Thermal-Inertial SLAM for the Environments With Challenging Illumination

Jiajun Jiang, Xingxin Chen, Weichen Dai, Zelin Gao, and Yu Zhang

***Abstract*—In recent years, longwave infrared (LWIR) cameras have become potential in visual simultaneous localization and mapping (SLAM) research since the delivered thermal images can provide information beyond the visible spectrum and are robust to environment illumination. However, due to modality differences, SLAM methods designed for visible cameras cannot be directly applied to thermal data. In this paper, we propose a thermal-inertial SLAM method for all-day autonomous systems. To overcome the challenge of the thermal data association, the proposed method represents several improvements, including singular-value-decomposition-based (SVD-based) image processing and ThermalRAFT tracking methods. Based on the characteristics of the thermal images, the SVD-based image processing method can exploit the fixed noise pattern of thermal images and enhance the image quality to improve the performance of subsequent steps, including thermal feature extraction and loop detection. To achieve real-time and robust feature tracking, we develop ThermalRAFT, an efficient optical flow network with iterative optimization. Moreover, the system introduces a bag-of-words-based loop detection method to maintain global consistency in long-term operation. The experimental results demonstrate that the proposed method can provide competitive performance in indoor and outdoor environments and is robust under challenging illumination conditions.**

***Index Terms*—SLAM, localization, visual-inertial SLAM.**

I. Introduction

SIMULTANEOUS localization and mapping localization methods play an essential role for autonomous mobile robots. To navigate in GPS-denied environments with onboard sensing, much of robotic research has focused on visible cameras and inertial sensors due to their portable size and low power requirements. However, traditional visual-inertial odometry and SLAM systems hinder robots from working in environments with poor illumination, such as underground tunnels and caves. In contrast, LWIR cameras are independent of illumination, raising high attention in recent SLAM studies.

Despite the advantages compared to visible cameras, realizing a thermal SLAM is challenging in several aspects. (1) LWIR cameras can capture thermal data in a high dynamic range (HDR), which saves in more than 8-bit formats (e.g., 14-bit or 16-bit) [1], increasing the compatible difficulty of traditional vision approaches, e.g., feature detection and association methods. Many methods directly scale thermal data into 8-bit using Automatic Gain Control (AGC) for contrast enhancement. However, AGC operation cannot cope with the significant photometric change brought by re-scaling in successive frames and may amplify the stripe noise generated by the thermal imagery mechanism [2]–[4]. (2) LWIR cameras will make noise reduction during operation by performing Non-Uniformity Correction (NUC), which periodically suspends the camera operation for approximately 500 milliseconds. The sudden interruption leads to significant viewpoint change and makes the odometry prone to track loss. Therefore, as described by the above two factors, the thermal image processing method and data association between frames are critical to SLAM.

In this paper, we propose a novel thermal-inertial SLAM method named TI-SLAM, as shown in Fig. 1. To address the issues mentioned above, SVD-based image processing and ThermalRAFT tracking are introduced to improve the robustness of the data association. SVD-based image processing can discover the inherent noise pattern and enhance the image quality of thermal images by SVD investigation, ensuring feature extraction without being affected by the noise and AGC operation. To overcome the challenge of the poor texture and photometric inconsistency of thermal images, ThermalRAFT can compute the optical flow by multiple optimization iterations to complete the feature tracking. Moreover, the inertial information is fused to further improve the tracking performance when NUC happens. In the back-end, loop closure is essential for a SLAM system. Therefore, a bag-of-words method optimized for thermal images is proposed to detect loop candidates and find the loop association. The main contributions of this work are as follows:

1) We propose a practical thermal-inertial SLAM system using a novel image processing algorithm for feature detection and a lightweight optical flow network for feature association.

Manuscript received 24 February 2022; accepted 6 June 2022. Date of publication 23 June 2022; date of current version 18 July 2022. This letter was recommended for publication by Associate Editor X. Zuo and Editor J. Civera upon evaluation of the reviewers' comments. This work was supported in part by the National Key Research and Development Program of China under Grants 2021ZD0201400 and 2022ZD0208800, in part by the National Natural Science Foundation of China under Grant 62088101, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LQ22F030022, in part by the Project of State Key Laboratory of Industrial Control Technology, Zhejiang University, China under Grant ICT2021A10, and in part by the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China under Grant ICT2022B04. *(Corresponding author: Yu Zhang.)*

Jiajun Jiang, Xingxin Chen, Zelin Gao, and Yu Zhang are with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310000, China (e-mail: elkulasjiang@zju.edu.cn; chenxingxin@zju.edu.cn; 22132038@zju.edu.cn; zhangyu80@zju.edu.cn).

Weichen Dai is with the College of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: weichendai@hotmail.com).

Digital Object Identifier 10.1109/LRA.2022.3185385

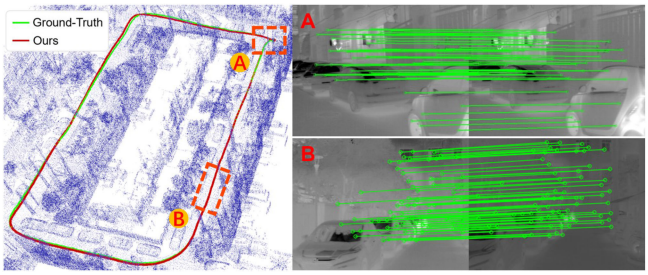


Fig. 1. The proposed thermal-inertial SLAM method conducts robust pose estimation in a residential area. The left figure shows the trajectory comparison of our method (red) and ground-truth (green). Marker A represents loop closure, and the top right image shows feature association using the proposed Thermal-RAFT between loop frames. Besides, a feature association example with NUC (B) is shown in the lower right image. ThermalRAFT can establish robust and accurate matches despite significant viewpoint and contrast changes.

