HOI-LCD: Leveraging Humans as Dynamic Landmarks Toward Thermal Loop Closing Even in Complete Darkness

Tatsuro Sakai, Yanshuo Bai, Kanji Tanaka, Wuhao Xie, Jonathan Tay Yu Liang and Daiki Iwata
University of Fukui, 3-9-1 Bunkyo, Fukui City, Fukui 910-0017, Japan

Keywords: Thermal Loop Closing, Human as Landmark, HOI Features.

Abstract: Visual SLAM (Simultaneous Localization and Mapping) is a foundational technology for autonomous navigation, enabling simultaneous localization and mapping in diverse indoor and outdoor environments. Among its components, loop closure plays a vital role in maintaining global map consistency by recognizing revisited locations and correcting accumulated localization errors. Conventional SLAM methods have primarily relied on RGB cameras, leveraging feature-based matching and graph optimization to achieve high-precision loop detection. Despite their success, these methods are inherently sensitive to illumination conditions and often fail under low-light or high-contrast scenes. Recently, thermal infrared cameras have gained attention as a robust alternative, particularly in dark or visually degraded environments. While various thermal-inertial SLAM approaches have been proposed, they still depend heavily on static structures and visual features, limiting their effectiveness in textureless or dynamic environments. To address this limitation, we propose a novel loop closure method that utilizes Human-Object Interaction (HOI) as dynamic-static composite landmarks in thermal imagery. Although humans are conventionally considered unsuitable as landmarks due to their motion, our approach overcomes this by introducing HOI feature points as landmarks. These feature points exhibit both a human attribute, characterized by stable detection across RGB and thermal domains via person tracking, and a static-object attribute, characterized by contact with visually consistent, semantically meaningful objects. This duality enables robust loop closure even in dynamic, low-texture, and dark environments, where traditional methods typically fail.

1 INTRODUCTION

Visual SLAM (Simultaneous Localization and Mapping) plays an essential role in both indoor and outdoor robotic navigation by enabling robots to simultaneously estimate their own position and construct a map of the surrounding environment. Within Visual SLAM, loop closing is a critically important function (Klein and Murray, 2007; Cummins and Newman, 2008; Mur-Artal et al., 2015a; Ali et al., 2022; Adlakha et al., 2020). It allows the system to recognize previously visited locations and integrate current observations with past map information, thereby mitigating the accumulation of localization errors and maintaining global map consistency. However, real-world environments are dynamic and often present significant challenges to loop closure due to factors such as lighting changes, viewpoint variations, occlusions, and differences in image features, all of which can significantly degrade the accuracy and robustness of the process.

Conventional loop closure methods in Visual SLAM have primarily been developed under the assumption of RGB cameras (visible light sensors). These approaches typically extract image features such as SIFT (Lowe, 2004) or ORB (Rublee et al., 2011) and estimate camera motion and environmental structure based on the geometric relationships among these features. Over time, the field has progressed from early filter-based techniques (Montemerlo et al., 2002) to more advanced graph optimization-based methods, such as ORB-SLAM (Mur-Artal et al., 2015b) and LSD-SLAM (Engel et al., 2014), achieving high-precision localization, robust mapping, and real-time performance. Nonetheless, RGB-based methods are fundamentally limited by their reliance on lighting conditions. In environments with low illumination or high dynamic range, extracting sufficient image features becomes difficult, leading to substantial degradation in overall system performance (Saputra et al., 2022).

To overcome the limitations of RGB cameras,

Sakai, T., Bai, Y., Tanaka, K., Xie, W., Liang, J. T. Y. and Iwata, D.
HOI-LCD: Leveraging Humans as Dynamic Landmarks Toward Thermal Loop Closing Even in Complete Darkness.
DOI: 10.5220/0013775900003982

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 2*, pages 105-115
ISBN: 978-989-758-770-2; ISSN: 2184-2809

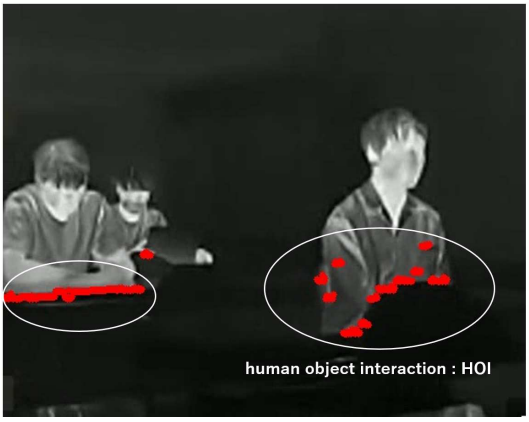

Figure 1: Humans are typically unsuitable as traditional landmarks due to their dynamic nature. However, this study focuses on Human-Object Interaction (HOI) with static objects, demonstrating their effectiveness as landmarks.

