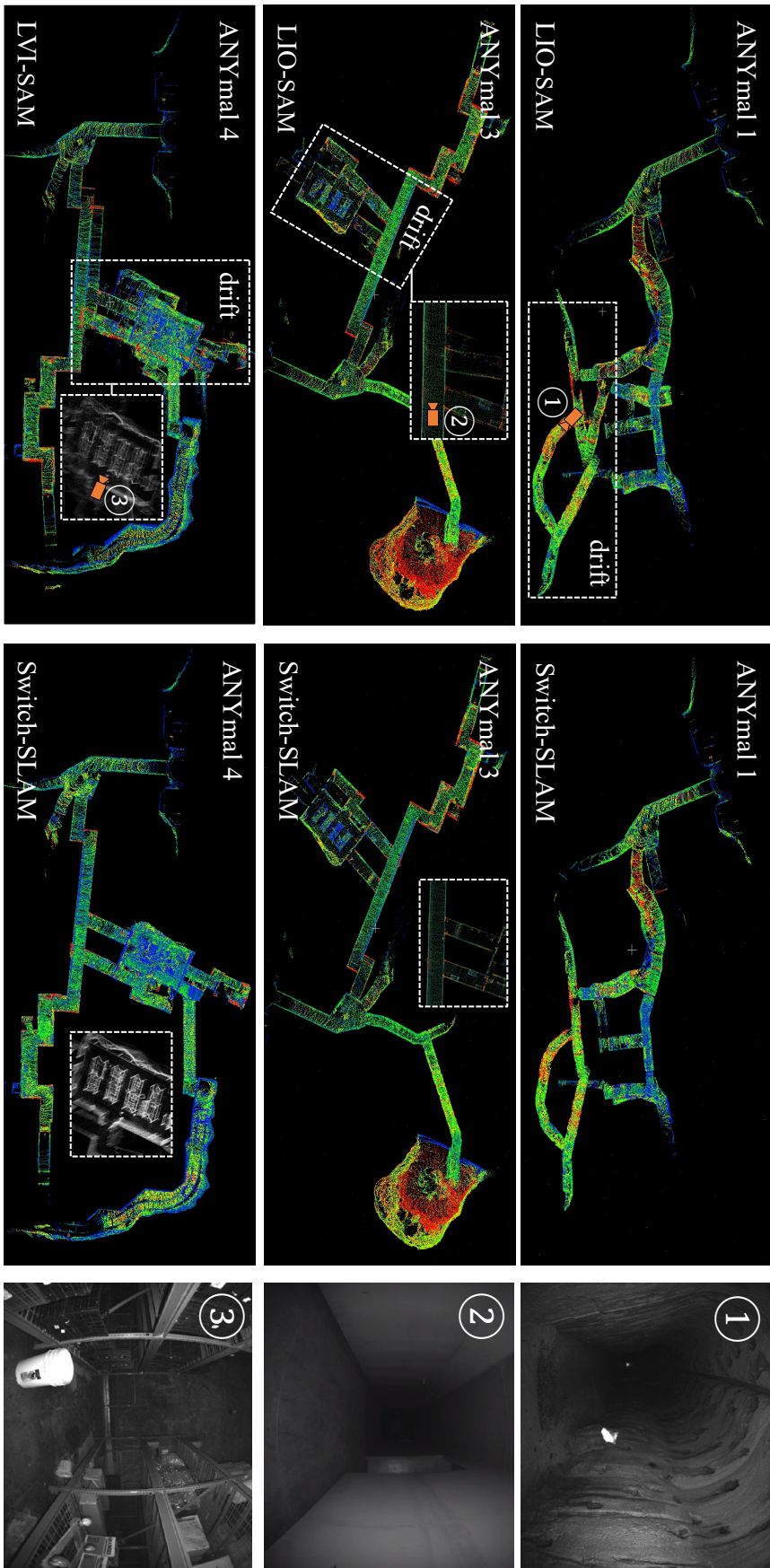


Fig. 3.9: Resulting maps from the compared methods and Switch-SLAM.

featuring both structure-less environments and visual degradation. By comparison, the proposed method deals with these challenges well using the switching-based optimization as expressed in eq. (3.19).

On the CERBERUS dataset, the proposed method demonstrates the best performance in ANYmal 1 and ANYmal 2. This result highlights the ability of Switch-SLAM to effectively address LiDAR degeneration, as illustrated in Fig. 3.9, even outperforming the compared LVIO methods. In ANYmal 3, which experiences a single camera interruption, VINS-MONO and LVI-SAM fail in mapping. Moreover, the corridor-like structure makes LOAM and LIO-SAM degenerate. Conversely, Switch-SLAM successfully conducts SLAM in these environments, owing to its switching structure. In ANYmal 4, the LIO-SAM and Switch-SLAM demonstrate superior performance to LVI-SAM. This result is because LVI-SAM relies on VINS-MONO as the initial guess for scan-to-map matching. A significant disparity in state estimation between VINS-MONO and scan-to-map matching lead to substantial drift. In contrast, owing to the status buffer and state interpolation method, the proposed method bridges the substantial gap between visual odometry and scan-to-map matching, leading to successful mapping.

3.3.3 Degeneracy Detection Evaluation

To evaluate the accuracy of degeneracy detection, the proposed method is compared with the state-of-the-arts [9], [68]. The ground truth is prepared by comparing GNSS data with scan-to-scan matching using ICP [34] at each keyframe. In the evaluation, the threshold for [9] ($= \lambda_1$) is set to 50, 100, and 200. Moreover, the threshold for [68] ($= \lambda_6/\lambda_1$) is set to 5, 10, and 15. Among them, the best accuracy and recall are obtained for a threshold of 100 for [9] and of 10 for [68].

The experimental results reveal an accuracy of 0.91 for [9], 0.96 for [68], and 0.96 for the proposed method. The recall is 0.91 for [9], 0.96 for [68], and 0.99 for the proposed method. These result show that the proposed method achieves 5.5% greater accuracy and 8.8% greater recall compared to [9]. The comparison of the proposed method with the state-of-the-arts is illustrated in Fig. 3.10. Notably, during the third phase of degeneracy, the proposed method successfully detects the degeneracy, which the state-of-the-art methods fail to identify. Note that compared methods are sensitive to threshold tuning, which is not required by proposed method. This detection is accomplished by normalizing the minimum eigenvalue using 3-DOF eigenvalues and applying a predefined threshold based on the Chi-squared test. Note that, in this scene, only 1-DOF is degenerate, whereas the Farm dataset, as noted in Table 3.1, achieves

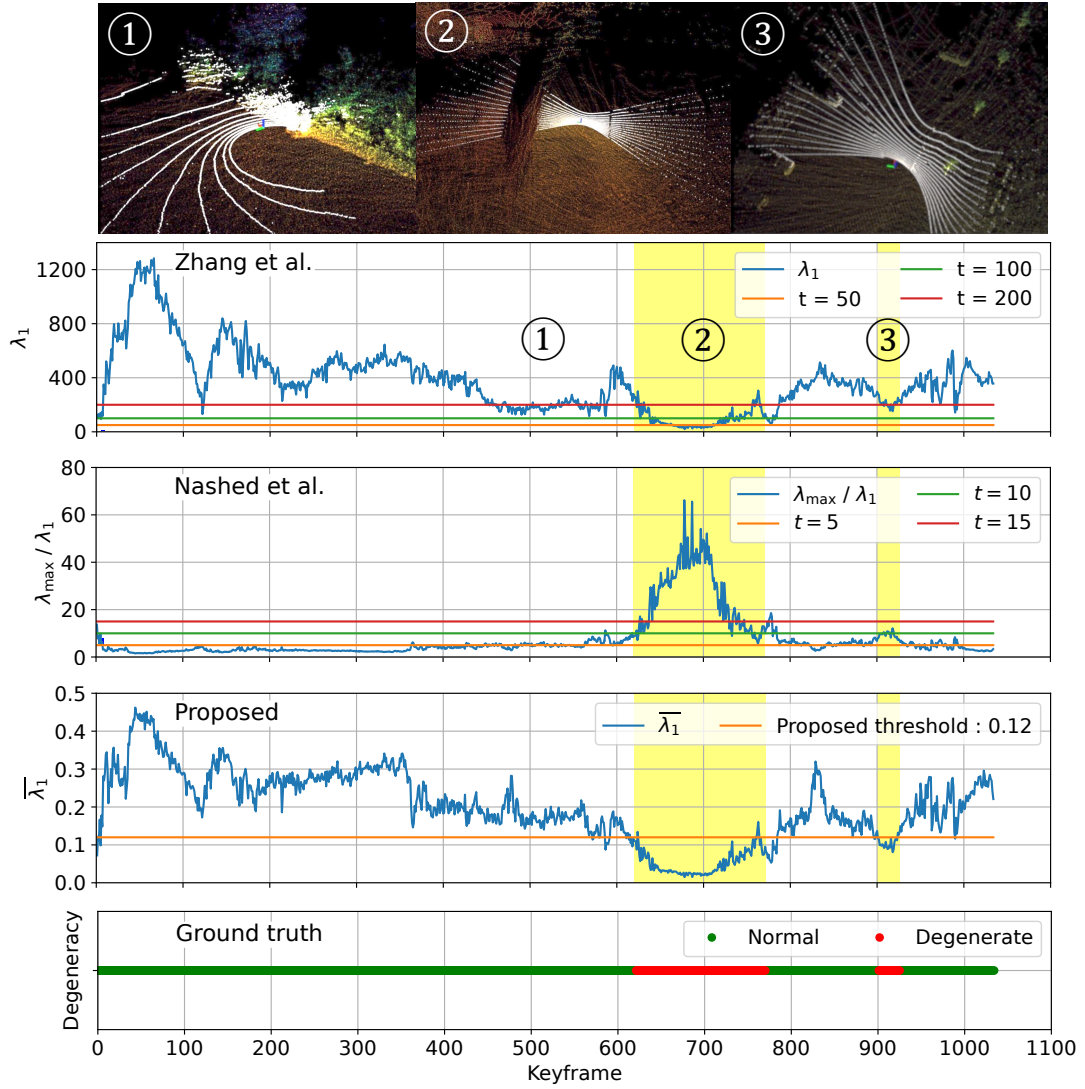


Fig. 3.10: Comparison of degeneracy detection of the state-of-the-art and proposed methods on a part of the Handheld dataset, with visualization of the map (color) and LiDAR scan (white). The bottom figure shows the ground truth. The yellow regions are the actual degenerate regions.

the best accuracy using well-defined parameters determined by the Chi-squared test, even in cases where more than 1-DOF is degenerate.

3.4 Summary

In this chapter, Switch-SLAM, specially designed to address degeneracy situations of individual sensor odometry, was proposed. To deal with the limitations of MAP-based sensor fusion, Switch-SLAM introduced a novel switching-based sensor fusion approach that utilizes a switching structure to effectively prevent failure information from propagating throughout the system, thereby enhancing robustness in degenerate situations. Furthermore, Switch-SLAM introduced non-heuristic degeneracy detection method, which eliminates the need for heuristic tuning. Experimental evaluations involving scenarios with degeneration in LiDAR or visual odometry revealed that Switch-SLAM outperformed the state-of-the-art LiDAR, visual, and LiDAR-visual SLAM methods in terms of accuracy and localizability.

Chapter 4

Self-TIO: Thermal-Inertial Odometry via Self-Supervised 16-bit Feature Extractor and Tracker

4.1	Introduction: Self-TIO	40
4.2	Proposed Method	42
4.2.1	ThermalLANet	42
4.2.2	Thermal Optical Flow.....	45
4.2.3	Hybrid Feature Tracker	48
4.2.4	Thermal Inertial Odometry Formulation.....	48
4.3	Experiments	50
4.3.1	Evaluation on Feature Point Detection.....	50
4.3.2	Evaluation on Feature Point Tracker.....	52
4.3.3	Evaluation on State Estimation Performance	54
4.4	Summary	60

4.1 Introduction: Self-TIO

Visual-inertial odometry has played a crucial role in state estimation for mobile robots and vehicles, enabling them to autonomously explore complex and extreme environments. However, visual odometry tends to degrade and fail in scenarios involving rapidly changing and poor lighting conditions [69]. To address these challenges, longwave infrared (LWIR) thermal cameras are employed because they are less affected by variations in visible light. These cameras primarily capture temperature information, which falls within the longwave infrared spectrum. Despite these advantages, thermal cameras have inherent limitations.

First, images captured by thermal cameras usually have lower contrast than that of images from standard visible light cameras. This occurs because thermal cameras primarily detect temperature radiation, making it difficult to capture the textural information that is predominantly present in the visible wavelength regions. Additionally, rescaling 16-bit thermal images, the standard data type for thermal cameras, to 8-bit formats further reduces contrast, leading to information loss, amplified noise, and image inconsistencies. Second, thermal cameras experience imaging interruptions, typically ranging from 0.1 to 1.0 second, during non-uniformity correction (NUC). NUC is an embedded sensor processing step designed to remove harsh fixed-pattern noise (FPN), also known as stripe noise [44]. These sudden frame drops can cause abrupt perspective shifts, resulting in tracking failures in visual odometry. Third, despite the NUC process, significant FPN and low-frequency noise (LPN) still persist in thermal images [70]. These factors make it challenging to accurately extract and track features in thermal images.

To address these issues, several learning-based methods for extracting or tracking visual features in thermal odometry have been proposed [44, 46]. However, these methods typically require labeled datasets, which are scarce for thermal vision applications [46], or involve converting 16-bit thermal images to 8-bit, which can result in data loss and reduced accuracy [44, 46]. Additionally, end-to-end learning-based methods for thermal odometry have also been developed [48]. These methods often require corresponding visible light image datasets for each thermal image [49] and may suffer from low generalizability [71] due to the inherent limitations of end-to-end learning.

To overcome the aforementioned issues, this research proposes a thermal-inertial odometry, Self-TIO, which meets real-time requirements and employs a fully self-supervised, lightweight

network for both feature extraction and tracking. The main contributions of the proposed method are as follows:

- **Self-supervised 16-bit feature extractor and tracker:** The proposed learning-based feature extractor and tracker are trained using fully self-labeled data, addressing the scarcity of thermal datasets. Moreover, proposed network is exclusively trained on original 16-bit thermal images, eliminating the need to convert them to 8-bit, thereby preventing potential information loss.
- **ThermalNet feature extractor:** This research introduces a self-supervised feature extractor based on LNet [72]. To ensure real-time execution and enhance baseline performance, this research proposes a gradient filter-fused feature extractor. The gradient filter reduces the load on the neural network, helping to make the entire network lighter and more efficient.
- **Hybrid feature tracker:** This research proposes a hybrid feature tracker that combines a learning-based optical flow [47] with a 16-bit level KLT tracker [44, 73]. Given that most optical flow networks [47] struggle with severe noise in thermal images and cannot ensure sub-pixel accuracy, this research refines the optical flow results using the optimization-based 16-bit KLT method. By integrating the learning-based optical flow and 16-bit KLT, proposed tracker achieves robustness in noisy thermal images and enhances sub-pixel accuracy.

4.2 Proposed Method

The proposed method, shown in Fig. 4.1, consists of three main components: a feature extractor (**ThermalLANet**), a feature tracker (**Hybrid Tracker**), and a thermal-inertial odometry module (**TIO**). The key contributions, the feature extractor and tracker, are fully self-trained on 16-bit radiometric images, enabling robustness in texture-less and noisy images and improving generalizability by addressing dataset limitations and rescaling issues in thermal images.

4.2.1 ThermalLANet

I. Network Architecture

To extract robust feature points, this research proposes ThermalLANet, inspired by the self-supervised feature extractor LANet [72]. The network overview is shown in Fig. 4.2. The input to the network is a raw 16-bit radiometric image, and the outputs include feature locations, confidence scores, and descriptors that are used only during training. The feature points for the TIO are selected based on the confidence scores. Moreover, the confidence score plays an important role in weighting each thermal landmark during bundle adjustment, which improves robustness in noisy and low-contrast scenes.

Note that while thermal images often have significant noise and low texture as shown in Fig. 1.3 and [70], the edges remain useful for feature point extraction as noted in [43]. Additionally, fusing handcrafted filters before the convolutional neural network (CNN) layers can help to make the entire network both lighter and more accurate as noted in [74]. Consequently, ThermalLANet incorporates gradient filters \mathbf{G} , which consists of five gradient filters, before the CNN feature extractor, as follows:

$$\mathbf{G}(\mathbf{I}) = [\mathbf{I}_x, \mathbf{I}_y, \mathbf{I}_{xx}, \mathbf{I}_{yy}, \mathbf{I}_{xy}], \quad (4.1)$$

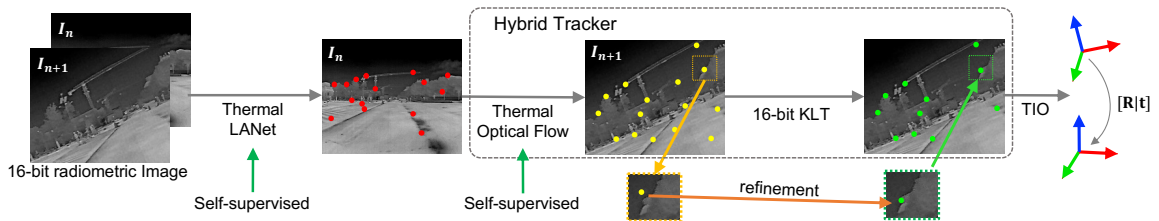


Fig. 4.1: System overview of Self-TIO.