2) We propose an SVD-based image processing method, improving image quality on low contrast thermal images for SLAM by singular value reallocation.
3) We propose ThermalRAFT, a real-time optical flow network architecture with a lightweight design. Thermal-RAFT produces accurate optical flow estimation on thermal images and significantly reduces the inference time.
4) We perform experiments in both indoor and outdoor environments and the public dataset. The results demonstrate that the proposed method can provide remarkable performance even under poor illumination.

This paper is organized as follows: Section-II gives a brief review on related works. Section-III introduces the details of the proposed method. Then extensive evaluations are presented in Section-IV. Section-V draws a conclusion. Data of our experiments is available online.[1]

II. Related Work

A. Thermal SLAM

Thermal image-based SLAM is much smaller literature than its visible spectrum counterpart. Some existing works combine thermal information with other sensors, such as lidar [5], [6], radar [7], visible camera [8], inertial [4], [9]. The inertial sensor is not affected by environmental changes, making it ideal for fusing the thermal modality. Khattak *et al.* [1] proposed a keyframe-based thermal-inertial odometry tracking with direct methods. It achieves good performance in visually degraded scenes, but the algorithm highly depends on the initialization process since the fixed initial depth may not converge. Zhao *et al.* [10] developed a thermal-inertial odometry method, adopting a deep learning method for thermal feature detection with real-time performance. However, the KLT-based feature tracking method is prone to fail due to stripe noise and periodic NUC interruption. Recently, Saputra *et al.* [11] designed DeepTIO, a deep neural network model for thermal-inertial odometry. DeepTIO shows better performance in the featureless scene, while conventional feature-based approaches lose track. Benefiting from the selective fusion of different modalities, DeepTIO is robust against sensor alignment issues. The researches above have accumulated drift problem and cannot maintain global consistency, especially in long-term navigation. Saputra *et al.* [12] realized a learning-based thermal-inertial system with loop closure, inheriting the DeepTIO framework in the front-end. The system achieves promising results in various indoor scenes. DeepTIO [11] and [12] operate in offline fashion and have high computation cost, limiting the real-time applications. Moreover, it is challenging for deep models to generalize in different scenarios and platforms. The models need to be trained for new datasets. Fully trusting deep learning modules has high potential risks due to the network's inherent limitations, such as generalization and computation cost. Taking advantage of traditional and deep learning-based methods seems a wiser choice.

B. Thermal Image Processing

In recent years, numerous image processing algorithms have been proposed to remove the pattern noise from thermal image [13], [14]. However, most traditional thermal image denoising algorithms ignore the time consumption and cannot cope with the photometric problem of re-scaled thermal images. To solve the photometric change caused by AGC operation, Mouats *et al.* [9] proposed to set an appropriate AGC threshold to smooth the variation of illumination of the images. However, this process merely delays the illumination change, and the problem remains unsolved fundamentally. Vidas [2] raised a histogram normalization method on 14-bit thermal images to obtain illumination-stable images. Papachristos *et al.* [3] set a fixed interval re-scaling in known environments. Both methods are highly relied on the temperature prior to a given environment, resulting in limitations in practice. In recent years, thermal image processing methods with the deep neural network have shown promising results [10]. To the best of our knowledge, few traditional image process methods based on re-scaled low contrast thermal images are available to solve noise and AGC problems simultaneously.

C. Feature Association on Thermal Images

Classical feature association methods, like BRIEF [15], LK sparse optical flow [16] have been proven effective in odometry and SLAM research. Mouats *et al.* [17] evaluated different feature descriptors on re-scaled thermal images. The results show that the matching methods for visible images get a lower matching performance on thermal images. The bad results may owe to poor image quality (e.g., noise, low contrast). Learning-based feature matching approaches have already been utilized in some SLAM systems, achieving competitive performance compared to classical methods and even obtaining better results in challenging conditions. GCNv2 [18] and DXSLAM [19] use deep neural networks to extract features and produce descriptors. Their experiments show that deep descriptors can improve the robustness in poor texture scenes. Zhan *et al.* [20] developed monocular visual odometry using deep optical flow to match two consecutive frames. The above methods are designed for visible

[1][Online]. Available: https://github.com/NGCLAB/multi-spectral-dataset/blob/master/VTI/TI.md

images and cannot be directly applied in the infrared domain. SuperThermal [21] proposed a deep neural network that detects features and computes descriptors simultaneously, showing the advantages of a deep model for thermal feature matching. Nevertheless, the performance of this network incorporating a complete SLAM system is unknown, and the computation cost is not presented.

III. Thermal-Inertial SLAM

The overview of our thermal-inertial SLAM is illustrated in Fig. 2. The input images are in 8-bit format, provided by the camera driver, or processed by min-max re-scale strategy from raw data. In the front-end, ThermalRAFT is adopted to track the features between the latest keyframe and the current frame, using the original 8-bit input images. Creating a new keyframe is based on the number of successfully tracked features. If the number of tracked features is less than a threshold, the current frame is set as a new keyframe. After keyframe creation, we perform SVD-based image processing and detect features to ensure a stable feature number on the keyframe. Then, the new keyframe is added to the backend and loop closure module. We maintain a sliding window to perform thermal-inertial optimization.

A. Thermal Image Processing

1) Thermal Imagery Model: Since the stripe noise shares the exact mechanism and affects the same, the column pattern will be illustrated as a representative in the following. Consider a $M \times N$ thermal image $\mathbf{X}$, with $M > N$, which suffers from the non-uniformity (stripe noise) problem. A commonly used approximate linear model for Focal Plane Array (FPA) sensor output is given by:

$$x(i,j) = a(i,j)y(i,j) + b(i,j) \tag{1}$$

where $x(i,j)$ denotes the actual output of the (i,j)th sensor, $a(i,j)$ and $b(i,j)$ are the gain and bias, and $y(i,j)$ is the real incident infrared radiation obtained by the sensor.