thermal infrared (IR) cameras have attracted increasing attention for loop closure applications (Saputra et al., 2022; Shin and Kim, 2019; van de Molengraft et al., 2023; Li et al., 2025; Xu et al., 2025). As they do not require visible light, thermal cameras offer a promising solution for maintaining localization and map consistency under visually degraded conditions, such as darkness, smoke, or dust. Several studies have explored this direction, including Graph-Based Thermal-Inertial SLAM (Saputra et al., 2022), which combines thermal imagery with IMU data, and FirebotSLAM (van de Molengraft et al., 2023), which targets disaster scenarios with poor visibility due to smoke. Other approaches include WTI-SLAM (Li et al., 2025), designed for weakly textured thermal images, and SLAM in the Dark (Xu et al., 2025), which employs self-supervised learning to achieve accurate loop closure. Additionally, Sparse Depth Enhanced SLAM (Shin and Kim, 2019) leverages sparse depth from external sensors to improve localization accuracy. While these works demonstrate the potential of thermal-based approaches for loop closure in harsh conditions, many of them still heavily rely on static structures and feature points in the environment. Thus, robustness remains limited in scenes that are weakly textured or feature-sparse.

To address these challenges, this study proposes a novel loop closure method that utilizes humans—a small set of object categories commonly present in human environments—as dynamic landmarks (Fig. 1). While humans are inherently dynamic and may seem unsuitable as conventional landmarks, this work focuses on human-object interactions (HOIs) (Antoun and Asmar, 2023), leveraging the static objects involved in such interactions as reliable cues. In thermal imagery, human features are especially salient (Teixeira et al., 2010), making them highly effective for detecting HOIs. The proposed method operates in three main steps. First, a thermal-domain-specific human tracker is trained to accurately localize human regions within thermal images. Second, HOI features are extracted from the detected human regions. Third, loop closure is performed by matching these HOI features between query and reference images using RANSAC (Random Sample Consensus) (Chum et al., 2003). This approach aims to enable stable loop closure even in challenging environments with limited static features or dynamic elements, where conventional feature-based methods often struggle.

2 RELATED WORK

Loop closing has developed as one of the core challenges in SLAM for robotics. In the early stages, Klein and Murray introduced PTAM (Parallel Tracking and Mapping), which separated real-time camera tracking from mapping, enabling high-precision mapping in small-scale environments (Klein and Murray, 2007). Subsequently, FAB-MAP, proposed by Cummins and Newman (Cummins and Newman, 2008), adopted a Bayesian approach based on the co-occurrence probability of visual features, significantly improving the reliability of loop detection. FAB-MAP 2.0 extended this framework to enable loop closure based solely on visual appearance in large-scale environments (Cummins and Newman, 2010). In 2015, ORB-SLAM by Mur-Artal et al. (Mur-Artal et al., 2015a) gained widespread adoption by employing the lightweight and high-precision ORB descriptor, demonstrating robust performance in real-world applications. More recently, advances in deep learning have led to learning-based approaches from the feature extraction stage, as seen in NetVLAD (Arandjelovic et al., 2018), a CNN-based place recognition method. Furthermore, Bi-directional Loop Closure (Ali et al., 2022) considers temporal context in both forward and backward directions, improving robustness. To ensure performance under conditions where visible light is unavailable, such as in darkness or smoke-filled environments, studies like DeepTIO (Adlakha et al., 2020) have explored the integration of thermal imagery with inertial measurements, indicating that multi-modal approaches adapted to specific sensing domains will be crucial in the future.

Recent loop closure techniques have achieved rapid innovation through advances in deep learning, neural rendering, semantics, and self-supervised learning. For instance, GLC-SLAM (Chen et al., 2024) integrates loop closure with a 3D scene rep-

resentation based on Gaussian Splatting, achieving both photorealistic rendering and precise localization. SGLC (Wang et al., 2024) targets loop closure in LiDAR SLAM by introducing a coarse-fine-refine strategy using semantic graphs, enabling accurate pose correction and map consistency in large-scale environments. DK-SLAM (Qu et al., 2024) presents a unified deep learning-based pipeline from keypoint detection to loop closure, significantly improving localization accuracy and generalization with monocular cameras. Loopy-SLAM (Liso et al., 2024) incorporates loop closure into dense NeRF-based scene representations, allowing relocalization and map consistency without relying on voxel grids. A novel trend is introduced by AutoLoop (Lahiany and Gal, 2025), which automates the fine-tuning process of existing SLAM models via agent-based curriculum learning, enabling fast and autonomous adaptation for loop closure. Additionally, 2GO (Lim et al., 2025) proposes an extremely efficient approach capable of detecting loops from just two viewpoints, dramatically reducing the computational cost of SLAM systems through lightweight multi-view inference. These cutting-edge studies contribute to enhancing the robustness, scalability, and autonomy of loop closure, each leveraging different sensors, representations, and learning strategies.