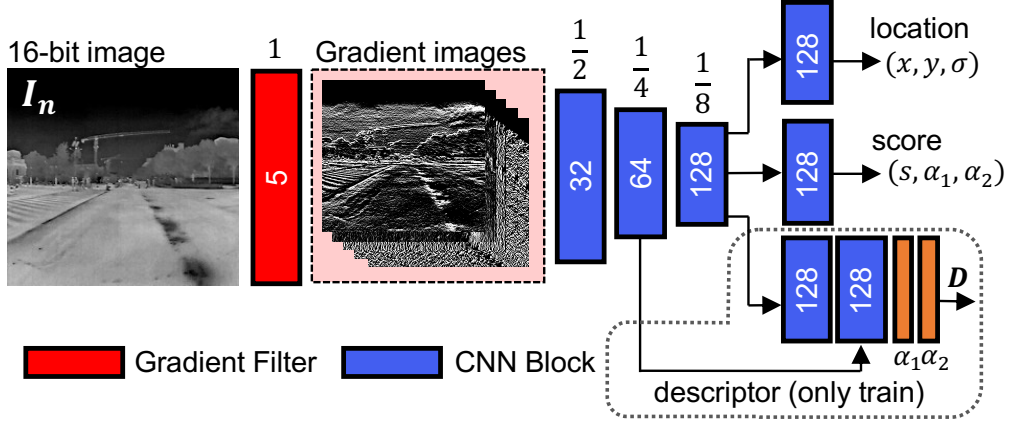


Fig. 4.2: Overview of the ThermalLANet network.

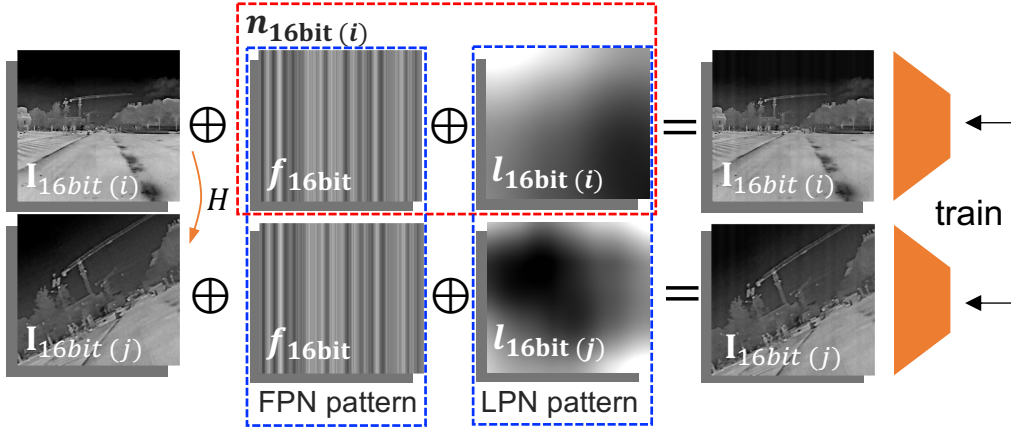


Fig. 4.3: Noise augmentation to train proposed network using a transformed 16-bit radiometric image.

where the subscripts x and y denote $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$, respectively. Note that this research uses the Sobel operator to implement the image gradient.

II. Training Details

To train the network in a self-supervised manner, image pairs and pseudo-ground truth, which are randomly masked to account for occlusions, are generated using homography and thin plate spline transformation \mathbf{H} . However, the FPN, which typically displays a vertical pattern, is also transformed according to the specific transformation. To address this issue, this subsection proposes a novel thermal augmentation method for 16-bit images \mathbf{I}_{16bit} , as shown in Fig. 4.3. Note that, unlike 8-bit visible images, 16-bit radiometric thermal images utilize only a small

portion of the entire 16-bit data range (0–65535), even though this portion is generally much larger than the 8-bit data range (0–255). For this reason, at the first, a pattern \mathbf{f} of FPN is generated by merging randomly generated vertical line ranges with values from 0 to 1 and applying blur noise. Then, the generated noise pattern is scaled and shifted to create a 16-bit noise pattern $\mathbf{f}_{16\text{bit}}$ using a scale parameter α and a shift parameter β , as follows:

$$\mathbf{f}_{16\text{bit}} = \text{blur}(\alpha\mathbf{f}) + \beta, \quad (4.2)$$

where \mathcal{U} denotes uniform distribution, and P_θ denotes a percentile from the top $\theta\%$. The parameters α and β are defined as follows,

$$\alpha = P_{\mathcal{U}(0,\theta)}(\mathbf{I}_{16\text{bit}}) - P_{\mathcal{U}(100-\theta,100)}(\mathbf{I}_{16\text{bit}}), \quad (4.3)$$

$$\beta = \mathcal{U}(P_{50-\pi}(\mathbf{I}_{16\text{bit}}), P_{50+\pi}(\mathbf{I}_{16\text{bit}})), \quad (4.4)$$

where θ and π are pre-defined parameters, set to 15 and 10, respectively, in this section. To apply a random thermal noise, LPN [70], a pattern \mathbf{l} is prepared and random transformations are applied to \mathbf{l} . This pattern is then converted into a 16-bit LPN pattern augmentation $\mathbf{l}_{16\text{bit}}$, as described in eq. (4.2). Finally, $\mathbf{l}_{16\text{bit}}$ and $\mathbf{f}_{16\text{bit}}$ are added to the original 16-bit image pairs.

Using pseudo-ground truth generated by \mathbf{H} , and noise augmented image pairs, ThermalLNet is trained in a self-supervised manner using four loss functions: location loss, score loss, point-wise triplet loss, and correspondence loss. The location loss, score loss, and triplet loss are calculated by comparing the source image points with the corresponding warped target image points, while the correspondence loss is calculated by comparing the predicted matches with the pseudo-ground truth. The location loss L_{loc} is defined as follows:

$$d(p_s^i, p_t^i) = \|\mathbf{H}p_s^i - p_t^i\|_{L2}, \quad (4.5)$$

$$L_{\text{loc}} = \frac{1}{N} \sum_i^N d(p_s^i, p_t^i), \quad (4.6)$$

where p_s and p_t denote the pixel positions in the source and target images, respectively, and N denotes the total number of pixels. The score loss L_{score} is defined as follows:

$$L_{\text{score}} = \frac{1}{N} \sum_i^N \left[\frac{(s_s^i + s_t^i)}{2} (d(p_s^i, p_t^i) - \bar{d}) + (s_s^i - s_t^i)^2 + (\alpha_s^{1,i} - \alpha_t^{1,i})^2 + (\alpha_s^{2,i} - \alpha_t^{2,i})^2 \right], \quad (4.7)$$

where \bar{d} is the average distance of the nearest point pairs as calculated with eq. (4.5), and s , α denote the confidence score and the descriptor-aware score, respectively, as described in Fig. 4.2. The triplet loss L_{tri} is defined as follows:

$$L_{\text{tri}} = \frac{1}{N} \sum_i^N [\|\mathbf{D}p^i - \mathbf{D}p_+^i\|_{L2} - \|\mathbf{D}p^i - \mathbf{D}p_-^i\|_{L2} + \beta], \quad (4.8)$$

where $\mathbf{D}p^i$, $\mathbf{D}p_+^i$, and $\mathbf{D}p_-^i$ denote the anchor, positive, and negative samples of the triplet, respectively, and β is a learning parameter. Note that the triplet loss is designed to ensure that the anchor $\mathbf{D}p^i$ is closer to a positive sample $\mathbf{D}p_+^i$ (a similar point) than to a negative sample $\mathbf{D}p_-^i$ (a dissimilar point) by a margin of at least β [75]. The correspondence loss L_{cor} is defined as follows:

$$L_{\text{cor}} = - \left[\frac{1}{|\mathbf{M}_+|} \sum_{(i,j) \in \mathbf{M}_+} \log \mathbf{C}_{s,t}^{i,j} + \frac{\gamma}{|\mathbf{M}_-|} \sum_{(i,j) \in \mathbf{M}_-} \log(1 - \mathbf{C}_{s,t}^{i,j}) \right], \quad (4.9)$$

where $\mathbf{C}_{s,t}$ denotes the predicted matching matrix, \mathbf{M}_+ and \mathbf{M}_- are the sets of positive and negative matches, respectively, and γ is a learning parameter used to balance the positive and negative terms. The further details of these loss functions and training details can be found in [72]. Meanwhile, the main differences from the original LANet are: (1) a lightweight network design, achieved through the fusion of handcrafted filters; (2) improved robustness to noise through noise augmentation that considers both FPN and LPN; and (3) enhanced accuracy by using raw 16-bit thermal images.

4.2.2 Thermal Optical Flow

I. Network architecture

To track feature points robustly in the face of aggressive scene changes, This research proposes a thermal Optical Flow, inspired by RAFT [47]. Unlike the previous raft-based network [46, 47], the proposed tracking network is fully self-supervised and operates with 16-bit radiometric image pairs. The network overview is shown in Fig. 4.4. Inspired by [46], this research use a lightweight context generator derived from a feature encoder and a single-layer correlation volume. Moreover, the thermal optical flow network employs a lightweight transformer [76, 77] to handle aggressive motion and occlusions caused by NUC. This research uses motion features as queries and keys, and context features as value vectors. This approach helps the model interpret the relationship between motion and the current state (context). Furthermore, unlike [46], proposed network maintains a gated recurrent unit (GRU) [78], which directly

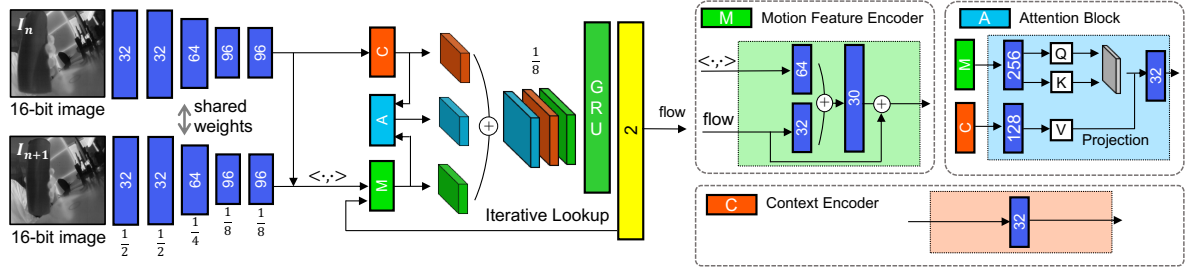


Fig. 4.4: Overview of the thermal optical flow network.

influences the motion feature encoder and assists the attention block in interpreting scene differences between normal and NUC-affected scenes. Additionally, for fast inference, the number of iterative lookups is limited to 3 during the inference stage, compared to 10 during the training stage.

II. Training Details

The training strategy for proposed optical flow network, using a self-supervised approach, is shown in Fig. 4.5. This research utilizes two types of datasets. First, multiple image pairs with self-labeled pseudo-ground truth are used, generated through homography and thin plate spline transformation \mathbf{H} , with noise augmentation shown in Fig. 4.3, similar to ThermalLANet. This provides supervisory signals and prevents the model from getting stuck in local minima. Second, consecutive image frames without ground truth are used to improve adaptability to natural movements and occlusions, where an occlusion mask \mathbf{O}_1 is computed using a forward backward consistency check [79].

The loss function includes photometric loss L_{photo} , smoothness loss L_{smooth} , self-supervision loss L_{self} , and L1 loss L_{L1} . The photometric loss is defined as

$$L_{\text{photo}} = \sum_{i=1}^H \sum_{j=1}^W \mathbf{O}_1 \odot \rho(\mathbf{I}_1, \omega(\mathbf{I}_2, \mathbf{F}_N)), \quad (4.10)$$

where \mathbf{F}_N is the predicted optical flow, ω is the warping function, \mathbf{I}_1 and \mathbf{I}_2 are the image pairs, and ρ represents the photometric error between the two images, as described in [80]. The smoothness loss is defined as

$$L_{\text{smooth}} = \sum_{i=1}^H \sum_{j=1}^W (\exp(-\|\mathbf{I}_x\|_2) \odot \|\mathbf{F}_{Nx}\|_2 + \exp(-\|\mathbf{I}_y\|_2) \odot \|\mathbf{F}_{Ny}\|_2 + \exp(-\|\mathbf{I}_{xy}\|_2) \odot \|\mathbf{F}_{Nxy}\|_2), \quad (4.11)$$

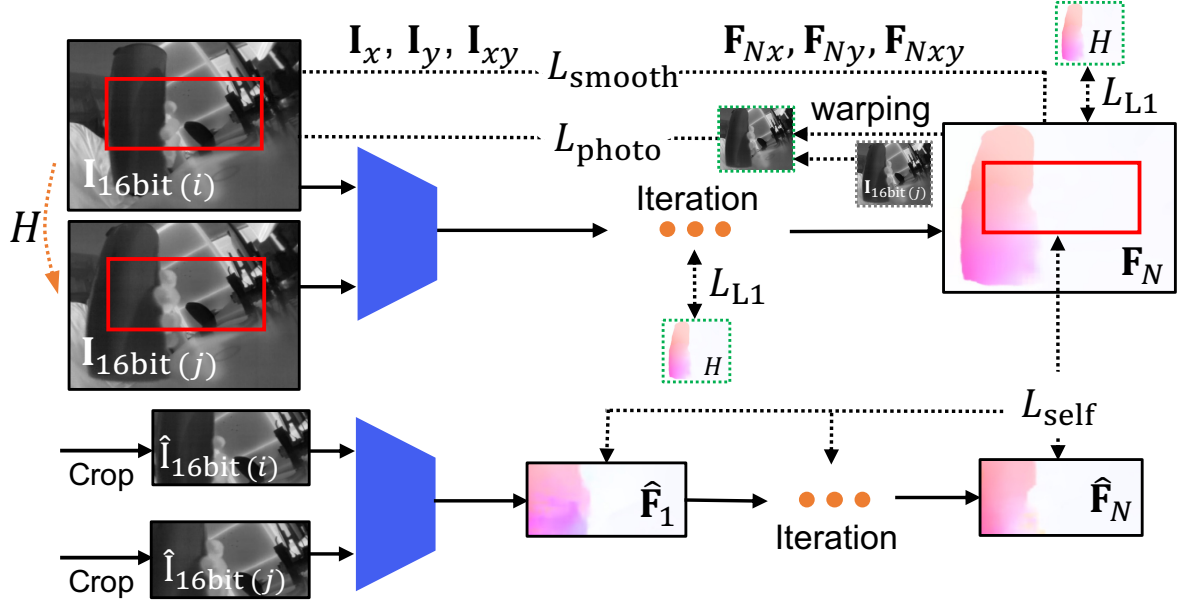


Fig. 4.5: Self-supervised training strategies for proposed thermal optical flow network.

where the subscripts x and y denote $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$, respectively. The self-supervision loss is defined as

$$L_{\text{self}} = \sum_{i=1}^N \sum_{j=1}^h \sum_{k=1}^w \gamma^{N-i} c(\hat{\mathbf{F}}_i, \mathbf{F}_N), \quad (4.12)$$

where N denotes the total iteration number, \odot is the Hadamard product, and γ denotes a discount factor, which is set to 0.8 in this section. Moreover, the function c is the generalized Charbonnier function [81], and $\hat{\mathbf{F}}_i$ represents the flow estimation from a randomly cropped and augmented image of size off (w, h) in each iteration [82, 83]. The L1 loss is defined in each iteration as

$$L_{\text{L1}} = \sum_{i=1}^N \gamma^{N-i} \|(\mathbf{F}_{\text{gt}} - \mathbf{F}_i)\|_1. \quad (4.13)$$

where the pseudo-ground truth flow \mathbf{F}_{gt} is calculated using \mathbf{H} , and \mathbf{F}_i denotes the predicted flow in each iteration. Note that the L1 loss is only used when training with the self-labeled data, as shown in Fig. 4.3.

4.2.3 Hybrid Feature Tracker

Due to the specific noise patterns in thermal images, the learning-based optical flow in these images tends to be unstable, especially in stationary scenes, and achieving subpixel accuracy is challenging. Meanwhile, optimization-based optical flow methods, such as 16-bit KLT [44], perform well even in noisy thermal images but struggle with rapid scene changes caused by frame drops due to NUC.

To address these issues, this section proposes a hybrid feature tracker that combines learning-based and optimization-based optical flow methods. In the hybrid tracker, features extracted using ThermalLANet are first tracked using the proposed learning-based thermal optical flow. Then, 16-bit KLT tracking is performed between two image patches, \mathbf{I}_n and \mathbf{I}_{n+1} , using the Newton–Raphson update [73]. The initial guess for this update is provided by the result of the learning-based thermal optical flow \mathbf{F}_1 , as follows:

$$\begin{cases} \mathbf{F}_0(\mathbf{x}) & \leftarrow \mathbf{F}_1(\mathbf{x}) \\ \mathbf{F}_{k+1}(\mathbf{x}) & \leftarrow \mathbf{F}_k(\mathbf{x}) + \boldsymbol{\mu}_k(\mathbf{x}), \end{cases} \quad (4.14)$$

where the update term $\boldsymbol{\mu}_k(\mathbf{x})$ is

$$\boldsymbol{\mu}_k(\mathbf{x}) = \frac{\sum_{\mathbf{x}} \mathbf{I}'_n(\mathbf{x})^2 \mathbf{I}'_n(\mathbf{x} + \mathbf{F}_k) \delta_k(\mathbf{I}_n, \mathbf{I}_{n+1})}{\sum_{\mathbf{x}} \mathbf{I}'_n(\mathbf{x})^2 \mathbf{I}'_n(\mathbf{x} + \mathbf{F}_k)^2}, \quad (4.15)$$

$$\delta_k(\mathbf{I}_n, \mathbf{I}_{n+1}) = \mathbf{I}_{n+1}(\mathbf{x}) - \mathbf{I}_n(\mathbf{x} + \mathbf{F}_k), \quad (4.16)$$

where 16-bit KLT is implemented using image pyramids to improve accuracy and robustness, as noted in [84]. By utilizing learning-based optical flow as the initial guess for KLT, proposed tracker can achieve subpixel accuracy and robustness while also effectively handling the NUC scenes.