For FPA sensors, the bias non-uniformity is in domination and the gain can be ignored [22]. The sensors in the same column share the same bias, and the biases of different columns are independent of each other. Thus, the observation model becomes

$$x(i,j) = y(i,j) + b(j) \tag{2}$$

For thermal image $\mathbf{X}$, the formulation is

$$\mathbf{X} = \mathbf{Y} + \mathbf{B} \tag{3}$$

where the bias matrix $\mathbf{B}$ is defined by

$$\mathbf{B} = \begin{bmatrix} b(1) & b(2) & \cdots & b(N) \\ b(1) & b(2) & \cdots & b(N) \\ \vdots & \vdots & \vdots & \vdots \\ b(1) & b(2) & \cdots & b(N) \end{bmatrix} \tag{4}$$

The bias affects the image structure significantly in the column and row direction for thermal imagery, resulting in stripe noise on the image.

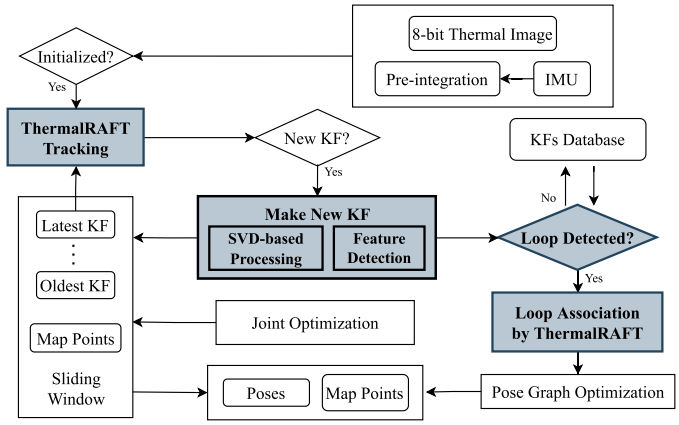


Fig. 2. Overview of proposed thermal-inertial SLAM. The bold blocks are our main contributions to this work.

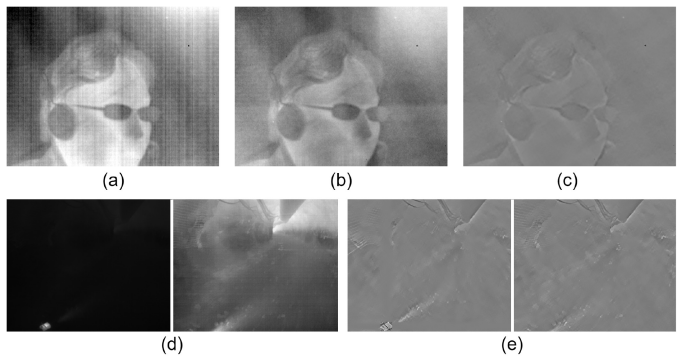


Fig. 3. Results of SVD-based thermal image processing method. The method shows the ability of anti-noise (top row) and anti-photometric (bottom two rows). Original images (a)(d). A denoised image whose largest SV is set to zero (b). Images after SV reallocation (c)(e).

2) SVD-Based Processing: Singular value decomposition (SVD) is adopted in thermal image processing for stripe noise suppression. SVD is a stable and effective method to decompose an image into a set of linearly independent components, commonly used in many image processing applications. A thermal image $\mathbf{X}$ could be represented by its SVD as follows:

$$\mathbf{X} = [\mathrm{U}][\mathrm{S}][\mathrm{V}]^{\mathrm{T}} = \sum_{i=1}^{N} s_i u_i v_i^T \tag{5}$$

where the columns of matrix U and V are called left singular vectors and right singular vectors, respectively. S is $M \times N$ matrix with the diagonal elements represent the singular values (SVs), s_i of $\mathbf{X}$, which are ordered in descent [23]. Singular values specify the luminance of the image, while the singular vectors specify the image structure [24].

The largest singular value packs most of the energy contained in the image [23], while singular vectors describe the global shape of the image in the vertical and horizontal direction [25]. According to the above, the largest singular value represents most of the vertical and horizontal information. By setting the largest singular value to zero, the stripe noise significantly decreased. Fig. 3(a) is a classic thermal image

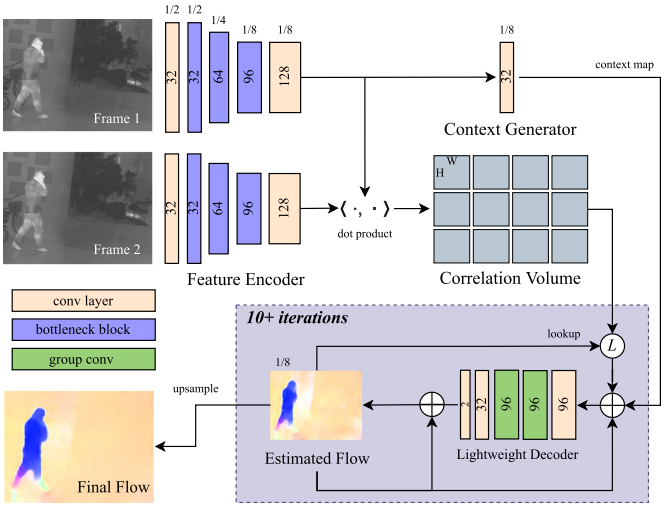


Fig. 4. Network architecture of proposed ThermalRAFT.

with stripe noise from the thermal denoising dataset [26], and the result of setting the largest singular value zero is shown in Fig. 3(b).

As mentioned before, each singular value of an image specifies the intensity of the image layer, and the respective pair of singular vectors specify the image structure or topology. After eliminating the first largest singular value, the image becomes dimmer, but the structure remains. Strengthening the structure of the thermal image is beneficial for feature detection. We reallocate singular values in ascending order on the base of the mean original singular value. Reallocation suppresses the stripe noise, emphasizes the detail of the image, and is an effective way to solve low contrast and photometric change problems. Non-local means filter and median filter are utilized for further image smoothing. The results are shown in Fig. 3(c). An example is given to further illustrate anti-photometric performance. Fig. 3(d) shows two successive images from the cave dataset, and we can observe that introducing a hot object (the torch on the ground) into the scene induces an instantaneous intensity change. In contrast, the images processed by the proposed method are virtually unaffected and maintain the key structure, as shown in Fig. 3(e).

Since the proposed method strengthens the structure of images, we adopt a gradient-based feature detection algorithm to extract robust features. The feature detection strategy is mainly inspired by DSO [27]. We select the pixels with a large gradient, and Non-Maximum Suppression is applied to avoid clustered features.

B. Feature Tracking With ThermalRAFT

1) Network Architecture: Due to the problems mentioned above, tracking feature points on thermal images is challenging. Inspired by RAFT [28], we designed ThermalRAFT, an accurate and lightweight deep network for optical flow estimation. The network architecture is illustrated in Fig. 4. Since real-time performance is vital in SLAM systems, we mainly focus on building a lightweight network architecture.

We first extract high-level features for the input image pair using a ResNet-Style encoder with bottleneck blocks, which output feature maps at 1/8 resolution. Then the context generator is applied to the feature map of frame1. The context generator only contains one 1×1 convolution layer, which reduces the feature map dimension and produces a context map.

The 4D correlation volume, which measures the visual similarity of two feature maps, is constructed by the dot product between all pairs of feature vectors. Unlike RAFT, we only keep one layer of the correlation volume pyramid to reduce computation and memory. To keep a large receptive field, we still use all-pairs correlation, and the lookup radius is set to 4. To this end, our method is robust to significant viewpoint changes.

We design a lightweight decoder to estimate the residual flow iteratively. The decoder concatenates the indexed correlation volume, context features, and the previously estimated flow as input and regresses the residual flow using five convolutional layers. Inspired by the famous ShuffleNet [29], we use group convolution with channel shuffle operation in the second and third convolution layers, reducing the parameters and making the network more efficient. Finally, the estimated flow is upsampled to the full resolution by adopting bilinear interpolation.

2) Train Details: Most deep optical flow networks lack generalization to the infrared domain because their training is mainly on visible images in a supervised manner. For better performance in thermal data, a training process on thermal images is required. However, we have to train on thermal images in a self-supervised manner since it is inaccessible to thermal datasets that provide ground-truth flow.

Following the training strategy in [30], we generated image pairs and the corresponding ground-truth flow by artificial warping, using a randomly generated geometric transformation, including affine, homography, and Thin-Plate Spline (TPS) transformation. We adopt the same L1 loss in RAFT [28] for training

$$L = \sum_{i=1}^{N} \gamma^{N-i} \left\| \mathbf{F}_{gt} - \mathbf{F}_i \right\|_1 \tag{6}$$

where $\mathbf{F}_{gt}$ is the ground-truth flow. $\mathbf{F}_i$ is the estimated flow at ith iteration. N is the iteration number and γ is the weight factor in each iteration, we set $\gamma = 0.8$ in the first training stage, then set $\gamma = 0.85$ in the following training stages.

We train ThermalRAFT using one NVIDIA RTX 3090 GPU. Following previous works, we first train on FlyingChairs [31] for 100 k iterations with a batch size of 16, then we finetune for 100 k iterations on the dataset composing of Sintel [32], KITTI-2012 [33], KITTI-2015 [34], and HD1K [35], batch size is set to 16. The first two training stages on visible datasets make it easier for the network to learn the basic concept of optical flow. Lastly, we finetune our network for another 100 k iterations on the artificial warped thermal image pairs, and the original images are from KAIST trainset [36] and KTIO [1] dataset. The thermal datasets contain images with sufficient Fixed Pattern Noise, contributing to the anti-noise performance of the network.

TABLE I
AVERAGE TIME OF FEATURE ASSOCIATION (MILLISEC)

Methods	Run-time
FlowNet2	296.03
RAFT	276.01
RAFT-small	100.37
ThermalRAFT	49.99
ThermalRAFT with TensorRT	8.12

C. Loop Closing

We introduce a loop closing module to detect and correct loops. The loop closing thread tries to find loops and perform global optimization after the loop association is established.

1) Loop Detection: Accumulated drift during long-term operation deteriorates the performance of the thermal-inertial SLAM system. Thus, loop closure is commonly used to keep global consistency. However, the traditional loop closing strategy [37], [38] cannot be directly applied to the infrared domain for the following problems.

Since the LWIR cameras capture the radiation emitted by surroundings, the appearance of the captured thermal images may change a lot, even in the same place [6]. The problem leads to missing loop candidates during robot operation and is even worse with re-scaled images because of the AGC problem. The specific stripe noise on the thermal images disturbs the visual similarity of the two images, making the retrieval of candidates difficult.

Benefiting from our proposed image processing algorithm, we detect loop candidates on the processed images, which are anti-noise and illumination-invariant. We detect a revisited place utilizing DBoW2 [39], a commonly used place recognition approach based on bag-of-words. The processed images from the front-end are reused to save computation. We extract 300 feature points using our gradient-based method and compute the BRIEF descriptor for each feature. The descriptors are further converted to visual words and added into the keyframe database with corresponding frames. We perform loop detection by retrieving frames with similar visual words in the keyframe database.

2) Loop Correction: Once the current keyframe K_{cur} successfully detects a loop candidate K_{old}, the next step is establishing the feature association between K_{cur} and K_{old}. The relative pose between K_{cur} and K_{old}, denoted as $\mathbf{T}_{\text{cur}}^{\text{old}}$, can be calculated by the established feature correspondences. As mentioned before, ThermalRAFT is a reliable feature association method compared to its traditional counterparts. Therefore, we adopt ThermalRAFT to obtain 2D-2D matches. The loop candidate is decided by the number of established matches. Since the map points and 2D-3D matches of the current keyframe are known, we can establish 2D-3D correspondences for the loop candidate and the current keyframe. An EPnP algorithm with RANSAC is utilized to calculate $\mathbf{T}_{\text{cur}}^{\text{old}}$ and returns the number of inliers. Only if the relative yaw angle and relative translation are smaller than a threshold will the loop candidate be treated as correct loop detection.

We establish a pose graph including the loop constrains and sequential keyframe constrains and refine the keyframe poses in the entire map, maintaining the global consistency of the estimated trajectory. Roll and pitch angles are observable for visual-inertial systems. To this end, we only perform four degrees-of-freedom pose graph optimization during loop correction.

IV. EXPERIMENTS

A. Dataset

1) Self-Collected Dataset: Our handheld device is equipped with an LWIR camera, an Xsens IMU, and a standard camera. The LWIR camera is Optris PI 640, which outputs thermal images at 32 Hz with a size 640×480 pixels. The standard camera is ImageSource DFK 22BUC03, which captures 640×480 RGB images at 32 Hz. We recorded several sequences in our lab, an underground parking lot, and an outdoor residential area containing various illumination conditions. For lab sequences, the ground truth is captured by the OptiTrack motion capture system. For other sequences, the ground truth is provided by lidar odometry LOAM [40]. It should be noted that since trees and buildings in the residential area block the GPS signal, we use lidar odometry as our outdoor ground truth for relatively high accuracy compared with visual ones. For a fair comparison, We ensure that every sequence begins with movements in a small range for algorithms that need initialization.

2) Public Dataset: ViViD++ dataset [41] is a public dataset collected to tackle poor illumination conditions in robotic applications. The dataset contains RGB, thermal, and IMU data and is divided into handheld and driving sequences. For our experiment, we use the handheld sequences, which contain both indoor and outdoor environments. The ground truth is provided by Vicon or LOAM [40].

The experiments were carried out using a laptop with an Intel i7-10870H processor and an NVIDIA GeForce RTX 3060 laptop GPU. The software environment is Ubuntu 18.04 system with Robot Operating System(ROS). We implement our ThermalRAFT network in C++ using the LibTorch library and NVIDIA TensorRT to reduce memory usage and inference time.

B. Run-Time Performance

To verify the real-time performance of our ThermalRAFT, we measure the run-time of different optical flow networks, such as FlowNet2 [42] and RAFT [28]. We test the algorithms with default configuration on an RTX 3060 laptop GPU. The iteration number is fixed to 20 for RAFT and ThermalRAFT.

The results are the average time of 100 executions with image input size 640×480 pixels, as shown in Table I. The proposed ThermalRAFT runs $5.5\times$ faster than RAFT and $2\times$ faster than RAFT small model. With the acceleration of TensorRT, the inference time of our ThermalRAFT can be further reduced by 85% to 8.12 ms (123 FPS). Owing to the improvements in network architecture and the use of TensorRT, our ThermalRAFT can work efficiently, ensuring the real-time performance of our entire SLAM system.

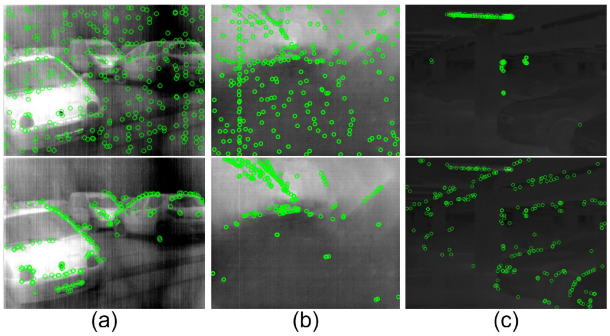


Fig. 5. Feature extraction performance on both re-scaled infrared image (top row) and processed image (bottom row). Features extracted from processed images are drawn on original images. (a) The urban scene, (b) the cave scene shows the anti-noise performance. The scenes with (c) high temperature object show the performance on extreme low contrast thermal images.

C. Image Processing Evaluation

In this part, we evaluate the performance of the proposed thermal image processing method by adopting the feature detection algorithm. We choose images from the thermal denoising dataset [26] and underground mine dataset [1] with typical stripe noise for anti-noise evaluation. We also evaluate the performance of extremely low contrast thermal images from our self-collected datasets. We extract gradient-based features from both original and processed images. The gradient-based feature is used in the proposed system, and the detection strategy is the same as DSO [27]. In this test, we compute a maximum of 300 points for both original and processed images. It is noted that in Fig. 5, the feature extracted from the processed images are drawn on the origin images for a clear illustration.

Fig. 5(a) and (b) show the anti-noise performance of the proposed image processing method. We can easily find that the feature points extracted from the processed images are mainly on the structure of the image, which is advantageous to further tracking. The original images have numerous wrong feature detection due to severe vertical stripe noise. Fig. 5(c) shows the performance on low contrast images. A fluorescent tube in Fig. 5(c) with relatively high temperature causes the rescaled thermal image low contrast. Thus, for the original image, the extracted features are clustered on relatively hot objects.

