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

Junwoon Lee[1*], Taisei Ando[1], Mitsuru Shinozaki[2], Toshihiro Kitajima[2], Qi An[1], and Atsushi Yamashita[1]

[1]Department of Human and Engineered Environmental Studies, The University of Tokyo, 277-8563, Japan ({leejunwoon, ando, anqi, yamashita}@robot.t.u-tokyo.ac.jp)
[2]Technology Innovation R&D Dept.Ⅱ, Research & Development Headquarters, KUBOTA Corporation, 590-0908, Japan ({mitsuru.shinozaki, toshihiro.kitajima}@kubota.com) * Corresponding author

Abstract: We propose tightly-coupled LiDAR thermal inertial odometry for LiDAR and visual odometry degraded environments to deal with LiDAR and RGB-based visual odometry degenerate environments. Our approach exhibits high robustness and accuracy by tightly-coupled sensor fusion between LiDAR edge/planar features and thermal camera point features. Furthermore, to mitigate challenges inherent in thermal vision such as low contrast, we employ a learning-based optical flow trained on both synthetic thermal images generated from RGB images and real-world thermal images. Experimental results demonstrate that our method effectively handles not only the degradation of LiDAR and visual odometry but also challenges inherent in thermal vision.

Keywords: Sensor fusion, LiDAR-Thermal odometry, Learning-based optical flow, Zero velocity detection

1. INTRODUCTION

Autonomous mobile robots play an important role in the automation of the delivery, manufacturing, farming, mining, and space exploration. While these robots traditionally rely on their localization with the GNSS/INS system [1], challenges arise in GNSS-denied environments such as indoors, areas with roofing, or dense vegetation, where signal loss occurs.

To overcome this limitation, simultaneous localization and mapping (SLAM) [2] methods have been proposed. SLAM is typically categorized into light detection and ranging (LiDAR) SLAM 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 [3]. 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 [4]. 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 [5]. 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 [6–8]. However, as most of these methods rely on loosely-coupled manner (inter-system fusion) [6, 7], 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 [8]. 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 [9], 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 [10] or a visual feature-based solution [11].

To address these limitations, we propose a tightly-coupled LiDAR thermal inertial odometry for LiDAR and visual odometry degraded environments (TC-LTIO). The main contributions of our work are as follows:

- **Tightly-coupled sensor fusion**: To address SLAM degeneration, we integrate LiDAR edge/planar and thermal point features in a tighly-coupled manner. Our 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, we leverage a learning-based optical flow approach. This method excels in handling low-contrast scenarios and aggressive motion. Our 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 our method, thermal features are excluded from the entire optimization process when zero velocity is detected based

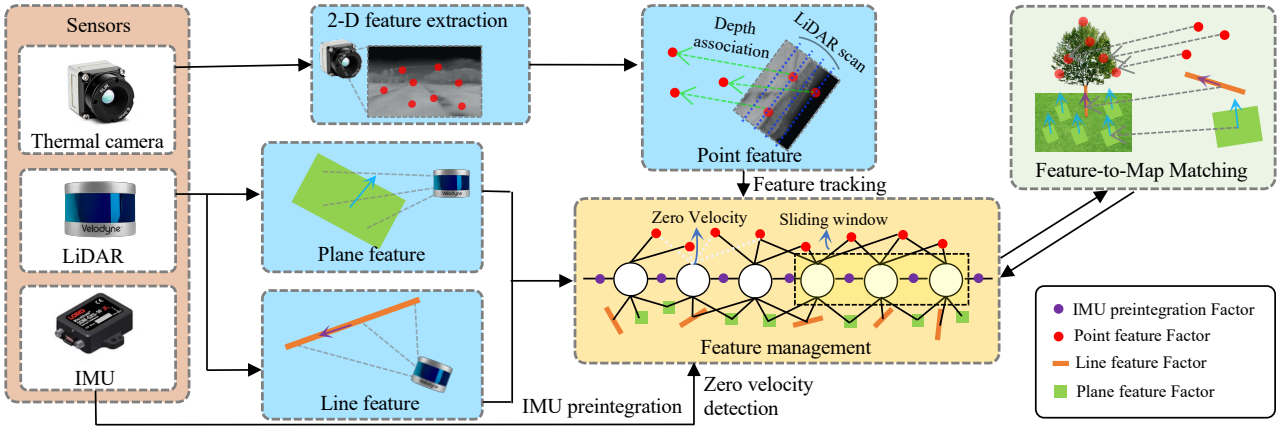


Fig. 1.: System overview of TC-LTIO.

on the values of LiDAR odometry and IMU.