4.2.4 Thermal Inertial Odometry Formulation

Using the tracked feature points and IMU, the k -th state $\mathbf{T}_k \in SE(3)$ can be estimated through tightly-coupled optimization. The sliding window optimization for thermal-inertial odometry is formulated by simultaneously minimizing both the reprojection residuals from consecutive thermal images and the IMU residuals from preintegration measurements. First, the reprojection residuals $\mathbf{r}_{P_i}(\mathbf{T}_k)$ can be formulated using the rigid transformation from the body frame to the camera frame $\mathbf{T}_c \in SE(3)$, the projection matrix of a camera $\pi_c : \mathbb{R}^3 \mapsto \mathbb{R}^2$, the tracked i -th

points in the k -th image frame $\mathbf{p}_i^k \in \mathbb{R}^2$, and the corresponding thermal landmark $\mathbf{l}_i \in \mathbb{R}^3$ in world frame as follows:

$$\mathbf{r}_{P_i}(\mathbf{T}_k) = s_i(\mathbf{p}_i^k - \pi_c(\mathbf{T}_c^{-1}\mathbf{T}_k^{-1}\mathbf{l}_i)), \quad (4.17)$$

where s_i is the confidence score extracted from ThermalLNet, which is used to weight each reprojection error according to its confidence. Additionally, the scale of each thermal landmark can be determined using the visual-inertial alignment strategies described in [14]. Second, the IMU residuals $\mathbf{r}_I(\mathbf{T}_k)$ to regulate the pose, velocity, and biases are formulated using IMU preintegration, which are defined in [17] as follows:

$$\mathbf{r}_I(\mathbf{T}_k) = [\mathbf{r}_{\Delta\mathbf{p}_k}, \mathbf{r}_{\Delta\mathbf{R}_k}, \mathbf{r}_{\Delta\mathbf{v}_k}, \mathbf{r}_{\Delta\mathbf{b}_k}], \quad (4.18)$$

where \mathbf{p} is the translation, \mathbf{R} is the rotation, \mathbf{v} is the velocity, and \mathbf{b} is the bias. With the IMU residuals and reprojection residuals, the sliding window optimization based on the MAP estimate to solve for the optimal state \mathbf{T}^* can be formulated as follows:

$$\mathbf{T}^* = \underset{\mathbf{T}}{\operatorname{argmin}} \sum_{k \in \mathcal{W}} (\|\mathbf{r}_I(\mathbf{T}_k)\|_2 + \sum_{i \in \mathcal{P}_k} \|\mathbf{r}_{P_i}(\mathbf{T}_k)\|_2) + \|\mathbf{r}_0\|_2, \quad (4.19)$$

where \mathcal{P} is a set of thermal features, \mathcal{W} is a set of frames in the sliding window, and \mathbf{r}_0 is a prior marginalized state.

4.3 Experiments

In this section, This research evaluated proposed system based on the performance of feature point detection, tracking, and state estimation. In the experiments, the proposed ThermalLANet and thermal optical flow network are trained using the FLIR thermal image train dataset [85], the VIVID++ driving dataset [86], the STheReO KAIST Valley dataset [87], the SubT tunnel and urban circuit dataset [50], and a custom dataset captured with the FLIR Boson+640. Note that the SubT dataset is used only for training due to missing intrinsic calibration. All networks were implemented using PyTorch and quantized with TensorRT for the inference stage. The training hyperparameters for ThermalLANet are unchanged from those in [72], while those for the proposed thermal optical network follow [83], except for the L1 loss defined in eq. (4.13), which follows the hyperparameters in [47]. All experiments, including network training and inference, were conducted on a laptop equipped with an Intel i9-12950HX CPU, an RTX-A4500 laptop GPU, and 64 GB RAM.

4.3.1 Evaluation on Feature Point Detection

To validate the accuracy and robustness of ThermalLANet, a quantitative evaluation was conducted. The image pairs for evaluation were obtained by applying a random homography transformation to proposed custom thermal dataset, as inspired by [44, 88]. The evaluation metrics include repeatability, localization error, homography estimation accuracy (H-1, H-3, H-5), matching score, model parameters. LANet and proposed method is trained with either 8-bit or 16-bit images.

Table 4.1: Comparison of feature point detection algorithms on the FLIR thermal image test dataset [85]. The image pairs are generated with a perspective transform with a random homography matrix. The metrics include repeatability, localization error, homography estimation accuracy, matching score, model parameters. LANet and proposed method is trained with either 8-bit or 16-bit images.

Methods	RE \uparrow	LE \downarrow	H-1 \uparrow	H-3 \uparrow	H-5 \uparrow	MS \uparrow	Param. \downarrow
ORB	0.611	1.159	0.001	0.033	0.101	0.212	-
SIFT	0.411	1.103	0.094	0.493	0.671	0.201	-
SuperPoint	0.577	1.058	0.094	0.510	0.703	0.473	1.30M
LANet (8-bit)	0.612	1.150	0.195	0.646	0.787	0.472	4.03M
LANet (16-bit)	0.643	0.984	0.279	0.704	0.827	0.545	4.03M
Proposed (8-bit)	0.602	1.030	0.194	0.642	0.785	0.490	0.398M
Proposed (16-bit)	0.641	0.983	0.282	0.694	0.822	0.546	0.398M

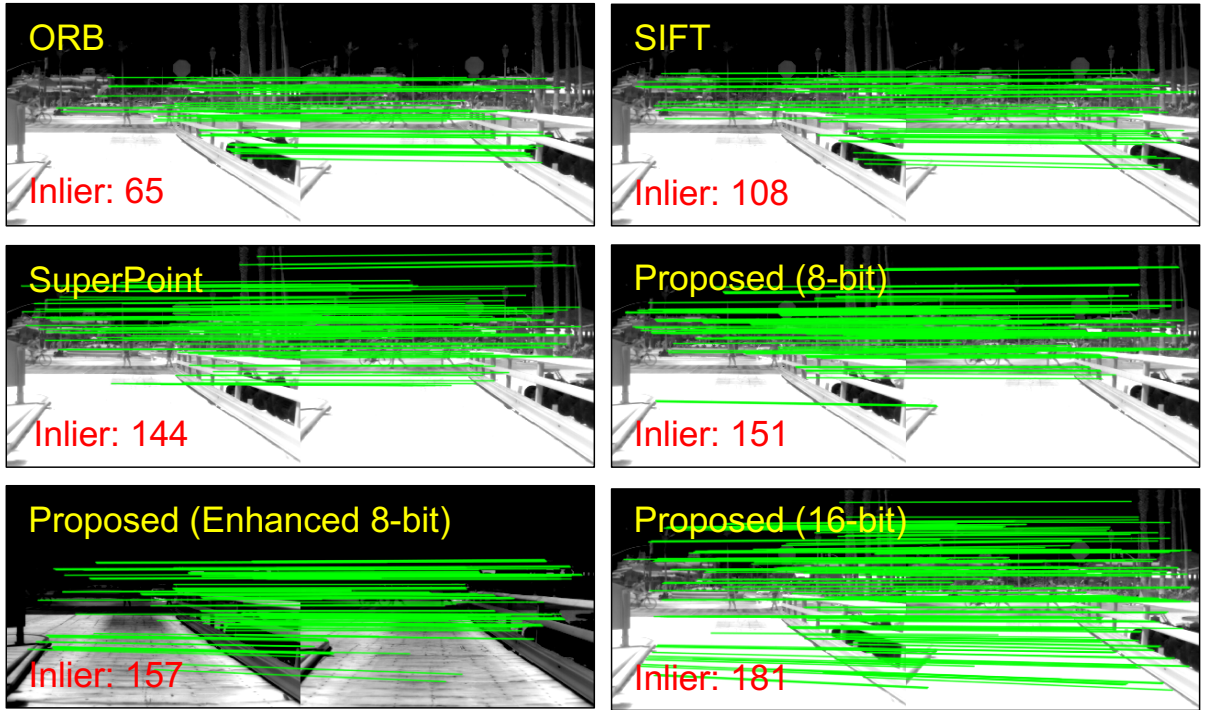


Fig. 4.6: Feature point extraction results and inliers when using different methods on the FLIR thermal image train dataset. A sequential image pair is prepared, with 300 points extracted and matched using each method, and inliers estimated through RANSAC.

H-5), and matching score, as in [45]. This research compared proposed method against the state-of-the-art methods [33, 45, 72, 89]. For [45], the non-maximum suppression radius was set to 8. In the overall evaluation, a maximum of 500 feature points were extracted from each image with a resolution of 640×512 using each method, and the homography matrix was estimated using RANSAC. The complete results are presented in Table 4.1. Overall, ThermalLANet outperforms the compared methods, with the fusion of gradient filters effectively reducing parameters while maintaining accuracy compared to LANet.

Furthermore, to evaluate the performance of ThermalLANet in sequential image frames with natural ego-motion and varying viewpoints, This research conducted an evaluation on the FLIR thermal validation dataset [85]. In this evaluation, a maximum of 300 feature points were extracted from each image with a resolution of 640×512 using each method, and RANSAC was applied to estimate the inliers. Moreover, ThermalLANet trained with 8-bit images was also included for comparison, alongside ORB, SIFT, and SuperPoint. For the 8-bit-based methods, the original 16-bit images were linearly rescaled using specific min/max values to produce 8-bit images. For the 8-bit-based LANet, two different 8-bit image pairs in each sequence

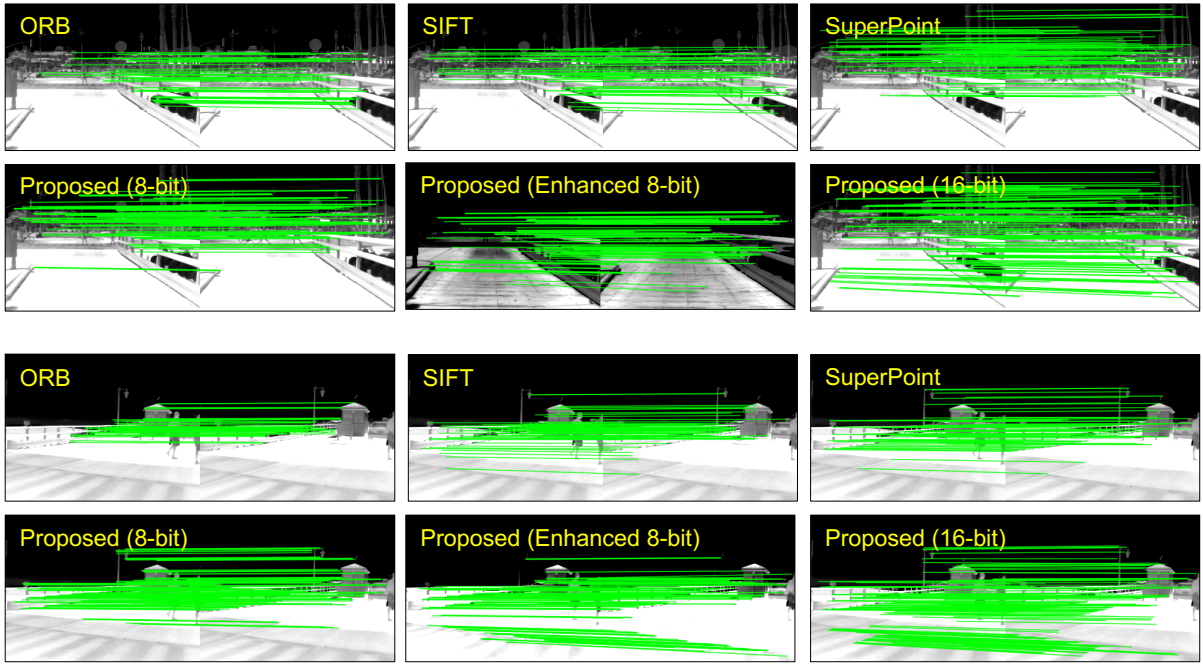


Fig. 4.7: Examples of feature point extraction results using different methods on the FLIR thermal image train dataset. A sequential image pair is prepared, with 300 points extracted and matched using each method, and inliers estimated through RANSAC.

are prepared, rescaled with different min/max values, to demonstrate the effectiveness of the proposed 16-bit-based training. The results are presented in Fig. 4.6 and Fig. 4.7. As can be observed, proposed method outperforms ORB, SIFT, and SuperPoint. Furthermore, the ablation study comparing the rescaled 8-bit and original 16-bit training strategies highlights that training with 16-bit images effectively addresses issues of low contrast and information loss, which are prevalent in rescaled 8-bit images and lead to extraction failures in 8-bit-based methods.

4.3.2 Evaluation on Feature Point Tracker

This research evaluated the effectiveness of the proposed hybrid tracker using two image pairs: one captured during a stationary phase and another during a phase with frame drops caused by NUC. As shown in Fig. 4.8, when the robot is stationary, the learning-based optical flow incorrectly indicates movement in the tracked features. This mistracking arises because the noise pattern in thermal images changes slightly from sequence to sequence, adversely affecting the tracking quality of the optical flow. Conversely, when using KLT, most features fail to be tracked in scenes with NUC. However, proposed hybrid method demonstrates robustness

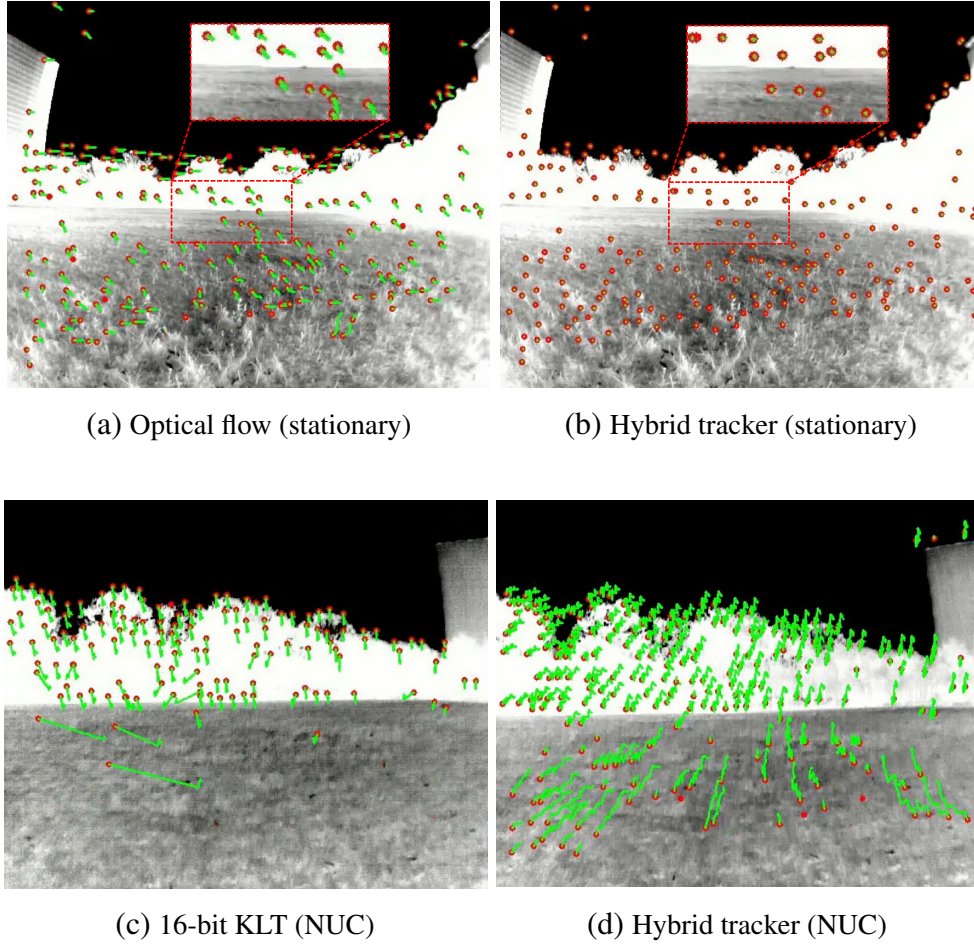


Fig. 4.8: Comparison results between the hybrid track and compared methods (16-bit KLT and thermal optical flow) in NUC and stationary scenes. The red points and green lines denote the tracked points and their past trajectory, respectively. The only inlier points estimated using RANSAC are described.

in handling both the NUC and stationary scenarios by integrating the strengths of KLT and learning-based optical flow.

Furthermore, to evaluate the performance of the hybrid tracker, a quantitative evaluation was conducted using state estimation performance on proposed real-world grassland dataset. Further details about this dataset can be found in Section 4.3.3. This research compared proposed method with LightGlue [90], RAFT [47], RAFT with 16-bit KLT, 16-bit KLT, proposed thermal optical flow network, and the hybrid tracker without attention. For RAFT, the public pre-trained weights were used. The results are presented in Table 4.2. As can be observed, proposed method outperforms the compared method.

Table 4.2: Comparison of ATE [m] when using each feature tracking algorithms in grassland dataset. ThermalLANet is used for feature point extraction, with 500 points. “AN” denotes the attention network.