D. Full System Evaluation

In this subsection, we thoroughly evaluate our entire system. We compare our method with state-of-the-art visual odometry or SLAM systems, i.e. VINS-Mono [38], ORB-SLAM3 [37], ROVIO [43], in all dataset sequences. Besides, We compare the proposed method with DeepTIO [11] in the public dataset. We evaluate the accuracy with RMS Absolute Trajectory Error (ATE), aligning the estimated trajectory with ground-truth. We present the ATE with loop closure for all algorithms except otherwise noted. Odometry results are presented for odometry-only algorithms. If the estimated trajectory is too short or divergent, the result will be marked as failed $(\times)$.

1) Indoor Experiments: The indoor experiments were conducted in a laboratory and an underground parking lot. The results are presented in Table II. The reported values show that our method outperforms competing systems by a wide margin. In lab sequences, ORB-SLAM3 has a high failure rate, mainly caused by the poor feature detection performance on thermal images, and descriptor-based methods are hard to find feature correspondences. Semi-direct methods (i.e., ROVIO and VINS-Mono) may be affected by sudden photometric changes due to AGC. For this reason, the ATE results become worse. With anti-photometric image processing and deep optical flow estimation, our method shows more precise ATE. In parking lot sequences, the scale becomes larger, and the camera motion is faster. The competing algorithms quickly accumulate drift in large-scale environments and are prone to track loss due to the significant viewpoint change caused by the NUC operation. Despite these, our SLAM system still shows a big leap in trajectory accuracy.

TABLE II
INDOOR EVALUATION RESULTS OF ATE(M) OF DIFFERENT METHODS ON THERMAL IMAGES

Sequence	Length(m)	ORB3	ROVIO	VINS	OURS
Lab					
xyz1	13.00	0.9392	0.2500	0.1083	**0.0294**
xyz2	11.92	0.3442	0.4752	0.1934	**0.0356**
xyz3	12.44	1.2267	0.7401	0.2933	**0.0634**
halfsphere1	13.75	5.6110	0.5851	0.0970	**0.0555**
halfsphere2	14.01	4.4261	0.2336	0.1256	**0.0678**
halfsphere3	9.49	×	0.3766	0.1220	**0.0888**
circle1	13.83	×	0.8032	0.2230	**0.0598**
circle2	14.90	0.3058	0.1220	0.2193	**0.0518**
circleN1[1]	35.86	×	1.6274	0.5470	**0.1312**
circleN2	46.17	×	1.1645	0.6048	**0.1538**
circleN3	23.84	×	0.6256	0.3178	**0.0889**
Parking lot					
parklot1	127.13	15.1138	4.2762	1.9213	**0.3653**
parklot2	126.31	7.7978	17.6889	1.8137	**0.3745**
parklot3	94.28	11.3250	3.5819	4.0364	**0.4258**
parklot4	181.71	×	3.7729	2.0632	**0.6706**
parklot5	352.30	×	14.3645	3.4360	**0.6327**

[1] circleN denotes that we walk multiple circles.

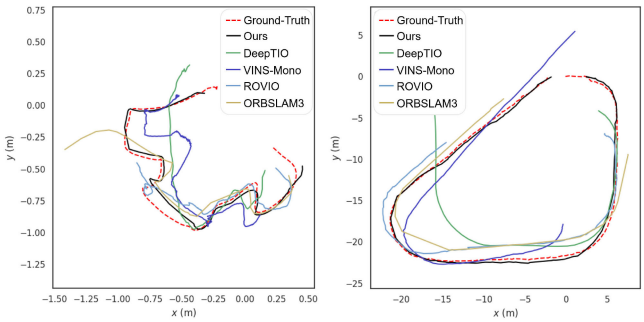
TABLE III
OUTDOOR EVALUATION RESULTS OF ATE(M) OF DIFFERENT METHODS ON THERMAL IMAGES

Sequence	Length(m)	ORB3	ROVIO	VINS	OURS
residential1	363.34	10.8636	18.3750	6.7905	**1.0374**
residential2	184.87	16.3574	11.8133	7.7007	**1.4374**
residential3	190.85	4.1689	16.2387	3.4904	**1.2626**

2) Outdoor Experiments: The outdoor environment is challenging since the tracked feature points may be too far, thus introducing drift in the estimated pose. Moreover, the radiation of the sky is much lower than the surroundings, resulting in extremely low contrast on re-scaled 8-bit thermal images. Since competing methods fail in such conditions, we re-scale the raw thermal images in a fixed temperature range. The experimental data in Table III indicates that our method achieves more accurate results compared to others. ThermalRAFT plays a vital role in accuracy improvement since it has learned how

TABLE IV
EVALUATION RESULTS OF ATE(M) IN PUBLIC DATASET

Sequence	ORB3	ROVIO	VINS	DeepTIO	OURS
Indoor-slow					
global	×	0.5391	0.5945	0.3109	**0.2697**
local	0.1545	0.1351	0.2089	0.2456	**0.0591**
varying	×	0.5624	0.1438	0.4485	**0.1144**
dark	0.1868	0.1480	0.1179	0.3581	**0.0851**
Outdoor-slow					
day1	4.6296	2.8706	2.7805	5.5819	**0.3552**
day2	5.1138	3.5077	1.5285	6.8662	**0.6447**
night1	3.2801	3.2054	3.3742	5.1786	**0.2329**
night2	4.8020	1.9814	5.3312	6.2689	**1.1135**



(a) (b)

Fig. 6. Estimated trajectory aligned with Ground-Truth. (a) Trajectory in indoor-slow-local. (b) Trajectory in outdoor-slow-night1.

to obtain correspondences in the thermal domain during the training process. The loop closing module also contributes to the accumulated drift reduction. The analysis above and the numerical data confirm that our SLAM system can be employed in large-scale scenes.

3) Evaluation in Public Dataset: In this experiment, we test various algorithms in the ViViD++ dataset. For DeepTIO, we use a pre-trained handheld model as the base model and fine-tune DeepTIO in ViViD++ handheld sequences with the default configuration. Table IV shows the numerical evaluation results in terms of ATE. It is noted that all the algorithms are evaluated on thermal images. Compared to the state-of-the-art methods, our system produces better results. ORB-SLAM3 loses tracks in two indoor sequences due to insufficient feature matches. The failure case also occurs in our self-collected indoor sequences, discussed in previous experiments. Compared with feature-based systems, the end-to-end method DeepTIO has great potential in thermal data association, even in the featureless area. The possible reason for limiting DeepTIO to better results is the inadequate training data in the ViViD++ dataset. Further, DeepTIO focuses more on indoor sequences under various conditions [11], which is probably the reason for the poor generalization to outdoor scenes. The trajectory examples are shown in Fig. 6.

4) Loop Closure Evaluation: To further illustrate the effectiveness of our loop closure, we compare the pose estimation results with another SLAM system with loop closure, VINS-Mono. Table V presents the ATE results. As seen in the table, the

TABLE V
THE LOOP CLOSURE EVALUATION RESULTS BETWEEN OURS AND VINS-MONO IN ATE(M)

Sequence	Length(m)	VINS (no loop)	VINS (loop)	OURS (no loop)	OURS (loop)
circleN2	46.17	0.6048	0.6155	0.1898	**0.1538**
parklot2	126.31	1.8137	1.8023	0.6175	**0.3745**
parklot5	352.30	3.4360	3.4524	2.0497	**0.6327**
residential1	363.34	6.7905	6.7856	2.5728	**1.0374**

TABLE VI
ATE(M) OF DIFFERENT METHODS UNDER VARIOUS ILLUMINATION

			Visible			Thermal
Sequence	Illumination	Length(m)	ORB3	ROVIO	VINS	OURS
xyz4	bright	13.02	**0.0851**	0.1460	0.0772	0.0927
half4	bright	10.01	0.1101	0.1604	0.0866	**0.0828**
circle4	bright	14.04	**0.0219**	0.0625	0.0904	0.1010
xyz5	varying	11.11	×	0.5015	2.8603	**0.0386**
half5	varying	10.22	×	0.5264	0.7058	**0.0408**
circle5	varying	23.84	×	0.2711	1.2183	**0.0889**
xyz6	dark	13.06	×	×	×	**0.0387**
half6	dark	14.60	×	×	×	**0.0805**
circle6	dark	14.14	×	×	×	**0.1126**

loop closure module of VINS-Mono fails to detect a loop candidate. Therefore, the results of VINS-Mono with and without the loop module are very close. During our experiment, we hypothesized that the low recall rate of VINS-Mono is mainly caused by the photometric change and the severe stripe noise. Even if it can detect a loop candidate, a few matches between the current keyframe and the loop candidate can be established. In contrast, our loop closure module shows more robustness and accuracy in closing the loop. From the data in Table V, our method with loop closure notably reduces the accumulated trajectory error. As the length of the trajectory grows, the advantages of loop closure become more significant.

5) Evaluations Under Different Illumination: In this experiment, we compare our method with the state-of-the-art visible visual SLAM systems to verify that LWIR cameras are a promising sensor in various illuminations. Table VI presents the ATE results. Varying denotes we turn on or turn off the light randomly. It can be seen that our method can obtain accurate pose estimation in different illumination conditions, while visible SLAM systems all fail in dark environments. In environments with varying illumination, ORB-SLAM3 fails in all sequences due to the sudden light change. With the aid of an inertial sensor, VINS-Mono and ROVIO can operate, but still have a large trajectory error. Even under bright illumination, our method also outperforms visible SLAM systems in the half4 sequence and achieves competitive performance in xyz4 and circle4. Results show that the proposed thermal-inertial SLAM has great potential to realize day and night operation in robotic applications.

V. CONCLUSION

In this work, we propose TI-SLAM, a novel and robust thermal-inertial SLAM system. SVD-based thermal image processing and ThermalRAFT methods overcome the difficulties

of data association in the infrared domain. The presented results demonstrate that the proposed system is a potentially reliable solution for robot navigation in challenging environments, especially under poor illumination conditions. However, the proposed system still has failure cases, such as operation in thermally texture-less environments and conditions with infrared reflections. In future research, we will focus on the cases above and experiment under various conditions, like conditions with smoke. Exploring the applications in monocular thermal SLAM research is also valuable since the monocular thermal system is more challenging when facing issues like NUC interruption.

REFERENCES

[1] S. Khattak, C. Papachristos, and K. Alexis, "Keyframe-based thermal-inertial odometry," *J. Field Robot.*, vol. 37, no. 4, pp. 552–579, 2020.

[2] S. Vidas and S. Sridharan, "Hand-held monocular SLAM in thermal-infrared," in *Proc. IEEE 12th Int. Conf. Control Automat. Robot. Vis.*, ICARCV, 2012, pp. 859–864.

[3] C. Papachristos, F. Mascarich, and K. Alexis, "Thermal-inertial localization for autonomous navigation of aerial robots through obscurants," in *Proc. IEEE Int. Conf. Unmanned Aircr. Syst.*, ICUAS, 2018, pp. 394–399.

[4] S. Khattak, C. Papachristos, and K. Alexis, "Keyframe-based direct thermal-inertial odometry," in *Proc. IEEE Int. Conf. Robot. Automat.*, ICRA, 2019, pp. 3563–3569.

[5] W. Chen, Y. Wang, H. Chen, and Y. Liu, "EIL-SLAM: Depth-enhanced edge-based infrared-lidar SLAM," *J. Field Robot.*, vol. 39, no. 2, pp. 117–130, 2022.

[6] Y.-S. Shin and A. Kim, "Sparse depth enhanced direct thermal-infrared SLAM beyond the visible spectrum," *IEEE Robot. Automat. Lett.*, vol. 4, no. 3, pp. 2918–2925, Jul. 2019.

[7] C. Doer and G. F. Trommer, "Radar visual inertial odometry and radar thermal inertial odometry: Robust navigation even in challenging visual conditions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, IROS, 2021, pp. 331–338.

[8] L. Chen, L. Sun, T. Yang, L. Fan, K. Huang, and Z. Xuanyuan, "RGB-T SLAM: A flexible slam framework by combining appearance and thermal information," in *Proc. IEEE Int. Conf. Robot. Automat.*, ICRA, 2017, pp. 5682–5687.

[9] T. Mouats, N. Aouf, L. Chermak, and M. A. Richardson, "Thermal stereo odometry for UAVs," *IEEE Sensors J.*, vol. 15, no. 11, pp. 6335–6347, Nov. 2015.

[10] S. Zhao, P. Wang, H. Zhang, Z. Fang, and S. Scherer, "TP-TIO: A robust thermal-inertial odometry with deep thermalpoint," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, IROS, 2020, pp. 4505–4512.

[11] M. R. U. Saputra *et al.*, "DeepTIO: A deep thermal-inertial odometry with visual hallucination," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1672–1679, Apr. 2020.

[12] M. R. U. Saputra, C. X. Lu, P. P. B. de Gusmao, B. Wang, A. Markham, and N. Trigoni, "Graph-based thermal–inertial SLAM with probabilistic neural networks," *IEEE Trans. Robot.*, vol. 38, no. 3, pp. 1875–1893, Jun. 2022.

[13] C. Lu, "Stripe non-uniformity correction of infrared images using parameter estimation," *Infrared Phys. Technol.*, vol. 107, 2020, Art. no. 103313.

[14] X. Chen, L. Liu, J. Zhang, and W. Shao, "Infrared image denoising based on the variance-stabilizing transform and the dual-domain filter," *Digit. Signal Process.*, vol. 113, 2021, Art. no. 103012.

[15] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 778–792.

[16] B. D. Lucas *et al.*, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, 1981, pp. 24–28.

[17] T. Mouats, N. Aouf, D. Nam, and S. Vidas, "Performance evaluation of feature detectors and descriptors beyond the visible," *J. Intell. Robotic Syst.*, vol. 92, no. 1, pp. 33–63, 2018.

[18] J. Tang, L. Ericson, J. Folkesson, and P. Jensfelt, "GCNv2: Efficient correspondence prediction for real-time slam," *IEEE Robot. Automat. Lett.*, vol. 4, no. 4, pp. 3505–3512, Oct. 2019.

[19] D. Li *et al.*, "DXSLAM: A robust and efficient visual SLAM system with deep features," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, IROS, 2020, pp. 4958–4965.

[20] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt," in *Proc. IEEE Int. Conf. Robot. Automat.*, ICRA, 2020, pp. 4203–4210.

[21] Y. Lu and G. Lu, "SuperThermal: Matching thermal as visible through thermal feature exploration," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 2690–2697, Apr. 2021.

[22] B. M. Ratliff, M. M. Hayat, and R. C. Hardie, "An algebraic algorithm for nonuniformity correction in focal-plane arrays," *J. Opt. Soc. Amer. A*, vol. 19, no. 9, pp. 1737–1747, 2002.

[23] J.-F. Yang and C.-L. Lu, "Combined techniques of singular value decomposition and vector quantization for image coding," *IEEE Trans. Image Process.*, vol. 4, no. 8, pp. 1141–1146, Aug. 1995.

[24] E. Ganic, N. Zubair, and A. M. Eskicioglu, "An optimal watermarking scheme based on singular value decomposition," in *Proc. IASTED Int. Conf. Commun. Netw. Inf. Secur.*, 2003, vol. 85, pp. 85–90.

[25] R. A. Sadek, "SVD based image processing applications: State of the art, contributions and research challenges," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 7, pp. 26–34, 2012.

[26] Y. Tendero and J. Gilles, "Admire: A locally adaptive single-image, non-uniformity correction and denoising algorithm: Application to uncooled Ir camera," in *Proc. Infrared Technol. Appl. XXXVIII*, 2012, pp. 580–595.

[27] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2017.

[28] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 402–419.

[29] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[30] P. Truing, M. Danelljan, R. Timofte, and L. Van Gool, "Learning accurate dense correspondences and when to trust them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5714–5724.

[31] A. Dosovitskiy *et al.*, "Flownet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2758–2766.

[32] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.

[33] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[34] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3061–3070.

[35] D. Kondermann *et al.*, "The HCI benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 19–28.

[36] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baselines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.

[37] 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 Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

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

[39] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[40] J. Zhang and S. Singh, "LOAM: Lidar odometry and mapping in real-time," in *Proc. Robotics: Sci. Syst.*, 2014, vol. 2, pp. 1–9.

[41] A. Lee, Y. Cho, Y.-S. Shin, A. Kim, and H. Myung, "ViViD++: Vision for visibility dataset," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 6282–6289, Jul. 2022.

[42] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2462–2470.

[43] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, IROS, 2015, pp. 298–304.