Loop closure based on thermal infrared cameras has recently gained attention as a robust solution for maintaining localization and map consistency in dark, smoky, or dusty environments where visible light cannot be used. Graph-Based Thermal-Inertial SLAM (Saputra et al., 2022) integrates thermal imagery with IMU data and applies probabilistic neural pose graph optimization to achieve both accuracy and robustness, demonstrating effectiveness across various scenarios including indoor and outdoor handheld cameras and SubT tunnels. Sparse Depth Enhanced SLAM (Shin and Kim, 2019) improves loop consistency by supplementing direct thermal SLAM with sparse depth measurements from external sensors, demonstrating the effectiveness of multimodal fusion. Firebot-SLAM (van de Molengraft et al., 2023), designed for smoke-obscured disaster environments, significantly improves situational awareness by generating maps and loop closing using thermal imagery alone in conditions of zero visibility. WTI-SLAM (Li et al., 2025) addresses the difficulty of loop detection in weakly textured thermal images by introducing a specialized feature extraction and tracking algorithm, enabling loop closure in cases where traditional visible-light SLAM methods fail. SLAM in the Dark (Xu et al., 2025), proposed by Xu et al., introduces a unified deep model that learns pose, depth, and loop closure entirely from thermal imagery in a self-supervised manner, achieving high-accuracy loop closure using only thermal sensing. These studies suggest new directions for robust loop closure in extreme environments through sensor fusion and learning-based approaches grounded in thermal infrared imaging.

3 PROBLEM FORMULATION

Following prior work, we formulate loop closure as an image retrieval problem. Given a query image captured at the current location, the objective is to find a matching image from a database of previously observed images that corresponds to the same physical location. This formulation casts loop closure as a place recognition task, where a successful match indicates a loop has been detected. This perspective enables the integration of image retrieval techniques—such as feature extraction, similarity computation, and geometric verification—into the SLAM pipeline to achieve robust loop detection. Under this formulation, the performance of loop closure is typically assessed using the metrics of precision and recall. Precision indicates the proportion of correctly identified loop closures among all retrieved results, reflecting the system's ability to minimize false positives. Recall measures the proportion of actual loop closures that are successfully detected, indicating the system's ability to avoid false negatives. These metrics often present a trade-off, and the design of the retrieval system must balance them appropriately depending on the characteristics of the target environment and the requirements of the task. To capture both precision and recall characteristics in a single metric, we adopt a ranking-based evaluation criterion, namely the Mean Reciprocal Rank (MRR), in our experiments. MRR evaluates the rank position of the first correct match in the retrieval results, thereby providing a balanced view of detection accuracy and reliability.

It should be noted, however, that loop closure differs from general image retrieval in several ways. First, in practice, loop closure is not triggered for every incoming image. Instead, a pre-processing step ensures that only sufficiently feature-rich query images are selected, avoiding loop detection attempts on low-texture or ambiguous inputs. Second, a post-processing step often evaluates the confidence of the retrieved result, and if the confidence is low, the result is discarded and not used in the subsequent SLAM optimization. These pre- and post-processing steps are crucial for maintaining the stability and reliability of SLAM in real-world conditions. For simplicity, the

proposed method in this work does not explicitly include such pre- or post-processing steps. We assume that a query image is already suitable for loop detection and that the retrieved results are directly used for matching and verification.

Traditional loop closure approaches have primarily used static objects in the environment as landmarks. This strategy offers advantages due to the fixed nature of such landmarks, allowing for consistent and reliable feature extraction. Additionally, decades of research have contributed to the maturity and robustness of this approach. However, it also has several limitations. Variations in lighting conditions, viewpoints, seasons, and time of day can significantly alter the visual appearance of scenes, introducing ambiguity and leading to false positives or missed detections. Moreover, in appearance-based SLAM systems that lack absolute distance measurements, aligning local maps built at different scales becomes difficult due to scale ambiguity. In large-scale environments, the growing size of the map increases computational costs, and the presence of moving objects or temporally varying structures can further destabilize the system and lead to incorrect loop closures.

To address these limitations, we propose a novel loop closure method that employs dynamic entities, specifically humans, as landmarks. One notable advantage of this approach is that humans exhibit clear thermal signatures in thermal images, making them relatively easy to detect compared to surrounding environments. This feature opens up the possibility of effective loop detection in dark environments where humans are prominent landmarks. Using dynamic humans as landmarks also enhances the adaptability of SLAM systems to dynamic environments, which are challenging for traditional static-object-based methods. Furthermore, leveraging human motion patterns and behaviors may offer additional cues for more complex scene understanding and localization.