Methods	Grassland01	Grassland02	Grassland03
LightGlue	fail	fail	fail
RAFT	fail	fail	fail
16-bit KLT	2.70	0.74	0.71
RAFT with 16-bit KLT	3.00	0.98	1.00
Thermal optical flow	15.8	0.94	0.95
Hybrid Tracker (w/o AN)	1.73	0.83	0.83
Hybrid Tracker (with AN)	1.53	0.62	0.70

4.3.3 Evaluation on State Estimation Performance

I. Public dataset

In this section, This research evaluates proposed full system, Self-TIO, using a public dataset. For comparison, this research includes state-of-the-art visual and thermal odometry methods such as ORB SLAM3 [16], ROVTIO [91], OpenVINS [15], SuperPoint [45] + OpenVINS, and Deep-TIO [48]. Note that, except for ROVTIO, which uses 16-bit raw thermal images, the other compared methods use rescaled 8-bit thermal images. For all the methods, a maximum of 500 feature points are used for tracking. The evaluation metric is the root mean squared error (RMSE) of the absolute trajectory error (ATE).

The evaluation results on the VIVID++ dataset are shown in Table 4.3. Overall, proposed method outperforms the state-of-the-art methods on the VIVID++ dataset. In this dataset, the compared methods often experience significant drift or failure due to frame drops caused by NUC and low contrast in the images. Notably, ORB SLAM3, which relies on a grid-based ORB algorithm, struggles in indoor environments where the scarcity of objects leads to a limited number of feature points and potentially low-textured features. In outdoor environments, although relatively more feature points exist than the indoor dataset, tracking becomes challenging due to harsh noise and low-contrast areas such as building edges and road curbs. This is why methods combining learning-based feature extraction with KLT tracking often fail in these scenarios, except in the Night1 sequence, where features from classical methods remain relatively stable. In contrast, proposed method shows significant advantages in both indoor and outdoor environments by leveraging the combination of learning-based feature extraction and tracking.

Table 4.3: Comparison of ATE [m] in the VIVID++ dataset [86].

	ORB SLAM3	ROVTIO	DeepTIO	VINS	SuperPoint + VINS	Self-TIO
Indoor-robust						
Dark	0.22	0.17	0.41	0.091	0.11	0.085
Varying	fail	0.24	0.46	0.10	0.16	0.080
Local	0.33	0.11	0.39	0.086	0.12	0.076
Global	fail	0.24	0.64	fail	fail	0.15
Outdoor-robust						
Day1	8.20	4.10	7.58	1.73	2.91	1.13
Day2	5.10	6.72	9.03	10.6	fail	1.79
Night1	1.01	4.71	7.80	1.24	fail	1.36
Night2	7.74	5.62	7.75	3.43	4.83	2.25

Table 4.4: Comparison of ATE [%] in the SNU sequence of the STheReO dataset [87].

Sequence	Length [m]	ORB SLAM3	ROVTIO	DeepTIO	VINS	SuperPoint + VINS	Self-TIO
SNU morning	7920	fail	fail	fail	1.00	fail	0.87
SNU afternoon	7920	fail	fail	fail	3.44	2.20	1.23
SNU evening	7920	fail	fail	fail	1.60	1.25	0.52

The evaluation results on the STheReO SNU dataset are shown in Table 4.4. The STheReO dataset presents significant challenges due to not only the inherent issues of thermal cameras, but also the relatively high velocities and many linear-only motions. These factors cause failures in most of the compared methods. In contrast, proposed method demonstrates robustness in these challenging conditions because it can extract feature points even in low-contrast scenes and track them effectively even during NUC events. As a result, proposed system is successful in the state estimation for all the given scenarios and outperforms the compared methods. Examples of the trajectory in public scenes are presented in Fig. 4.9.

Moreover, in public datasets, this research measures the execution time of each network used in the proposed method. Overall, ThermalLANet takes an average of 7.9 ms, and the thermal optical flow network takes an average of 14.0 ms per frame. Note that all the networks in proposed method are quantized using TensorRT. In total, proposed TIO system achieves an average of 32.2 FPS.

II. Real-world Dataset

The proposed method was evaluated on real-world datasets, as shown in Fig. 4.10. The datasets were captured using a Clearpath Jackal UGV equipped with a FLIR Boson+640 thermal camera, a MicroStrain 3DM-GX5-25 IMU, and a Kubota RTK-GNSS unit. The Grassland dataset

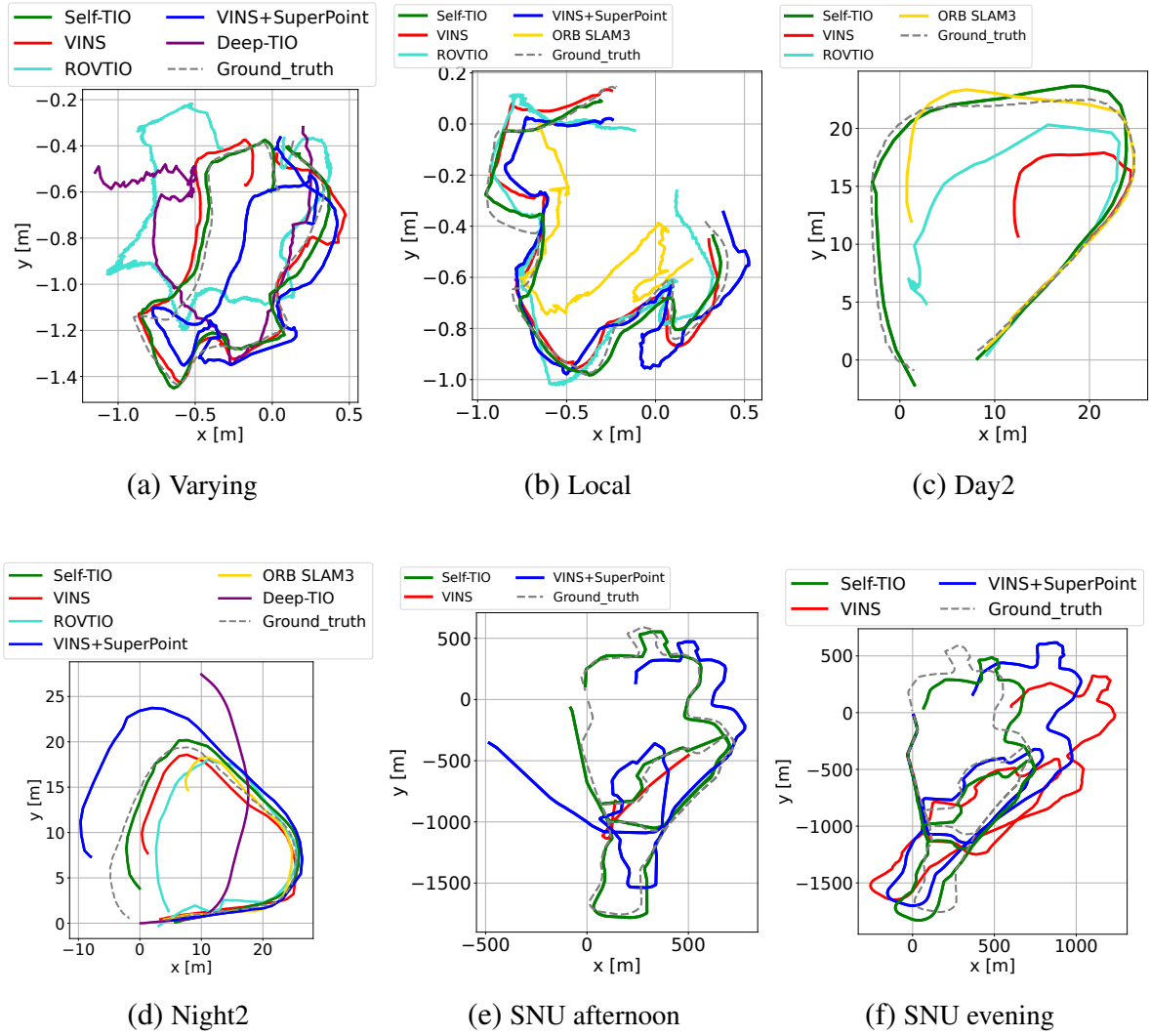


Fig. 4.9: Trajectories in public dataset (VIVID++ [86], STheReO SNU [87]), compared with VINS, VINS+SuperPoint, and ROVTIO.

includes scenes with NUC and significant vibrations due to the grassland environment. The Street dataset features linear-only motion and was captured primarily on paved roads with minimal vibration. Both datasets involve aggressive motions, including fast rotations, and were recorded during midnight, as shown in Fig. 4.10. For comparison, along with state-of-the-art visual and thermal odometry methods [15, 16, 48, 91], methods for ablation studies of ThermalLANet and the hybrid tracker are also implemented. Moreover, the combination of the 8-bit-based ThermalLANet and 8-bit-based hybrid tracker is also implemented to demonstrate the effectiveness of using raw 16-bit radiometric images. For all the methods, a maximum

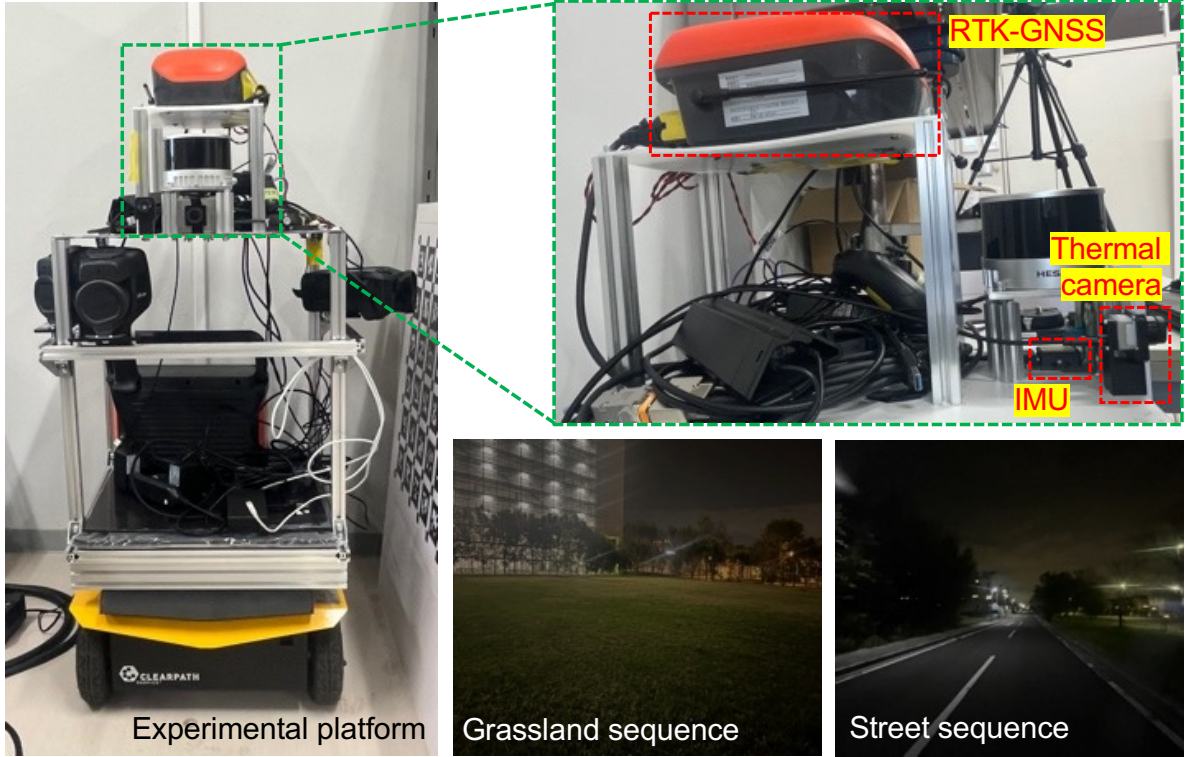


Fig. 4.10: Experimental platform and environment snapshots. RTK-GNSS is only used for ground truth.

of 500 feature point are used for tracking. The evaluation metrics used are RMSE and the maximum error of ATE.

The complete results are detailed in Table 4.5, and the overall trajectories are illustrated in Fig. 4.11. Proposed method consistently outperforms the compared approaches, which struggle with aggressive motion, NUC, and low texture, leading to feature tracking failures. In the Grassland01 and Grassland03 sequences, classical feature extraction methods, such as those used in ORB SLAM3, VINS, and FAST+hybrid tracker, face challenges in extracting feature points. In the Street02 sequence, feature tracking is particularly challenging due to the presence of road curbs and low texture of the paved road, which cause significant drift in KLT-based methods such as VINS and ThermalLANet+KLT. Finally, the results of the ablation study underscore the effectiveness of combining ThermalLANet and the hybrid tracker, along with training strategies using radiometric images, demonstrating their critical role in achieving accurate and robust state estimation in thermal-inertial odometry.

Table 4.5: Comparison of ATE [m] in proposed real-world dataset.

Sequence	Length [m]	Previous method										Ablation study					
		ORB SLAM3		ROV-TIO		DepthTIO		VINS		TL ₁₆ + KLT ₁₆		FAST ₈ + HT ₁₆		Self-TIO ₈		Self-TIO ₁₆	
		MAX	RMSE	MAX	RMSE	MAX	RMSE	MAX	RMSE	MAX	RMSE	MAX	RMSE	MAX	RMSE	MAX	RMSE
Grassland01	132	fail	fail	17.0	7.07	49.2	23.8	16.3	8.73	6.95	3.04	13.2	5.49	9.39	4.30	3.14	1.53
Grassland02	43.0	fail	fail	4.36	2.59	22.3	12.9	fail	fail	2.03	0.76	2.06	0.46	1.97	0.70	1.75	0.62
Grassland03	79.9	fail	fail	4.46	1.47	21.1	11.4	fail	fail	1.47	0.75	12.1	7.90	1.64	1.01	2.00	0.70
Street01	71.9	30.6	15.0	49.7	15.0	27.1	10.5	fail	fail	11.1	5.14	11.8	5.47	7.10	4.37	2.06	1.07
Street02	93.2	20.1	6.79	62.0	24.9	28.0	10.1	40.5	14.0	96.4	31.7	18.7	6.00	38.6	15.7	6.77	2.04

“TL” and “HT” denote ThermalANet and hybrid tracker, respectively. The subscripts indicate the bit size of the thermal images used for training (8-bit or 16-bit).

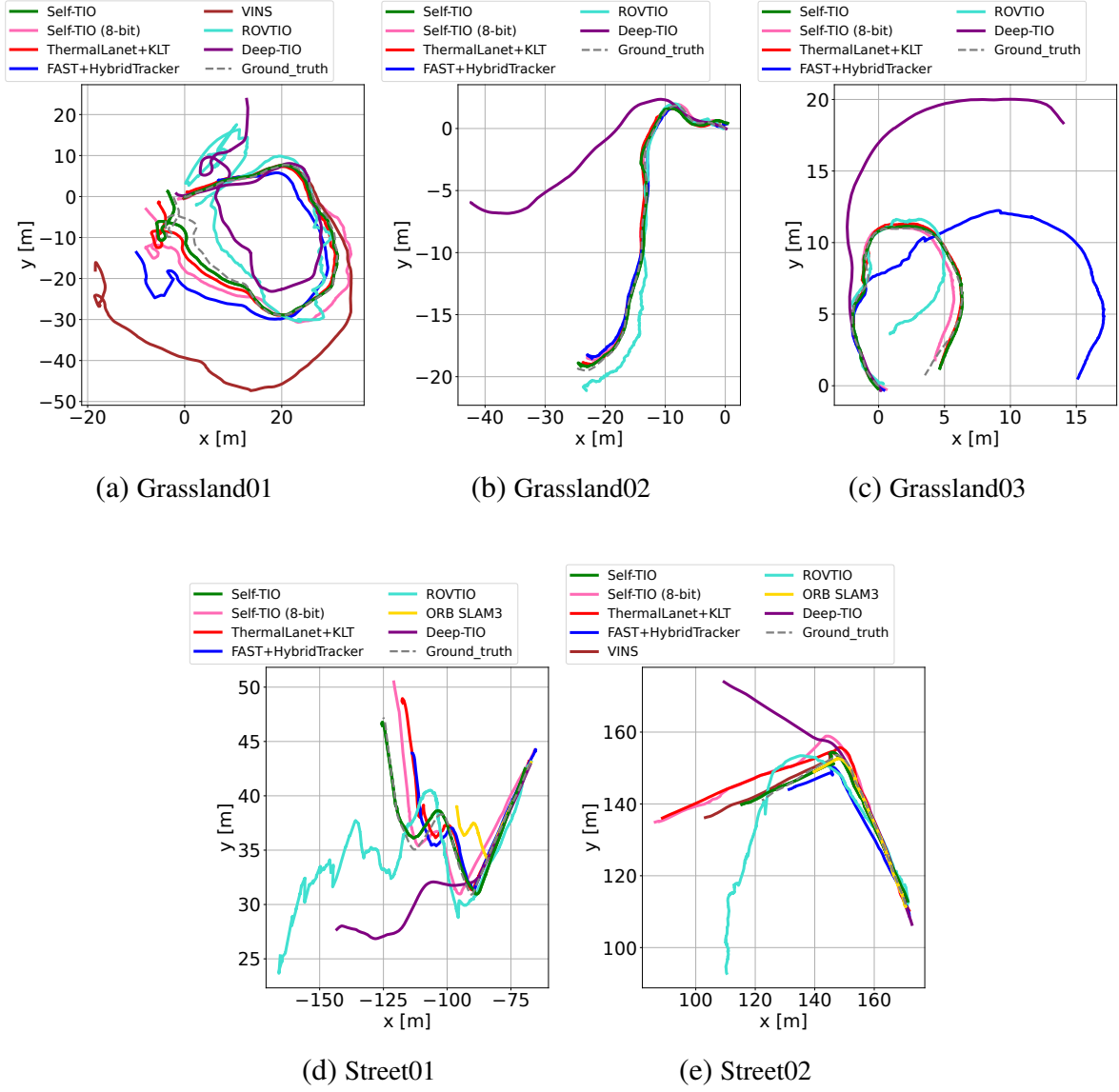


Fig. 4.11: Trajectories in real-world dataset. The proposed method is compared with VINS, ThermalLanet+KLT, FAST+hybrid tracker, ROVTIO, Deep-TIO, ORB SLAM3, Self-TIO trained with 8-bit images.