2. TIGHTLY-COUPLED LIDAR THERMAL INERTIAL ODOMETRY

2.1 System Overview

An overview of the proposed SLAM system, TC-LTIO, is shown in Fig. 1. Three sensors, a thermal camera, LiDAR, and an inertial measurement unit (IMU), are used for TC-LTIO. Our 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 [12]. 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.

2.2 Feature Extraction

2.2.1 LiDAR Feature Extraction

Line and plane features are extracted following [3], 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

$$\sigma^{(m,n)} = \frac{1}{\left|\mathcal{S}^{(m,n)}\right|} \sum_{\mathbf{p}^{(m,j)} \in \mathcal{S}^{(m,n)}} (\|\mathbf{p}^{(m,j)} - \mathbf{p}^{(m,n)}\|). \tag{1}$$

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

2.2.2 Visual Feature Extraction

To extract visual feature points from the 8-bit thermal image, we adapt grid-based FAST as proposed in [13]. 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, we completely follow [8], which projects LiDAR scans onto the 2-D image plane for association. Note that our 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.

2.3 Visual Feature Tracking

Optimization-based feature tracking methods such as [14] and feature matching-based methods such as [15,16] are prone to failure, especially due to low contrast problem of the thermal image. To robustly track extracted visual features, we propose learning-based optical flow for thermal image based on [17] and its lightweight variant [18]. The proposed optical flow network is illustrated in Fig. 2, leveraging not only a lightweight CNN feature encoder and 3-D correlation volume for efficient computation but also a GRU-based update operator [19] for the accurate estimation of residual flow.

To effectively train our optical flow network for the thermal domain with a focus on noise suppression, we introduce novel training strategies. Initially, we train with synthetic thermal optical flow datasets originating from RGB image optical flow datasets such as [20] and [21]. As referred to in [22], the differences between thermal and RGB images 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 $\mathbf{n}_{\text{lfn}}$ and FPN noise $\mathbf{n}_{\text{fpn}}$ pattern are added to the grayscale image as shown in Fig. 3. For FPN noise, we randomly generate an image by rearranging various vertical stripes. For LFN noise, the prepared image is

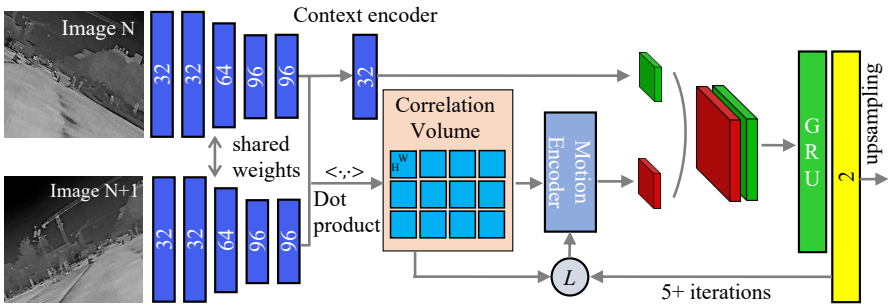


Fig. 2.: Proposed optical flow network.

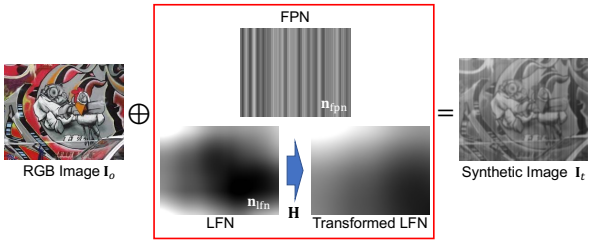


Fig. 3.: Method to generate synthetic thermal image from RGB image.

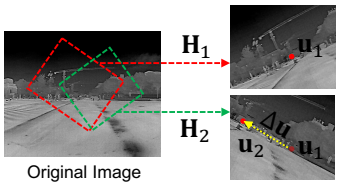


Fig. 4.: Labeling for self-supervised learning from single thermal image.

randomly transformed with random homography transformation $\mathbf{H}$. Consequently, the formulation of the synthetic thermal image is as follows:

$$\mathbf{I}_t = w_1\mathbf{I}_o + w_2\mathbf{n}_{\text{fpn}} + w_3\mathbf{H}\mathbf{n}_{\text{lfn}}, \tag{2}$$

where w_1, w_2, and w_3 are weighted parameters such that $w_1 + w_2 + w_3 = 1$. 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, we also train our network with real-world thermal images to enhance robustness against actual real-world noise. Given the scarcity of optical flow datasets based on thermal images, we adopt a self-supervised training strategy. Inspired by [23], 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:

$$\Delta\mathbf{u} = \mathbf{u}_2 - \mathbf{u}_1 = \mathbf{u}_2 - \mathbf{H}_2{\mathbf{H}_1}^{-1}\mathbf{u}_1. \tag{3}$$

Here, $\mathbf{u} = [u, v, 1]^{\mathrm{T}}$ is a pixel location in each warped image. The detail is shown in Fig. 4.

2.4 Tightly-Coupling formulation

2.4.1 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, we solve the current optimal robot state $\mathbf{X}^*$ using the least squares minimization problem, where costs from each factor within a sliding window are minimized, as depicted below:

$$\begin{aligned}\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmin}} \sum_{i\in\mathcal{K}} (\|\mathbf{r}_{I_i}\|^2_{\Sigma_{I_i}} + \sum_{j\in\mathcal{P}} \|\mathbf{r}_{p_{(i,j)}}\|^2_{\Sigma_p} \\ + \sum_{j\in\mathcal{L}} \|\mathbf{r}_{l_{(i,j)}}\|^2_{\Sigma_l} + \sum_{j\in\mathcal{D}} \|\mathbf{r}_{d_{(i,j)}}\|^2_{\Sigma_d}) + \|\mathbf{r}_0\|^2_{\Sigma_0},\end{aligned} \tag{4}$$

where $\|\mathbf{x}\|^2_{\mathbf{\Sigma}} = \mathbf{x}\mathbf{\Sigma}^{-1}\mathbf{x}^\top$ 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 [24].

However, the tracking quality of the learning-based optical flow deteriorates in scenarios where the camera experiences zero velocity. Consequently, Eq. (4) 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, we exclude point features from the MAP fusion process to prevent the propagation of mistracking results to the entire system.

2.4.2 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 our factor graph, we follow the IMU preintegration method proposed in [25]. Therefore, The IMU preintegration residual $\mathbf{r}_{I_i}$ in Eq. (4) can be defined as:

$$\mathbf{r}_{I_i} = [\mathbf{r}_{\Delta\mathbf{p}_i}, \mathbf{r}_{\Delta\mathbf{v}_i}, \mathbf{r}_{\Delta\mathbf{R}_i}, \mathbf{r}_{\mathbf{b}^a_i}, \mathbf{r}_{\mathbf{b}^g_i}]^\top. \tag{5}$$

Here, $\mathbf{r}_{\Delta\mathbf{p}_i}$, $\mathbf{r}_{\Delta\mathbf{v}_i}$, and $\mathbf{r}_{\Delta\mathbf{R}_j}$ are position, linear velocity, and orientation residual between prior and present

Table 1.: Comparison of Absolute Translational Errors (Maximum, RMSE) on Prepared Datasets.

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$_{\text{RGB}}$	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$_{\text{Thermal}}$	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**

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 [25].

2.4.3 LiDAR Feature Factors

LiDAR plane and line features, which are extracted according to the smoothness value of Eq. (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 [26]. 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:

$$\mathbf{r}_{l_{(i,j)}} = \frac{(\mathbf{p}^l_{(i,j)} - \hat{\mathbf{p}}^l_1) \times (\mathbf{p}^l_{(i,j)} - \hat{\mathbf{p}}^l_2)}{\|\hat{\mathbf{p}}^l_1 - \hat{\mathbf{p}}^l_2\|}, \tag{6}$$

$$\mathbf{r}_{p_{(i,j)}} = \frac{(\mathbf{p}^p_{(i,j)} - \hat{\mathbf{p}}^p_1)((\hat{\mathbf{p}}^p_1 - \hat{\mathbf{p}}^p_2) \times (\hat{\mathbf{p}}^p_1 - \hat{\mathbf{p}}^p_3))}{\|(\hat{\mathbf{p}}^p_1 - \hat{\mathbf{p}}^p_2) \times (\hat{\mathbf{p}}^p_1 - \hat{\mathbf{p}}^p_3)\|}, \tag{7}$$

given a line feature $\mathbf{p}^l_{(i,j)} \in \mathbb{R}^3$ and the corresponding nearest line feature $\hat{\mathbf{p}}^l_1$ and second one $\hat{\mathbf{p}}^l_2$ on the map. Moreover, given a plane feature $\mathbf{p}^p_{(i,j)} \in \mathbb{R}^3$ and the corresponding nearest plane feature $\hat{\mathbf{p}}^p_1$, second one $\hat{\mathbf{p}}^p_2$, and third one $\hat{\mathbf{p}}^p_3$.