However, this approach also presents significant challenges. Thermal images are often noisy, and human appearance can vary substantially depending on posture, clothing, and carried items, making detection and re-identification difficult. Humans constantly move, appear, and disappear, introducing high variability and instability when used as landmarks. As a result, achieving high-precision loop detection remains extremely challenging, and a single misidentification can severely compromise the entire system. Limitations of thermal cameras further complicate the problem. Thermal imagery lacks color information and often provides limited shape or semantic detail, making it difficult to extract meaningful features compared to RGB imagery. Additionally, variations in body temperature can alter thermal signatures over time, hindering consistent feature extraction.

4 APPROACH

Figure 2 illustrates the system architecture of the proposed method. As shown, the system consists of three main components: static landmark-based loop closure detection (SLCD), dynamic landmark-based loop closure detection (DLCD), and an information fusion module that integrates the outputs of these two LCD modules.

The SLCD component can employ any existing method such as conventional thermal SLAM. In this study, as described in Section 4.1, we utilize a classical approach based on SuperPoint matching. The DLCD component is our newly introduced dynamic landmark-based LCD, detailed in Section 4.2. The information fusion module fuses the outputs from the two LCD modules; while it is not the main focus of this study, we provide a simple implementation example in Section 4.3.

4.1 Static Landmark-Based LCD

As a conventional static landmark-based LCD (SLCD) method, this study adopts a feature matching approach using only 2D coordinates extracted by the SuperPoint feature extractor (DeTone et al., 2018). The detailed procedure is as follows.

4.1.1 Mutual Nearest-Neighbor Matching

For each feature point in the query image, the nearest feature point in the reference image is found, and vice versa. Only feature point pairs mutually nearest in both directions are retained as matches, effectively reducing false matches.

4.1.2 Geometric Consistency Verification

Matches are filtered by constraining the vertical coordinate difference based on the image height to be within a predefined threshold, enhancing the geometric validity of the matching results.

4.1.3 Homography Estimation via RANSAC

Random samples of matched feature pairs are selected to estimate a homography (planar projective transform). All matched points are then projected by this homography, and points with reprojection error below a threshold are considered inliers. This process is re-

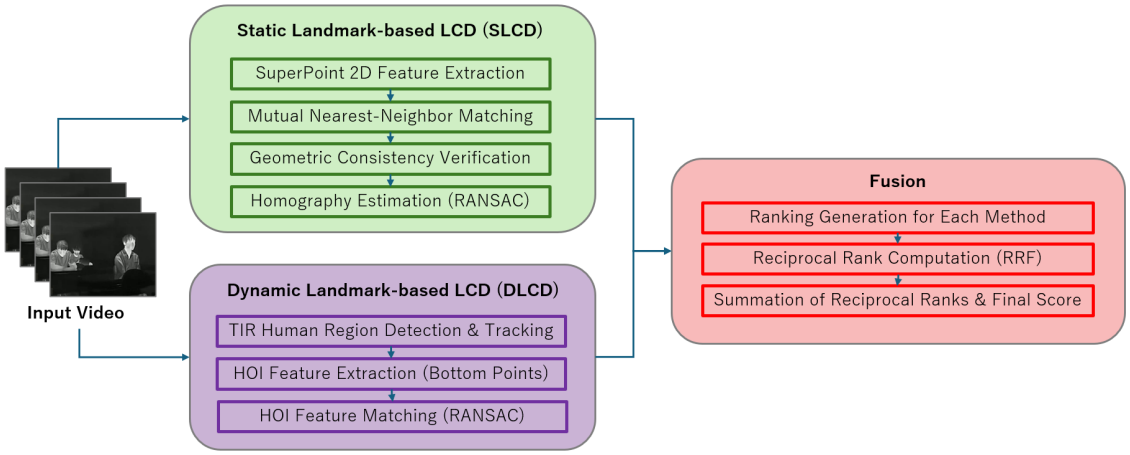

Figure 2: The SLCD module performs loop closure detection using static landmarks based on conventional methods like SuperPoint matching. The DLCD module, newly proposed in this research, performs loop closure detection using dynamic landmarks. The Information Fusion module integrates the results from both the SLCD and DLCD modules.

peated multiple times, and the homography with the largest inlier set is chosen as the final model.

