Robustness Enhancement of Self-Localization for Drone-View Mixed Reality via Adaptive RGB-Thermal Integration

Ryuto Fukuda and Tomohiro Fukuda *

Posted Date: 8 January 2026

doi: 10.20944/preprints202601.0625.v1

Keywords: adaptive tracking; drone-view mixed reality; effective inlier count; robust self-localization; sensor fusion; spatial filtering; thermal-RGB integration; texture-less environments

 doi:10.20944/preprints202601.0625.v1

Article

Robustness Enhancement of Self-Localization for Drone-View Mixed Reality via Adaptive RGB-Thermal Integration

Ryuto Fukuda and Tomohiro Fukuda *

Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, The University of Osaka, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan

* Correspondence: fukuda.tomohiro.see.eng@osaka-u.ac.jp

Abstract

Drone-view mixed reality (MR) in the Architecture, Engineering, and Construction (AEC) sector faces significant self-localization challenges in low-texture environments, such as bare concrete sites. This study proposes an adaptive sensor fusion framework integrating thermal and visible light (RGB) imagery to enhance tracking robustness for diverse site applications. We introduce the Effective Inlier Count (N_{eff}) as a lightweight gating mechanism to evaluate the spatial quality of feature points and dynamically weight sensor modalities in real-time. By employing a 20×16 grid-based spatial filtering algorithm, the system effectively suppresses the influence of geometric burstiness without significant computational overhead on server-side processing. Validation experiments across various real-world scenarios demonstrate that the proposed method maintains high geometric registration accuracy where traditional RGB-only methods fail. In texture-less and specular conditions, the system consistently maintained an average Intersection over Union (IoU) above 0.72, while the baseline suffered from complete tracking loss or significant drift. These results confirm that thermal-RGB integration ensures operational availability and improves long-term stability by mitigating modality-specific noise. This approach offers a reliable solution for various drone-based AEC tasks, particularly in GPS-denied or adverse environments.

Keywords: adaptive tracking; drone-view mixed reality; effective inlier count; robust self-localization; sensor fusion; spatial filtering; thermal-RGB integration; texture-less environments

1. Introduction

In recent years, consumer drones have proliferated rapidly, bringing about fundamental changes in business processes across diverse industrial fields. Applications have expanded beyond hobbyist use to a wide range of areas, including entertainment and aerial photography [1], geological surveying and mapping, traffic monitoring [2], disaster prevention, search and rescue [3], and agriculture [4]. In parallel, digital transformation is progressing in the Architecture, Engineering, and Construction (AEC) industry, and the utilization of drones as a core data acquisition platform is expanding rapidly [5]. Since drones possess advantages in terms of time efficiency, safety, and cost compared to conventional manual work [6], their utilization is expected in various processes, ranging from the construction planning stage to progress management and post-completion maintenance and inspection [7].

One factor supporting this proliferation is the integration with diverse onboard sensors. While visible-light RGB cameras remain fundamental sensors, drones equipped with thermal cameras are already becoming common in applications such as the detection of structural cracks [8] and tile detachment [9], as well as damage assessment at disaster sites [10].

Furthermore, attempts to integrate drones and mixed reality (MR) technology [11] have recently attracted significant attention. MR is a technology that superimposes virtual 3D information onto real

 doi:10.20944/preprints202601.0625.v1

space in real-time, enabling intuitive visualization and interaction. In the AEC field, various applications have been proposed, such as visualization to confirm landscapes by superimposing future buildings onto vacant land [12], comparison between Building Information Modeling (BIM) and actual structures [13], and superimposed display of inspection results onto facility walls [9,14]. When applying MR to large-scale spaces such as high-rise buildings or urban blocks, or to locations inaccessible to human hands, it is often difficult to grasp the entire target solely from a ground-level viewpoint (e.g., via handheld terminals). In such situations, realizing MR from a drone-view, which moves freely within 3D space, is extremely effective for understanding the entire site comprehensively and intuitively [13].

However, robustly realizing drone-view MR is challenging. The primary technical hurdle is the requirement for high-precision and stabilized self-localization. While it is possible to satisfy this required accuracy by integrating a Global Navigation Satellite System (GNSS) or Real-Time Kinematic GNSS (RTK-GNSS) with the Inertial Measurement Unit (IMU) built into the airframe, accessing position and attitude information typically requires coordination with model-specific Software Development Kits (SDKs) or dedicated airframes. In fact, drone-view MR systems developed by Wen and Kang [12] and Botrugn et al. [15] utilize airframes designed specifically for the project or model-specific SDKs; this reduces system versatility and significantly constrains use cases. Furthermore, the use of GNSS is often restricted in building canyons between urban high-rise buildings or at indoor/semi-indoor construction sites.

To circumvent such dependence on proprietary SDKs and dedicated airframes, methods realizing MR using only general-purpose technologies such as video streaming and screen sharing have been proposed. For example, Kikuchi et al. [16] integrated commercially available drones and MR using general technologies; however, to align the real and virtual spaces, it was necessary to pre-define the flight route and orientation, resulting in the constraint that the trajectory could not be freely changed during MR execution.

Against this background, vision-based approaches using only camera images are attracting attention as a highly versatile solution. For example, Kinoshita et al. [17] realized MR with free flight paths without relying on specific SDKs, using only RGB images and a pre-constructed 3D map. However, a significant limitation exists: methods relying solely on RGB images are vulnerable to texture-less environments, which are ubiquitous at AEC sites [18]. On homogeneous surfaces such as exposed concrete, painted walls, and metal panels, tracking often fails due to a lack of visual feature points, causing significant drift in position estimation.

For this vulnerability in texture-less environments, infrared (thermal) thermography provides a promising physical solution. Unlike RGB cameras that capture visible light reflected from surfaces, thermal sensors detect infrared radiation emitted by materials based on their emissivity and heat capacity [19]. This fundamental difference allows thermal cameras to perceive minute temperature unevenness—referred to as “thermal texture”—even on visually homogeneous surfaces. However, to integrate thermal information into MR systems, two essential challenges must be addressed. The first is temporal variance; since thermal distribution changes drastically due to solar radiation and weather conditions, utilizing a pre-constructed thermal map as a permanent coordinate reference is inappropriate for MR. The second is the computational cost. Many existing RGB-thermal fusion methods require complex feature extraction via deep learning [20] or dense optimization using all pixels. In the context of MR superimposition, where low-latency processing is critical, the computational load of these methods is excessive for real-time execution on limited hardware resources.

Therefore, this study proposes a novel localization method for drone-view MR that adaptively integrates RGB and thermal images. This method primarily utilizes an RGB map-based Visual Positioning System (VPS) for absolute position estimation under general conditions and adaptively fuses RGB and thermal data for relative motion estimation. Through this adaptive sensor fusion, we realize high-precision and seamless position estimation that serves as the foundation for drone-view MR systems, even in challenging environments where conventional RGB-only methods fail.

 doi:10.20944/preprints202601.0625.v1

The main contributions of this paper are as follows:

- Decoupled and Adaptive Fusion Architecture: We propose a hybrid architecture that separates absolute position correction using RGB and relative tracking using RGB-thermal fusion. This enables the system to utilize the robustness of thermal images for MR tracking while avoiding the instabilities associated with their temporal variability.
- Empirical Validation in AEC Scenarios: We implemented the proposed system and demonstrated that stable MR superimposition is possible in proximity scenarios involving texture-less building exterior walls, even in environments where conventional methods fail to maintain tracking.

The structure of this paper is as follows. Section 2 provides an overview of related work. Section 3 describes the details of the proposed method, and Section 4 explains the implementation of the prototype. Section 5 reports the experimental results. Following the discussion in Section 6, the conclusion is presented in Section 7.

2. Literature Review

2.1. Drone-Based Mixed Reality in AEC

Research on the integration of drones and AR/MR is diverse, but it can be broadly classified into two categories: data collection for MR content creation and MR utilization for drone operations. In this study, we focus on the latter, specifically drone-viewpoint MR that adds virtual information to drone camera footage. The basic requirement for such systems is registration that matches the coordinates of the real camera and the virtual camera with high precision.

Early approaches tended to sacrifice system flexibility to ensure accuracy. Wen and Kang [12] realized aerial MR for construction site simulation using the Vuforia SDK. While this method enabled free flight paths, it required building a project-specific airframe to access the necessary sensor data, which reduced system versatility. Similarly, the method by Botrugno et al. [15] also depends on specific hardware configurations, limiting deployment to diverse AEC projects.

To enhance versatility, attempts were made in subsequent research to eliminate dependence on dedicated SDKs. Raimbaud et al. [13] implemented an interactive MR application that superimposes BIM models onto real video for quality control. While operating without a dedicated SDK, the registration of reality and virtuality required manual adjustment of rotation and translation, resulting in a high operational load. Kikuchi et al. [16] realized occlusion-aware MR using commercially available drones and general streaming technologies. Although hardware versatility was high, it was necessary to strictly define the drone's start time, flight route, and orientation in advance for registration. Due to this constraint, the pilot could not freely change the path during operation, making application to dynamic inspection tasks difficult.

More recently, VPS and Simultaneous Localization and Mapping (SLAM) technologies have been applied to balance hardware independence and flight flexibility. Kinoshita et al. [17] combined absolute position estimation via VPS and natural feature point tracking to realize drone MR capable of free flight without a specific SDK. However, a significant limitation exists in this method. It is the reliance solely on RGB images. At AEC sites, there are many texture-less surfaces, such as concrete walls, roofs and roads, where RGB-based tracking fails. This study aims to solve this issue through the integration of thermal information; however, simply adding a sensor is insufficient, and an integration method that satisfies MR-specific requirements (permanence of the coordinate system) is required.

 doi:10.20944/preprints202601.0625.v1

2.2. Visual Localization Challenges: Texture-less & Geometric Burstiness

2.2.1. VPS and SLAM

In recent years, VPS and Visual SLAM have been widely studied as image-based localization methods for achieving high-precision spatial registration in MR applications. VPS is a method that estimates the 6 degrees of freedom (6DoF) camera pose by pre-constructing a three-dimensional (3D) map of the target environment and associating monocular camera images acquired during operation with 3D points in the known map [21,22]. In this framework, a method is generally used that obtains the camera position and orientation by extracting local features from a query image, estimating the correspondence with feature points on the 3D map, and then solving the Perspective-n-Point (PnP) problem [23]. It has been reported that when sufficiently distributed correspondence points are obtained, high-precision pose estimation on the order of centimeters is possible, making it suitable for high-precision MR superimposition. Such a VPS framework is considered to have high practicality in MR applications in real environments because it can be realized with a relatively lightweight configuration without requiring additional sensors. In fact, it is widely deployed as commercial services such as Immersal VPS [24], Lightship VPS [25], and ARCore Geospatial API [26], and is utilized as a global position estimation infrastructure regardless of whether it is outdoors or indoors.