2.4.4 Visual Feature Factors

The tracked point features $\mathbf{p}^d_j \in \mathbb{R}^3$, acquired through visual feature point and depth association, are stored as the map (visual landmarks) in the world frame. Subsequently, $\mathbf{p}^d_j$ 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:

$$\mathbf{r}_{d_{(i,j)}} = \mathbf{u}_i - \pi(\mathbf{T}_i \mathbf{p}^d_j), \tag{8}$$

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.

2.5 Zero Velocity Detection

Zero velocity detection is critical to our 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, we simultaneously utilize LiDAR feature residuals and IMU preintegration results, which are independent of visual features. Leveraging Eq. (7), and (5), we can compute the positional or rotational differences between the first keyframe and the last keyframe within the sliding window. Consequently, zero velocity detection can be formulated as follows:

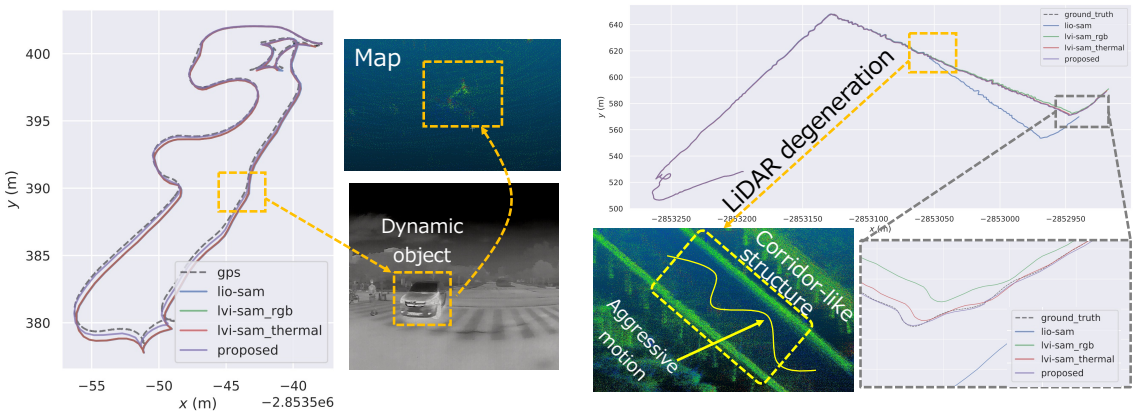
$$[\sum_{i\in\mathcal{K}} \|\mathbf{r}_{\Delta\mathbf{p}_i}\|^2, \sum_{i\in\mathcal{K}} \|\mathbf{r}_{\Delta\mathbf{R}_i}\|^2, \sum_{i\in\mathcal{K}}\sum_{j\in\mathcal{P}} \|\mathbf{r}_{p_{(i,j)}}\|^2] < [\delta_{\Delta\mathbf{p}}, \delta_{\Delta\mathbf{R}}, \delta_p]. \tag{9}$$

Here, where $\delta_{\Delta\mathbf{p}}$, $\delta_{\Delta\mathbf{R}}$, and δ_p are heuristic threshold for each criterion. In Eq. (9), if all the criteria are smaller than their respective thresholds, we consider the current state as a zero velocity state.