4.2 Dynamic Landmark-Based LCD

To overcome the limitations discussed previously, several prior works offer important insights. In the domain of loop closure and image change detection (LCD-ICD), studies show that if the relative position between landmark A and object B remains invariant, they can be considered "pseudo-static," allowing their use as features in dynamic environments over short SLAM durations. Research on Human-Only SLAM (Tanaka, 2002) treats occlusion boundaries between humans and occluding objects as landmarks, mitigating false negatives by exploiting intermediate properties between dynamic and static objects. Building on these ideas, HO3-SLAM (Human-Object Occlusion Ordering SLAM) (Liang and Tanaka, 2024) effectively utilizes occlusion boundary points as keypoints for loop closure (see Fig. 1). HO3-SLAM notably exploits the stability of human attributes across RGB and thermal (T) domains and the reliability of static object attributes in appearance, shape, and semantics, providing valuable guidance for robust loop closure in dynamic environments.

4.2.1 HOI Feature Extraction

The proposed dynamic landmark-based loop closure detection proceeds as follows. Given each image frame, human regions are detected from thermal images. Detected bounding boxes are tracked and assigned unique human IDs. Pixel-wise AND operation between bounding boxes and human regions identifies precise human areas with IDs.

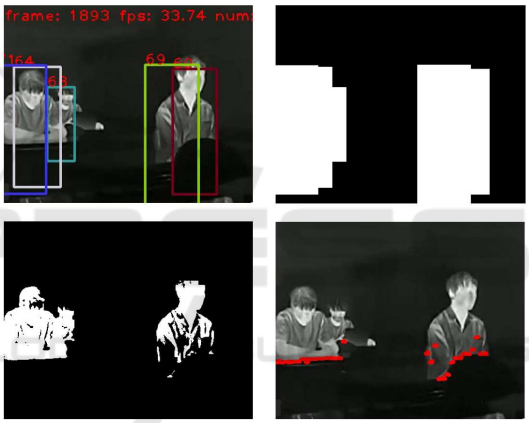

Figure 3: As a simple example of an HOI feature, this experiment utilizes the bottom-most point formed at the boundary between the human and the object. Top-left: Human tracking result; Top-right: Temperature thresholding; Bottom-left: AND operation; Bottom-right: HOI feature point.

For each human ID, at every horizontal pixel coordinate x within the bounding box, the pixel with the largest vertical coordinate y (closest to the image bottom) belonging to the human region is selected (Fig. 3). These "bottom points" serve as Human-Object Interaction (HOI) feature points, possessing two key properties: (1) human attributes that are stably detectable via tracking in both RGB and thermal domains, and (2) static object attributes that are invariant in appearance, shape, and semantics, functioning as reliable landmarks. The set of HOI feature points is defined as landmarks and recorded in the map along with frame IDs. These landmarks extracted from the current frame are matched against previously recorded landmarks. The matching score

is computed between landmark sets of current and past frames using RANSAC (Section 4.2.2) to evaluate matching reliability. Loop closure candidates are ranked based on these matching scores.

4.2.2 RANSAC Feature Matching

Because outliers are inevitable in feature matching, we apply the RANSAC algorithm to increase the reliability of the estimated geometric transform. The algorithm starts by randomly selecting four non-repeating correspondence pairs (the minimum required to estimate a homography) from the initial matching set between the query and database images. A tentative homography matrix is computed from these samples. Linear algebraic issues such as singularities invalidate the trial. The homography model projects points from the reference image to the query image; correspondences with reprojection error below the inlier threshold ($T_{\text{inlier}} = 5.0$ pixels) are considered inliers. The process iterates up to a maximum number of trials ($N_{\text{trials}} = 100$) to find the homography with the largest inlier count. If fewer than four initial matches exist, homography estimation is aborted and RANSAC terminated.

4.2.3 Training Thermal Infrared (TIR) Human Tracker

Tracking humans in darkness is critical for applications including surveillance, security, and disaster rescue. Conventional RGB cameras fail under low illumination, necessitating thermal infrared (TIR) cameras. Developing a TIR human tracker requires extensive annotated data, but manual labeling is costly. We propose a training method that reduces this burden by synchronously capturing RGB and TIR videos from co-located cameras. A pretrained high-performance RGB tracker generates bounding boxes and human IDs automatically on RGB videos, creating pseudo-labels. These pseudo-labels paired with corresponding TIR frames train the TIR tracker, enabling robust human tracking in darkness without manual annotations. Both RGB and TIR trackers use ByteTrack (Zhang et al., 2022) as their backbone.

4.3 Information Fusion

To leverage complementary information and improve robustness in image retrieval, this study integrates the baseline static landmark-based LCD (Section 4.1) and the proposed dynamic landmark-based LCD (Section 4.2). This integration aims to maximize the strengths of both methods and enhance retrieval accuracy.

For effective fusion of the two retrieval results, we adopt Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), a widely used technique to combine rankings from heterogeneous sources. RRF fairly balances the contribution of each method's top results, improving overall performance.