On the other hand, Visual SLAM, represented by ORB-SLAM [27], is an approach that performs localization and mapping simultaneously, and is characterized by the fact that it does not require prior map construction [27,28]. In Visual SLAM, since the environment map is generated and updated online as the camera moves, there is no need for the prior cost of map construction, and it has the advantage of being able to flexibly adapt to changes such as aging of the environment and the installation of temporary structures. For this reason, it has been widely used in fields that emphasize grasping relative positional relationships, such as robotics and navigation.

Thus, VPS and Visual SLAM are technologies with different premises and characteristics, and selective use according to the application is required. It has been pointed out that VPS, which provides a globally stable coordinate system, is effective for systems that perform MR display based on a consistent world coordinate system over a long period [29,30]. In VPS, since position estimation is performed within a pre-constructed global map coordinate system, there is an advantage that once the correspondence between the MR coordinate system and the map coordinate system is defined, that correspondence can be reused repeatedly.

Based on the above previous studies, in this study, we focus on a position estimation method based on VPS from the viewpoint of emphasizing high-precision consistency with design data and long-term stable MR display.

2.2.2. Visual Localization Challenges in AEC

As previously mentioned, both VPS and Visual SLAM are based on feature point extraction and matching from images. However, at AEC sites, there are widely existing texture-less surfaces that lack the high-contrast visual patterns necessary for robust feature detection, such as freshly poured concrete walls, unpainted gypsum boards, smooth metal panels, and glass curtain walls [18]. Furthermore, even in areas where feature points are detectable, their spatial distribution may compromise the reliability of position estimation. Sattler et al. [21] defined the phenomenon where feature points are detected concentrated in local regions of an image as Geometric Burstiness and pointed out that this is a serious challenge for visual localization. In construction sites, this phenomenon becomes pronounced when textures such as vents, stains, or background vegetation exist locally within vast texture-less wall surfaces. Under such circumstances, even if algorithms such as Random Sample Consensus (RANSAC) detect an apparently large number of inliers (matching points), the configuration becomes geometrically degenerate, causing visual odometry to lose tracking or the PnP solver to become unstable, leading to significant drift or complete failure of position estimation during close-range inspection [31].

 doi:10.20944/preprints202601.0625.v1

As one direction to address this problem, methods utilizing geometric primitives such as line segments or planes instead of point features have been proposed. Structure-based methods using PL-SLAM or the Manhattan World assumption (an assumption modeling the environment as a set of orthogonal planes) are representative examples, and improved robustness in architectural environments has been reported [18]. While these methods are effective for corridors and facades where clear edges and contours exist, they remain vulnerable in situations where a drone navigates the center of vast flat surfaces where geometric primitives themselves are scarce or ambiguous.

As another approach, methods estimating unconstrained Optical Flow integrating rigid body motion and general object motion through learning have been proposed [32]. Such machine learning-based Optical Flow estimation is reported to be able to achieve relatively good motion estimation even in low-texture regions by utilizing learned priors and spatial smoothness. However, many of them require large-scale neural networks, and the computational load is extremely high. In general GPU environments, it is not easy to achieve real-time processing at the resolution and frame rate required for MR applications.

Based on the above, a fundamental trade-off exists in RGB-based position estimation methods for AEC environments. That is, methods such as VPS, which balance the accuracy and operational cost required for MR, are prone to fatal failure on texture-less surfaces, while advanced methods such as learning-based optical flow, which are robust to texture-less conditions, have high computational loads and are not necessarily suitable from the viewpoints of real-time performance for MR and global registration.

2.3. Thermal Imaging for Visual Localization

Thermal imaging has attracted attention in recent years as a localization method in environments where vision is obstructed by smoke, such as fire scenes, or in low-light environments such as nighttime. Under such conditions, while conventional RGB cameras are affected by low contrast or strong illumination fluctuations, thermal cameras can stably acquire patterns based on temperature differences [33,34].

In early research, the feasibility of thermal inertial odometry was demonstrated by directly utilizing radiometric information. For example, Borges et al. [35] proposed a method combining semi-dense optical flow of monocular thermal images with an IMU, recovering scale via road surface detection. By leveraging edge information in thermal images, low-cost and robust pose estimation was realized, demonstrating effectiveness in ground vehicles.

On the other hand, feature-point-based approaches are also evolving. Li et al. [33] proposed a new infrared SLAM algorithm integrating feature point methods and optical flow technology, addressing the problem of cumulative errors caused by weak textures in thermal images.

Furthermore, in recent years, deep learning-based approaches have also been investigated. In this domain, Transformers have been applied to enhance feature robustness. Zhang et al. [20] proposed template-guided low-rank adaptation (TGATrack) to improve RGB-T tracking. While these data-driven methods demonstrate high performance in recognition tasks, they often entail high computational costs for real-time processing.

In addition to system-level research, fundamental performance evaluations of thermal features have also been conducted. Johansson et al. [36] and Mouats et al. [37] conducted comprehensive benchmarks of feature point detectors and descriptors for thermal images. According to their evaluations, it is reported that combinations of Oriented FAST and Rotated BRIEF (ORB) and Fast Retina Keypoint (FREAK) or Binary Robust Invariant Scalable Keypoints (BRISK), or combinations of Speeded Up Robust Features (SURF) and FREAK, show relatively good performance. However, they also concluded that matching using visible light images was still superior under conditions with good visibility and sufficient texture. This suggests that while thermal features are robust against illumination changes, they have lower discriminative power compared to RGB features under standard environments, supporting the view that thermal images should not simply replace RGB but adaptively complement it.

 doi:10.20944/preprints202601.0625.v1

However, when viewing existing research from the perspective of realizing high-precision MR in the AEC field, there are decisive mismatches in both the application domain and system configuration. First, many existing RGB-T frameworks focus primarily on autonomous robot navigation in nighttime or smoke environments [38], and do not focus on the drift problem caused by photometric homogeneity, such as concrete and metal panels, which is ubiquitous in daytime environments at AEC sites. Second, existing Thermal SLAM methods [33] construct maps using thermal images, but this ignores the fatal problem of the lack of map reusability due to the temporal variance of thermal information mentioned in the Introduction. Since a consistent coordinate reference is required throughout the process in MR applications, unstable thermal maps are unacceptable.

Therefore, the unresolved challenge lies in the establishment of Decoupled Adaptive Fusion, which uses thermal information to complement only local tracking robustness in texture-less regions while maintaining a high-precision global coordinate reference via RGB. By proposing this asymmetric hybrid architecture, this study overcomes the practical challenges faced by existing Thermal SLAM and Deep Learning methods and establishes the reliability of drone MR at AEC sites.

2.4. Multi-Modal Sensor Fusion and Uncertainty Estimation Strategies

2.4.1. Tightly-Coupled vs. Loosely-coupled Architectures

In recent research on RGB-Thermal Odometry and SLAM, tightly-coupled approaches—represented by VINS-Fusion [39] and LVI-SAM [40]—have become the dominant paradigm due to their mathematical optimality under ideal conditions. However, recent studies focusing on extreme environments have revealed that tight coupling possesses structural vulnerabilities in dynamic and severely degraded settings [41]. The primary risk associated with tight coupling is the "Ripple Effect" of sensor failure [42]. In such systems, where all sensors contribute to a single state vector, outliers—arising from phenomena such as gradient vanishing due to "Thermal Equilibrium" in thermal cameras or "Washout" in RGB cameras—can propagate through the system, leading to Catastrophic Failure. Furthermore, since the computational complexity of nonlinear optimization scales cubically with the number of feature points ($O(N^3)$) [43], the high computational load poses a significant challenge for real-time robotic applications.

In contrast, loosely-coupled approaches adopt a decentralized architecture that estimates the odometry of each modality independently. As demonstrated in disaster response missions such as the DARPA SubT Challenge, this modularity provides "Fault Tolerance" [41]. By functioning as a "bulkhead" that immediately gates (isolates) malfunctioning sensor inputs, this architecture allows the system to survive using the remaining healthy sensors. In tasks within complex AEC environments, this survivability—the ability to maintain estimation even during partial sensor blackouts—is a critical attribute that must take precedence over the theoretical maximization of accuracy.

2.4.2. Uncertainty Estimation: Deep Learning vs. Geometric Approaches

For a loosely-coupled system to function robustly, it is necessary to dynamically and accurately estimate the reliability (Uncertainty/Weight) of each sensor. Ideally, the observation noise covariance matrix should be dynamically estimated in a Extended Kalman Filter (EKF); however, in texture-less environments, the problem of Over-confidence has been reported, where erroneous correspondence points (Outliers) are misidentified as true values, causing the covariance to converge excessively [44]. Furthermore, Deep Learning-based methods such as DROID-SLAM [45] can estimate per-pixel uncertainty, but this consumes a large amount of GPU resources. Additionally, the risk of exhibiting unpredictable behavior in unknown environments not included in the training data cannot be ignored.

 doi:10.20944/preprints202601.0625.v1

3. Proposed Methods

3.1. System Overview

The method by Kinoshita et al. [17] is an effective framework that combines RGB-based VPS and monocular tracking; however, in texture-less construction sites, RGB tracking frequently fails, leading to the challenge that position estimation is interrupted between VPS updates. To overcome this challenge, this study proposes an adaptive fusion architecture that adds a parallel tracking layer using thermal images, enabling continuous position estimation even in environments with scarce visual features.

The flow diagram of the processing of the proposed system is shown in Figure 1.

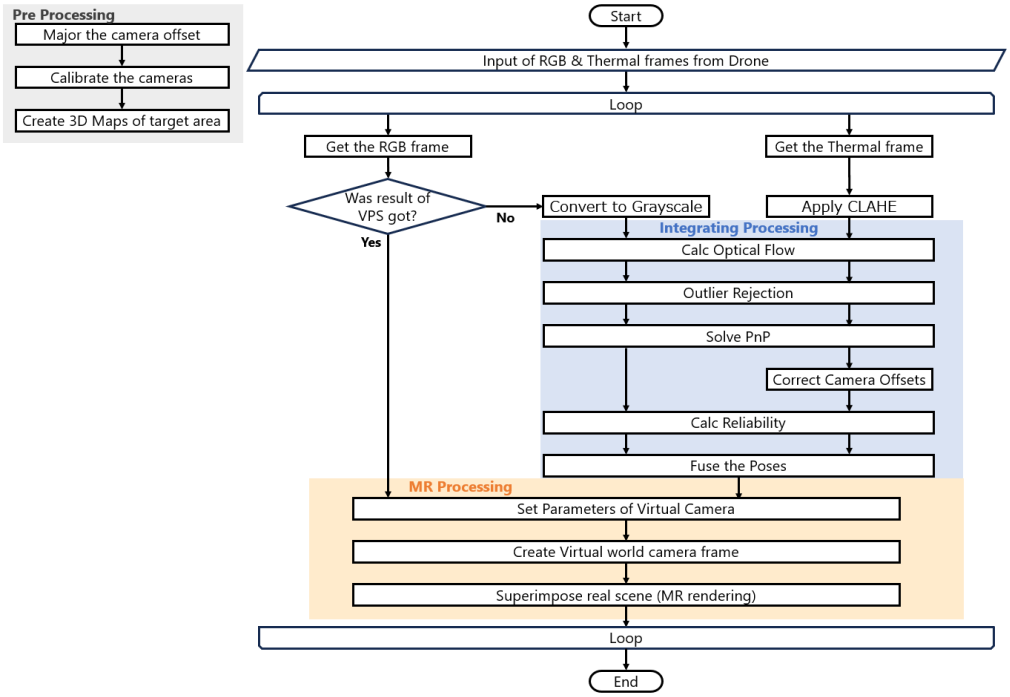


Figure 1. Flow diagram of the proposed system.