4.4 Summary

In this chapter, a novel thermal-inertial odometry system, Self-TIO was proposed. To address the inherent limitations of thermal cameras, Self-TIO introduced a fully self-supervised 16-bit feature extractor, ThermalLANet, and a feature tracker, thermal optical flow. Furthermore, Self-TIO introduced a hybrid tracker, which fuses a learning-based optical flow with a 16-bit KLT. Experimental results showed that proposed method outperformed the state-of-the-art visual and thermal inertial odometry methods and efficiently addresses scenes containing NUC and low-contrast.

Chapter 5

TC-LTIO: Tightly-coupled LiDAR Thermal Inertial Odometry for LiDAR and Visual Odometry Degraded Environments

5.1	Introduction: TC-LTIO	62
5.2	Tightly-coupled LiDAR Thermal Inertial Odometry	64
	5.2.1 System Overview	64
	5.2.2 Feature Extraction	64
	5.2.3 Visual Feature Tracking	65
	5.2.4 Tightly-Coupled optimization formulation.....	68
	5.2.5 Zero Velocity Detection	69
5.3	Experiments	71
5.4	Summary	75

5.1 Introduction: TC-LTIO

As mentioned in Section 2, SLAM is typically categorized into LiDAR and visual SLAM, depending on the primary sensor used. LiDAR SLAM offers high accuracy and robustness in scenarios involving aggressive motion and complexly structured environments, owing to its capability to directly measure distances between objects and the sensor using multiple rays [5]. However, as LiDAR SLAM performs the localization by matching each structural scan, LiDAR SLAM can degenerate in structure-less scenes such as tunnels, vast planes, and corridors [10]. On the other hand, visual SLAM, harnessing textural information from RGB images, can work in structure-less environments due to its reliance on texture-based features, which can be extracted even in scenes lacking clear structural elements [9]. However, visual SLAM has weaknesses in scale estimation and is susceptible to rapid changes in lighting conditions.

To address the limitations of both LiDAR and visual SLAM, various LiDAR visual SLAM methods, which simultaneously integrate information from both LiDAR and visual sensors, have been proposed [18, 22, 92]. However, as most of these methods rely on loosely-coupled manner (inter-system fusion) [18, 22], the failure in either system can lead to overall SLAM failure. To tackle the weakness of the loosely-coupled manner, tightly-coupled methods (inter-feature fusion) have been proposed [92]. These methods can effectively deal with structurally and visually degraded scenes by simultaneously incorporating LiDAR and visual features into the maximum a posteriori (MAP) formulation. On the other hand, thermal SLAM methods such as [46], which utilize thermal infrared cameras capable of capturing temperature variations in a scene, are proposed to address the limitations of visual SLAM arising from lighting conditions. However, these methods still encounter challenges such as scale uncertainty and handling aggressive motion. Consequently, LiDAR-thermal SLAM systems have been proposed. Nevertheless, existing LiDAR-thermal SLAM systems predominantly focus on either a loosely-coupled approach [93] or a visual feature-based solution [94]. To address these limitations, a tightly-coupled LiDAR thermal inertial odometry for LiDAR and visual odometry degraded environments (TC-LTIO) is proposed in this research. The main contributions of proposed work are as follows:

- **Tightly-coupled sensor fusion:** To address SLAM degeneration, this research integrates LiDAR edge/planar and thermal point features in a tightly-coupled manner. Proposed method shows better accuracy and robustness in its localization when compared to other LiDAR/visual/LiDAR-visual SLAMs in various experiments.

- **Learning-based optical flow for thermal point features:** To accurately and robustly track point features derived from consecutive thermal images, this research leverages a learning-based optical flow approach. This method excels in handling low-contrast scenarios and aggressive motion. Proposed learning-based optical flow is designed for real-time operation and trained not only on thermal images but also on synthetic thermal images generated from RGB images.
- **Thermal point rejection based on velocity:** The performance of learning-based optical flow can be compromised by zero velocity situations caused by low contrast and harsh noise in thermal images. Therefore, in proposed method, thermal features are excluded from the entire optimization process when zero velocity is detected based on the values of LiDAR odometry and IMU.

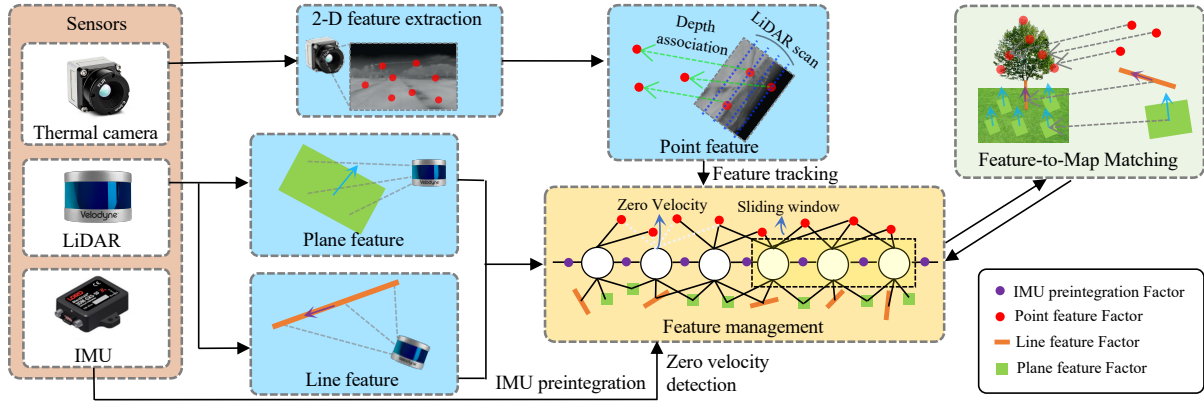


Fig. 5.1: System overview of TC-LTIO.

5.2 Tightly-coupled LiDAR Thermal Inertial Odometry

5.2.1 System Overview

An overview of the proposed SLAM system, TC-LTIO, is shown in Fig. 5.1. Three sensors, a thermal camera, LiDAR, and an inertial measurement unit (IMU), are used for TC-LTIO. Proposed framework is composed of three key subsystems: feature extraction, feature management, and feature-to-map matching. In the feature extraction subsystem, point, line, and plane features are extracted. After the point features are extracted from the thermal image, the depth of the point features is assigned using a LiDAR scan. Line features and edge features are extracted from LiDAR scans based on structural smoothness. In the feature management subsystem, entire features are managed with a factor graph structure [95]. When zero velocity is detected using IMU and LiDAR data, point features are excluded from the graph to prevent the inclusion of degenerate visual features. In the feature-to-map matching subsystem, a cost function between extracted features and corresponding map features is formulated based on a sliding window approach. The final pose and map are updated by minimizing the cost function.

5.2.2 Feature Extraction

I. LiDAR Feature Extraction

Line and plane features are extracted as following [5], which relies on the structural smoothness of each LiDAR scan. The structural smoothness $\sigma^{(m,n)}$ of an n -th point along the m -th scan

line can be defined as

$$\sigma^{(m,n)} = \frac{1}{|\mathcal{S}^{(m,n)}|} \sum_{\mathbf{p}^{(m,j)} \in \mathcal{S}^{(m,n)}} (\|\mathbf{p}^{(m,j)} - \mathbf{p}^{(m,n)}\|), \quad (5.1)$$

where $\mathcal{S}^{(m,n)}$ is the set of adjacent points (denoted as $\mathbf{p}^{(m,j)}$) from the same scan line, and $|\mathcal{S}^{(m,n)}|$ denotes the number of points in $\mathcal{S}^{(m,n)}$.

II. Visual Feature Extraction

Firstly, as thermal images are usually derived as 14-bit or 16-bit scale, 16/14-bit thermal images is chaged to 8-bit images using clipping method with median value as follows

$$\begin{cases} \bar{\mathbf{I}}(x) = 0 & \text{if } \mathbf{I}(x) < I_{\min} \vee \mathbf{I}(x) > I_{\max} \\ \bar{\mathbf{I}}(x) = \frac{\mathbf{I}(x) - I_{\min}}{I_{\max} - I_{\min}} & \text{otherwise} \end{cases}, \quad (5.2)$$

where $\mathbf{I}(x)$ is original pixel value from the image and $\bar{\mathbf{I}}(x)$ is remapped pixel value. Also I_{\max} , I_{\min} are defined as follows:

$$\begin{cases} I_{\max} &= m + \delta_{\max} \\ I_{\min} &= m - \delta_{\min} \end{cases}, \quad (5.3)$$

where m is a median value of the entire image pixel, and $\delta_{\max/\min}$ is a threshold of the clipping. Note that while raw 14/16-bit data is used in Section 4, 8-bit conversion is used in this section for simplicity, particularly in the feature point extraction process, to enable simultaneous processing of thermal images and LiDAR scans while ensuring real-time execution.

To extract visual feature points from the 8-bit thermal image, this research adapts grid-based FAST as proposed in [15]. This grid-based approach ensures that feature points are evenly distributed across all regions of the image. Subsequently, the 2-D visual feature point obtains its depth from LiDAR scan. For depth association, this research completely follows [92], which projects LiDAR scans onto the 2-D image plane for association. Note that proposed system only uses depth-associated visual features among all visual features because depth-associated features are less affected by the robot's ego motion compared to triangulated features from sequential frames, which can easily diverge in scenarios involving linear-only robot motion.

5.2.3 Visual Feature Tracking

Optimization-based feature tracking methods such as [73] and feature matching-based methods such as [33, 89] are prone to failure, especially due to low contrast problem of the thermal image. To robustly track extracted visual features, this research proposes learning-based optical

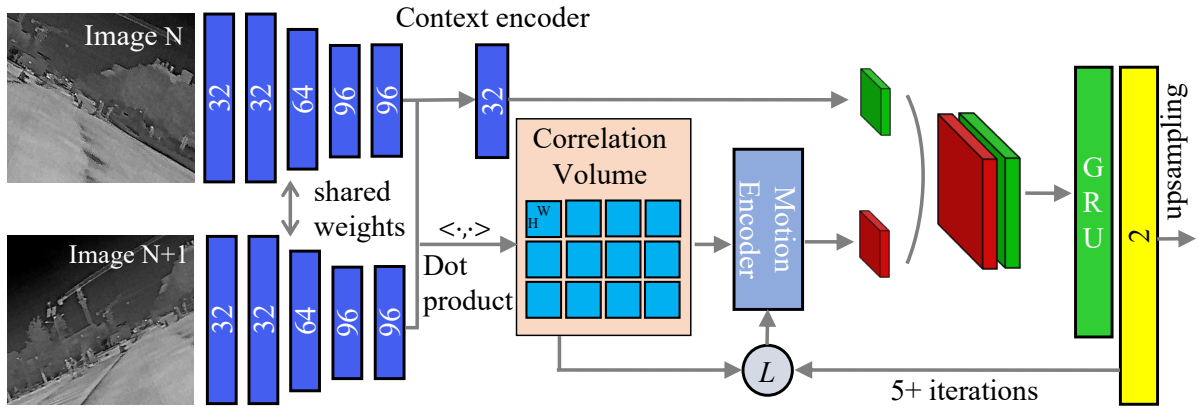


Fig. 5.2: Proposed optical flow network.

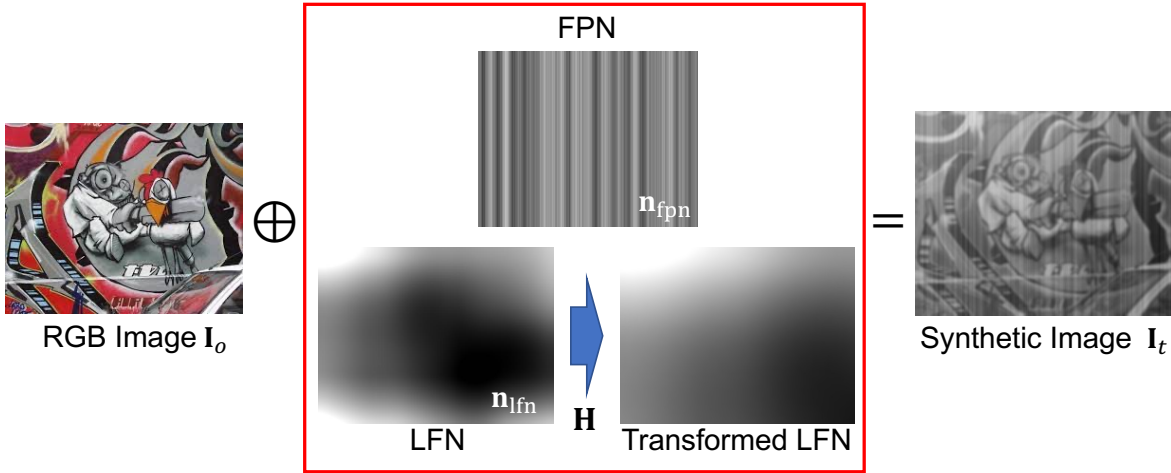


Fig. 5.3: Method to generate synthetic thermal image from RGB image.

flow for thermal image based on [47] and its lightweight variant [46]. The proposed optical flow network is illustrated in Fig. 5.2, leveraging not only a lightweight CNN feature encoder and 3-D correlation volume for efficient computation but also a GRU-based update operator [78] for the accurate estimation of residual flow.

To effectively train proposed optical flow network for the thermal domain with a focus on noise suppression, this research introduces novel training strategies. Initially, synthetic thermal optical flow datasets originating from RGB image optical flow datasets such as [96] and [97] are used for training. As referred to in [70], the differences between thermal and RGB images

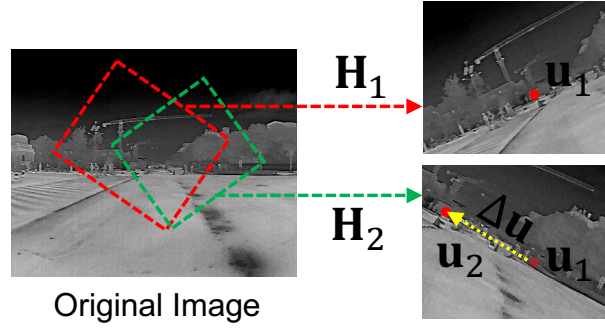


Fig. 5.4: Labeling for self-supervised learning from single thermal image.

mainly stem from two types of noise, low frequency noise (LFN): randomized noise and fixed pattern noise (FPN): striped pattern noise. Therefore, to generate synthetic thermal image \mathbf{I}_t , the target RGB image is firstly translated to grayscale image \mathbf{I}_o , and, LFN noise \mathbf{n}_{lfn} and FPN noise \mathbf{n}_{fpn} pattern are added to the grayscale image as shown in Fig. 5.3. For FPN noise, an image by rearranging various vertical stripes is randomly generated. For LFN noise, the prepared image is randomly transformed with random homography transformation \mathbf{H} . Consequently, the formulation of the synthetic thermal image is as follows:

$$\mathbf{I}_t = w_1 \mathbf{I}_o + w_2 \mathbf{n}_{\text{fpn}} + (1 - w_1 - w_2) \mathbf{H} \mathbf{n}_{\text{lfn}}, \quad (5.4)$$

where w_1 and w_2 are weighted parameters. Moreover, to deal with the aggressive motion of the sensor, the network is trained using image pairs with a certain number of frame intervals, which are also randomized between 0 to 10 frames.

Secondly, proposed network is trained with real-world thermal images to enhance robustness against actual real-world noise. Given the scarcity of optical flow datasets based on thermal images, this research adopt a self-supervised training strategy. Inspired by [45], the warping technique is used to make image pairs from a single thermal image. Therefore, using the warped image pair $\mathbf{I}_1, \mathbf{I}_2$ and corresponding homography matrix $\mathbf{H}_1, \mathbf{H}_2$, the optical flow label $\Delta \mathbf{u}$ can be made as follows:

$$\Delta \mathbf{u} = \mathbf{u}_2 - \mathbf{u}_1 = \mathbf{u}_2 - \mathbf{H}_2 \mathbf{H}_1^{-1} \mathbf{u}_1, \quad (5.5)$$

where $\mathbf{u} = [u, v, 1]^T$ is a pixel location in each warped image. The detail is shown in Fig. 5.4.

5.2.4 Tightly-Coupled optimization formulation

I. Maximum a Posteriori Estimation

TC-LTIO consists of factor-based integration to estimate the optimal robot state. The entire factor graph encompasses IMU preintegration factors, LiDAR feature factors (line/plane), and visual feature factors (point) with an initial guess of each state. As the robot state can be optimized by solving the MAP problem according to the measurements of each sensor, the current optimal robot state \mathbf{X}^* is solved using the least squares minimization problem, where costs from each factor within a sliding window are minimized, as depicted below:

$$\begin{aligned} \mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmin}} & \sum_{i \in \mathcal{K}} (\|\mathbf{r}_{I_i}\|_{\Sigma_{I_i}}^2 + \sum_{j \in \mathcal{P}} \|\mathbf{r}_{p(i,j)}\|_{\Sigma_p}^2 \\ & + \sum_{j \in \mathcal{L}} \|\mathbf{r}_{l(i,j)}\|_{\Sigma_l}^2 + \sum_{j \in \mathcal{D}} \|\mathbf{r}_{d(i,j)}\|_{\Sigma_d}^2) + \|\mathbf{r}_0\|_{\Sigma_0}^2, \end{aligned} \quad (5.6)$$

where $\|\mathbf{x}\|_{\Sigma}^2 = \mathbf{x}\Sigma^{-1}\mathbf{x}^\top$ and \mathcal{K} is a set of all the keyframe indices within the sliding window. Moreover, \mathcal{P} , \mathcal{L} , and \mathcal{D} denote set of plane, line, and point features, where I , p , l , and d are indices about measurements of IMU preintegration, plane, line, and point features. Note that the factor graph is optimized with a fixed lag smoothing method based on iSAM2 [62].