The fusion process proceeds as follows.

4.3.1 Ranking Generation for Each Method

SLCD and DLCD produce separate ranked lists of database images in descending order of matching scores for a given query.

4.3.2 Reciprocal Rank Computation

For each database image, the reciprocal rank is calculated for its rank r in each method by:

$$\text{RRF} = \frac{1}{k+r}$$

where k is a constant parameter. Based on preliminary experiments, we set $k = 0$ in this study. Lower ranks yield higher reciprocal rank values, giving more weight to more relevant images.

4.3.3 Summation of Reciprocal Ranks and Final Score

For each database image, reciprocal ranks from SLCD and DLCD are summed to obtain a final fusion score:

$$\text{Fusion Score} = \text{RRF}_{\text{SLCD}} + \text{RRF}_{\text{DLCD}}.$$

Sorting database images by these fusion scores produces the final retrieval ranking.

This reciprocal rank fusion allows images ranked highly by either method to be properly emphasized, thus enhancing overall retrieval performance.

5 EXPERIMENTS

The primary objective of this experiment is to evaluate the loop closing performance of a mobile robot operating in low-light indoor environments. Specifically, we assess the effectiveness of the proposed method using thermal images captured by a thermal camera in the presence of dynamic elements such as humans.

5.1 Setup

The experimental platform is a tricycle-drive mobile robot equipped with an onboard computer. A

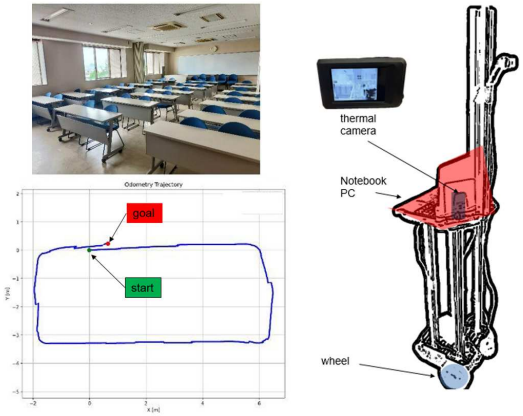

Figure 4: Experimental environment, robot, and robot trajectory.

monocular thermal camera (HIKMICRO Pocket2) is mounted on the front of the robot. This camera has a resolution of 256×192 pixels; thermal images are continuously recorded during motion. The data are collected in video mode, with an empirically measured average frame rate of approximately 25 Hz. Each frame is timestamped and precisely synchronized with odometry data (position: x, y; orientation: θ) estimated from the robot's encoders. These data are used to construct the ground-truth dataset for loop closure evaluation. The temperature range of the thermal camera is configured to 20-30°C to enhance contrast between human bodies and the background. As a result, human subjects (with body temperatures of approximately 36°C) appear as high-intensity regions, enabling clear identification as dynamic landmarks.

Experiments were conducted in an indoor environment measuring approximately 7.0 m by 3.0 m (Fig. 4). Six rectangular tables were arranged throughout the space, with four individuals randomly seated around each table. This layout was designed to reproduce conditions in which humans function as prominent high-temperature regions in thermal imagery. The robot autonomously navigated a clockwise path around the room while continuously recording thermal images. The experiment was conducted to evaluate loop closing performance in this dynamic environment. Figure 5 shows thermal images captured by the robot's on-board camera.

The collected dataset (1996 frames) includes seated individuals who appear as dynamic high-intensity regions. Odometry and timestamps were recorded alongside each frame to serve as input for loop closure evaluation.

Loop closing performance was evaluated using the number of inliers from homography estimation as the key metric. Specifically, the mean reciprocal rank (MRR) score was calculated based on the ranking

Figure 5: Input images.

Figure 6: Image matching process. For each row, from left to right, the images represent the following: Input Image, Feature Map, Top-1 Ranked Reference Image, and Ground-Truth Reference Image.

of ground-truth loop-closure pairs using RANSAC-based inlier counts. In addition, the average processing time per query frame was measured to assess computational efficiency.

All thermal images were synchronized with odometry data via linear interpolation based on acquisition timestamps. High-intensity regions at the lower part of each image were extracted via thresholding to obtain bottom-edge feature points, which serve as dynamic landmarks. Ground-truth loop closures were determined by identifying database frames within 0.2 m of the query frame's odometric position. Figure 6 shows the image matching process in the proposed method, DLCD.

Table 1: Performance results.