This system is an extension of the MR framework established by Kinoshita et al. [17]. In the method of Kinoshita et al., absolute position (Localization) is periodically estimated by RGB-based VPS using a pre-constructed 3D map, and the gaps between VPS updates are complemented by natural feature point tracking (Tracking) of RGB images. While following this basic structure, the proposed method extends the local tracking process from monocular RGB to RGB-Thermal parallel tracking. The system acquires synchronized footage from RGB and Thermal cameras mounted on the drone and performs processing in the following two loops:

- Global Initialization Loop: To reset drift errors, VPS is executed asynchronously using RGB keyframes to estimate the 6DoF pose in the absolute coordinate system.
- Adaptive Dual-Tracking Loop: To fill the update interval (latency) of VPS, relative pose displacement between frames is estimated in two independent pipelines, RGB and Thermal, and fused adaptively based on reliability.

3.2. Geometric and Radiometric Pre-Processing

3.2.1. Geometric Rectification and Calibration

In this system, the RGB camera and the thermal camera are arranged with a physically fixed baseline (offset). To maximize the accuracy of optical flow and ensure pixel correspondence between heterogeneous sensors, the following geometric corrections are performed. First, using pre-acquired

 doi:10.20944/preprints202601.0625.v1

intrinsic parameters and distortion coefficients, lens distortion of each image is corrected (undistorted). This ensures compatibility with the pinhole model where straight lines are projected as straight lines, reducing tracking errors in optical flow estimation [46].

Next, to integrate the estimated pose $T_{thermal}$ of the thermal camera into the RGB coordinate system, the pre-calibrated extrinsic parameter (rigid body transformation matrix) T_{offset} is applied.

3.2.2. Asymmetric Radiometric Enhancement

To maximize feature tracking capabilities in each modality while minimizing the influence of sensor-specific noise, the following asymmetric image processing is applied:

- RGB Stream: Under general daytime environments, RGB sensors have a sufficient dynamic range. Since excessive enhancement processing may compromise the brightness constancy assumption in optical flow, only standard grayscale conversion is performed to preserve the raw luminance distribution.
- Thermal Stream: Thermal images have the characteristic that histograms concentrate in a narrow range for isothermal building materials such as concrete walls, resulting in extremely low contrast. To address this, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied. Unlike standard histogram equalization (HE), CLAHE effectively reveals only the thermal gradients (thermal texture) within the material while suppressing the amplification of particle noise characteristic of thermal sensors by setting a limit (Clip Limit) on the degree of contrast enhancement [47].

3.3. Independent Dual-Pipeline Tracking

The core of this method is a parallel pipeline structure that independently performs pose estimation for each of the RGB and thermal images. By performing geometric verification individually for each sensor rather than using simple image composition (early fusion), the influence when one sensor fails is eliminated. The processing for each sensor is conducted according to the following steps:

1. Optical Flow Estimation: For the undistorted images, feature points from the previous frame are tracked to the current frame using the Lucas-Kanade method [48] or similar.
2. Outlier Rejection: Geometric verification using RANSAC is performed to remove outliers due to moving objects or mismatches.
3. Pose Estimation (PnP): The PnP problem is solved using the feature points remaining as inliers and the intrinsic parameter matrix K specific to each camera, and the camera movement amount (relative pose) is calculated.

3.4. Adaptive Pose Fusion

This section describes the method for adaptively integrating the individual pose estimation results obtained from RGB and Thermal images in an environment-adaptive manner.

3.4.1. Reliability Metric based on Geometric Consistency

The design of the reliability metric is a crucial element that determines the robustness of the fusion. Conventionally, image gradient intensity (Gradient Magnitude) and the number of inliers by RANSAC have been used as reliability metrics [27]. However, in texture-less environments such as construction sites, these metrics often have the problem of outputting erroneously high reliability. For example, when a phenomenon occurs where feature points concentrate in a narrow area (Geometric Burstiness), such as on fences, vegetation, or stains on concrete walls, the apparent number of inliers increases despite the low geometric constraint [21].

If pose estimation is performed without considering such biased distribution, it leads to drift in a specific direction or instability in estimation. Therefore, in this study, to evaluate not only the number of feature points but also the spatial distribution simultaneously, the effective inlier count

 doi:10.20944/preprints202601.0625.v1

(N_{eff}) proposed by Sattler et al. [21] is adopted as the reliability metric. The image area is divided into an $N \times N$ grid, and only if at least one inlier exists within each grid cell $C_{i,j}$, that cell is counted as effective:

$$N_{eff,s} = \sum_{i=1}^{M}\sum_{j=1}^{N} \min(C_{i,j}, 1)$$
$$Rel_s = \begin{cases} N_{eff,s} & (N_{eff,s} > \tau) \\ 0 & (N_{eff,s} < \tau) \end{cases} \quad (1)$$

where τ is the confidence threshold for guaranteeing geometric stability, and the subscript $s \in \{RGB, IR\}$ denotes the sensor modality. By using the metric, it becomes possible to eliminate mis-evaluations due to local bursts and selectively utilize only high-quality frames where feature points are dispersed throughout the entire screen.

3.4.2. Spatial Pose Fusion Strategy

Integration is performed according to the following procedure. First, the estimated pose $P_{thermal}$ of the thermal camera is transformed into the RGB coordinate system using the offset T_{offset} defined in Section 3.2.1 to obtain the transformed pose $P'_{thermal}$. Next, an adaptive fusion weight α_s for each sensor modality based on the number of effective inliers for each sensor is calculated. This is an approximation of Inverse Variance Weighting based on the assumption that the uncertainty of pose estimation is inversely proportional to the number of effective geometric constraints.

$$\alpha_s = \frac{Rel_s}{Rel_{RGB} + Rel_{Thermal}}$$
$$\alpha_{RGB} + \alpha_{Thermal} = 1 \quad (2)$$

The fused pose P_{fused} and R_{fused} are defined as follows.

- Translation: Since this is between two points in Euclidean space, linear interpolation (Lerp) is used.

$$P_{fused} = \alpha_{RGB} P_{RGB} + \alpha_{Thermal} P'_{Thermal} \quad (3)$$

- Rotation: In a unit quaternion space representing 3D rotation, simple linear interpolation leads to distortion of the rotation speed and norm failure. Therefore, Spherical Linear Interpolation (Slerp), which is a mathematically rigorous rotation interpolation, is adopted [49].

$$R_{fused} = Slerp(R_{rgb}, R'_{Thermal}, \alpha_{Thermal}) \quad (4)$$

By this logic, when both RGB and Thermal are valid, they are optimally blended according to their respective geometric strengths; if RGB is lost and only Thermal survives, $\alpha_{\text{Thermal}} = 1.0$ and the system automatically transitions seamlessly to Thermal-only tracking.

4. Implementation of prototype system

This section describes the hardware configuration, software environment, and detailed implementation parameters of the prototype system used for the validation experiments of the proposed method. Details of the equipment used for the implementation of the prototype system are shown in Table 1, details of the software and services are shown in Table 2, and a conceptual image of the development environment integration is shown in Figure 2.

 doi:10.20944/preprints202601.0625.v1

Table 1. Environment of the prototype system.

Purpose of use in the prototype	Name	Device information	
Server PC	Home-built desktop PC	OS	Microsoft Windows 11 Education
		CPU	Intel(R) Core(TM) i7-9700 CPU @ 3.00GHz
		RAM	32.0 GB
		GPU	GeForce RTX 2070
Drone	DJI Mavic 3T	Weight	920 g
		Focal length	24 mm (RGB) 40 mm (Thermal)
		Max resolution	3840 × 2160 (RGB) 640 × 512 (Thermal)

Table 2. Software/services used in the prototype system.

Software/service	Purpose of use in the prototype	General usage
Unity 2022.3.34f1	MR scene rendering and visualization in a virtual environment	Game engine / real-time 3D development platform
Immersal SDK 2.2.1	VPS execution	VPS / localization SDK
OpenCV for Unity 3.0.1	Image processing and feature tracking within the Unity runtime	Computer vision library integration for Unity
Agisoft Metashape	SfM-based camera poses estimation and 3D map reconstruction	Structure-from-Motion (SfM) / photogrammetry software

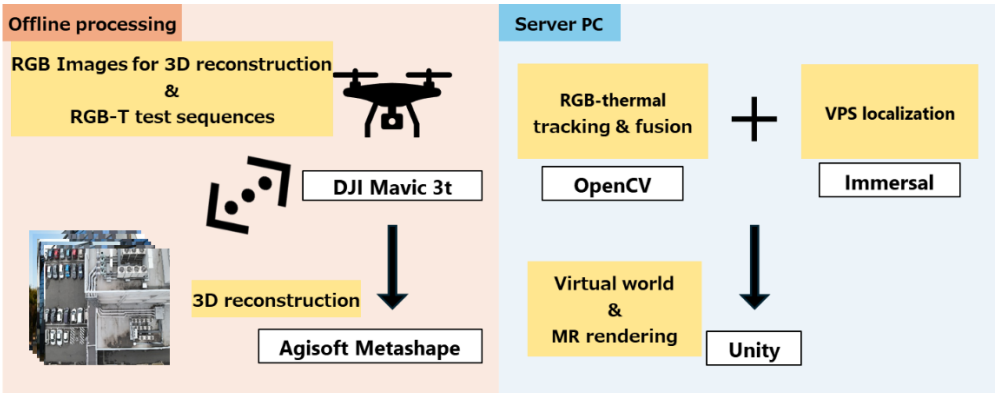


Figure 2. Development Environment Integration Concept.

4.1. Implementation of Prototype System

The basic architecture for global self-localization follows the asynchronous remote rendering framework established by Kinoshita et al. [17]. Specifically, the processing flow of VPS requests, the latency compensation logic, and the registration of virtual cameras for MR on the Unity scene using these methods follow their approach. The system operates in the following two parallel threads:

Global localization thread: Keyframes are sent asynchronously to the Immersal VPS server. Since a delay of 1 to 3 seconds occurs due to network transmission and server processing, directly applying the returned pose would result in visual inconsistency with the current frame.

Real-time compensation thread: To compensate for this latency, the system continuously tracks the relative movement of the camera using 2D-3D point matching and optical flow. Upon receiving the delayed VPS pose P_{vps} at time t_{now}, the system calculates the accumulated displacement ΔP from the tracking since the transmission time t_{sent} to the present. The corrected global pose P_{global} is calculated as $P_{global} = P_{vps} \oplus \Delta P$. The unique contribution of this study lies in implementing RGB-Thermal adaptive tracking in this thread to enhance robustness in texture-less environments.

 doi:10.20944/preprints202601.0625.v1

Through this approach, the position and orientation of the virtual camera in Unity maintain constant synchronization with the live video, regardless of the VPS response time.

4.2. Dual-Sensor Pre-Processing

In heterogeneous sensor fusion, ensuring geometric consistency between the RGB camera and the thermal camera is indispensable. In particular, the thermal camera of the DJI Mavic 3T has a low resolution (640x512), and since sensor cropping occurs in the dual-screen video mode with RGB, it possesses intrinsic parameters that differ from the nominal parameters (Table 3).

Table 3. Comparison of native and effective sensor resolutions in dual-stream visualization mode.

	Native Resolution [px]	**Dual-Stream Resolution [px]**
Thermal camera	640×518	960×768
RGB camera	3840×2160	960×768

To address this, the following calibration methods were implemented.

- RGB camera: Using Zhang's method [50] with a standard chessboard pattern, intrinsic parameters and distortion coefficients were calculated using the OpenCV library.
- Thermal camera: Thermal cameras face a challenge where the detection accuracy of chessboard corners significantly decreases due to low resolution and the influence of thermal diffusion [51]. Therefore, in this study, as a more practical and high-precision method, self-calibration applying Structure-from-Motion (SfM) technology was adopted. Specifically, aerial photography of outdoor structures rich in thermal texture was conducted using only the thermal camera, and intrinsic parameters were retrieved via inverse estimation through a 3D reconstruction process using Agisoft Metashape. This approach yielded high-precision parameters suitable for the actual operational environment (Table 4).
- Extrinsic parameters: The extrinsic parameters (rigid body transformation matrix) between both cameras were fixed with a translation vector $T_{offset} = [-0.018,0,0]^T\ [m]$ based on the physical design blueprints of the gimbal mechanism.

During online processing, lens distortion correction and parallax correction for both images are applied in real-time using these parameters.

Table 4. Estimated intrinsic parameters for the RGB and thermal cameras.

	Focal Length $(f_x = f_y)$ **[px]**	**Principal Point** (c_x, c_y) **[px]**	**Distortion** $(k_1, k_2, p_1, p_2, k_3)$
RGB Camera	1145.785	$(476.151, 384.036)$	$(-0.3548, 0.1350, -0.0001, -0.0002, -0.1049)$
Thermal Camera	1047.238	$(500.845, 383.183)$	$(0.3277, -1.3227, -0.0008, -0.0004, 0.8015)$

4.3. Implementation Details of Adaptive Tracking

This section details the specific engineering parameters and optimization methods applied to realize the adaptive tracking logic described in Section 3.

- Radiometric Enhancement: The CLAHE algorithm defined in Section 3 was implemented using the Imgproc.createCLAHE function in OpenCV. Following the standard configuration recommended in the literature [47], the Clip Limit was set to 2.0, and the tile grid size was set to 8×8.
- Tracking Parameters: The Lucas-Kanade tracker was configured with a search window size of 71×71 pixels and a pyramid level of 5 to accommodate rapid drone maneuvers. Furthermore, the RANSAC reprojection error threshold for outlier rejection was empirically set to 5.0 pixels.
- Fusion Logic Execution: For reliability evaluation, the Effective Inlier Count (N_{eff}) defined in Section 3 is used. Following the methodology of Sattler et al. [21] and, which suggests a cell size of approximately 50×50 pixels to ensure spatial diversity, we partitioned our 960×768

 doi:10.20944/preprints202601.0625.v1

resolution frames into a 20×16 grid. This results in an exact cell size of 48×48 pixels, providing a balanced spatial constraint across the entire field of view. In this implementation, considering that the number of feature points required for robust initialization in standard Visual SLAM, such as ORB-SLAM, is 30 to 50 [27], a conservative value of $\tau = 50$ was adopted. If the N_{eff} of either sensor falls below 50, the system immediately sets the reliability of that sensor to zero, thereby excluding unstable estimation results that cause drift from the fusion.

4.4. Experimental Setup Using Pre-Recorded Datasets

While the proposed architecture supports real-time wireless video transmission, offline evaluation using pre-recorded datasets was adopted for the experimental validation in this study.

The reason for adopting this protocol is to eliminate variability in flight trajectories and environmental conditions (wind, illumination changes, etc.) and to verify the proposed method against conventional methods (RGB-only, Thermal-only) under Identical Conditions. By inputting lossless recorded RGB and thermal video streams on the drone into the server pipeline, it was possible to purely evaluate the processing performance and robustness of the algorithm. Note that the influence of latency caused by network transmission has already been quantified and verified by Kinoshita et al. [17] and is therefore excluded from the scope of this experiment.

5. Verification Test

In this section, using the prototype system constructed in Section 4, we demonstrate that the proposed method achieves comparable performance in general environments and significantly improved robustness in challenging scenarios over the existing method [17] in diverse environments. The validation was conducted in two major phases. First is the quantitative comparison of MR registration accuracy between the proposed system and the existing system. Both algorithms were applied to the same recorded video to verify the fundamental accuracy in general flight environments. Second is the validation of robustness in environments that are harsh for RGB cameras (texture-less). Here, a durability test was performed where initialization by VPS was successful only once, after which registration was continued using only tracking, and the transition of the accuracy was analyzed chronologically. In this experiment, the drone functioned as an image acquisition device, and all computational processing was performed on a server PC after video transmission.

5.1. Comparison of Positioning Accuracy

In MR, geometric registration between the real space and the virtual space is one of the most important requirements [52]. Therefore, it is essential to verify the fundamental registration accuracy of the proposed system. In this validation, IoU (Intersection over Union) was adopted as a quantitative index for registration accuracy. IoU is an evaluation scale generally used in benchmarks for object detection and similar tasks, and is defined by the following equation [53]:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

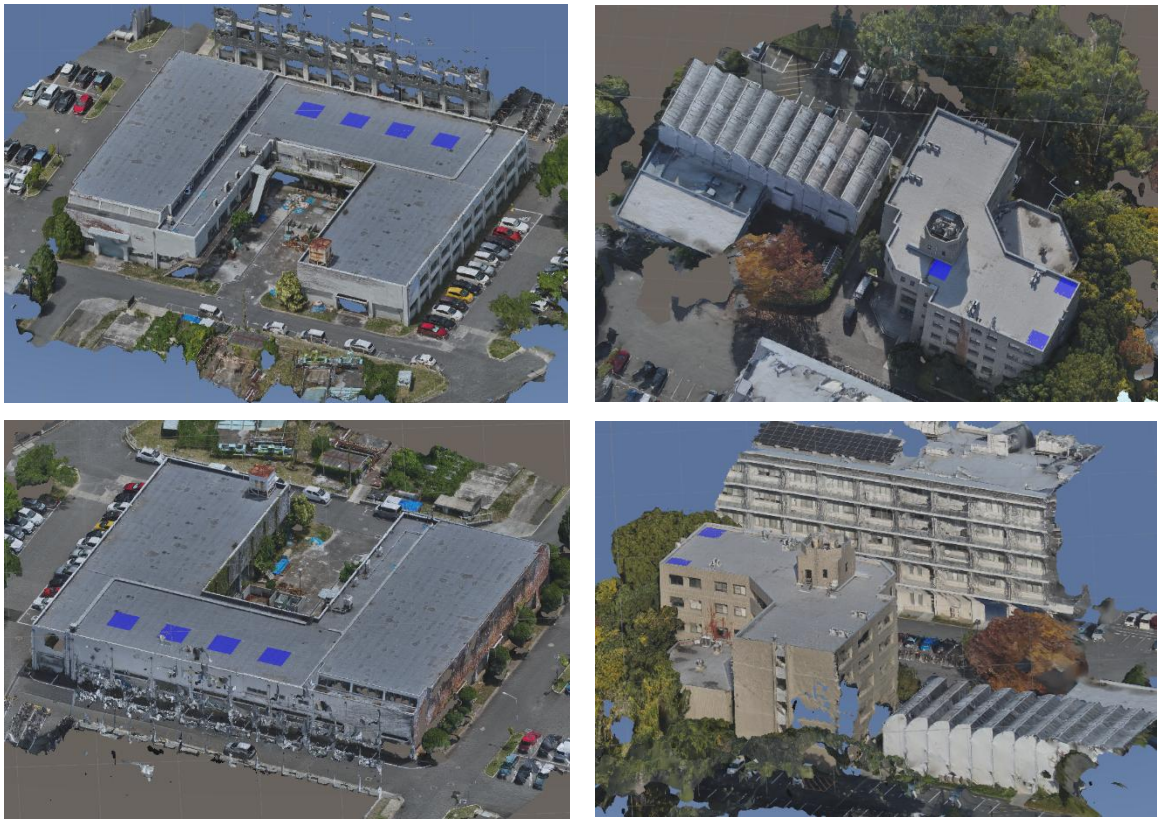
where, TP (True Positive) represents the number of pixels in correctly detected regions of interest (ROI), FP (False Positive) represents erroneously detected non-interest regions, and FN (False Negative) represents regions of interest that were missed. In MR, geometric registration can be indexed by setting an ROI on an object in the real space and evaluating the degree of overlap with the corresponding virtual object (semi-transparent monochromatic texture). In this validation, an IoU closer to 1 is judged as higher registration accuracy.

As the validation dataset, video was recorded using manual drone piloting at the Suita Campus of The University of Osaka (2-1 Yamadaoka, Suita, Osaka 565-0871, Japan). The data acquisition was conducted specifically at Building S2 for Routes A and B, and Building M4 for Route C, along the following three types of routes:

 doi:10.20944/preprints202601.0625.v1

- Route A (Texture-Rich): An overhead viewpoint from above Building S2. A general environment includes diverse objects such as buildings, roads, and vegetation. It follows an elliptical orbit after translation.
- Route B (Texture-less / Bursty): An environment in close proximity to the rooftop of Building S2. Homogeneous concrete surfaces occupy most of the field of view, accompanied by rapid frame changes due to drone movement.
- Route C (Long-term / Specular): An environment involving close-up photography of the rooftop and walls of Building M4. Although it presents low-texture characteristics similar to Route B, the visibility of tiled walls and surrounding vegetation renders it less challenging than the completely homogeneous environment of Building S2. However, it introduces other difficulties, such as specular reflections from water puddles and repetitions of similar shapes.

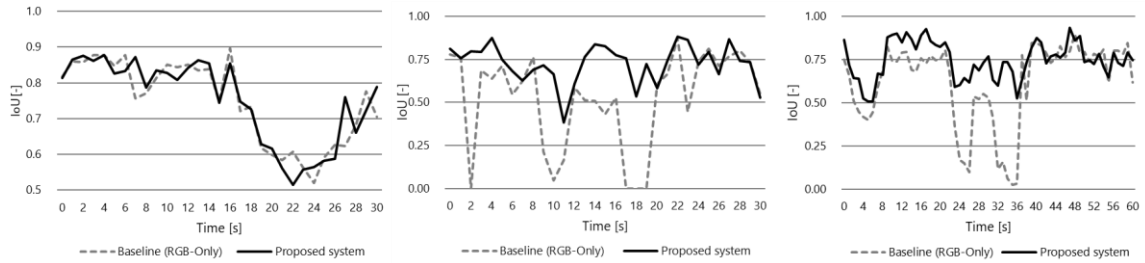
Route A is for control experiments where tracking is easy even with RGB alone, while Routes B and C assume environments that are harsh for RGB. For the recorded video of each route, MR superimposition was performed using the proposed system and the existing system, and the average IoU was calculated approximately every second. For validation, virtual 3D models (blue semi-transparent) arranged at equal intervals were used (see Figure 3), and the IoU with mask images created using the image editing software LabelMe was measured. Figure 4 and Table 5 show the measured IoU results. While both methods showed high accuracy in Route A, the proposed method recorded accuracy significantly exceeding that of the existing method in Routes B and C.



(a) 3D Map used for Routes A and B (b) 3D Map used for Route C

Figure 3. 3D maps used for (a) Routes A and B, (b) Route C. The semi-transparent blue squares placed on the building rooftops are the objects for MR display.

 doi:10.20944/preprints202601.0625.v1



(a) Route A (b) Route B (c) Route C

Figure 4. Comparison of the proposed system and the baseline method in Routes A–C (IoU).

Table 5. Average IoU in each route.

	Baseline Method (IoU)	Proposed system (IoU)
Route A	0.747	0.750
Route B	0.510	0.725
Route C	0.62	0.735

5.2. Durability Analysis in Communication-Denied or VPS Lost Environments

The proposed system typically compensates for cumulative error (drift) through periodic VPS requests. However, in real-world environments such as infrastructure inspections, it is expected that scenarios where VPS itself fails due to insufficient texture, or where verification with the server is impossible due to communication delays or disruptions, will occur frequently. Therefore, in this section, a durability test was conducted on Route B and Route C, where initialization by VPS was performed only once at the beginning, after which registration was maintained solely through pure tracking. This evaluates the geometric robustness of the system during VPS failure or in communication-denied environments.

5.2.1. Analysis of Route B: Mutual Rescue in Geometric Burstiness

The time-series data for Route B (effective inlier count, fusion weight, IoU) is shown in Figure 5, and a comparison of the tracking status is shown in Figure 6.

Of particular note is the behavior near $t \approx 1[s]$ immediately after the start of the experiment. At this point, the number of RGB effective inliers dropped to 46 (Figure 5 (a)). This is a value below the reliability threshold ($\tau = 50$) set for this prototype system (Section 4.3), suggesting that geometric information from the RGB image is insufficient for position estimation. The proposed system responded immediately to this critical situation. The thermal weight at $t = 1[s]$ reached 1.0 (Figure 5 (b)). Given the normalization constraint $\alpha_{RGB} + \alpha_{Thermal} = 1$, this saturation implies that the system decided to completely cut off the RGB information and rely solely on the thermal camera. In contrast, the existing method (baseline method), which relies solely on RGB, could not withstand this moment of texture loss, and its IoU plummeted to nearly 0 at approximately $t \approx 4[s]$, falling into a lost state (Figure 5 (c)). Subsequently, as the drone moved and the number of RGB effective inliers increased, the system immediately adjusted the weights for RGB and thermal again, returning to cooperative tracking. This result demonstrates that in environments where visual texture is missing, thermal information prevents system loss and improves availability.

 doi:10.20944/preprints202601.0625.v1

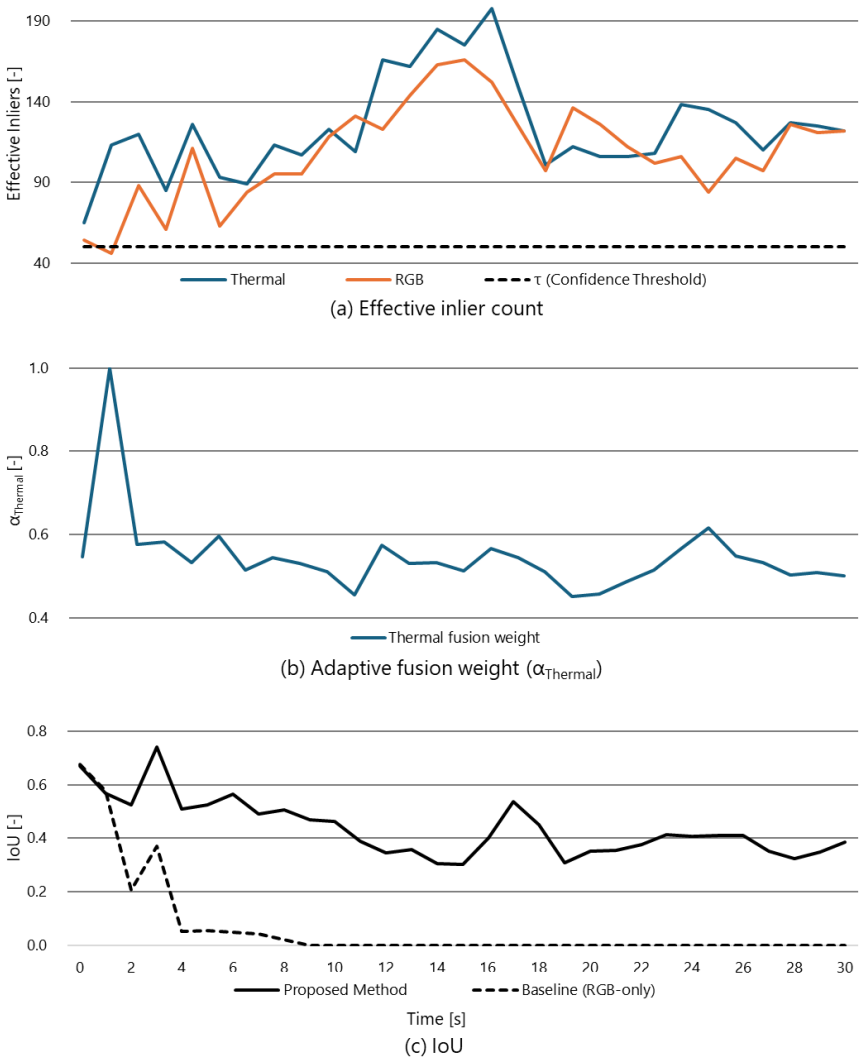
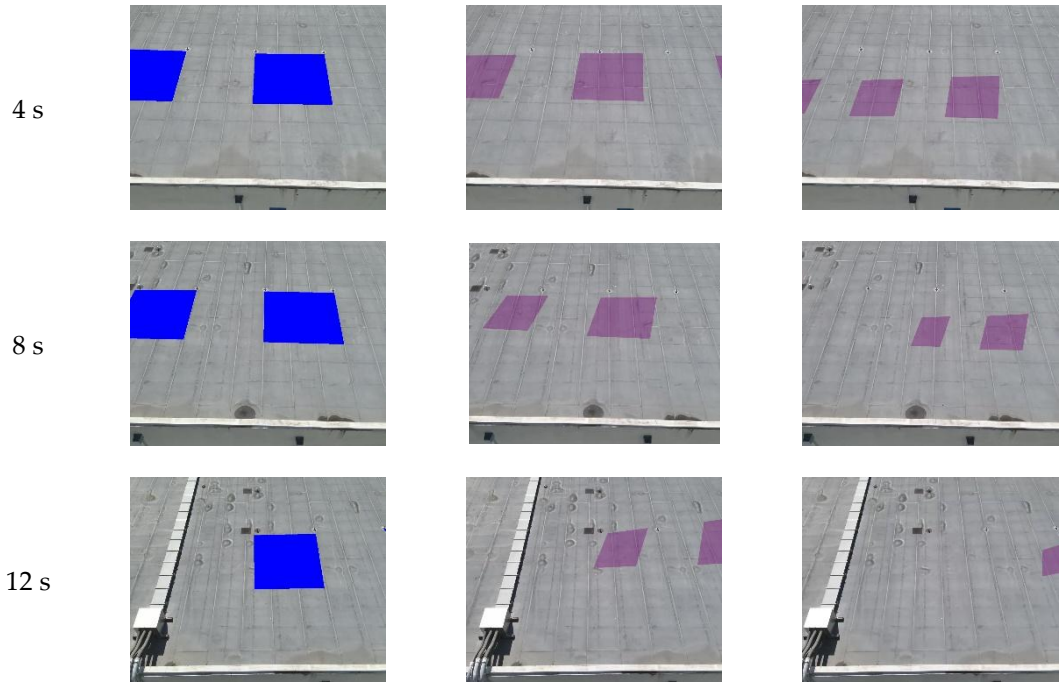


Figure 5. (a) Effective inlier count, (b) fusion weight, and (c) IoU in Route B after operating VPS only once.



 doi:10.20944/preprints202601.0625.v1

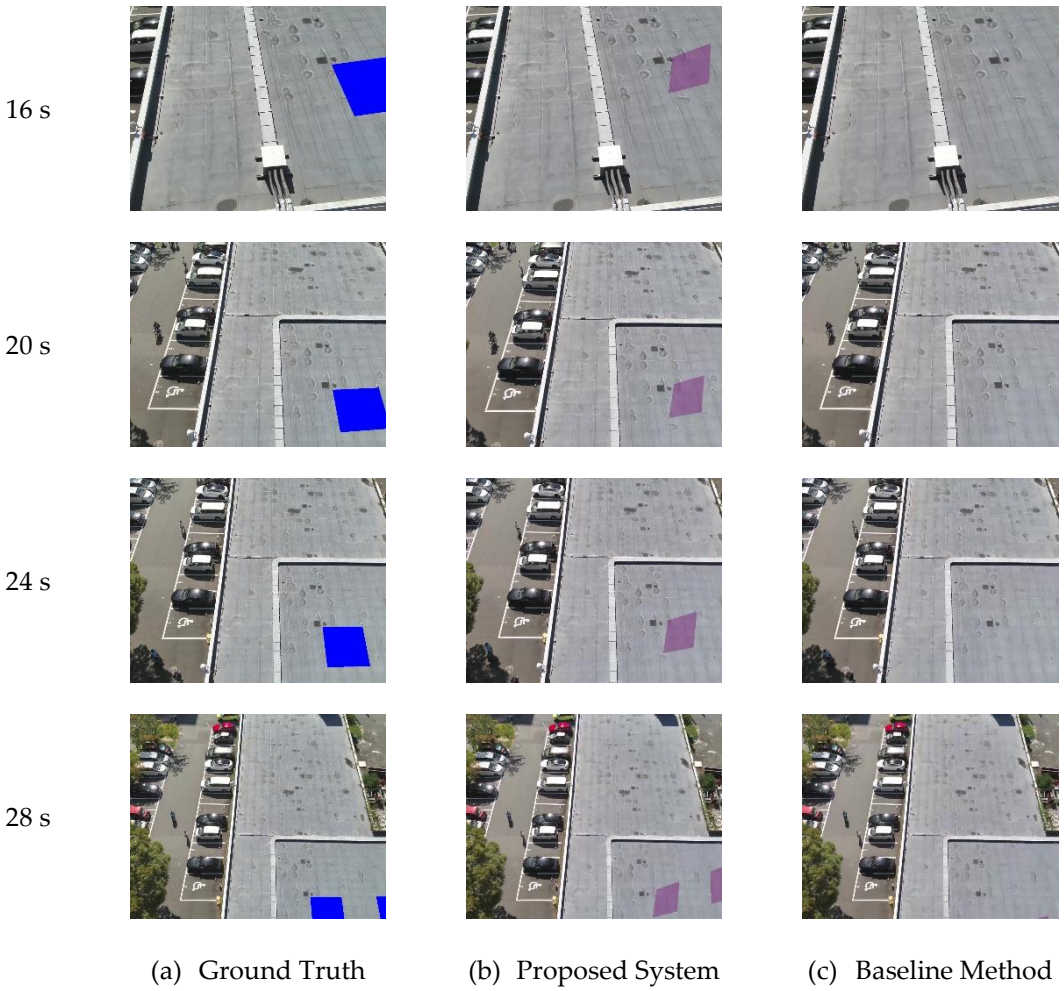


Figure 6. Comparison of tracking status in Route B over time (4 s to 28 s). The blue overlays represent the Ground Truth (GT) positions, while the purple overlays indicate the MR registration results estimated by the respective methods.

5.2.2. Analysis of Route C: Drift Suppression via Consensus

The time-series data for Route C (effective inlier count, fusion weight, IoU) is shown in Figure 7, and a comparison of the tracking status is shown in Figure 8.

The results for Route C demonstrate the effect of raising the baseline accuracy in long-term tracking. In this route, although the existing method avoided complete loss, a trend was observed where the IoU gradually decreased over time. This is a drift phenomenon that occurs as a result of the accumulation of minute matching errors characteristic of a single sensor.

On the other hand, the proposed method maintained a high accuracy of $IoU > 0.6$ for 30 seconds. In this interval, it is noteworthy that both RGB and thermal maintained a sufficient number of effective inliers. Under this condition, the system exhibited behavior where the fusion weight was allocated approximately at $0.5 : 0.5$.

Statistically, integrating these two independent observation sources (consensus) -based on the assumption that the measurement noises of the two sensors are independent since RGB and thermal have different physical characteristics (visible light reflection and thermal radiation) - has the effect of reducing the variance of accidental noise from individual sensors. Since RGB and thermal modalities possess distinct physical properties (visible light reflection vs. thermal radiation), the factors causing drift are independent, allowing them to mitigate each other's biases. This confirms that multimodal integration not only prevents loss but actively improves tracking precision even in feature-rich environments.

 doi:10.20944/preprints202601.0625.v1

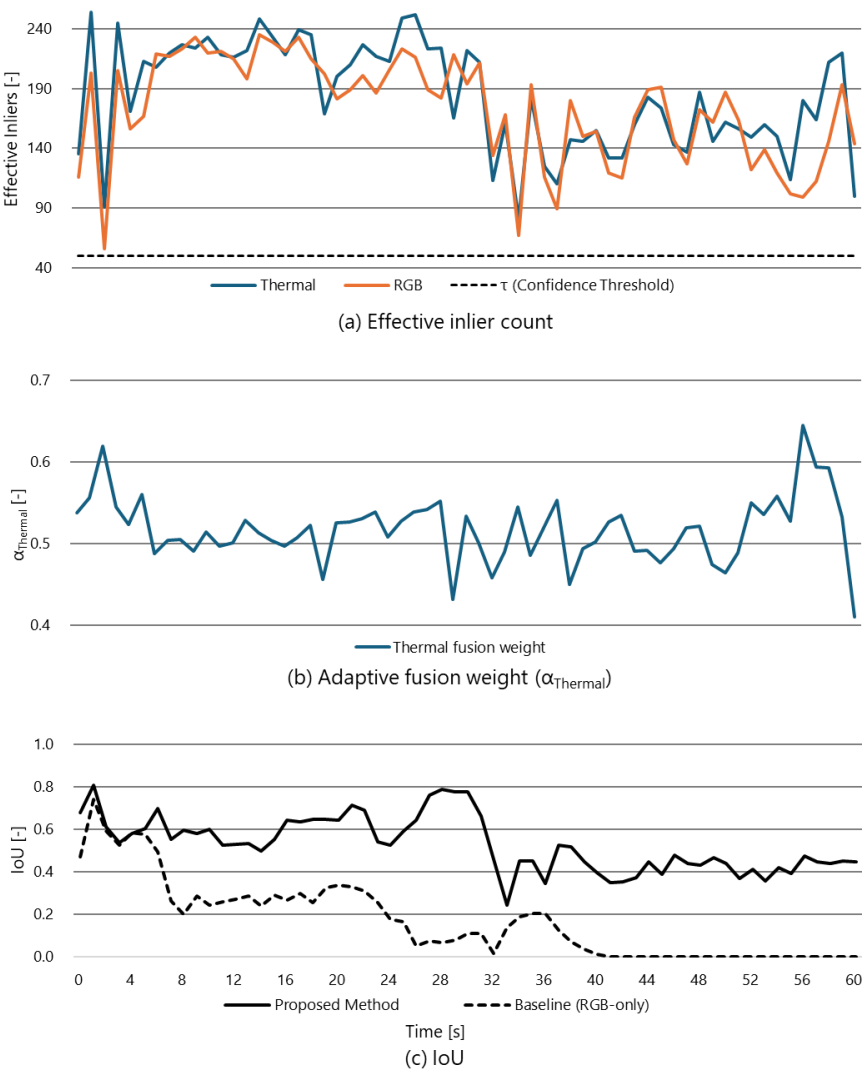
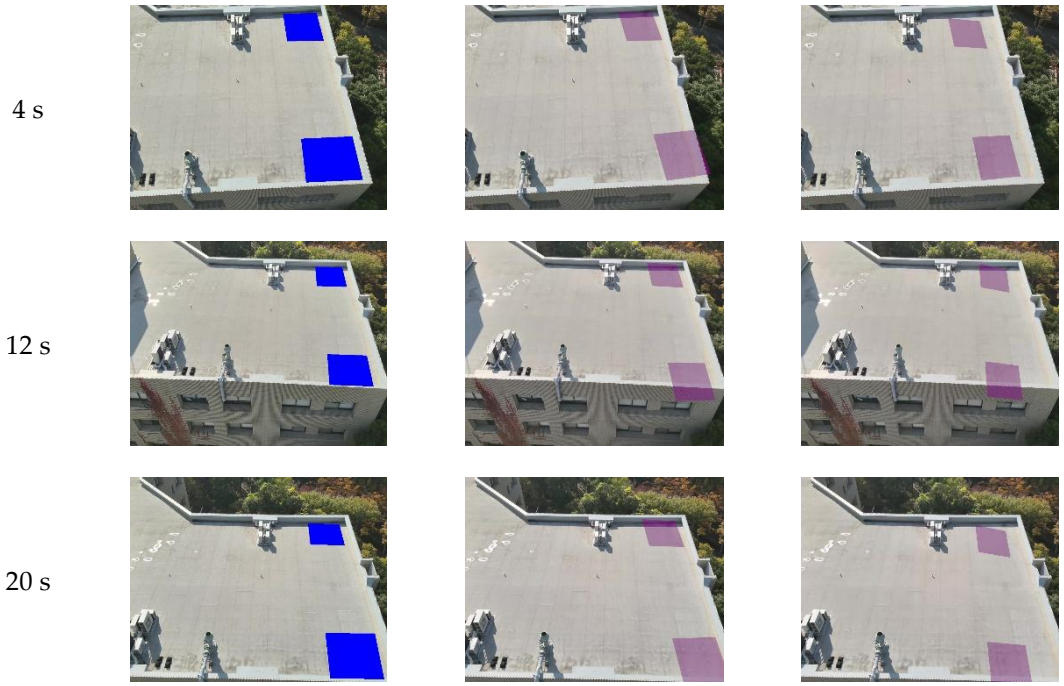


Figure 7. (a) Effective inlier count, (b)fusion weight, and (c) IoU in Route C after operating VPS only once.



 doi:10.20944/preprints202601.0625.v1

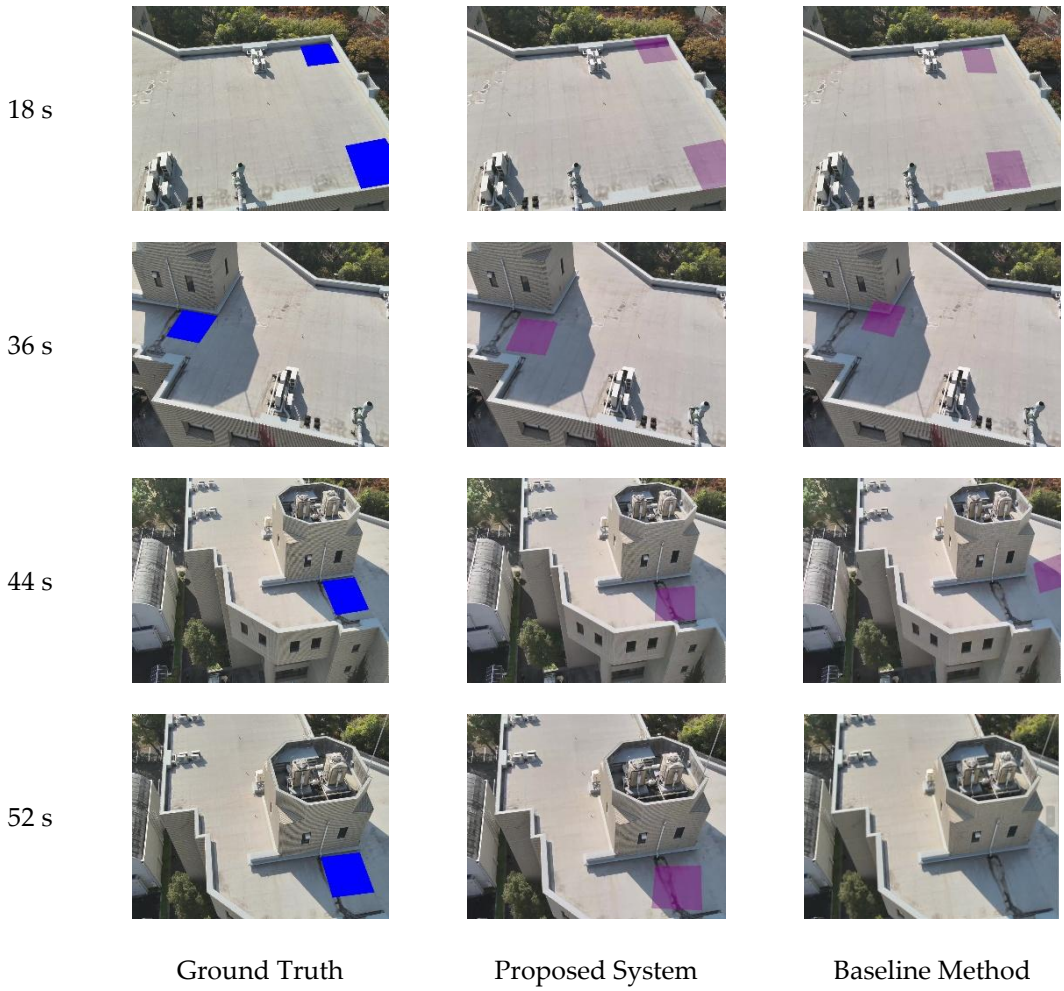


Figure 8. Comparison of tracking status in Route C over time. The blue overlays represent the Ground Truth (GT), and the purple overlays indicate the MR registration results estimated by the respective methods.

6. Discussion

In this section, we discuss the theoretical significance and practical value of the proposed method based on the findings obtained from the experimental results.

6.1. Lightweight Spatial Filtering for Real-time Multimodal Fusion

The technical core of this study lies in the introduction (Section 1) of the Effective Inlier Count (N_{eff}) as a metric to dynamically compare and integrate the reliability of heterogeneous sensors (RGB and Thermal). While state-of-the-art deep learning-based reliability estimators can provide robust results, the proposed grid-based N_{eff} offers an exceptionally low computational footprint. Even with the high-performance computational resources of a server-side PC, minimizing the per-frame processing overhead is critical for reducing end-to-end latency in MR applications. As shown by the results of experimental Route B, N_{eff} accurately detected the decrease in inliers of the RGB sensor and functioned as a trigger to switch the system to thermal-led tracking. By employing a lightweight spatial filtering algorithm, we effectively suppressed the influence of geometric burstiness without sacrificing the real-time responsiveness required for synchronized MR visualization. Even with the high-performance computational resources of a server-side PC, minimizing the per-frame processing overhead is critical for reducing end-to-end latency in MR applications.

A noteworthy point of N_{eff} in the proposed method is that it effectively suppresses the influence of geometric burstiness while using an extremely lightweight algorithm based on grid-based spatial filtering. As shown by the results of experimental Route B, N_{eff} accurately detected

 doi:10.20944/preprints202601.0625.v1

the decrease in inliers of the RGB sensor and functioned as a trigger to switch the system to thermal-led tracking. This metric provides a practical solution for evaluating sensors with different physical characteristics on the same scale and achieving adaptive sensor fusion.

6.2. Achieving Robust Drone-based MR via Thermal Integration

The achievement of this study lies in the improvement of the robustness of drone-view MR by integrating Thermal images into an RGB map-based self-localization system in low-texture environments such as construction sites.

In construction sites where bare concrete walls or frequently changing lighting conditions exist, there are limits to maintaining tracking with RGB alone, which has been a factor hindering the practical application of drone MR in the AEC field (the loss of the Baseline in Route B supports this). The results of this experiment demonstrated that thermal information is one of the effective modalities for complementing this weakness of RGB. In situations where visible light texture is missing (Route B), the thermal distribution provided stable geometric features and prevented system loss.

Furthermore, in long-term flights (Route C), although the environment was non-ideal for RGB—including specular reflections and weak-textured planes—the integration effect between heterogeneous sensors mitigated the drift error characteristic of a single sensor and contributed to maintaining the geometric consistency of the MR display. The proposed method is an approach that complements vulnerabilities through Thermal fusion while taking advantage of high-precision foundational technology that is RGB map-based. This has shown the possibility of providing a highly reliable MR experience that does not depend on the SDK or model, even in GPS-denied environments or under adverse conditions.

6.3. Limitations and Future Work

The constraints and future challenges of the proposed system in this study are described.

First, there are inherent physical limitations to the sensor modalities. While RGB and thermal cameras are generally complementary, tracking remains challenging in extreme environments where both lose discernible features. A prime example is facing polished metal surfaces in a state of thermal equilibrium; such surfaces exhibit extremely low emissivity, resulting in specular thermal reflections of the surrounding environment rather than providing usable temperature gradients for feature matching.

Second, there is a dependency on pre-existing maps and adaptation to environmental changes. Since this method requires pre-created 3D maps, its application is restricted in environments where the structure changes daily, such as buildings under construction, or for objects that are too vast to make map creation difficult.

Third, there is a dependency on the initialization process (VPS). The system depends on VPS for the initial alignment of tracking and the correction of cumulative errors. Therefore, if VPS does not succeed even once in a texture-less environment or similar, MR superimposition cannot be started.

Finally, there is a lack of a recovery function for VPS false positives. Although VPS is a highly accurate technology, its performance can deteriorate not only in environments with repetitive patterns but also in texture-less environments—such as the concrete surfaces investigated in this study—where the lack of distinct visual features can lead to incorrect initializations or significant pose errors. Since the current architecture does not have a mechanism to detect or correct errors if the initialization is incorrect, there is a risk that tracking will continue with the initial position shifted. Future challenges include adding a function to verify geometric consistency and outlier removal through feedback from the fusion system side.

 doi:10.20944/preprints202601.0625.v1

7. Conclusions

In this study, to address the vulnerability of self-localization in low-texture environments—a challenge for drone-view MR systems in the AEC industry—we proposed an adaptive sensor fusion method that integrates Thermal and RGB images. The primary conclusions of this study are as follows.

- Establishment of a lightweight geometric gating mechanism: We introduced the Effective Inlier Count N_{eff} as a metric to evaluate the reliability of heterogeneous sensors in real-time and on a unified scale. In drone systems with limited computational resources, this metric effectively eliminated erroneous reliability judgments caused by geometric burstiness (local concentration of feature points) without the need for computationally expensive covariance estimation. This enabled robust sensor selection based on spatial quality rather than mere quantity of feature points.
- Simultaneous achievement of availability and stability: Validation experiments in real-world environments demonstrated that the proposed method exhibits robustness through two distinct mechanisms. First, in critical situations where visible light texture is missing, thermal information functioned as a fail-safe modality to prevent system loss, ensuring operational availability. Second, during long-term flights, the integration effect between sensors with different physical characteristics mitigated drift errors inherent to single modalities, thereby maintaining geometric stability.
- Contribution to Digital Transformation in the AEC industry: The proposed system provides a practical solution for the full automation of infrastructure inspection and construction management by offering an MR experience that remains continuous and accurate even in GPS-denied environments or under adverse conditions. By leveraging existing high-precision RGB map-based foundational technology while complementing its primary weakness through thermal fusion, this approach significantly enhances the reliability of drone utilization within the AEC sector.

Supplementary Materials: The following supporting information can be downloaded at: https://doi.org/10.5281/zenodo.18087871. The materials are organized into the following folders: Video Folder: Input footage for Routes A–C and output result videos under various experimental conditions; Data Folder: result.xlsx containing detailed time-series IoU measurement data; Unity Project: Source code files for the simulation.

Author Contributions: Conceptualization, R.F. and T.F.; methodology, R.F.; software, R.F.; validation, R.F.; formal analysis, R.F.; investigation, R.F.; resources, **T.F.**; data curation, R.F.; writing—original draft preparation, R.F.; writing—review and editing, R.F. and T.F.; visualization, R.F.; supervision, T.F.; project administration, **T.F.** All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by JSPS KAKENHI, grant number 23K11724.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code, experimental data (**Data Folder**), and video materials (**Video Folder**) presented in this study are openly available in **Zenodo** at https://doi.org/10.5281/zenodo.18087871. The specific 3D map data of the experimental site used for the VPS is not publicly available due to privacy and security restrictions.

Acknowledgments: During the preparation of this work, the authors used ChatGPT and Gemini to improve the readability and language of the text. After using these tools/services, the authors reviewed and edited the content as necessary and took full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202601.0625.v1

Abbreviations

The following abbreviations are used in this manuscript:

6DoF	6 degrees of freedom
AEC	Architecture, Engineering, and Construction
BIM	Building Information Modeling
CLAHE	Contrast Limited Adaptive Histogram Equalization
GNSS	Global Navigation Satellite System
IMU	Inertial Measurement Unit
IoU	Intersection over Union
Lerp	Linear interpolation
MR	Mixed Reality
PnP	Perspective-n-Point
RANSAC	Random Sample Consensus
ROI	regions of interest
RTK-GNSS	Real-Time Kinematic GNSS
SDKs	Software Development Kits
SLAM	Simultaneous Localization and Mapping
SfM	Structure-from-Motion
Slerp	Spherical Linear interpolation
VPS	Visual Positioning System

References

1. Bonatti, R.; Ho, C.; Wang, W.; Choudhury, S.; Scherer, S. Towards a Robust Aerial Cinematography Platform: Localizing and Tracking Moving Targets in Unstructured Environments. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 3-8 Nov. 2019, 2019; pp. 229-236. doi:10.1109/ IROS40897.2019.8968163.
2. Giordan, D.; Adams, M.S.; Aicardi, I.; Alicandro, M.; Allasia, P.; Baldo, M.; De Berardinis, P.; Dominici, D.; Godone, D.; Hobbs, P.; et al. The use of unmanned aerial vehicles (UAVs) for engineering geology applications. *Bulletin of Engineering Geology and the Environment* 2020, *79*, 3437-3481, doi:10.1007/s10064-020-01766-2.
3. Outay, F.; Mengash, H.A.; Adnan, M. Applications of unmanned aerial vehicle (UAV) in road safety, traffic and highway infrastructure management: Recent advances and challenges. *Transportation Research Part A: Policy and Practice* 2020, *141*, 116-129, doi:/10.1016/j.tra.2020.09.018.
4. Mogili, U.M.R.; Deepak, B.B.V.L. Review on Application of Drone Systems in Precision Agriculture. *Procedia Computer Science* 2018, *133*, 502-509, doi:/10.1016/j.procs.2018.07.063.
5. Shakhatreh, H.; Sawalmeh, A.H.; Al-Fuqaha, A.; Dou, Z.; Almaita, E.; Khalil, I.; Othman, N.S.; Khreishah, A.; Guizani, M. Unmanned Aerial Vehicles (UAVs): A Survey on Civil Applications and Key Research Challenges. *IEEE Access* 2019, *7*, 48572-48634, doi:10.1109/ACCESS.2019.2909530.
6. Guan, S.; Zhu, Z.; Wang, G. A Review on UAV-Based Remote Sensing Technologies for Construction and Civil Applications. *Drones* 2022, *6*, 117, doi:10.3390/drones6050117.
7. Videras Rodríguez, M.; Melgar, S.G.; Cordero, A.S.; Márquez, J.M. A Critical Review of Unmanned Aerial Vehicles (UAVs) Use in Architecture and Urbanism: Scientometric and Bibliometric Analysis. *Applied Sciences* 2021, *11*, 9966, doi:10.3390/app11219966.
8. Tomita, K.; Chew, M.Y. A Review of Infrared Thermography for Delamination Detection on Infrastructures and Buildings. *Sensors* 2022, *22*, doi:10.3390/s22020423.
9. Fukuda, R.; Fukuda, T.; Yabuki, N. Advancing Building Facade Inspection: Integration of an Infrared Camera-Equipped Drone and Mixed Reality. In Proceedings of the 42nd Conference on Education and Research in Computer Aided Architectural Design in Europe (eCAADe 2024), Nicosia, Cyprus, 11–13 September 2024; Volume 2, pp. 139–148, doi:10.52842/conf.ecaade.2024.2.139
10. Yeom, S. Thermal Image Tracking for Search and Rescue Missions with a Drone. *Drones* 2024, *8*, doi:10.3390/drones8020053.

11. Milgram, P.; Kishino, F. A Taxonomy of Mixed Reality Visual Displays. *IEICE Trans. Information Systems* 1994, *vol. E77-D, no. 12*, 1321-1329. doi:10.1016/j.procs.2018.07.063.
12. Wen, M.-C.; Kang, S.-C. Augmented Reality and Unmanned Aerial Vehicle Assist in Construction Management. In *Computing in Civil and Building Engineering (2014)*; Proceedings; 2014; pp. 1570-1577. doi:10.1061/9780784413616.195.
13. Raimbaud, P.; Lou, R.; Merienne, F.; Danglade, F.; Figueroa, P.; Hernández, J.T. BIM-based Mixed Reality Application for Supervision of Construction. In Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 23-27 March 2019, 2019; pp. 1903-1907. doi:10.1109/VR.2019.8797784.
14. Cao, Y.; Dong, Y.; Wang, F.; Yang, J.; Cao, Y.; Li, X. Multi-sensor spatial augmented reality for visualizing the invisible thermal information of 3D objects. *Optics and Lasers in Engineering* 2021, *145*, 106634, doi:/10.1016/j.optlaseng.2021.106634.
15. Botrugno, M.C.; D'Errico, G.; De Paolis, L.T. Augmented Reality and UAVs in Archaeology: Development of a Location-Based AR Application. In Proceedings of the Augmented Reality, Virtual Reality, and Computer Graphics, Cham, 2017//, 2017; pp. 261-270. doi:10.1007/978-3-319-60928-7_23.
16. Kikuchi, N.; Fukuda, T.; Yabuki, N. Future landscape visualization using a city digital twin: integration of augmented reality and drones with implementation of 3D model-based occlusion handling. *Journal of Computational Design and Engineering* 2022, *9*, 837-856, doi:10.1093/jcde/qwac032.
17. Kinoshita, A.; Fukuda, T.; Yabuki, N. Drone-Based Mixed Reality: Enhancing Visualization for Large-Scale Outdoor Simulations with Dynamic Viewpoint Adaptation Using Vision-Based Pose Estimation Methods. *Drone Systems and Applications* 2024, doi:10.1139/dsa-2023-0135.
18. Xie, X.; Yang, T.; Ning, Y.; Zhang, F.; Zhang, Y. A Monocular Visual Odometry Method Based on Virtual-Real Hybrid Map in Low-Texture Outdoor Environment. *Sensors* 2021, *21*, 3394, doi:10.3390/s21103394.
19. Qin, Y.W.; Bao, N.K. Infrared thermography and its application in the NDT of sandwich structures. *Optics and Lasers in Engineering* 1996, *25*, 205-211, doi:/10.1016/0143-8166(95)00066-6.
20. Wu, Y.; Wang, L.; Zhang, L.; Chen, M.; Zhao, W.; Zheng, D.; Cai, Y. Monocular thermal SLAM with neural radiance fields for 3D scene reconstruction. *Neurocomputing* 2025, *617*, 129041, doi:/10.1016/j.neucom.2024.129041.
21. Sattler, T.; Havlena, M.; Schindler, K.; Pollefeys, M. Large-Scale Location Recognition and the Geometric Burstiness Problem. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016, 2016; pp. 1582-1590. doi: 10.1109/CVPR.2016.175
22. Sarlin, P.; Cadena, C.; Siegwart, R.Y.; Dymczyk, M. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019; pp. 12708–12717. doi : 10.1109/CVPR.2019.01300
23. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision* 2009, *81*, doi:10.1007/s11263-008-0152-6.
24. Immersal Ltd. Immersal SDK. Available online: https://immersal.com/ (accessed on 26 December 2025).
25. Niantic, I. Niantic Lightship VPS. Available online: https://lightship.dev/ (accessed on 26 December 2025).
26. Google Developers. ARCore Geospatial API. Available online: https://developers.google.com/ar/develop/geospatial (accessed on 26 December 2025).
27. Mur-Artal, R.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics* 2015, *31*, 1147-1163, doi:10.1109/TRO.2015.2463671.
28. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, D.; Neira, J.; Reid, I.; Leonard, J. *Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age*; 2016. doi:10.1109/TRO.2016.2624754
29. Zollmann, S.; Hoppe, C.; Kluckner, S.; Poglitsch, C.; Bischof, H.; Reitmayr, G. Augmented Reality for Construction Site Monitoring and Documentation. *Proceedings of the IEEE* 2014, *102*, 137-154, doi:10.1109/JPROC.2013.2294314.
30. Behzadan Amir, H.; Kamat Vineet, R. Georeferenced Registration of Construction Graphics in Mobile Outdoor Augmented Reality. *Journal of Computing in Civil Engineering* 2007, *21*, 247-258, doi:10.1061/(ASCE)0887-3801(2007)21:4(247).

31. Zhuang, L.; Zhong, X.; Xu, L.; Tian, C.; Yu, W. Visual SLAM for Unmanned Aerial Vehicles: Localization and Perception. *Sensors* 2024, *24*, 2980, doi:10.3390/s24102980.
32. Sun, D.; Yang, X.; Liu, M.Y.; Kautz, J. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18-23 June 2018, 2018; pp. 8934-8943. doi:10.1109/CVPR.2018.00931
33. Li, S.; Ma, X.; He, R.; Shen, Y.; Guan, H.; Liu, H.; Li, F. WTI-SLAM: a novel thermal infrared visual SLAM algorithm for weak texture thermal infrared images. *Complex & Intelligent Systems* 2025, *11*, 242, doi:10.1007/s40747-025-01858-0.
34. Zhao, X.; Luo, Y.; He, J. Analysis of the Thermal Environment in Pedestrian Space Using 3D Thermal Imaging. *Energies* 2020, *13*, doi:10.3390/en13143674.
35. Borges, P.V.K.; Vidas, S. Practical Infrared Visual Odometry. *IEEE Transactions on Intelligent Transportation Systems* 2016, *17*, 2205-2213, doi:10.1109/TITS.2016.2515625.
36. Johansson, J.; Solli, M.; Maki, A. An Evaluation of Local Feature Detectors and Descriptors for Infrared Images. In Proceedings of the Computer Vision – ECCV 2016 Workshops, Cham, 2016//, 2016; pp. 711-723. doi:/10.1007/978-3-319-49409-8_59.
37. Mouats, T.; Aouf, N.; Nam, D.; Vidas, S. Performance Evaluation of Feature Detectors and Descriptors Beyond the Visible. *Journal of Intelligent & Robotic Systems* 2018, *92*, 33-63, doi:10.1007/s10846-017-0762-8.
38. Delaune, J.; Hewitt, R.; Lytle, L.; Sorice, C.; Thakker, R.; Matthies, L. Thermal-Inertial Odometry for Autonomous Flight Throughout the Night. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 3-8 Nov. 2019, 2019; pp. 1122-1128, doi:10.1109/IROS40897.2019.8968238.
39. Deng, Z.; Zhang, Z.; Ding, Z.; Liu, B. Multi-Source, Fault-Tolerant, and Robust Navigation Method for Tightly Coupled GNSS/5G/IMU System. *Sensors* 2025, *25*, 965, doi:10.3390/s25030965.
40. Shan, T.; Englot, B.; Ratti, C.; Rus, D. LVI-SAM: Tightly-coupled Lidar-Visual-Inertial Odometry via Smoothing and Mapping. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), 30 May-5 June 2021, 2021; pp. 5692-5698. doi:10.48550/arXiv.2104.10831
41. Ebadi, K.; Bernreiter, L.; Biggie, H.; Catt, G.; Chang, Y.; Chatterjee, A.; Denniston, C.E.; Deschênes, S.-P.; Harlow, K.; Khattak, S.; et al. Present and Future of SLAM in Extreme Underground Environments. *IEEE Transactions on Robotics* **2024**, *40*, 936-959. doi:10.1109/TRO.2023.3323938
42. Sun, Y.; Wang, Q.; Yan, C.; Feng, Y.; Tan, R.; Shi, X.; Wang, X. D-VINS: Dynamic Adaptive Visual–Inertial SLAM with IMU Prior and Semantic Constraints in Dynamic Scenes. *Remote Sensing* **2023**, *15*, 3881, doi:10.3390/rs15153881.
43. Xu, X.; Zhang, L.; Yang, J.; Cao, C.; Wang, W.; Ran, Y.; Tan, Z.; Luo, M. A Review of Multi-Sensor Fusion SLAM Systems Based on 3D LIDAR. *Remote Sensing* **2022**, *14*, 2835, doi:10.3390/rs14122835.
44. Shetty, A.; Gao, G. Adaptive covariance estimation of LiDAR-based positioning errors for UAVs. *Navigation* **2019**, *66*, doi:10.1002/navi.307.
45. Teed, Z.; Deng, J. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In Proceedings of the Advances in Neural Information Processing Systems, 2021; pp. 16558-16569. doi:10.48550/arXiv.2108.10869.
46. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2 ed.; Cambridge University Press: Cambridge, 2004.
47. Zuiderveld, K. VIII.5. - Contrast Limited Adaptive Histogram Equalization. In *Graphics Gems*, Heckbert, P.S., Ed.; Academic Press: 1994; pp. 474-485. doi:10.1109/VBC.1990.109340.
48. Lucas, B.D.; Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI), Vancouver, BC, Canada, 24–28 August 1981; pp. 674–679.
49. Shoemake, K. Animating rotation with quaternion curves. Proceedings of the 12th annual conference on Computer graphics and interactive techniques - SIGGRAPH '85 1985, 245-254, doi:10.1145/325334.325242.
50. Zhang, Z. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2000, *22*, 1330-1334, doi:10.1109/34.888718.

 doi:10.20944/preprints202601.0625.v1

51. ElSheikh, A.; Abu-Nabah, B.A.; Hamdan, M.O.; Tian, G.-Y. Infrared Camera Geometric Calibration: A Review and a Precise Thermal Radiation Checkerboard Target. *Sensors* 2023, *23*, doi:10.3390/s23073479.
52. Neumann, U.; You, S. Natural feature tracking for augmented reality. *IEEE Transactions on Multimedia* 1999, *1*, 53-64, doi:10.1109/6046.748171.
53. Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision* 2010, *88*, 303-338, doi:10.1007/s11263-009-0275-4.