3. EXPERIMENTS

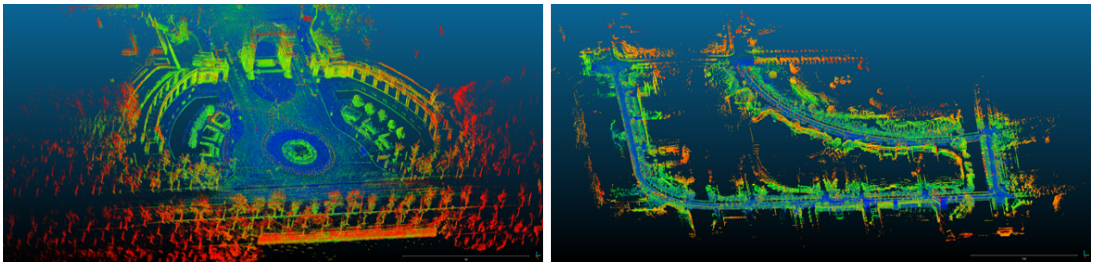
To evaluate the proposed method, we compare our method with the state-of-the-art LiDAR-inertial SLAM, LIO-SAM [3] and LiDAR-visual-inertial SLAM, LVI-SAM [6]. In case of LVI-SAM, the visual odometry submodule is executed using RGB camera (denoted as LVI-SAM$_{\text{RGB}}$) or thermal camera (denoted as LVI-SAM$_{\text{Thermal}}$). M2DGR dataset [27] is used for evalutation, 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 represent the positional difference between the ground truth points and the estimated trajectories. As shown in Table 1, the proposed method consistently achieves lower ATE compared to the compared 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 [3] scan-to-map optimization.

In gate01, the dynamic object caused the drift of both LIO-SAM and LVI-SAM as shown in Fig. 5(a). This drift occurs because the LOAM-based scan-map matching algorithm matches all line and plane features with the map. In contrast, our 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(b), LIO-SAM and LVI-SAM exhibited significant drift. Conversely, our proposed method produced accurate localization results by leveraging a learning-based feature tracker for robustness



(a) gate01 (b) street01

Fig. 5.: Experimental result on gate01 and street01.



(a) gate02 (b) street02

Fig. 6.: Resulting map in gate02 and street02. The point cloud is colored according to LiDAR intensity.

in visually-degraded situations and tightly-coupled sensor fusion for robustness in LiDAR-degraded situations. In gate03, although our method showed slightly lower accuracy compared to LVI-SAM 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 tightly-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. 6, showcase the 3-D mapping potential of our method.

4. CONCLUSIONS

In this paper, we propose TC-LTIO: tightly-coupled LiDAR thermal inertial odometry for LiDAR and visual odometry degraded environments, 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, we employed factor graph-based tightly-coupled approach. To construct the factor graph accurately, we track point features extracted from thermal images 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, we employ zero velocity detection based on the reults 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 evalutations, TC-LTIO demonstrated high robustness and accuracy in such environments compared to state-of-the-art methods. For future research, our aim is to explore learning-based robust feature extraction from low-contrast thermal images at the raw 16-bit level.

ACKNOWLEDGEMENT

This research was conducted under the university-corporate collaboration agreement between Kubota Corporation and the University of Tokyo.

REFERENCES

[1] 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.

[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J.J. Leonard, “Past, present, and future of simultaneous local-

ization and mapping: Toward the robust-perception age," *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309-1332, 2016.

[3] 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 *Proc. of 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 5135-5142, 2020.

[4] 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, 2024.

[5] J. Zhang, M. Kaess, and S. Singh, "On degeneracy of optimization-based state estimation problems," in *Proc. of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 809-816, 2016.

[6] T. Shan, B. Englot, C. Ratti, and D. Rus, "Lvi-sam: Tightly-coupled lidar-visual-inertial odometry via smoothing and mapping," in *Proc. of 2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5692-5698, 2021.

[7] J. Zhang and S. Singh, "Visual-lidar odometry and mapping: Low-drift, robust, and fast," in *Proc. of 2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2174-2181, 2015.

[8] 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.

[9] 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.

[10] 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.

[11] Y. 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.

[12] 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.

[13] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "Openvins: A research platform for visual-inertial estimation," in *Proc. of 2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4666-4672, 2020.

[14] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, pp. 674-679, 1981.

[15] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. of 2011 International Conference on Computer Vision (ICCV)*, pp. 2564-2571, 2011.

[16] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. of the Seventh IEEE international Conference on Computer Vision (ICCV)*, vol. 2, pp. 1150-1157, 1999.

[17] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Proc. of 16th European Conference on Computer Vision (ECCV)*, pp. 402-419, 2020.

[18] 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.

[19] K. Cho, B. V. 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.

[20] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. of 12th European Conference on Computer Vision (ECCV)*, pp. 611-625, 2012.

[21] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3061-3070, 2015.

[22] 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, pp. 104888, 2023.

[23] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 224-236, 2018.

[24] 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.

[25] 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, 2017.

[26] 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.

[27] 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.