	MRR
DLCD	0.222
SLCD	0.124
DLCD+SLCD	0.224

5.2 Baseline and Proposed Methods

The evaluated methods detect loop closures between thermal images using feature point matching followed by homography estimation via RANSAC. Image features were precomputed and loaded as sets of 2D coordinates. Homographies were estimated using the Direct Linear Transform (DLT) algorithm applied to homogeneous coordinates, with RANSAC employed for outlier rejection. In each RANSAC iteration, four randomly selected correspondences were used to compute a homography hypothesis. Inliers were determined as correspondences with projection errors below a 5.0-pixel threshold. The model with the maximum number of inliers over 100 trials was selected as the final result. If fewer than four correspondences were available, homography estimation was skipped.

We compare the conventional SLCD, the proposed DLCD, and their fusion as described in Section 4.3. Each method is evaluated to assess its performance and robustness in a dynamic environment.

5.3 Results

The experimental results report the loop closing accuracy, specifically the mean reciprocal rank (MRR) score. The average processing time per query frame, including homography estimation and inlier computation, is summarized in Table 1. These results clarify the loop closing performance and highlight the computational efficiency of the proposed approach in a dynamic environment. Figure 7 shows matching results for both of the SLCD and DLCD methods.

We can see that the proposed DLCD method clearly outperforms the baseline SLCD method. In addition, the fusion method SLCD+DLCD performs slightly better than the proposed method, demonstrating the effectiveness of the approach of fusing the two methods.

6 CONCLUSIONS & FUTURE WORK

This study introduced a novel loop closure method for Visual SLAM that leverages human-object interactions (HOIs) as dynamic landmarks in challenging environments. Unlike conventional approaches that rely on static features or are limited by illumination conditions, our method utilizes the salient features of humans in thermal imagery to detect HOIs, treating the static objects involved in these interactions as reliable cues for loop closure. The proposed pipeline involves a thermal-domain-specific human tracker, HOI feature extraction from detected human regions, and RANSAC-based matching for robust loop closure. Our approach aims to enhance the stability of loop closure, particularly in scenarios characterized by limited static features or significant environmental dynamics where traditional methods often fail.

Building upon the insights and methodology presented in this study, future research will explore several promising directions:

- Quantitative Evaluation in Diverse Real-World Scenarios: While the theoretical framework for utilizing HOIs has been established, comprehensive quantitative evaluation in a wider array of real-world environments is crucial. This includes datasets with varying degrees of human activity, diverse object categories, and more complex environmental changes (e.g., severe occlusions, extreme temperature variations).
- Integration with Existing SLAM Frameworks: Our current work focuses on the loop closure module. Integrating this HOI-based loop closure method into a full-fledged Visual SLAM system (e.g., ORB-SLAM, LSD-SLAM) and evaluating its end-to-end performance would be a significant next step. This would involve assessing the impact on global map consistency, localization accuracy, and real-time performance.
- Robustness to Ambiguous HOI Detections: The accuracy of HOI detection is paramount for the effectiveness of our method. Future work will investigate techniques to improve the robustness of HOI detection, especially in cases of partial occlusions, unusual human poses, or ambiguous interactions. This could involve exploring more advanced deep learning architectures or incorporating temporal reasoning.
- Scalability for Large-Scale Environments: For applications in large-scale environments, managing and matching a potentially vast number of HOI features could become computationally intensive. Research into efficient data structures, indexing methods, and feature aggregation techniques will be necessary to ensure scalability.
- Extension to Other Dynamic Elements: While this study focuses on humans, the concept of leveraging dynamic entities interacting with static objects

Figure 7: Image matching results. Left: DLCD. Right: SLCD.

could be extended to other categories, such as vehicles or other mobile robots. Investigating the applicability and benefits of such extensions would broaden the scope of this approach.

- Fusion with Complementary Sensors: Combining thermal imagery with data from other sensor modalities (e.g., event cameras for high dynamic range, LiDAR for precise 3D geometry) could further enhance the robustness and accuracy of HOI-based loop closure, especially in highly challenging conditions.

7 LIMITATIONS

While the proposed method demonstrates promising results, it is important to acknowledge several limitations:

- **Robustness of HOI Detection:** The accuracy of human-object interaction (HOI) detection directly impacts the performance of our method. Detecting HOIs can be challenging under complex poses, partial occlusions, or extreme lighting conditions. Although human features are salient in

thermal images, these challenges still exist.

- **Nature of Dynamic Landmarks:** Due to the dynamic nature of humans, the reliability of loop closure might be affected if HOIs are transient. While our focus is on interactions with static objects, the duration and stability of the interaction itself can influence the overall robustness of the system.
- **Sparsity of Features:** While our method is effective in environments with few static features, there might be situations where HOIs themselves are very rare, leading to an insufficient number of landmarks. This becomes particularly evident in environments with minimal human presence or infrequent interactions with specific objects.
- **Computational Cost:** Detecting and tracking HOIs, and subsequently performing feature matching based on them, can be computationally more intensive compared to traditional static feature-point-based methods. Efficient algorithms and optimization will be crucial to maintain real-time performance.
- **Dataset Diversity:** Current evaluations might be dependent on specific datasets. A comprehensive quantitative evaluation across a wider range of real-world scenarios, especially those involving diverse human activities, object categories, and complex environmental changes (e.g., severe occlusions, extreme temperature variations), is necessary.