However, the tracking quality of the learning-based optical flow deteriorates in scenarios where the camera experiences zero velocity. Consequently, eq. (5.6) cannot be appropriately determined during zero velocity situations due to the presence of mistracked visual features. To mitigate this issue, when zero velocity of the camera is detected, point features are excluded from the MAP fusion process to prevent the propagation of mistracking results to the entire system.

II. IMU Preintegration Factors

IMU can effectively deal with both aggressive motion and short-term degeneration in terms of localization. To integrate an IMU factor into the factor graph, this research follows the IMU preintegration method proposed in [17]. Therefore, The IMU preintegration residual \mathbf{r}_{I_i} in eq. (5.6) can be defined as:

$$\mathbf{r}_{I_i} = [\mathbf{r}_{\Delta \mathbf{p}_i}, \mathbf{r}_{\Delta \mathbf{v}_i}, \mathbf{r}_{\Delta \mathbf{R}_i}, \mathbf{r}_{\mathbf{b}_i^a}, \mathbf{r}_{\mathbf{b}_i^g}]^\top, \quad (5.7)$$

where $\mathbf{r}_{\Delta \mathbf{p}_i}$, $\mathbf{r}_{\Delta \mathbf{v}_i}$, and $\mathbf{r}_{\Delta \mathbf{R}_i}$ are position, linear velocity, and orientation residual between prior and present keyframe. $\mathbf{r}_{\mathbf{b}_i^a}$ and $\mathbf{r}_{\mathbf{b}_i^g}$ denote bias residual of the accelerometer and gyroscope on the consecutive keyframe. Further details are described in [17].

III. LiDAR Feature Factors

LiDAR plane and line features, which are extracted according to the smoothness value of eq. (5.1), are tracked using a k-d tree-based nearest neighbor search after the coordinates of consecutive frames are adjusted using IMU preintegration results. Then, mistracked features are removed using RANSAC [98]. The criteria for RANSAC is the angle between two direction vectors for the line features and two normal vectors for plane features.

The residual of the plane and line features can be formulated with feature-to-map matching cost as follows:

$$\mathbf{r}_{l(i,j)} = \frac{(\mathbf{p}_{(i,j)}^l - \hat{\mathbf{p}}_1^l) \times (\mathbf{p}_{(i,j)}^l - \hat{\mathbf{p}}_2^l)}{\|\hat{\mathbf{p}}_1^l - \hat{\mathbf{p}}_2^l\|}, \quad (5.8)$$

$$\mathbf{r}_{p(i,j)} = \frac{(\mathbf{p}_{(i,j)}^p - \hat{\mathbf{p}}_1^p)((\hat{\mathbf{p}}_1^p - \hat{\mathbf{p}}_2^p) \times (\hat{\mathbf{p}}_1^p - \hat{\mathbf{p}}_3^p))}{\|(\hat{\mathbf{p}}_1^p - \hat{\mathbf{p}}_2^p) \times (\hat{\mathbf{p}}_1^p - \hat{\mathbf{p}}_3^p)\|}, \quad (5.9)$$

given a line feature $\mathbf{p}_{(i,j)}^l \in \mathbb{R}^3$ and the corresponding nearest line feature $\hat{\mathbf{p}}_1^l$ and second one $\hat{\mathbf{p}}_2^l$ on the map. Moreover, given a plane feature $\mathbf{p}_{(i,j)}^p \in \mathbb{R}^3$ and the corresponding nearest plane feature $\hat{\mathbf{p}}_1^p$, second one $\hat{\mathbf{p}}_2^p$, and third one $\hat{\mathbf{p}}_3^p$.

IV. Visual Feature Factors

The tracked point features $\mathbf{p}_j^d \in \mathbb{R}^3$, acquired through visual feature point and depth association, are stored as the map (visual landmarks) in the world frame. Subsequently, \mathbf{p}_j^d is projected onto each image plane within the sliding window to compute the reprojection error. Therefore, the residual of the point features can be formulated as follows:

$$\mathbf{r}_{d(i,j)} = \mathbf{u}_i - \pi(\mathbf{T}_i \mathbf{p}_j^d), \quad (5.10)$$

where $\mathbf{u}_i \in \mathbb{R}^2$ is a tracked visual feature on i -th camera plane. $\pi : \mathbb{R}^3 \mapsto \mathbb{R}^2$ is a projection matrix of a camera. $\mathbf{T}_i \in SE(3)$ is a transformation matrix from the world to i -th camera frame.

5.2.5 Zero Velocity Detection

Zero velocity detection is critical to the proposed system, particularly because visual features, tracked using learning-based optical flow, are prone to mistracking during zero velocity conditions. To accurately detect the zero velocity state, this research simultaneously utilizes LiDAR feature residuals and IMU preintegration results, which are independent of visual

features. Leveraging Eqs. (5.9), and (5.7), the positional or rotational differences between the first keyframe and the last keyframe within the sliding window can be computed. Consequently, zero velocity detection can be formulated as follows:

$$[\sum_{i \in \mathcal{K}} \|\mathbf{r}_{\Delta \mathbf{P}_i}\|^2, \sum_{i \in \mathcal{K}} \|\mathbf{r}_{\Delta \mathbf{R}_i}\|^2, \sum_{i \in \mathcal{K}} \sum_{j \in \mathcal{P}} \|\mathbf{r}_{p(i,j)}\|^2] < [\delta_{\Delta \mathbf{P}}, \delta_{\Delta \mathbf{R}}, \delta_p], \quad (5.11)$$

where $\delta_{\Delta \mathbf{P}}$, $\delta_{\Delta \mathbf{R}}$, and δ_p are heuristic threshold for each criterion. In eq. (5.11), if all the criteria are smaller than their respective thresholds, I consider the current state as a zero velocity state.

5.3 Experiments

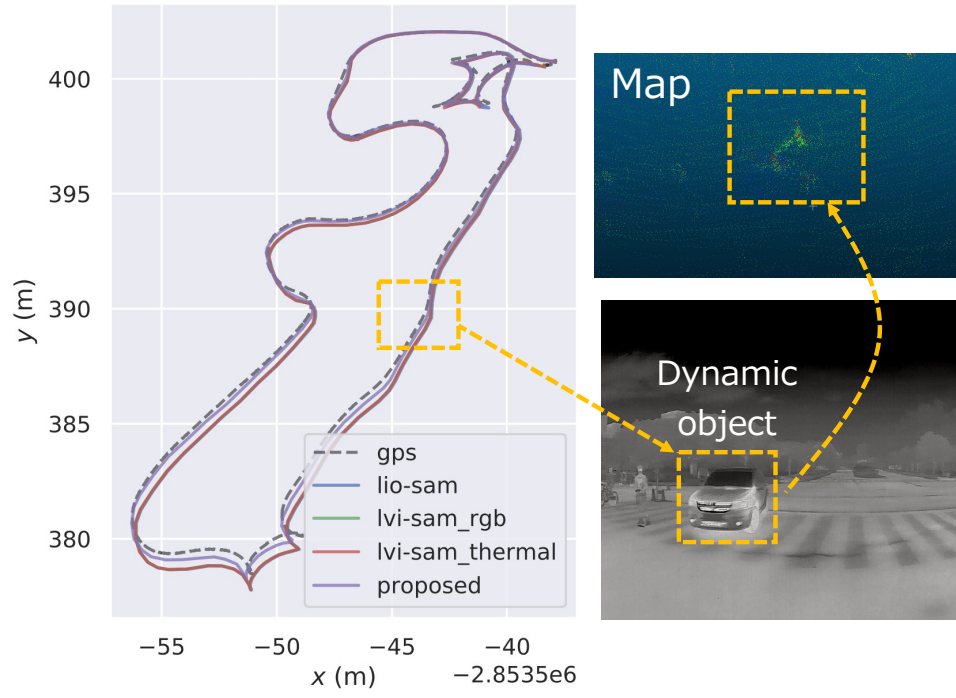
To evaluate the proposed method, this research compares proposed method with the state-of-the-art LiDAR-inertial SLAM, LIO-SAM [5] and LiDAR-visual-inertial SLAM, LVI-SAM [18]. In case of LVI-SAM, the visual odometry submodule is executed using RGB camera (denoted as LVI-SAM_{RGB}) or thermal camera (denoted as LVI-SAM_{Thermal}). M2DGR dataset [99] is used for evaluation, as it contains LiDAR, RGB camera, thermal camera, and IMU data with ground truth captured using RTK-GNSS. In M2DGR, the gate01/02 and street01/03 datasets were captured at night, leading to visual degeneration. The error matrix is measured using absolute trajectory errors (ATE), which represents the positional difference between the ground truth points and the estimated trajectories. As shown in Table 5.1, the proposed method consistently achieves lower ATE compared to the other methods, except in the case of gate03. Furthermore, the proposed method achieves an average processing speed of 23 FPS from feature extraction to optimization, due to the implementation of a fixed-lag smoothing method. In contrast, LIO-SAM and LVI-SAM achieve approximately 10 to 17 FPS, as they rely on LOAM-based [5] scan-to-map optimization.

In gate01, the dynamic object caused the drift of both LIO-SAM and LVI-SAM as shown in Fig. 5.5(a). This drift occurs because the LOAM-based scan-map matching algorithm matches all line and plane features with the map. In contrast, the proposed method demonstrated robustness to dynamic objects by sequentially tracking each feature and removing outliers with the RANSAC algorithm. In street01, where both corridor-like structures and aggressive motion degraded both LiDAR and visual sensors as shown in Fig. 5.5(b), LIO-SAM and LVI-SAM exhibited significant drift. Conversely, the proposed method produced accurate localization results by leveraging a learning-based feature tracker for robustness in visually-degraded situations and tightly-coupled sensor fusion for robustness in LiDAR-degraded situations. In gate03, although proposed method showed slightly lower accuracy compared to LVI-SAM

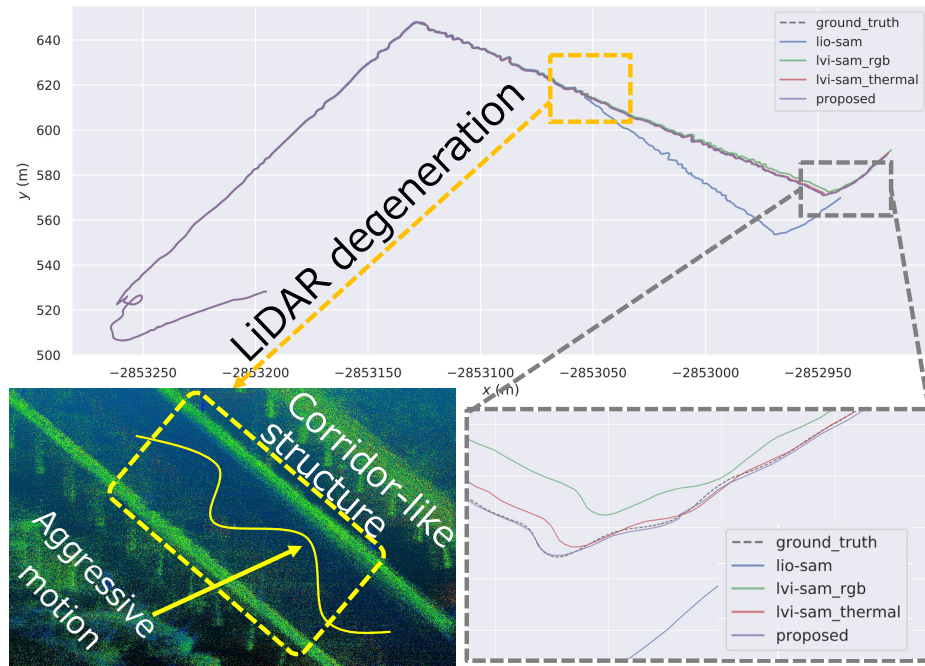
Table 5.1: Comparison of ATE [m] on M2DGR dataset.

Dataset	gate01		gate02		gate03		street01		street02		street03	
	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE	Max	RMSE
LIO-SAM	0.66	0.32	4.23	2.97	0.34	0.1126	30.0	10.9	16.5	7.22	0.42	0.15
LVI-SAM _{RGB}	0.64	0.31	4.01	2.85	0.30	0.1123	3.52	1.21	16.3	7.04	0.44	0.15
LVI-SAM _{Thermal}	0.64	0.31	3.99	2.83	0.26	0.1120	1.73	0.61	16.3	7.10	0.43	0.15
Proposed	0.51	0.24	3.62	2.58	0.34	0.1130	1.16	0.49	13.3	6.39	0.37	0.13

due to the absence of LiDAR/visual degeneration in the environment, the proposed method still outperformed LVI-SAM in terms of computational efficiency. This advantage stems from the tight-coupling, which manages all features with a graph and eliminates the need to simultaneously run both visual and LiDAR odometry nodes. The resulting maps, depicted in Fig. 5.6, showcase the 3-D mapping potential of proposed method.

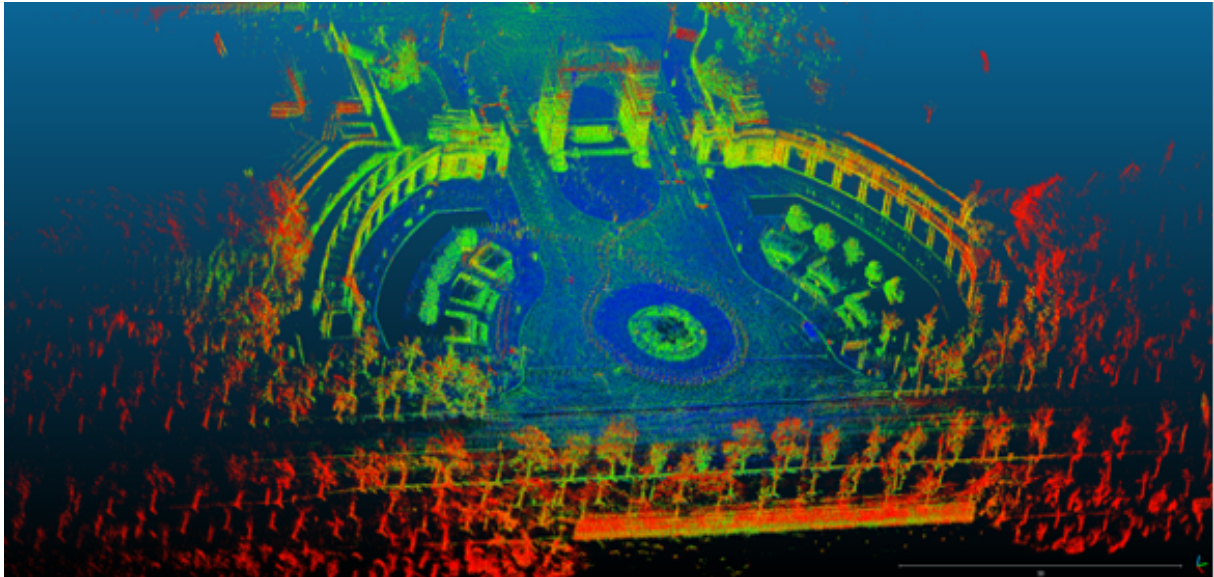


(a) gate01

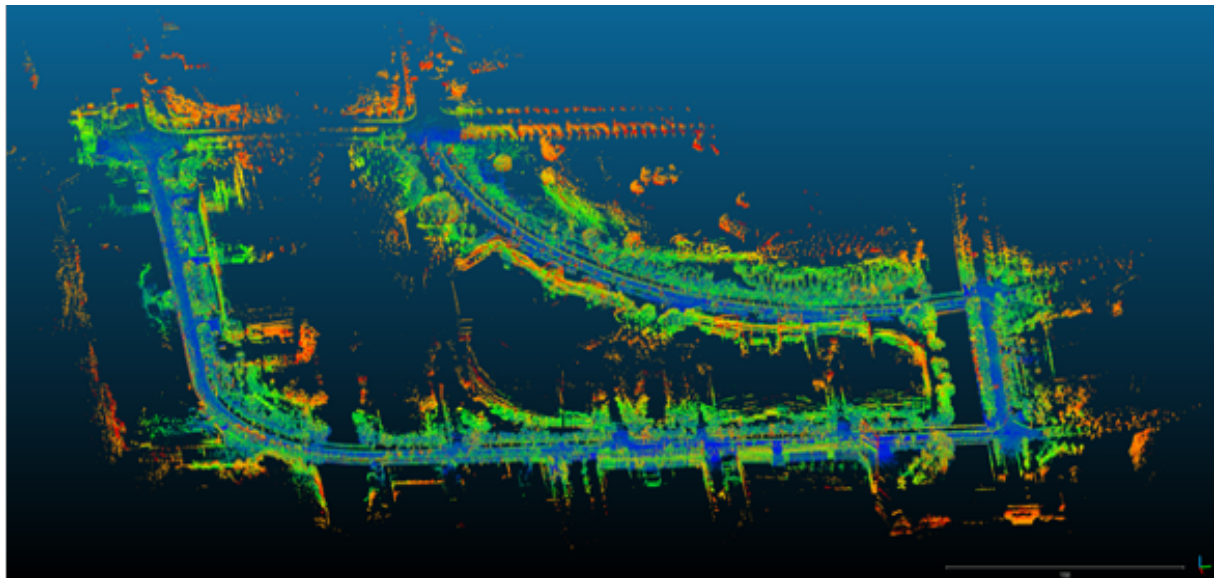


(b) street01

Fig. 5.5: Experimental result on gate01 and street01.



(a) gate02



(b) street02

Fig. 5.6: Resulting map in gate02 and street02. The point cloud is colored according to LiDAR intensity.

5.4 Summary

In this chapter, TC-LTIO: a tightly-coupled LiDAR thermal inertial odometry for LiDAR and visual odometry degraded environments was proposed, designed specifically to address the degradation of LiDAR odometry in structure-less scenes and visual odometry in harsh lighting conditions. To integrate the measurements of a LiDAR, thermal camera, and an IMU, factor graph-based tightly-coupled approach was employed. To construct the factor graph accurately, point features extracted from thermal images was tracked using a learning-based optical flow method implemented with a lightweight network trained on synthetic thermal images generated from RGB data. Moreover, to deal with mistracked point features caused by learning-based optical flow during zero velocity, zero velocity detection was employed based on the results of LiDAR odometry and IMU preintegration. The proposed method was evaluated on a public dataset containing scenarios with degradation in both LiDAR and visual odometry, including structure-less scans, aggressive motion, and dynamic objects. In experimental evaluations, TC-LTIO demonstrated high robustness and accuracy in such environments compared to state-of-the-art methods.

Chapter 6

Conclusion

6.1	Summary	78
6.2	Future Work	80

6.1 Summary

Localization and mapping in structurally and visually degraded environments remain critical challenges for SLAM systems. Conventional SLAM approaches struggle with degeneration in harsh environments, highlighting the necessity for robust, adaptive solutions. To address these issues, this thesis presented three innovative methods: Switch-SLAM, Self-TIO, and TC-LTIO, solving the degeneration problem of each sensor and contributing to the field of multi-sensor SLAM for extreme conditions.

In Chapter 2, after outlining the current challenges of LiDAR and visual SLAM degeneration, the main objective of this research is defined: to develop a LiDAR-thermal-inertial SLAM system robust in structurally and visually degenerate environments.

In Chapter 3, the conventional SLAM system was reviewed, covering traditional LiDAR, visual, and thermal SLAM approaches, along with their limitations in degenerate scenarios. Existing not only single but also multi sensor-based solutions have limitations in handling structural, visual degeneration, and challenges specific to thermal imaging, such as noise and frame drops. This Chapter highlighted the need for methods to tackle these issues effectively.

In Chapter 4, the proposed Switch-SLAM was introduced. This method uses a novel switching structure to alternate between LiDAR and visual odometry based on degeneracy detection. Key contributions include:

- A switching node for selecting optimal odometry results, ensuring consistent and robust pose estimation even in degenerate environments.
- Non-heuristic degeneracy detection using Chi-squared test-based thresholds, enhancing adaptability to diverse scenarios.
- Comprehensive experimental validation, demonstrating superior performance compared to state-of-the-art LiDAR-visual-inertial SLAM systems in challenging environments, effectively fulfilling our objective of accuracy and robustness.

In Chapter 5, Self-TIO was proposed to address visual degradation and the limitations of thermal imaging in SLAM. By introducing a self-supervised feature tracker tailored for 16-bit thermal images, the proposed method overcame challenges related to texturelessness, frame drops caused by NUC, and 8-bit conversion. The main contributions are:

- A ThermalLANet that maintains lightweight architecture while ensuring accuracy via gradient filters, grounded in self-supervised feature point extraction.

- A hybrid feature tracker that combines a learning-based optical flow network and an optimization-based KLT tracker, ensuring reliable tracking in thermal data.
- 16-bit level learning strategies to retain critical details in thermal images, significantly enhancing extraction and tracking robustness and accuracy.

In Chapter 6, TC-LTIO extended the insights from the Chapters 4 and 5 into a tightly-coupled LiDAR-thermal-inertial SLAM system. The proposed method addressed structural and visual degeneracy by combining thermal and LiDAR modalities with inertial measurements, tightly. Key contributions include:

- Tightly-coupled optimization framework integrating multi-modal sensor data, enabling robust pose estimation in complex and challenging environments.
- Zero-velocity detection and update, blocking the instability of optical flow networks for feature tracking in stationary scenes, thereby enhancing accuracy.
- Comprehensive experimental validation in real-world dataset, demonstrating TC-LTIO as a pioneering approach for environments previously unsuitable for current LiDAR-visual-inertial SLAM system, such as corridor structure in nighttime.

6.2 Future Work

While this thesis addresses significant challenges in structurally and visually degraded environments through the development of Switch-SLAM, Self-TIO, and TC-LTIO, critical issues remain in advancing SLAM systems for extreme and degraded environments. First, the proposed systems assume that each sensor is precisely synchronized at the hardware level. Second, the map representation in the proposed systems is limited to points and presented in a sparse manner, which cannot capture structural, textural, or semantic information. Third, adaptability to multi-robot systems is not considered, even though such systems could be highly effective in extreme environments by enabling exploration through collaborative robots. For these reasons, future research directions should include field experiments in diverse environments, such as large-scale, complex, foggy, and long-term settings, and focus on the following three domains:

- I. Suggesting continuous multi-sensor SLAM systems that eliminate the need for precise sensor synchronization.
- II. Developing dense map representation methods with learning-based approaches such as NeRF and Gaussian Splatting.
- III. Designing lightweight, scalable, and fully decentralized multi-robot SLAM systems.

I. Continuous SLAM

Continuous SLAM (C-SLAM) is an emerging paradigm for seamlessly integrating data from asynchronous, multi-modal sensor systems. Unlike traditional discrete-time SLAM, which estimates poses at discrete intervals, C-SLAM utilizes continuous models such as B- and Z-Splines. While C-SLAM does not require rigid synchronization among sensors, it heavily relies on nonlinear least squares optimization [18, 51], which is computationally expensive and limits scalability. To solve this issue, relaxation techniques, as well presented in [100], have been introduced to simplify the nonlinear least squares optimization problem. However, relaxation remains limited in large-scale and long-term environments, where the optimization residuals grow exponentially, significantly reducing computational efficiency. Continuous-time Gaussian belief propagation (GBP) [101, 102], a fully distributed, message-passing-based optimization method, has shown potential for application in C-SLAM systems. Recently, combinations of C-SLAM and GBP have been proposed by Hug *et al.* [103], demonstrating improved inference times compared to relaxation-based optimization. However, these

approaches are still limited in their applicability to large-scale and long-term environments in the real world. In future work, efforts should focus on improving the scalability and robustness of C-SLAM systems in large-scale and long-term environments, particularly within the context of multi-sensor SLAM. This would enable the proposed multi-sensor systems to operate without complicated setups and with minimal complexity.

II. Map representation

Traditional visual SLAM systems have limitations with discrete surface representations, such as point [8, 14], line [104], surfel [105], and mesh [106]. These representations cannot appropriately account for both structural and textural information in the resulting 3D map. Additionally, accurately estimating geometries in areas that are not directly observed remains a significant challenge. Recently, novel map representation methods, such as neural radiance fields (NeRF) [107] and 3D Gaussian splatting (3DGS) [108], have been proposed. These methods focus on representations for high-fidelity view synthesis, combined with learned models for capturing geometric fields. NeRF uses a multilayer perceptron (MLP) to map query points in 3D space along occupancy or color with the learned model, enabling accurate scene representation. However, NeRF is computationally inefficient when compared to classical methods like photogrammetry [109]. Therefore, sparse keyframe selection techniques are required for real-time performance, as demonstrated in NeRF-based SLAM systems [110, 111]. In contrast to NeRF, which usually uses voxel grids, 3DGS uses differentiable 3D Gaussian distributions that contain information such as location, color, and uncertainty, and renders this information with backpropagation. As a result, 3DGS achieves real-time performance, unlike NeRF, which is limited by volumetric ray sampling. Recently, SLAM systems using 3DGS have been proposed [112, 113], focusing on keyframe selection and map refinement techniques required for insufficiently observed scenes. These 3DGS-based SLAM systems demonstrate better computational efficiency and accuracy compared to NeRF-based SLAM systems. Note that current NeRF- or 3DGS-based SLAM systems have only been tested in indoor environments, as depth estimation outdoors is often unstable and the viewpoints are relatively monotonous. This makes the entire mapping process difficult to constrain in 3D space. Therefore, as future research, the integration of semantic information [114] and depth estimation [115] into NeRF and 3DGS should be explored to better constrain mapping processes, even in extreme and outdoor environments. This work would enhance the proposed visual or thermal SLAM system by incorporating spatial multi-modalities directly into the map, thereby improving both the diversity of information and the overall map quality.

III. Multi-robot SLAM

To explore extreme environments, multi-robot missions are inevitable. For multi-robot missions, a multi-robot SLAM system, which needs to be decentralized and lightweight, should be constructed. As noted in the Continuous SLAM subsection, current multi-robot SLAM systems also rely on (i) relaxation for nonlinear least squares optimization [116] and (ii) GBP-based distributed optimization [117] to solve the complicated pose graph optimization problem in a robust and lightweight manner. The work of Tian *et al.* [116] focused on a visual multi-robot SLAM application using relaxation techniques [118] for distributed pose graph optimization. This work also includes outlier rejection for loop closure, mesh map refinement, and real-world experiments of the full SLAM system. On the other hand, the work of McGann *et al.* [117] focused on the theoretical evaluation of GBP in the context of pose graph optimization, but it does not account for loop closure outliers, adaptability toward visual SLAM systems, and reliability in real-world scenarios. As future research, GBP-based multi-robot SLAM, building on the proposed LiDAR-thermal-inertial SLAM system from this work, should be developed with a focus on enhancing loop closure handling and ensuring robustness in real-world environments.

Acknowledgements

I would like to express my heartfelt gratitude to everyone who supported me throughout my master's journey. First and foremost, I extend my deepest and best appreciation to my supervisor, **Professor Atsushi Yamashita**, for his invaluable guidance, insightful feedback, and unwavering encouragement. This thesis would not have been possible without his support. I am also deeply grateful to **Professor Qi An** for his profound insights during our group meetings, critical feedback, and exceptional support during field experiments, all of which enriched my research with perspectives from a different field of study. I would like to express my sincere thanks to **Professor Hajime Asama** for his valuable advice and thought-provoking discussions during the first year of my master's degree. I am also thankful to **Dr. Ren Komatsu**, currently affiliated with Mujin Inc. and formerly at the University of Tokyo, for his direct mentorship, insightful feedback, and unwavering encouragement during the first year of my master's degree. I also wish to extend my special thanks to **Mr. Taisei Ando**, whom I had the privilege of mentoring. I am deeply grateful for his dedicated support and assistance during the field experiments. My gratitude extends to the current and former members of the Yamashita-An-Hamada Laboratory, as well as the lab secretaries, for their support and collaboration throughout my research journey. Also thanks to **Mr. Masaki Chino**, **Mr. Sanghyuk Lee**, and **Ms. Rukiye Aydin** for checking my thesis. Moreover, I would like to express my gratitude to **Dr. Shinichi Warisawa** for his commitment as a sub-referee for this thesis. I would like to give special thanks to **Dr. Toshihiro Kitajima** and **Mr. Mitsuru Shinozaki** of Kubota Corporation for fostering the collaborative work between the University of Tokyo and Kubota, and for their invaluable insights and support during research and field experiments. This research was conducted under the university-corporate collaboration agreement between Kubota Corporation and the University of Tokyo. Additionally, I am deeply appreciative of the Rotary Yoneyama Memorial Foundation for providing me with a full scholarship during my master's studies. Their support has been instrumental in enabling me to pursue my academic goals. Lastly, I would like to thank my friends for their unwavering support and encouragement. Above all, I am profoundly grateful to my family for their unconditional support throughout this journey.

February 2025

Junwoon Lee

Reference

- [1] J. Lee, T. Ando, M. Shinozaki, T. Kitajima, Q. An, and A. Yamashita, “Tc-ltio: Tightly-coupled lidar thermal inertial odometry for lidar and visual odometry degraded environments,” in *Proceedings of the 24th International Conference on Control, Automation and Systems (ICCAS)*, 2024, pp. 655–660.
- [2] J. Lee, R. Komatsu, M. Shinozaki, T. Kitajima, H. Asama, Q. An, and A. Yamashita, “Switch-slam: Switching-based lidar-inertial-visual slam for degenerate environments,” *IEEE Robotics and Automation Letters*, vol. 9, no. 8, pp. 7270–7277, 2024.
- [3] J. Lee, T. Ando, M. Shinozaki, T. Kitajima, Q. An, and A. Yamashita, “Self-tio: Thermal-inertial odometry via self-supervised 16-bit feature extractor and tracker,” *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 1003–1010, 2025.
- [4] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, “Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping,” in *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5135–5142.
- [5] J. Zhang and S. Singh, “Low-drift and real-time lidar odometry and mapping,” *Autonomous Robots*, vol. 41, pp. 401–416, 2017.
- [6] T. Shan and B. Englot, “Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain,” in *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 4758–4765.
- [7] J. Engel, T. Schöps, and D. Cremers, “Lsd-slam: Large-scale direct monocular slam,” in *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, 2014, pp. 834–849.
- [8] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “Orb-slam: a versatile and accurate monocular slam system,” *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [9] J. Zhang, M. Kaess, and S. Singh, “On degeneracy of optimization-based state estimation problems,” in *Proceedings of the 2016 IEEE International Conference on Robotics and*

- Automation (ICRA)*, 2016, pp. 809–816.
- [10] T. Tuna, J. Nubert, Y. Nava, S. Khattak, and M. Hutter, “X-icp: Localizability-aware lidar registration for robust localization in extreme environments,” *IEEE Transactions on Robotics*, vol. 40, pp. 452–471, 2023.
 - [11] J. Lin, C. Zheng, W. Xu, and F. Zhang, “R²live: A robust, real-time, lidar-inertial-visual tightly-coupled state estimator and mapping,” *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7469–7476, 2021.
 - [12] W. Xu and F. Zhang, “Fast-lio: A fast, robust lidar-inertial odometry package by tightly-coupled iterated kalman filter,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3317–3324, 2021.
 - [13] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, “Fast-lio2: Fast direct lidar-inertial odometry,” *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2053–2073, 2022.
 - [14] T. Qin, P. Li, and S. Shen, “Vins-mono: A robust and versatile monocular visual-inertial state estimator,” *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.
 - [15] P. Geneva, K. Eickenhoff, W. Lee, Y. Yang, and G. Huang, “Opencvins: A research platform for visual-inertial estimation,” in *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4666–4672.
 - [16] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, “Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam,” *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.
 - [17] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, “On-manifold preintegration for real-time visual-inertial odometry,” *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, 2016.
 - [18] T. Shan, B. Englot, C. Ratti, and D. Rus, “Lvi-sam: Tightly-coupled lidar-visual-inertial odometry via smoothing and mapping,” in *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 5692–5698.
 - [19] J. Lin and F. Zhang, “R³live: A robust, real-time, rgb-colored, lidar-inertial-visual tightly-coupled state estimation and mapping package,” in *Proceedings of the 2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 10 672–10 678.
 - [20] T. Wen, Y. Fang, B. Lu, X. Zhang, and C. Tang, “Liver: A tightly coupled lidar-inertial-visual state estimator with high robustness for underground environments,” *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2399–2406, 2024.
 - [21] J. Zhang, M. Kaess, and S. Singh, “Real-time depth enhanced monocular odometry,” in *Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and*

-
- Systems (IROS)*, 2014, pp. 4973–4980.
- [22] J. Zhang and S. Singh, “Visual-lidar odometry and mapping: Low-drift, robust, and fast,” in *Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 2174–2181.
- [23] S. Zhao, H. Zhang, P. Wang, L. Nogueira, and S. Scherer, “Super odometry: Imu-centric lidar-visual-inertial estimator for challenging environments,” in *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 8729–8736.
- [24] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, vol. 120, pp. 122–125, 2000.
- [25] Pillow documentation. [Accessed 6-January-2025]. [Online]. Available: <https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf>
- [26] J. J. Leonard and H. F. Durrant-Whyte, “Mobile robot localization by tracking geometric beacons,” *IEEE Transactions on robotics and Automation*, vol. 7, no. 3, pp. 376–382, 1991.
- [27] F. Lu and E. Milios, “Robot pose estimation in unknown environments by matching 2d range scans,” *Journal of Intelligent and Robotic Systems*, vol. 18, no. 3, pp. 249–275, 1997.
- [28] A. Macario Barros, M. Michel, Y. Moline, G. Corre, and F. Carrel, “A comprehensive survey of visual slam algorithms,” *Robotics*, vol. 11, no. 1, p. 24, 2022.
- [29] J. Shi *et al.*, “Good features to track,” in *Proceedings of the 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.
- [30] A. J. Davison, “Real-time simultaneous localisation and mapping with a single camera,” in *Proceedings of the 9th International Conference on Computer Vision (ICCV)*, vol. 3, 2003, pp. 1403–1403.
- [31] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, “Monoslam: Real-time single camera slam,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.
- [32] G. Klein and D. Murray, “Parallel tracking and mapping for small ar workspaces,” in *Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 2007, pp. 225–234.
- [33] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in *Proceedings of the 2011 International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.

- [34] P. Besl and N. D. McKay, “A method for registration of 3-d shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [35] A. Segal, D. Haehnel, and S. Thrun, “Generalized-icp,” in *Proceedings of the Robotics Science and Systems (RSS)*, vol. 2, no. 4, 2009, p. 435.
- [36] F. Moosmann and C. Stiller, “Velodyne slam,” in *Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV)*, 2011, pp. 393–398.
- [37] E. Mendes, P. Koch, and S. Lacroix, “Icp-based pose-graph slam,” in *Proceedings of the 2016 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 2016, pp. 195–200.
- [38] R. Kuramachi, A. Ohsato, Y. Sasaki, and H. Mizoguchi, “G-icp slam: An odometry-free 3d mapping system with robust 6dof pose estimation,” in *Proceedings of the 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2015, pp. 176–181.
- [39] F. Sun, Y. Zhou, C. Li, and Y. Huang, “Research on active slam with fusion of monocular vision and laser range data,” in *Proceedings of the 8th World Congress on Intelligent Control and Automation*, 2010, pp. 6550–6554.
- [40] C. Forster, M. Pizzoli, and D. Scaramuzza, “Svo: Fast semi-direct monocular visual odometry,” in *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 15–22.
- [41] R. Ren, H. Fu, H. Xue, X. Li, X. Hu, and M. Wu, “Lidar-based robust localization for field autonomous vehicles in off-road environments,” *Journal of Field Robotics*, vol. 38, no. 8, pp. 1059–1077, 2021.
- [42] J. Nubert, E. Walther, S. Khattak, and M. Hutter, “Learning-based localizability estimation for robust lidar localization,” in *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 17–24.
- [43] Y. Wang, H. Chen, Y. Liu, and S. Zhang, “Edge-based monocular thermal-inertial odometry in visually degraded environments,” *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2078–2085, 2023.
- [44] S. Zhao, P. Wang, H. Zhang, Z. Fang, and S. Scherer, “Tp-tio: A robust thermal-inertial odometry with deep thermalpoint,” in *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 4505–4512.
- [45] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superpoint: Self-supervised interest point detection and description,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 224–236.
- [46] J. Jiang, X. Chen, W. Dai, Z. Gao, and Y. Zhang, “Thermal-inertial slam for the

-
- environments with challenging illumination,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8767–8774, 2022.
- [47] Z. Teed and J. Deng, “Raft: Recurrent all-pairs field transforms for optical flow,” in *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020, pp. 402–419.
- [48] M. R. U. Saputra, P. P. De Gusmao, C. X. Lu, Y. Almalioglu, S. Rosa, C. Chen, J. Wahlström, W. Wang, A. Markham, and N. Trigoni, “Deeptio: A deep thermal-inertial odometry with visual hallucination,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1672–1679, 2020.
- [49] J. Hoffman, S. Gupta, and T. Darrell, “Learning with side information through modality hallucination,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 826–834.
- [50] J. G. Rogers, J. M. Gregory, J. Fink, and E. Stump, “Test your slam! the sub-tunnel dataset and metric for mapping,” in *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 955–961.
- [51] C. Zheng, Q. Zhu, W. Xu, X. Liu, Q. Guo, and F. Zhang, “Fast-livo: Fast and tightly-coupled sparse-direct lidar-inertial-visual odometry,” in *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 4003–4009.
- [52] C. Qin, H. Ye, C. E. Pranata, J. Han, S. Zhang, and M. Liu, “Lins: A lidar-inertial state estimator for robust and efficient navigation,” in *Proceedings of the 2020 IEEE international Conference on Robotics and Automation (ICRA)*, 2020, pp. 8899–8906.
- [53] H.-A. Loeliger, “An introduction to factor graphs,” *IEEE Signal Processing Magazine*, vol. 21, no. 1, pp. 28–41, 2004.
- [54] P. D. Groves, “Principles of gnss, inertial, and multisensor integrated navigation systems, [book review],” *IEEE Aerospace and Electronic Systems Magazine*, vol. 30, no. 2, pp. 26–27, 2015.
- [55] D. W. Marquardt, “An algorithm for least-squares estimation of nonlinear parameters,” *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.
- [56] N. Gelfand, L. Ikemoto, S. Rusinkiewicz, and M. Levoy, “Geometrically stable sampling for the icp algorithm,” in *Proceedings of the Fourth International Conference on 3-D Digital Imaging and Modeling (3DIM)*, 2003, pp. 260–267.
- [57] K. Pearson, “On the criterion that a given system of deviations from the probable in the

- case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.
- [58] F. Benaych-Georges and A. Knowles, “Lectures on the local semicircle law for wigner matrices,” *arXiv preprint arXiv:1601.04055*, 2016.
 - [59] E. P. Wigner, “On the distribution of the roots of certain symmetric matrices,” *Annals of Mathematics*, vol. 67, no. 2, pp. 325–327, 1958.
 - [60] C. Hertzberg, R. Wagner, U. Frese, and L. Schröder, “Integrating generic sensor fusion algorithms with sound state representations through encapsulation of manifolds,” *Information Fusion*, vol. 14, no. 1, pp. 57–77, 2013.
 - [61] N. Demmel, D. Schubert, C. Sommer, D. Cremers, and V. Usenko, “Square root marginalization for sliding-window bundle adjustment,” in *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13 260–13 268.
 - [62] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, “isam2: Incremental smoothing and mapping using the bayes tree,” *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 216–235, 2012.
 - [63] G. Kim and A. Kim, “Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map,” in *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 4802–4809.
 - [64] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng *et al.*, “Ros: an open-source robot operating system,” in *Proceedings of the 2009 IEEE International Conference on Robotics and Automation Workshop on Open Source Software*, vol. 3, no. 3.2, 2009, p. 5.
 - [65] N. Koenig and A. Howard, “Design and use paradigms for gazebo, an open-source multi-robot simulator,” in *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, 2004, pp. 2149–2154.
 - [66] M. Tranzatto, M. Dharmadhikari, L. Bernreiter, M. Camurri, S. Khattak, F. Mascarich, P. Pfreundschuh, D. Wisth, S. Zimmermann, M. Kulkarni *et al.*, “Team cerberus wins the darpa subterranean challenge: Technical overview and lessons learned,” *Field Robotics*, vol. 4, pp. 249–312, 2024.
 - [67] S. Zhao, D. Singh, H. Sun, R. Jiang, Y. Gao, T. Wu, J. Karhade, C. Whittaker, I. Higgins, J. Xu *et al.*, “Subt-mrs dataset: Pushing slam towards all-weather environments,” in *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern*

-
- Recognition (CVPR)*, 2024, pp. 22 647–22 657.
- [68] S. B. Nashed, J. J. Park, R. Webster, and J. W. Durham, “Robust rank deficient slam,” in *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 6603–6608.
- [69] S. Khattak, C. Papachristos, and K. Alexis, “Keyframe-based thermal-inertial odometry,” *Journal of Field Robotics*, vol. 37, no. 4, pp. 552–579, 2020.
- [70] K. Liu, H. Chen, W. Bao, and J. Wang, “Thermal imaging spatial noise removal via deep image prior and step-variable total variation regularization,” *Infrared Physics & Technology*, vol. 134, p. 104888, 2023.
- [71] S. Wang, R. Clark, H. Wen, and N. Trigoni, “Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks,” in *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2043–2050.
- [72] C. Wang, G. Zhang, Z. Cheng, and W. Zhou, “Rethinking low-level features for interest point detection and description,” in *Proceedings of the 2022 Asian Conference on Computer Vision (ACCV)*, 2022, pp. 2059–2074.
- [73] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2, 1981, pp. 674–679.
- [74] A. Barroso-Laguna and K. Mikolajczyk, “Key. net: Keypoint detection by handcrafted and learned cnn filters revisited,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 698–711, 2022.
- [75] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.
- [76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, no. 11, 2017, pp. 6000–6010.
- [77] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, “Learning to estimate hidden motions with global motion aggregation,” in *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9772–9781.
- [78] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.

- [79] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in *Proceedings of the 8th European Conference on Computer Vision (ECCV)*, 2004, pp. 25–36.
- [80] S. Meister, J. Hur, and S. Roth, “Unflow: Unsupervised learning of optical flow with a bidirectional census loss,” in *Proceedings of the 2018 AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [81] D. Sun, S. Roth, and M. J. Black, “Secrets of optical flow estimation and their principles,” in *Proceedings of the 2010 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2432–2439.
- [82] P. Liu, I. King, M. R. Lyu, and J. Xu, “Ddflow: Learning optical flow with unlabeled data distillation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8770–8777.
- [83] A. Stone, D. Maurer, A. Ayvaci, A. Angelova, and R. Jonschkowski, “Smurf: Self-teaching multi-frame unsupervised raft with full-image warping,” in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3887–3896.
- [84] J.-Y. Bouguet, “Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm,” *Intel Corporation*, vol. 5, no. 1-10, p. 4, 2001.
- [85] Free teledyne flir thermal dataset for algorithm training. [Accessed 7-August-2024]. [Online]. Available: <https://www.flir.com/oem/adas/adas-dataset-form/>
- [86] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung, “Vivid++: Vision for visibility dataset,” *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6282–6289, 2022.
- [87] S. Yun, M. Jung, J. Kim, S. Jung, Y. Cho, M.-H. Jeon, G. Kim, and A. Kim, “Sthereo: Stereo thermal dataset for research in odometry and mapping,” in *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 3857–3864.
- [88] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, “Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5173–5182.
- [89] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the seventh IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 1999, pp. 1150–1157.
- [90] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, “Lightglue: Local feature matching at light speed,” in *Proceedings of the IEEE/CVF International Conference on Computer*

-
- Vision (ICCV)*, 2023, pp. 17 627–17 638.
- [91] H. D. Flemmen, “Rovtio: Robust visual thermal inertial odometry,” Master’s thesis, Norwegian University of Science and Technology, 2021.
- [92] D. Wisth, M. Camurri, S. Das, and M. Fallon, “Unified multi-modal landmark tracking for tightly coupled lidar-visual-inertial odometry,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1004–1011, 2021.
- [93] W. Chen, Y. Wang, H. Chen, and Y. Liu, “Eil-slam: Depth-enhanced edge-based infrared-lidar slam,” *Journal of Field Robotics*, vol. 39, no. 2, pp. 117–130, 2022.
- [94] Y.-S. Shin and A. Kim, “Sparse depth enhanced direct thermal-infrared slam beyond the visible spectrum,” *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2918–2925, 2019.
- [95] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, “Factor graphs and the sum-product algorithm,” *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.
- [96] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, “A naturalistic open source movie for optical flow evaluation,” in *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, 2012, pp. 611–625.
- [97] M. Menze and A. Geiger, “Object scene flow for autonomous vehicles,” in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3061–3070.
- [98] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [99] J. Yin, A. Li, T. Li, W. Yu, and D. Zou, “M2dgr: A multi-sensor and multi-scenario slam dataset for ground robots,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2266–2273, 2021.
- [100] G. Grisetti, C. Stachniss, S. Grzonka, W. Burgard *et al.*, “A tree parameterization for efficiently computing maximum likelihood maps using gradient descent,” in *Proceedings of the Robotics: Science and Systems III*, 2008, pp. 65–72.
- [101] J. Ortiz, T. Evans, and A. J. Davison, “A visual introduction to gaussian belief propagation,” *arXiv preprint arXiv:2107.02308*, 2021.
- [102] R. Murai, J. Ortiz, S. Saeedi, P. H. Kelly, and A. J. Davison, “A robot web for distributed many-device localisation,” *IEEE Transactions on Robotics*, vol. 40, pp. 12–138, 2023.
- [103] D. Hug, I. Alzugaray, and M. Chli, “Hyperion—a fast, versatile symbolic gaussian belief

- propagation framework for continuous-time slam,” in *Proceedings of the 18th European Conference on Computer Vision*, 2024, pp. 215–231.
- [104] Q. Fu, J. Wang, H. Yu, I. Ali, F. Guo, Y. He, and H. Zhang, “Pl-vins: Real-time monocular visual-inertial slam with point and line features,” *arXiv preprint arXiv:2009.07462*, 2020.
- [105] T. Schops, T. Sattler, and M. Pollefeys, “Bad slam: Bundle adjusted direct rgb-d slam,” in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 134–144.
- [106] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, “Kimera: an open-source library for real-time metric-semantic localization and mapping,” in *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1689–1696.
- [107] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [108] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering,” *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 139–1, 2023.
- [109] A. Fisher, R. Cannizzaro, M. Cochrane, C. Nagahawatte, and J. L. Palmer, “Colmap: A memory-efficient occupancy grid mapping framework,” *Robotics and Autonomous Systems*, vol. 142, p. 103755, 2021.
- [110] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, “imap: Implicit mapping and positioning in real-time,” in *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6229–6238.
- [111] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, “Nice-slam: Neural implicit scalable encoding for slam,” in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 786–12 796.
- [112] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, “Gaussian splatting slam,” in *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 18 039–18 048.
- [113] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten, “Splatam: Splat track & map 3d gaussians for dense rgb-d slam,” in *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 21 357–21 366.
- [114] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao,

-
- S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.
- [115] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 270–279.
- [116] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, “Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems,” *IEEE Transactions on Robotics*, vol. 38, no. 4, 2022.
- [117] D. McGann and M. Kaess, “imesa: Incremental distributed optimization for collaborative simultaneous localization and mapping,” in *Proceedings of the Robotics: Science and Systems XX*, 2024.
- [118] Y. Tian, A. Koppel, A. S. Bedi, and J. P. How, “Asynchronous and parallel distributed pose graph optimization,” *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5819–5826, 2020.

Research Publications

Related to this Thesis

A. Journal Publication

- [1] **Junwoon Lee**, Taisei Ando, Mitsuru Shinozaki, Toshihiro Kitajima, Qi An, and Atsushi Yamashita, “Self-TIO: Thermal-Inertial Odometry via Self-Supervised 16-bit Feature Extractor and Tracker,” *IEEE Robotics and Automation Letters (RA-L)*, Vol. 10, No. 2, pp. 1003-1010, February 2025.
- [2] **Junwoon Lee**, Ren Komatsu, Mitsuru Shinozaki, Toshihiro Kitajima, Hajime Asama, Qi An, and Atsushi Yamashita, “Switch-SLAM: Switching-based LiDAR-Inertial-Visual SLAM for Degenerate Environments,” *IEEE Robotics and Automation Letters (RA-L)*, Vol. 9, No. 8, pp. 7270-7277, August 2024.

B. Peer-reviewed International Conference

- [1] **Junwoon Lee**, Taisei Ando, Mitsuru Shinozaki, Toshihiro Kitajima, Qi An, and Atsushi Yamashita, “TC-LTIO: Tightly-coupled LiDAR Thermal Inertial Odometry for LiDAR and Visual Odometry Degraded Environments,” *Proceedings of the 24th International Conference on Control, Automation and Systems (ICCAS2024)*, pp. 655-660, Jeju (Korea), October-November 2024. **(Best Paper Award)**
- [2] **Junwoon Lee**, Ren Komatsu, Mitsuru Shinozaki, Toshihiro Kitajima, Hajime Asama, Qi An, and Atsushi Yamashita “Switch-SLAM: Switching-based LiDAR-Inertial-Visual SLAM for Degenerate Environments,” *40th Anniversary of the IEEE Conference on Robotics and Automation (ICRA40)*, Rotterdam (Netherlands), September 2024. (Transferred from IEEE Robotics and Automation Letters, Vol. 9, No. 8, pp. 7270-7277, August 2024.)

C. Non-reviewed Domestic Conference

- [1] **Junwoon Lee**, Mitsuru Shinozaki, Toshihiro Kitajima, Qi An, and Atsushi Yamashita, “LiDAR-Visual-Inertial SLAM Robust in Structurally and Visually Degenerate Environments,” 第 42 回日本ロボット学会学術講演会予稿集 (RSJ2024), RSJ2024AC3K4-02, pp. 1-4, 大阪, September 2024.

D. Award

- [1] **Junwoon Lee**, Taisei Ando, Mitsuru Shinozaki, Toshihiro Kitajima, Qi An, and Atsushi Yamashita, Best Paper Award (ICCAS2024), October. 2024. (Top 1 out of 400 submitted papers)

Not related to this Thesis

E. Journal Publication

- [1] **Junwoon Lee**, Masamitsu Kurisu, and Kazuya Kuriyama, “Three-dimensionalized feature-based LiDAR-visual odometry for online mapping of unpaved road surfaces,” *Journal of Field Robotics*, Vol. 41, No. 5, pp. 1452-1468, August 2024.

F. Peer-reviewed International Conference

- [1] Taisei Ando, **Junwoon Lee**, Mitsuru Shinozaki, Toshihiro Kitajima, Qi An, and Atsushi Yamashita, “Highly Accurate and Fast Two-view Pose Estimation by Fast Reduction of Spherical Image Distortion Effects,” *Proceedings of the 24th International Conference on Control, Automation and Systems (ICCAS2024)*, pp. 774-779, Jeju (Korea), October-November 2024.

G. Non-reviewed Domestic Conference

- [1] 安藤 大生, 小松 廉, **Lee Junwoon**, 篠崎 充, 北島 利浩, 浅間 一, 安 琪, 山下 淳, “全天球画像の 2 視点間位置姿勢推定のための適応的閾値を用いた相互最近傍マッチング,” 2024 年度精密工学会春季大会学術講演会講演論文集, pp. 496-497, 東京, March 2024.
- [2] 安藤 大生, 小松 廉, **Lee Junwoon**, 篠崎 充, 北島 利浩, 浅間 一, 安 琪, 山下 淳, “適応的閾値を用いた相互最近傍マッチングによる全天球画像の 2 視点間位置姿勢推定精度向上,” 第 24 回計測自動制御学会システムインテグレーション部門講演会講演論文集 (SI2023), pp. 2448-2453, 新潟, December 2023.

H. Patent

- [1] 足立 馨, 栗栖 正充, イ・ジュンウォン, “地形検知システム、および地形検知方法,” 特願 2023-105215.