REFERENCES

Adlakha, D. et al. (2020). Deeptio: A deep thermal-inertial odometry with visual hallucination. *arXiv*.

Ali, I., Peltonen, S., and Gotchev, A. (2022). Bi-directional loop closure for visual slam. *arXiv*.

Antoun, M. and Asmar, D. (2023). Human object interaction detection: Design and survey. *Image and Vision Computing*, 130:104617.

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2018). Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):1168–1181.

Chen, Z., Song, Z., Pang, Z.-J., Liu, Y., Chen, Z., Han, X.-F., Zuo, Y.-W., and Shen, S.-J. (2024). Glc-slam: Gaussian splatting slam with efficient loop closure.

Chum, O., Matas, J., and Kittler, J. (2003). Locally optimized ransac. In Michaelis, B. and Krell, G., editors, *Pattern Recognition*, pages 236–243, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cormack, G. V., Clarke, C. L. A., and Buettcher, S. (2009). Reciprocal rank fusion. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 528–529. ACM.

Cummins, M. and Newman, P. (2008). Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665.

Cummins, M. and Newman, P. (2010). Appearance-only slam at large scale with fab-map 2.0. *International Journal of Robotics Research*, 29(8):943–959.

DeTone, D., Malisiewicz, T., and Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 337–33712.

Engel, J., Schöps, T., and Cremers, D. (2014). Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer.

Klein, G. and Murray, D. (2007). Parallel tracking and mapping for small ar workspaces. In *Proceedings of ISMAR 2007*. IEEE.

Lahiany, A. and Gal, O. (2025). Autoloop: Fast visual slam fine-tuning through agentic curriculum learning.

Li, S., Ma, X., He, R., Shen, Y., Guan, H., Liu, H., and Li, F. (2025). Wti-slam: a novel thermal infrared visual slam algorithm for weak texture thermal infrared images. *The Journal of Engineering*.

Liang, J. T. Y. and Tanaka, K. (2024). Robot traversability prediction: Towards third-person-view extension of walk2map with photometric and physical constraints. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024*, pages 11602–11609. IEEE.

Lim, T. Y., Sun, B., Pollefeys, M., and Blum, H. (2025). 2go: Loop closure from two views.

Liso, L., Sandstrom, E., Yugay, V., Gool, L. V., and Oswald, M. R. (2024). Loopy-slam: Dense neural slam with loop closures.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110.

Montemerlo, M., Thrun, S., Koller, D., Wegbreit, B., et al. (2002). Fastslam: A factored solution to the simultaneous localization and mapping problem. *Aaai/iaai*, 593598:593–598.

Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D. (2015a). Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163.

Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D. (2015b). Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163.

Qu, H., Tang, X., Liu, C.-T., Chen, X.-Y., Li, Y.-Z., Zhang, W.-K., Zhang, H.-T., and Li, J.-H. (2024). Dk-slam: Monocular visual slam with deep keypoint learning, tracking and loop-closing.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). Orb: An efficient alternative to sift or surf. In

2011 International Conference on Computer Vision, pages 2564–2571.

Saputra, M. R. U., Lu, C. X., de Gusmao, P. P., Wang, B., Markham, A., and Trigoni, N. (2022). Graph-based thermal–inertial slam with probabilistic neural networks. *IEEE Transactions on Robotics*, 38(3):1875–1893.

Shin, Y.-S. and Kim, A. (2019). Sparse depth enhanced direct thermal-infrared slam beyond the visible spectrum.

Tanaka, K. (2002). Detecting collision-free paths by observing walking people. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 55–60. IEEE.

Teixeira, T., Dublon, G., and Savvides, A. (2010). A survey of human-sensing: Methods for detecting presence, count, location, track, and identity. *ACM Computing Surveys*, 5(1):59–69.

van de Molengraft, S., Ferranti, L., and Dubbelman, G. (2023). Firebotslam: Thermal slam to increase situational awareness in smoke-filled environments. *Sensors*, 23(17):7611.

Wang, N., Chen, X., Shi, C., Zheng, Z., Yu, H., and Lu, H. (2024). Sglc: Semantic graph-guided coarse-fine-refine full loop closing for lidar slam.

Xu, Y., Hao, Q., Zhang, L., Mao, J., He, X., Wu, W., and Chen, C. (2025). Slam in the dark: Self-supervised learning of pose, depth and loop-closure from thermal images.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., and Wang, X. (2022). Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer.