Automation in Construction 187 (2026) 106889

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Mobile robot inspection of building plenum spaces via vision-based multi-sensor fault detection

Gao Lou [a], Shuai Liu [a], Yi Shen [a], Wing Yin Wong [b], Yanglong Lu [a,c], Molong Duan [a,c,*]

[a] Department of Mechanical and Aerospace Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[b] MTR Corporation Limited, Hong Kong SAR, China
[c] Center for Smart Manufacturing, The Hong Kong University of Science and Technology, Hong Kong SAR, China

ARTICLE INFO

Keywords:
Vision-based fault detection
Multi-modal data
Mobile robot inspection
Plenum space

ABSTRACT

Building plenum spaces house dense HVAC, electrical, and plumbing infrastructure where failures, such as structure deformation, water leakage, pose significant safety and service risks. Traditionally, labor-intensive inspections are inefficient and pose injury risks. To address this, autonomous solutions like mobile robots are emerging. These robots mainly rely on vision information. However, current robotic inspection methods struggle with environmental variability and temporal data reliance in crowded plenums, where the novelty affects the detection accuracy. The paper proposes STF-Inspector, integrating semantic information and object tracking to improve model generalization capability. Further refinements of point cloud distance and object thermal contrast ensure robust feature extraction. Experimental validation on a mobile robot confirms its superior performance and robustness compared to the traditional method.

1. Introduction

The plenum space is a place above the false ceiling that stores essential infrastructure, such as heating, ventilation ducts, electrical equipment and wiring, and plumbing systems [1,2]. This space must undergo routine maintenance procedures. Without regular inspection and maintenance activities, various health and safety risks will threaten building users, including water pooling, material corrosion, excessive sagging, and deterioration of concrete support structures [3]. However, this maintenance process is greatly dependent on human labor. The requirement for continuous and frequent maintenance is a considerable burden and raises the risk of injury to the workers.

Current inspection practices predominantly rely on manual workflows involving ladders and scaffolding. These methods are inherently labor-intensive, inefficient, and expose inspectors to safety risks, such as falls from height and prolonged exposure to dusty environments. While robotic inspection systems have been proposed to mitigate these labor and safety risks, existing solutions face distinct limitations. For instance, vertical wall-climbing robots [4–6] involve direct surface contact, posing risks of accidental drops or surface damage, while fixed-shape ground robots [7] often lack the necessary vertical reach to inspect embedded infrastructure above the false ceiling. Other robots, such as the cable driven robot [8] and boom lift robot [9] are not suitable for the inspection task in the crowded plenum space. Therefore, a mobile robotic platform capable of extending to surface levels for comprehensive inspection offers a practical and innovative solution to these challenges.

To address these issues, this study targets the improvement of the inspection workflow and the enhancement of detection accuracy. The paper proposes a mobile robotic system that automates access to elevated plenum spaces, thereby reducing human safety risks. However, enabling such a robot to perform effective automatic inspection introduces new challenges. Due to the difficulties in using the contact-based sensors in the crowded space on the moving robot, these robots primarily use visual inspection. There are numerous works involved in the visual inspection algorithms that can work in conjunction with robots. For example, Cai et al. [10] developed an autonomous robotic system that uses the You Only Look Once (YOLO) model on thermal images to detect heat leakage. Ge et al. [11] proposed a YOLO-based crack-detection method for inspection robots, additionally leveraging depth images for 3D fault localization. Semwal et al. [12] presented a plenum-space deterioration-detection framework based on Faster R-CNN. Atkinson et al. [13] addressed underfloor scene segmentation with Mask R-CNN, and Techasarntikul et al. [14] likewise applied Mask R-CNN for instance segmentation in plenum spaces to identify key objects.

* Corresponding author at: Department of Mechanical and Aerospace Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China.
E-mail address: duan@ust.hk (M. Duan).

https://doi.org/10.1016/j.autcon.2026.106889
Received 20 October 2025; Received in revised form 9 March 2026; Accepted 11 March 2026
Available online 8 April 2026

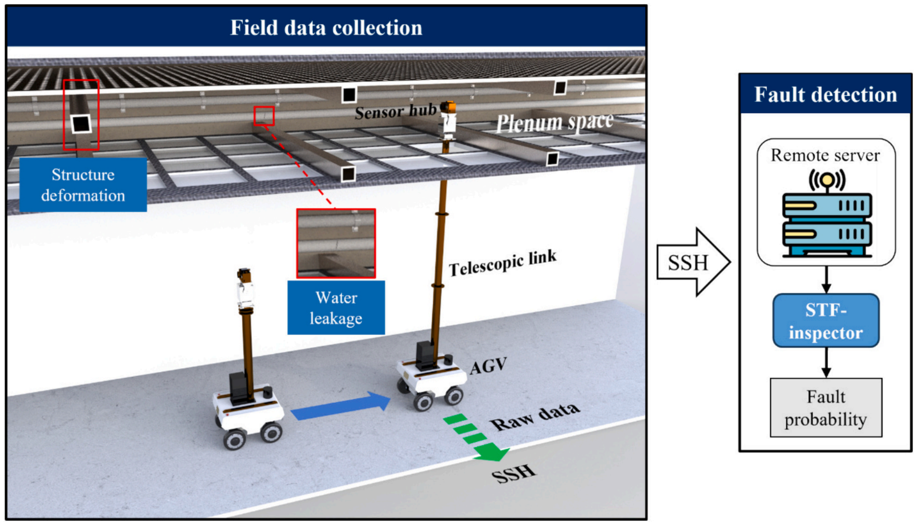


Fig. 1. Overview of the mobile robot plenum space inspection task.

Czerniawski and Leite [15] utilized semantic information to segment RGB-D images into 13 distinct building component classes. Menendez et al. [16] presented a tunnel inspection system that employs Convolutional Neural Networks (CNN) to identify surface defects and coordinates a robotic arm to place ultrasonic sensors for depth measurement. Jiao et al. [17] developed a real-time marine structural health monitoring system, employing a modified YOLO-Underwater model to identify concrete defects such as cracks and corrosion in complex underwater environments. Previtali et al. [18] presented a methodology for the automated generation of 3D building models from laser data and their integration with UAV-acquired thermal images to detect and localize thermal anomalies for energy efficiency retrofitting. Similarly, McLaughlin et al. [19] developed a framework for concrete bridge inspection using a mobile robot equipped with LiDAR, RGB, and thermal cameras, employing SLAM to fuse CNN-based defect predictions into a 3D map for precise area quantification. In a related multi-modal approach, Kim et al. [20] proposed a hybrid object recognition system using a mobile robot equipped with LiDAR scanners and an infrared camera. They developed a method to generate thermal-mapped point clouds, utilizing thermal data for segmentation and geometric features for classification. Palanisamy et al. [21] developed the reconfigurable robot 'Raptor', which fuses Deep Learning-based visual defect detection with LiDAR-SLAM mapping to accurately localize faults in complex drain environments. However, despite all these efforts, the methods they developed only use the single modality information for the fault detection, even though some of the robots are equipped with sensors in multiple modalities. For the plenum space, it is often difficult to detect using the information from a single sensor alone due to several inherent limitations. Use the color image-based methods as an example. Many subtle structural defects, such as slight deformations or concealed leakage, exhibit minimal visual contrast against the surrounding surfaces, making them virtually indistinguishable, but these faults can be easily detected by the LiDAR or the thermal camera. Hence, the fault detection method that utilizes data from different modalities is a more favored choice for the plenum space inspection.

Many papers have studied the vision-based fault detection method with multimodal data fusion, since multimodal sensors have been widely adopted on robots in different inspection tasks [22–25]. Han et al. [26] presented a multispectral water-leakage detection method for metro tunnels, using a Feature Pyramid Network (FPN) to fuse visual-optical (VIS) and thermal infrared (IR) imagery. Jalil et al. [27]. developed a transmission-line fault-detection framework that aligns color and infrared images captured by unmanned aerial vehicles. Li et al. [28] proposed a cross-modality learning architecture that extracts features from images and acceleration signals with CNNs and fuses them before the fault-identification module. Mussina et al. [29] developed a Fusion Convolutional Network (FCN) for insulator monitoring by combining surface-contamination images with leakage-current measurements. Mouzenidis et al. [30] proposed a Variational Faster R-CNN that fuses RGB and depth features in a discriminative latent space using Variational Auto-Encoders and channel attention, enhancing robustness for industrial object detection. Zhang et al. [31] proposed an end-to-end RGB-T network employing multi-branch group fusion and joint attention to integrate multi-level features. Liu et al. [32] developed a target-aware dual adversarial learning framework that jointly optimizes fusion and detection to distinguish thermal targets from visible textures. Grandio et al. [33] presented a multimodal deep learning approach for panoptic segmentation of railway infrastructure, utilizing rasterized point clouds to filter irrelevant data and efficiently identify pole-like and linear assets at high densities. However, there are several limitations when it comes to the plenum space inspection. First, the high variability of object shapes, complex distributions, and noisy data limits the generalization capability of end-to-end learning methods, making them prone to failure in unseen scenarios. This necessitates a greater reliance on generalizable object detection and robust feature engineering. Second, the identification of many faults relies on comparison with historical data. For instance, a duct that is normally hotter than others could be flagged as a false positive if the model lacks a historical baseline for its normal operating state.

This paper introduces the Segmentation-Tracking-Fusion Inspector (STF-Inspector), a framework that addresses the challenges outlined above. Following Liu et al. [34] and Lee et al. [20], who used the open-set vision language model Grounded-SAM [35] for 3D inventories of transportation infrastructure and automated inspection of bridge expansion joints, this paper likewise adopts the state-of-the-art open-set vision language model SAM3 [36] for object detection in RGB images. This paper further incorporates Associating Objects with Transformers (AOT) [37] to track and associate the same objects across time. STF-Inspector fuses color, depth, and thermal imagery with environmental temperature and humidity to detect structure deformation and water leakage. There are many related papers regarding these faults. For example, to detect structure deformation, Jafari et al. [38] presented a 3D point cloud change analysis method, using Iterative Closest Point (ICP) to align the point clouds of deformed and undeformed sample plates, then applying Cloud-to-Cloud (C2C) and Multiscale Model to Model Cloud Comparison (M3C2) methods to measure the difference between them. Graves et al. [39] use a laser scanner to measure the deformation of a bridge, and apply C2C, M3C2, and Point-to-Point (P2P)

Table 1
Sensor hub components specification.

Components	Function	Specification
L515	LiDAR and color camera	1920 × 1080 (color image) < 6 m (color image) 640 × 480 (depth image) 0.25 m to 4 m (depth image)
Lepton 3.5	Thermal camera	160 × 120 −10 °C to 140 °C
DHT11	Temperature and humidity sensor	0 °C to 50 °C (temperature) 20% and 90% (relative humidity)
Raspberry Pi	Data preprocessing and transmission	16GB RAM

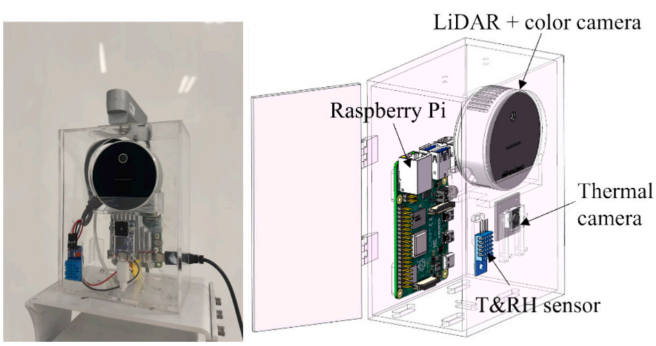


Fig. 2. Sensor hub comprising the color, depth, thermal camera, with a temperature and humidity sensor.

methods to calculate the deformation by comparing its point cloud to the reference surface. Kalasapudi et al. [40] proposed a multi-level 3D data registration approach that separates rigid body motions from element-level deformations to improve the reliability of spatial change classification in single-pier bridges. However, the existing point cloud comparison methods exhibit excessive sensitivity to partial data, which is a significant drawback since inspection robots operate in crowded spaces where structure elements frequently exceed the camera frame of view or suffer occlusion, resulting in incomplete point cloud data. Furthermore, for the detection of water leakage, Hang et al. [41] proposed a method for the detection of water leakage of hot water pipes through pipe segmentation and thermal value distribution analysis. However, this method only considered the thermal value on the pipe, whereas the thermal difference between the pipe and its surroundings is not included. The key contributions of this paper are:

- Introduces STF-Inspector, a vision-guided multimodal fusion framework that performs object-centric, temporal-difference fault detection in plenum spaces by combining RGB, depth, thermal, and environmental signals.
- Achieves robust feature extraction under mobile inspection constraints, such as handling incomplete point clouds, occlusions, and sensor noise, by using open-set segmentation (SAM3) and temporal association (AOT) to maintain consistent object identities over time.
- Validates the full pipeline on a mobile robot in real building plenum spaces, with experiments comparing alternative feature designs for detecting structure deformation and water leakage.

The paper is structured as follows. Section 2 details the hardware setup, data preparation, and the STF-Inspector framework. Section 3 validates the framework through experiments, comparative tests, and an ablation test. Section 4 concludes with technical contributions, model limitations, and future work.

2. Methodology

The overall pipeline of mobile robot plenum space inspection is shown in Fig. 1. An AGV-based robot with a telescopic link carries multimodal sensors to access the plenum space through a ceiling opening. Sensor observations and temporal features are streamed to a remote server. The streamed data is processed by the STF-inspector, which recognizes structure elements, performs multimodal data fusion, and detects faults. Target faults include structure deformation, water leakage, and loose bolts. Further information about the hardware setup and the STF-inspector will be detailed in the following subsections.

2.1. Sensor hub and Mobile platform for the inspection

The sensor hub is the critical component for plenum space inspection. It is responsible for collecting, preprocessing, and transmitting inspection data remotely. The components contained in the sensor hub are listed in Table 1. It integrates a LiDAR, a color camera, a thermal camera, and a temperature and humidity sensor, all connected to a Raspberry Pi. The Raspberry Pi handles data preprocessing and transmission, with all components housed within a compact box measuring 107 × 70 × 148 mm, as shown in Fig. 2. The Raspberry Pi runs the Robot Operating System (ROS), which provides a unified framework for managing the sensors. It is also equipped with a WiFi module, enabling remote communication with the mobile robotic platform's controller. Since fault detection requires Compute Unified Device Architecture (CUDA) for object detection, the Raspberry Pi transmits the collected data to a high-performance remote server for further processing. The LiDAR used is the Realsense L515, which includes an integrated color camera. This sensor has an ideal working range of 0.25 m to 9 m; To ensure high accuracy and dense point cloud data of the on-site environments, the optimal range is 0.25 m to 4 m. The built-in color camera has a resolution of 1920 × 1080 and supports accurate SAM3 detection. The thermal camera, Lepton 3.5, can detect thermal values ranging from −10 °C to 140 °C, with a resolution of 160 × 120. Additionally, the DHT11 temperature and humidity sensor measures temperatures from 0 °C to 50 °C and relative humidity levels between 20% and 90%.

A robotic platform is designed to provide mobility for the sensor hub, as shown in Fig. 3. The platform's core components include a four-wheel AGV, which serves as the system's base. The telescopic link with a rotational platform provides precise height adjustment and angle positioning for the sensor hub. The platform is designed for fully autonomous operation rather than relying on manual remote control. It operates on a pre-built occupancy grid map, utilizing Adaptive Monte

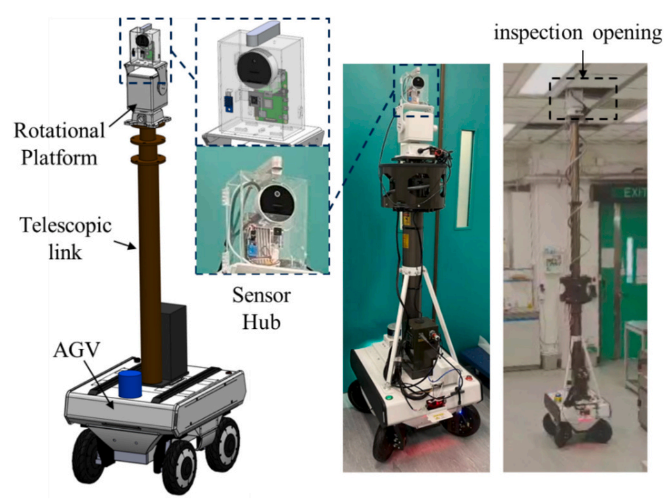
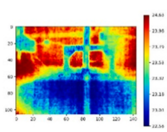
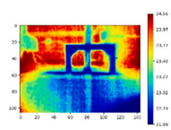
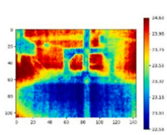
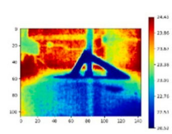


Fig. 3. Mobile platform setup and its operation at the ceiling inspection opening.

Table 2
Mobile robotic platform specification.

Overall system	Height range	2.0–3.75 m
AGV	Size	650 × 550 × 440 mm
	Load capacity	80 Kg
	Driving mode	Omnidirectional
Telescopic link	Load capacity	10 Kg
	Position Shift	≤ 5 mm
Rotation platform	Load capacity	5 Kg
	Rotation range	360°
	Rotation accuracy	0.1°
LED light	Power	5 w

Table 3
Demonstration of the raw data for the structure deformation and water leakage detection.

	Normal	Fault
Structure deformation		
Structure deformation (foreign objects)		
Water leakage		
Structure deformation, Water leakage		

Carlo Localization (AMCL) for precise positioning. Upon receiving a target coordinate, the system's global planner generates a collision-free path, and the AGV navigates to the vicinity of the inspection point. Once arrived, the robot switches to a fine-positioning mode, using ceiling-referenced LiDAR data to align the telescopic link vertically with the center of the opening before raising the sensor hub. The detailed specifications are shown in Table 2. Considering that plenum spaces are typically dark environments without external illumination, the mobile platform is equipped with a 5 W LED lighting module mounted on the sensor hub. This onboard lighting provides sufficient illumination for the color camera to capture clear images of objects within a 5-m range.

To ensure precise operation, the navigation targets are formally defined as Inspection Openings (IOPs), specific ceiling voids that offer sufficient vertical and lateral clearance for the sensor hub's safe insertion. These targets are identified through a dual-plane mapping framework where the system analyzes the ceiling-referenced point cloud to isolate height-constrained regions. A morphological erosion process is applied to filter out noise and identify valid openings that meet the collision-free clearance requirements. While the detection of candidate IOPs is automated via this geometric screening, the final selection of targets is confirmed by the user to ensure safety. Once the targets are set, the platform runs fully automatically in the inspection area. Regarding operational capacity, the platform is designed for 4 h of continuous inspection, capable of covering a 3500 m^2 area on a single charge. To mitigate potential sensor interference, the sensor hub employs a modular design that spatially integrates the LiDAR and environmental sensors. The inspection protocol dictates a 60-s stationary dwell time at each angle primarily to ensure stable data capture. Concerns regarding thermal accumulation are mitigated by the system's design: sensors operate sequentially rather than continuously to minimize heat generation, and the sensor hub's structural layout physically isolates the temperature and humidity sensor from internal heat sources with its Acrylic shell, ensuring accurate readings.

Despite these capabilities, the current system has specific operational limitations that define its applicability. First, the robot lacks a manipulation mechanism to physically open or remove ceiling tiles. Consequently, it requires pre-existing inspection openings or panels that have been manually removed prior to deployment to access the plenum space. Second, to ensure operational safety and data consistency, inspection missions are strictly scheduled during unoccupied hours (e.g., early morning). This timing avoids dynamic interference from building occupants and ensures the static environment necessary for accurate autonomous navigation and thermal data collection.

2.2. *Multimodal field-collected plenum space dataset*

A real-world plenum space environment was selected to construct the dataset for evaluating the proposed method. The site was a ceiling inspection opening providing access to overhead infrastructure such as support frames and water pipes. Because these systems were operational, modifications were not permitted. Therefore, the scene temporarily introduced fault-inducing objects to simulate failure cases.

To ensure realistic variability and operational relevance, a cyclic data acquisition protocol was employed. The process began with the AGV autonomously navigating to the vicinity of the ceiling inspection opening. Once positioned, a specific environmental state, including normal conditions, structure deformation, or water leakage, was physically staged within the plenum space. The sensor hub was then raised to capture synchronized multimodal data, including color, depth, thermal, temperature, and humidity readings. To simulate the natural positioning errors inherent in autonomous navigation, the AGV was translated slightly with randomized shifts before the next capture. This loop of setting up the environment, capturing data, and shifting the robot was repeated to build the dataset.

The dataset includes specific physical variations to test detector sensitivity. Structure deformation was simulated by displacing structural steel elements relative to their original position. These deformations were categorized into two severity levels, consisting of small displacements of approximately 10 to 20 mm and large displacements of approximately 40 to 50 mm. Additionally, foreign objects such as extra steel bars were introduced to simulate items that had fallen or been misplaced. For water leakage, the scenarios varied from localized partial wetting, where small droplets covered a specific section, to complete surface saturation, creating varying thermal signatures as the water cooled.

Ground truth annotation followed a multi-label binary classification strategy. Each sample was assigned a label vecto, where 1 indicates the presence of a fault and 0 indicates a normal state. This resulted in four distinct sample categories based on the combination of structure deformation and water leakage labels: normal (0,0), structure deformation only (1,0), water leakage only (0,1), and combined failures (1,1).

The total dataset comprises 127 samples. These were partitioned into a training set of 58 samples and a testing set of 69 samples. This specific split ratio was chosen based on the architectural design of the STF-Inspector. Since the final decision-making module is a lightweight Multi-Layer Perceptron (MLP) that processes low-dimensional feature vectors rather than raw high-dimensional images, it does not require massive datasets to converge. Consequently, a larger proportion of the data was reserved for testing to evaluate the model's reliability and generalization capability on unseen data. Table 3 demonstrates raw data variations. Furthermore, the statistical breakdown of this partition is visually detailed in Fig. 4. These distributions confirm that despite the inherent difficulty in capturing real-world faults, the dataset maintains a

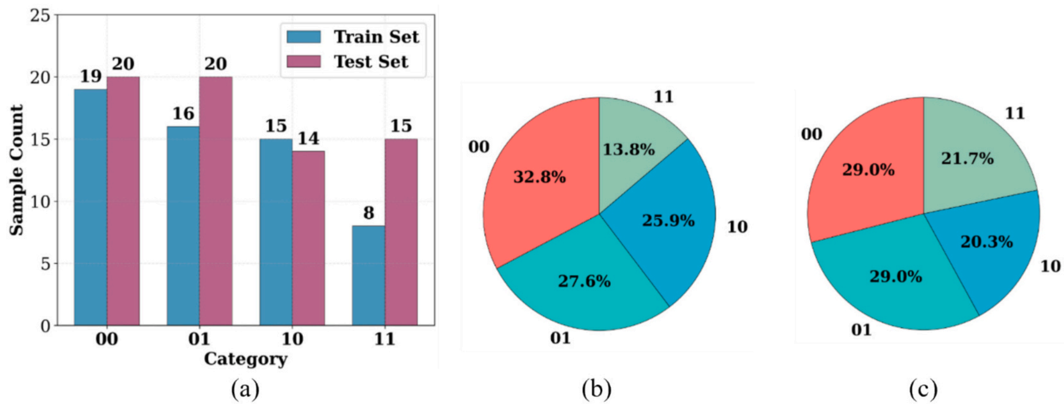


Fig. 4. Distribution of the constructed dataset.

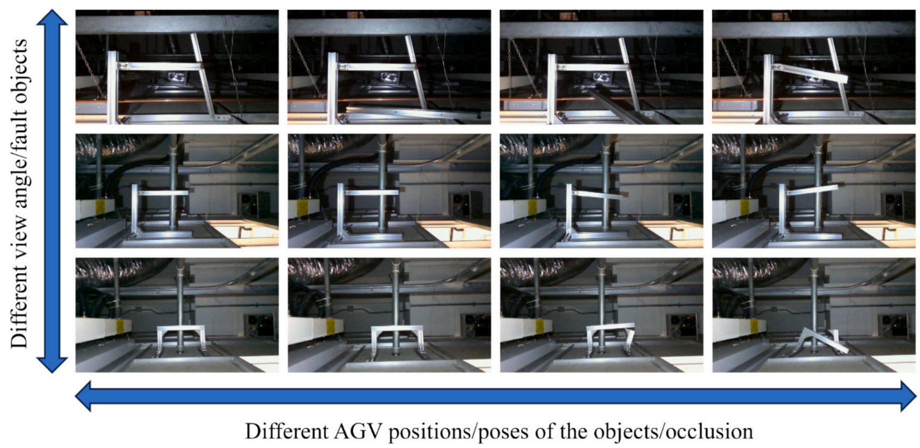


Fig. 5. Partly presented dataset, featuring different target objects, scenes, and AGV sampling locations.

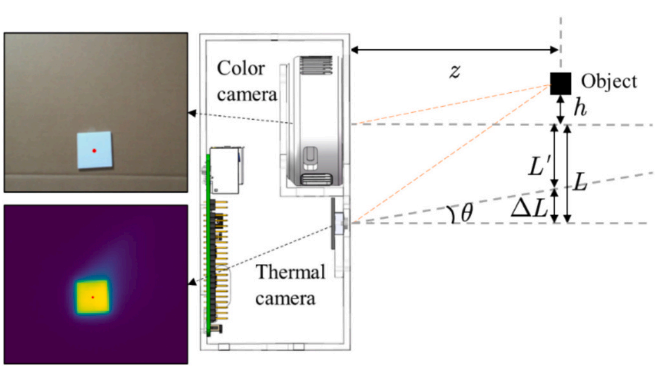


Fig. 6. Stereo system of the color camera and thermal camera.

sufficient representation of fault classes (10,01,11) relative to the normal class (00). This balanced distribution is critical for the learning process, as it prevents the classifier from becoming biased toward the non-fault state and ensures that the model is exposed to adequate variations of deformation and leakage scenarios during training.

Data collection in plenum spaces presents specific environmental challenges. These spaces are typically cluttered and unstructured, characterized by dense infrastructure that leads to frequent occlusion and complex background interference. While the AGV performs data acquisition in a static state, our dataset was constructed by sampling the same targets across different temporal instances. Due to the inherent variability in navigation, the robot's stopping position naturally differed slightly between these sessions. As illustrated in Fig. 5, the dataset introduces positional variations in the AGV's coordinates and different target objects and scenes. These natural deviations capture the operational reality where precise repeatability is rarely achieved, thereby enriching the dataset with a representative distribution of AGV positions and preventing the framework from overfitting to a single, fixed perspective.

2.2.1. *Multimodal images alignment*

The depth image and thermal image should be aligned with the color image frame. The L515 camera provides built-in alignment between its depth and color images. However, the thermal camera needs an additional calibration procedure.

The thermal camera and color camera are coplanar, so the multi-camera system can be treated as a simple stereo system as in Fig. 6, where the center of the target object is marked as the red point in different images for camera calibration. In the diagram, an object was placed at a distance of d. All images have undergone distortion removal processing, so they can be treated as the pinhole camera model. The mapping of the object's pixel position between the thermal camera and the color camera should be:

$$u_T = f_T\left(\frac{u_C - c_C}{f_C} + \frac{L}{z}\right) + c_T \quad (1)$$

In the equation, u_T and u_C are the pixel positions in the thermal camera and color camera sensor frame, L is the distance between the main axis of the color camera and the thermal camera, z is the object distance measured by the depth camera. Ideally, the Eq. (1) can be used for the calibration directly. However, there is a small installation error θ,

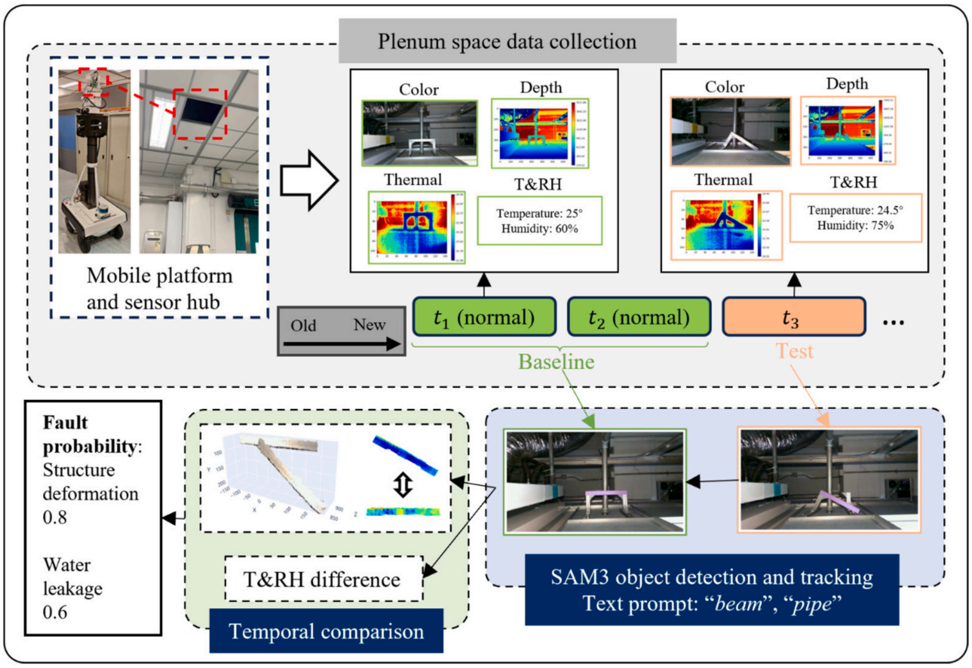


Fig. 7. Overview of the STF-Inspector.

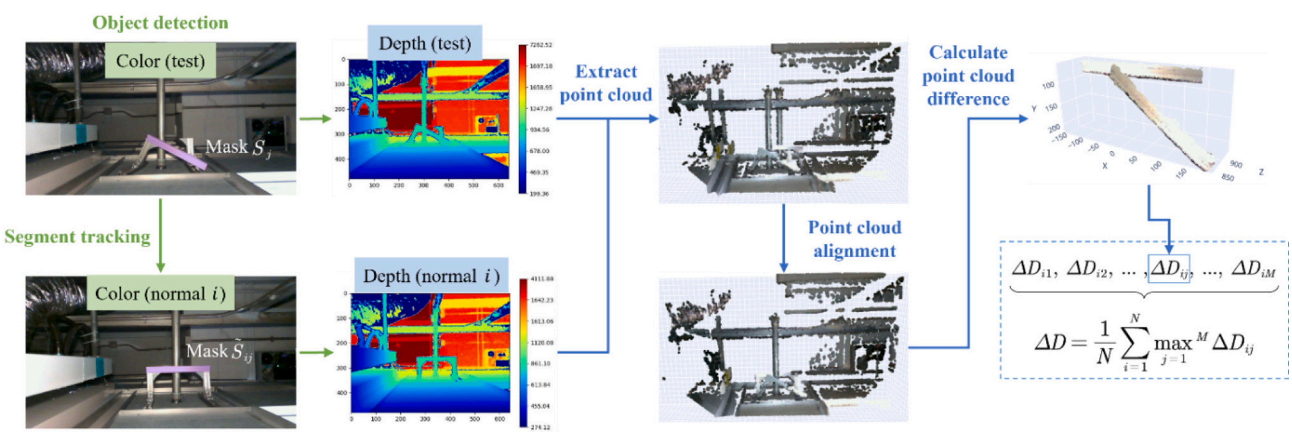


Fig. 8. Process to calculate the metric of the point cloud distance difference ΔD.

so the calibration process should be further refined.

For the condition where $h + L' \gg \Delta L$, the camera model can be written as:

$$\begin{cases} \dfrac{h}{z} = \dfrac{x_C - c_C}{f_C}, \\ \dfrac{h + L'}{z} \approx \dfrac{\overline{x}_T - c_T}{f_T}, \\ L' = L - \Delta L. \end{cases} \tag{2}$$

Hence, the actual mapping is:

$$\begin{aligned} \overline{u}_T &= f_T\left(\frac{u_C - c_C}{f_C} + \frac{L}{z}\right) - f_T\frac{\Delta L}{z} + c_T, \\ &= u_T - f_T \tan\theta = u_T + C. \end{aligned} \tag{3}$$

It shows that the actual mapping position $\overline{u}_T$ has an offset C on the ideal mapping position u_T. This offset can be obtained by identifying the same object in the color and thermal images of a heated object.

2.3. STF-inspector architecture

The overview of the STF-inspector is presented in Fig. 7. First, the data collected from the plenum space test area will be rearranged by time, each set will contain a color, depth, and thermal image, with the environmental temperature and humidity. The open-set object detection model SAM3 is incorporated to detect the target object following the given text prompt. After the object detection, the detected object should be compared with its baseline. The AOT is used to detect the same object in the baseline data. With the object detection and tracking, the temporal comparison pair is established and ready for the feature extraction. Before extracting features, the depth and thermal images are aligned with the color image, so the point cloud and thermal value inside the object mask can be extracted and compared. This paper developed robust, physically grounded feature extraction methods to eliminate the possible disturbance that affects the final decision-making process. The following subsections will discuss the details.

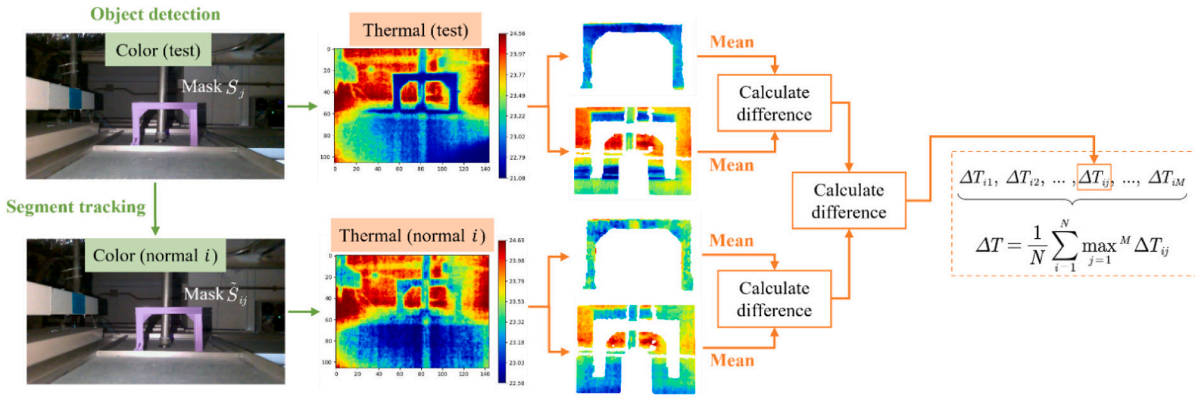


Fig. 9. Process of calculating the metric of the relative thermal value difference ΔT.

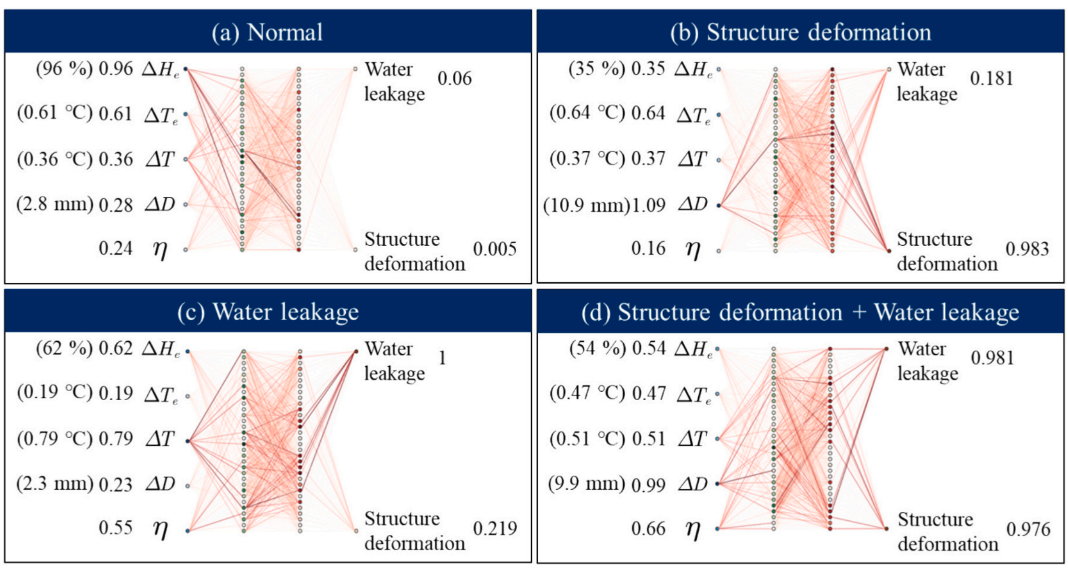


Fig. 10. Demonstration of the neural network data stream.

2.3.1. *Structure deformation and water leakage detection*

The framework begins with the color image of the test data. In this step, SAM3 detects objects for the "color (test)" in the object detection. Next, the identified masks are tracked in "color (normal i)" which are used as the baseline for comparison. With these two methods embedded, our framework can focus on the same objects between two images. The outputs of "color (test)", S_j, are the masks of the identified objects, where $j \in \{1, 2, \cdots, M\}$ represents the object IDs. The framework uses text prompts like "beams" and "pipes" for SAM3 to identify objects associated with structure deformation and water leakage. In our test, SAM3 demonstrated reliable performance in segmenting objects related to the prompts in images within novel environments featuring complex structural arrangements. Once the masks S_j are obtained, they are used to derive the tracking masks $\widetilde{S}_{ij}$ in the "color (normal i)" through segment tracking, where $i \in \{1, 2, \cdots, N\}$. Using the initial masks S_j, the AOT tracks the object masks and maintains their IDs in $\widetilde{S}_{ij}$. To address potential situations such as when objects appear in the test data but not in the normal data, the mask missing ratio is calculated during the tracking process. The expression for the mask missing ratio is:

$$\eta = \frac{1}{N}\sum_{i=1}^{N}\frac{m_i}{M} \tag{4}$$

where m_i is the count of the missing masks in the "color (normal i)".

Next, the object masks S_j and $\widetilde{S}_{ij}$ are used to extract the object's point cloud and mean thermal value from the depth and thermal images of the test and normal data. The point cloud distance ΔD_{ij} and the relative thermal value difference ΔT_{ij} are calculated to quantify structure deformation and water leakage. The process of calculating the point cloud distance ΔD_{ij} is illustrated in Fig. 8. It first aligns the global point cloud to eliminate the difference introduced by the robot movement and then measures the structure deformation by calculating the object's point cloud difference. And the process of calculating the relative thermal value difference ΔT_{ij} is illustrated in Fig. 9. It first compares the mean thermal value of the object and its peripheral area, and then calculates the difference between the relative thermal values of the test data and the normal data.

After this process, a list of ΔD_{ij} and ΔT_{ij} are obtained, which are further refined into two aggregated metrics, ΔD and ΔT, defined as:

$$\begin{cases} \Delta D = \dfrac{1}{N}\sum_{i=1}^{N}\max_j \Delta D_{ij}, \\ \Delta T = \dfrac{1}{N}\sum_{i=1}^{N}\max_j \Delta T_{ij}. \end{cases} \tag{5}$$

The environmental temperature and relative humidity (T&RH) differences are also incorporated into fault detection. They are denoted as ΔT_e and ΔH_e, with the expression:

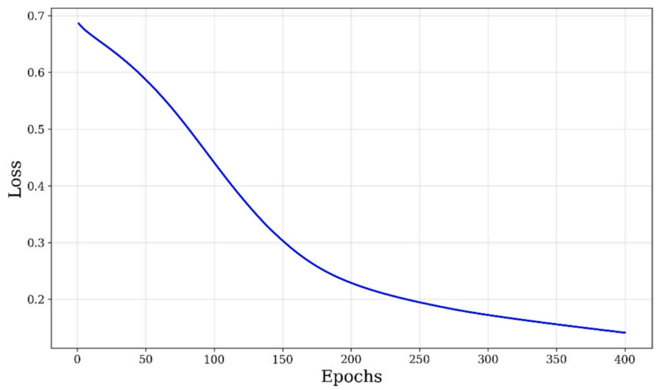


Fig. 11. Loss curve of the MLP training.

$$\begin{cases} \Delta T_e = T_e - \overline{T}_0, \\ \Delta H_e = H_e - \overline{H}_0. \end{cases} \tag{6}$$

In the equations, T_e and H_e are the environmental temperature and humidity of the test data, $\overline{T}_0$ and $\overline{H}_0$ are the average environmental temperature and humidity of the normal data.

Finally, the multi-layer perceptron (MLP) consists of two hidden layers and one output layer, as shown in Fig. 10. The network uses η, ΔD, ΔT, ΔT_e and ΔH_e to generate a probability prediction of the structure deformation and water leakage. The input passes through two fully connected layers, each with 32 neurons and ReLU activation function. The network has 2 outputs, and a Sigmoid to convert into fault probabilities of the structure deformation and water leakage.

2.3.2. Calculation of point cloud difference

Global and object point clouds are extracted from depth images for structure deformation detection. To address LiDAR measurement noise, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) filters outliers by clustering dense regions. For alignment, global point clouds (test vs. normal data) are registered using Fast Point Feature Histograms (FPFH) features with Random Sample Consensus (RANSAC) for coarse alignment, followed by Iterative Closest Point (ICP) with Point-to-Plane metric for refined transformation. Object point clouds are aligned using the derived global transformation matrix before computing deformation differences.

Structure deformation is detected by calculating differences between object point clouds. After aligning the global point clouds, these differences should stem solely from the object's movement relative to the environment. While Hausdorff distance might seem a natural choice for quantifying such differences, its application in the plenum space presents challenges: beams and pipes frequently extend beyond the frame and occlude each other in most configurations. Consequently, even undeformed objects exhibit different masks between "color (test)" and "color (normal i)" due to robot motion. The Hausdorff distance can be expressed as:

Table 4
Demonstration of the structure deformation and water leakage detection.

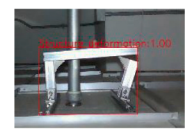
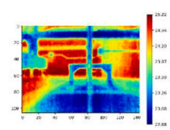
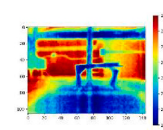
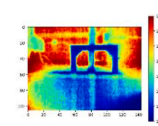
Normal	Structure deformation	Water leakage
		
		

Table 5
Demonstration of the structure deformation and water leakage detection.

Fault	Precision	Recall	F1-Score
Structure deformation	97.26%	93.55%	95.37%
Water leakage	100.00%	91.43%	95.52%

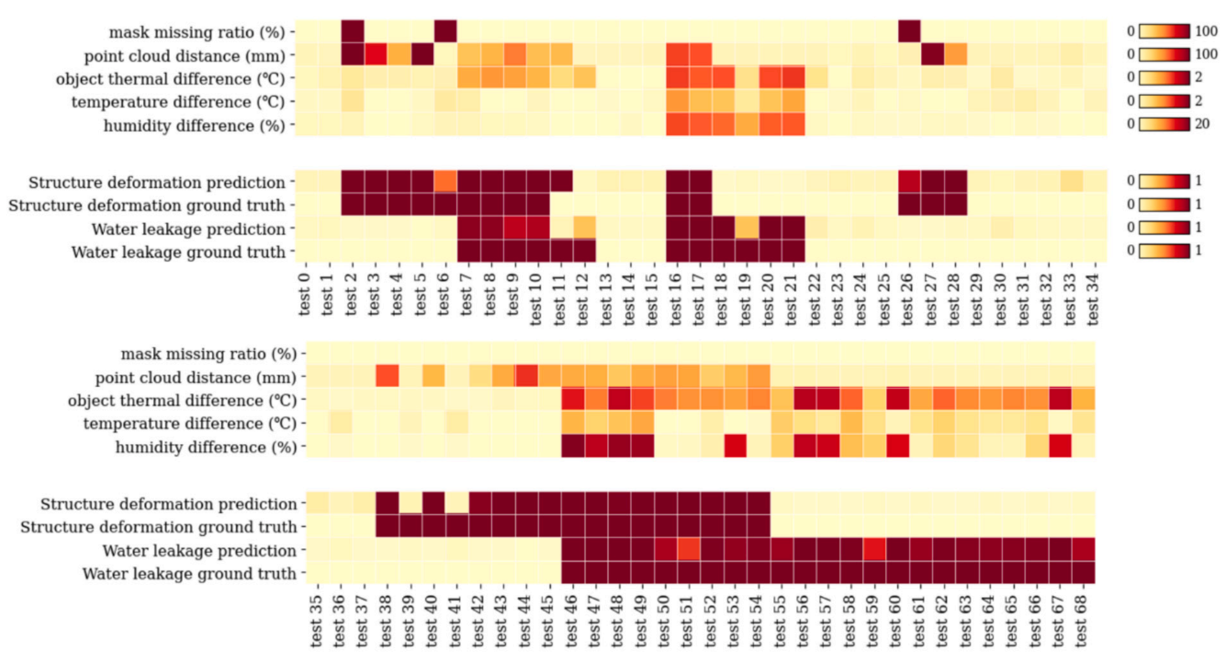


Fig. 12. MLP prediction evaluation: heatmap comparison of 69 structural monitoring test cases.

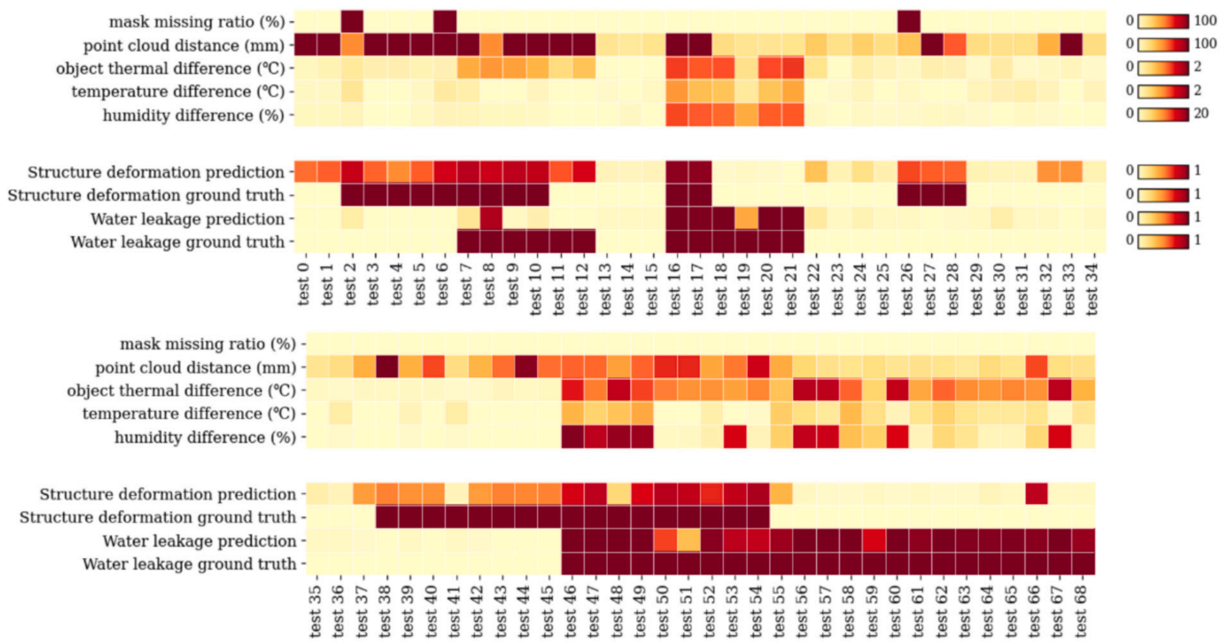


Fig. 13. MLP prediction evaluation (Hausdoff distance to measure the point cloud distance): heatmap comparison of 69 structural monitoring test cases.

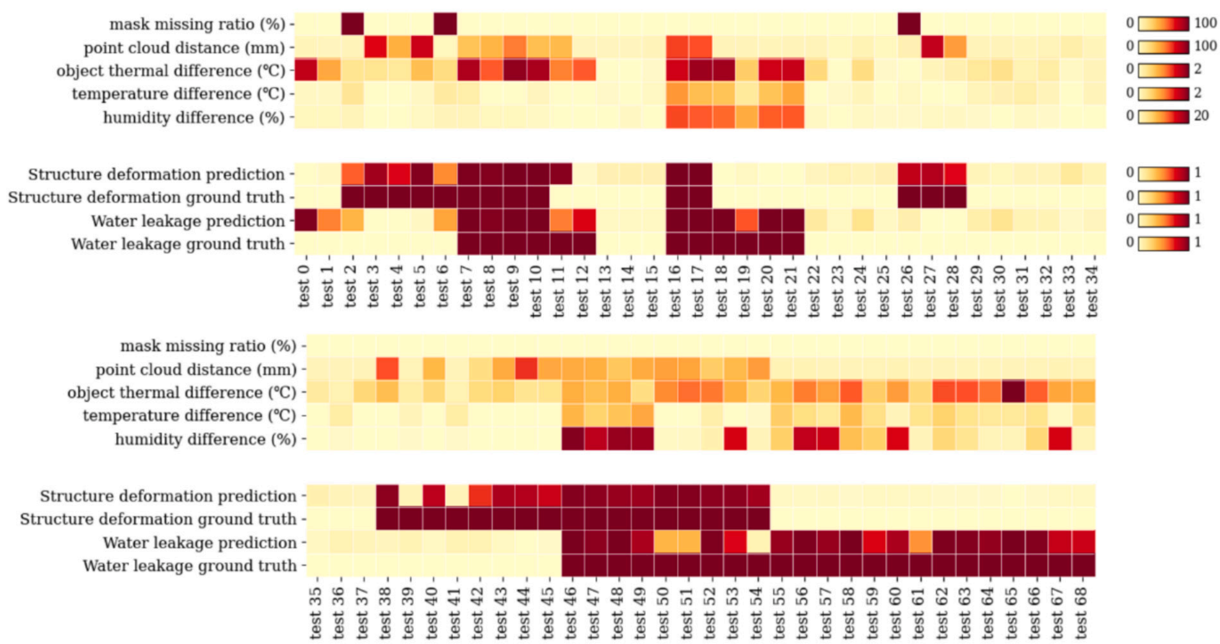


Fig. 14. MLP prediction evaluation (use absolute object temperature difference): heatmap comparison of 69 structural monitoring test cases.

$$H(A,B) = \max\left\{\max_{a\in A}\min_{b\in B} \| a-b \|, \max_{b\in B}\min_{a\in A} \| a-b \|\right\} \tag{7}$$

It takes the maximum distance between the point clouds A and B in both directions, which is very sensitive to the point cloud completeness. Hence, the point cloud comparison is further refined by using the percentile function:

$$D\left(A,B\right) = \min\left\{\max_{a\in A} Q_{0.9}\min_{b\in B} \| a-b \|, \max_{b\in B} Q_{0.9}\min_{a\in A} \| a-b \|\right\} \tag{8}$$

The notation $\max Q_{0.9}$ means that the function sorts the distances from small to large and takes the value at 90%. It takes the minimum rather than the maximum distance in both directions because the function should focus on the overlapping region. In our following test, this function shows better robustness against the mask being incomplete and occlusion.

2.3.3. Generalization capability through decoupled architecture

To address the challenge of limited real-world fault data and the constraints of artificial defect generation described in Section 2.2, the STF-Inspector employs a modular design that significantly enhances generalization capability. Unlike traditional end-to-end fault detection networks that are susceptible to Out-Of-Distribution (OOD) errors when facing unseen environments or textures, our framework relies on the pre-trained foundation model SAM3 and explicit physical features. SAM3 provides robust open-set segmentation capabilities derived from vast pre-training, while the extracted features, point cloud distance, and

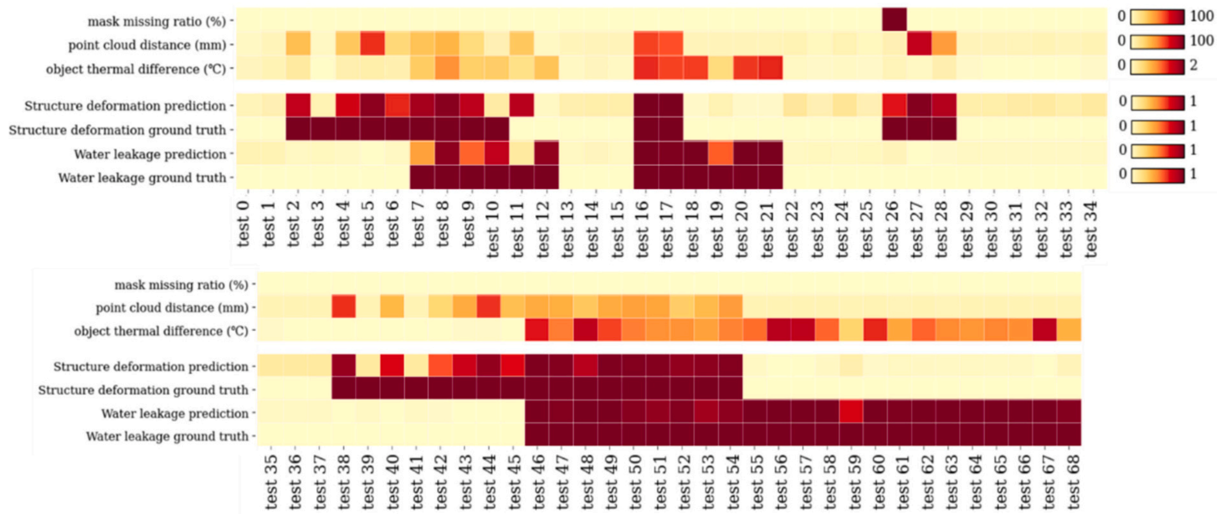


Fig. 15. MLP prediction evaluation (excluding environment temperature and humidity): heatmap comparison of 69 structural monitoring test cases.

Table 6
Structure deformation detection result comparison.

Method	True Positive	False positive	True Negative	False Negative
STF-Inspector	93.55%	2.63%	97.37%	6.45%
Hausdoff distance	83.87%	13.16%	86.84%	16.13%

Table 7
Water leakage detection result comparison.

Method	True Positive	False positive	True Negative	False Negative
STF-Inspector	91.43%	0.00%	100.00%	8.57%
Absolute thermal difference	88.57%	5.88%	94.12%	11.43%

relative thermal difference are physically grounded metrics independent of specific visual appearances. Consequently, the training process is confined solely to the lightweight MLP classifier based on these unified, low-dimensional feature vectors. This decoupling of feature extraction from decision-making ensures that the system maintains high performance across diverse plenum environments without requiring extensive retraining on massive, task-specific datasets.

3. Experiment

The fault classification model (a multi-layer perceptron) was implemented using the PyTorch framework on a workstation with an Intel Xeon w5–3425 CPU and a single NVIDIA GeForce RTX 4080 SUPER GPU (for the SAM3 model deployment). The model parameters are optimized by using the Adam optimizer with a fixed learning rate of 0.001. The binary cross-entropy loss function (BCELoss) was employed to supervise the multi-label classification task. The model was trained using full-batch gradient descent for 400 epochs. No early stopping criteria were applied as the loss curve demonstrated stable convergence without overfitting within this duration. The loss curve of the MLP training is provided in Fig. 11. The entire training pipeline required approximately 1 h. This duration primarily accounts for the computationally intensive extraction of geometric and thermal features from the raw multimodal dataset before the rapid MLP optimization phase.

The demonstration of the MLP's input-output relationships against ground truth labels of all test samples is shown in Fig. 12. The results reveal two distinct patterns: structure deformation probability increases proportionally with mask missing ratios and mean point cloud differences, while water leakage likelihood shows an inverse correlation with thermal variations and a positive correlation with humidity differences. These trends align with fundamental engineering failure principles. The output of the MLP model is shown in Table 4, and the model performance evaluation metrics: precision, recall, and F1-score are provided in Table 5. As described in the previous section, the trained MLP discriminator is an independent module decoupled from the feature extractor: it only receives feature vectors from the fusion network, thus remaining unaffected by image resolution or target quantity constraints. Leveraging this property, the MLP can operate directly on local features extracted from any candidate box/segmented region.

Comparative experiments are designed to evaluate critical components: 1) Replacing our point cloud distance method with Hausdorff distance for structure deformation detection, as shown in Fig. 13, and 2) Using the absolute thermal value difference rather than the relative thermal value difference to reveal its impact on leakage identification, as shown in Fig. 14. An ablation experiment is also designed to evaluate the impact of removing the environmental temperature and humidity from the system input, the result is shown in Fig. 15.

Comparative experiments validate the effectiveness of the proposed framework. For the structure deformation detection test, as shown in Table 6, the STF-Inspector achieves a true positive rate (TP) of 93.55% and true negative rate (TN) of 97.37%, outperforming the Hausdorff

Table 8
Ablation experiment result.

Fault	Method	True Positive	False positive	True Negative	False Negative
Structure deformation	STF-Inspector	93.55%	2.63%	97.37%	6.45%
	STF-Inspector (excluding T&RH)	87.10%	2.63%	97.37%	12.90%
Water leakage	STF-Inspector	91.43%	0%	100.00%	8.75%
	STF-Inspector (excluding T&RH)	94.29%	0%	100.00%	5.71%

distance method (TP: 83.87%, TN: 86.84%) with significantly lower false positives (2.63% vs. 13.16%) and false negatives (6.45% vs. 16.13%). In the water leakage detection test, as shown in Table 7, the STF-Inspector demonstrates higher robustness with 100% TN and an 8.57% false negative rate, surpassing the absolute thermal value difference method (TN: 94.12%, FN: 11.43%). These results confirm the framework's reliability in critical infrastructure monitoring, balancing sensitivity and specificity for safety-critical applications.

To assess the contribution of environmental conditioning, environmental temperature and relative humidity inputs are removed from STF-Inspector while keeping all other settings unchanged. As the result shown in Table 8, eliminating these signals has little effect overall. For water leakage detection, the specificity remains essentially unchanged, and the miss rate varies only marginally, indicating that the model relies primarily on relative thermal contrast rather than absolute ambient values. In contrast, structure deformation detection shows a slight drop in TP (with a corresponding increase in FN) after removing the environment cues. Because deformation cues are predominantly geometric-based, this can be attributed to minor fluctuations in feature extraction. In conclusion, with the object's thermal value difference provided, the

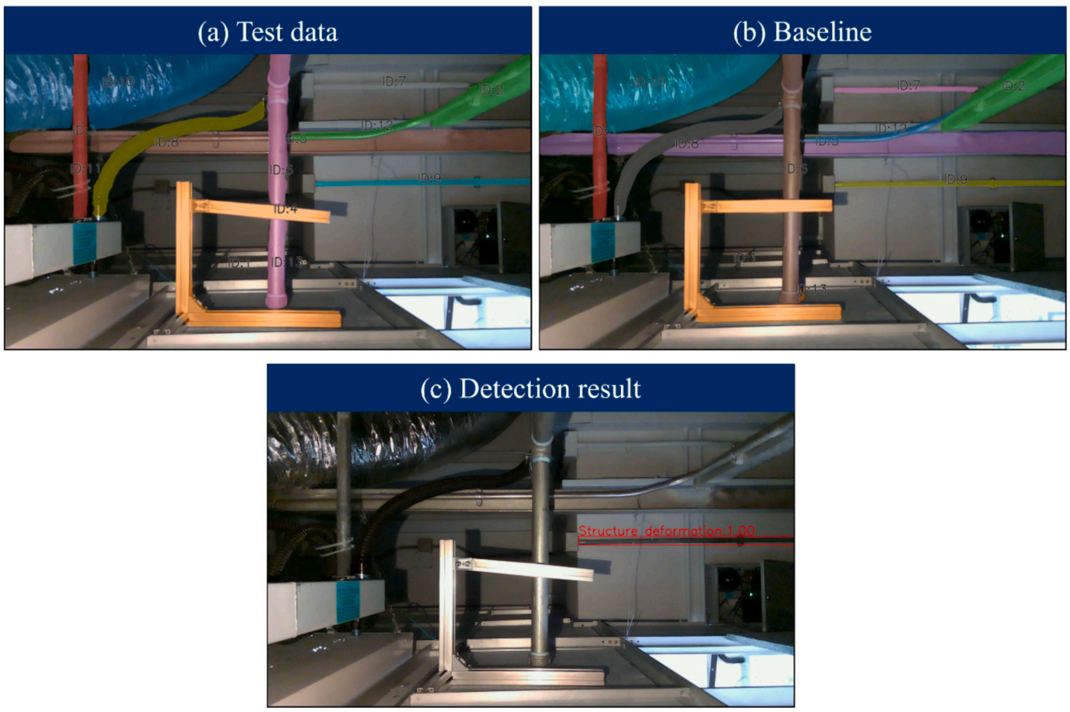


Fig. 16. A representative false prediction with 1 false positive and 1 false negative of the structure deformation.

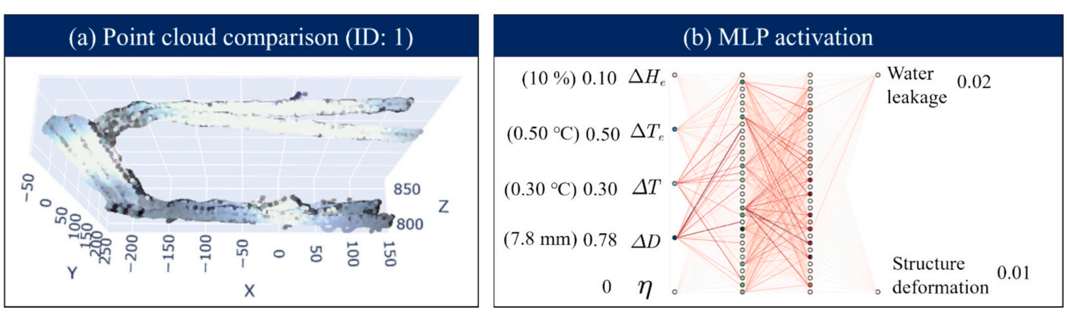


Fig. 17. Comparison of the point clouds of the fault-inducing object (ID: 1) in the test data and baseline.

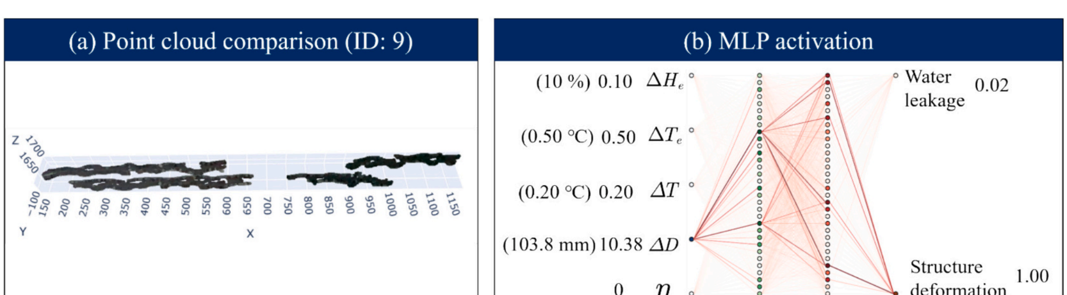


Fig. 18. Comparison of the point clouds of the false positive object (ID: 9) in the test data and baseline.

detector is insensitive to ambient temperature variations, as the thermal frames already encode the necessary temperature context.

3.1. Failure analysis

To provide a deeper understanding of the system's limitations, a granular analysis of the false predictions observed in the experimental results is conducted. Specifically, it isolates characteristic FP/FN samples to examine the feature vectors and the corresponding response of the MLP neurons.

Two distinct failure modes for structure deformation are analyzed: one related to sensitivity thresholds and the other to data quality. First, as shown in Fig. 16, an FN occurred with object ID 1. In this case, the fault-inducing object experienced a minor displacement primarily in the XY plane (horizontal shift). This detection failure stems from the single-viewpoint nature of the data acquisition. Since the sensor only captures the visible surface facing the robot, the resulting point clouds for beams and pipes appear as partial, planar patches rather than complete 3D volumes. Consequently, when such a "flat" surface undergoes a small lateral shift (sliding along its own plane), the computed point-to-plane distance remains negligible compared to shifts along the depth axis (Z-axis), causing the MLP to miss the fault, as shown in Fig. 17. Second,

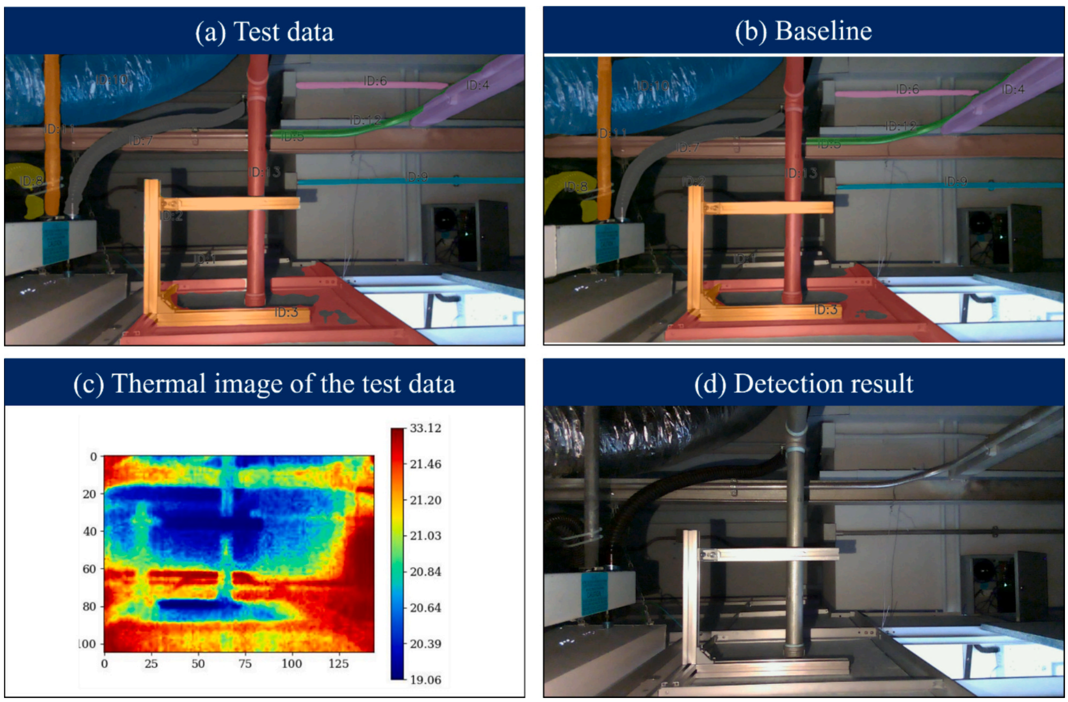


Fig. 19. Representative false prediction with a false negative of the water leakage.

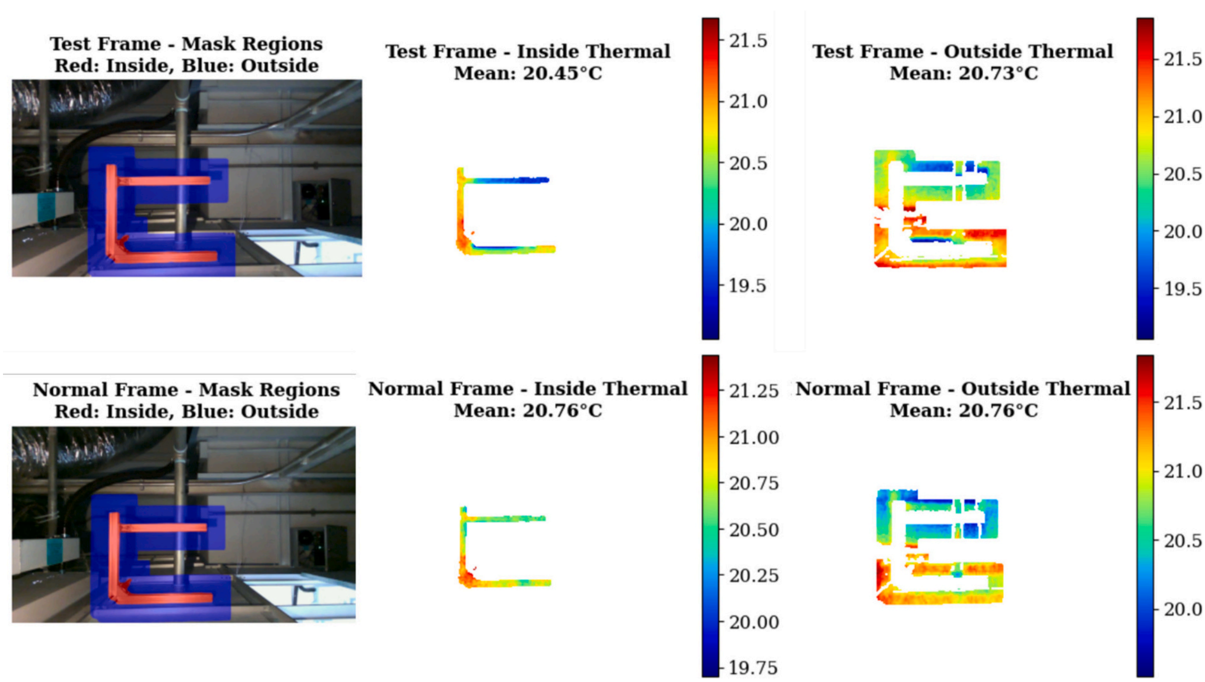


Fig. 20. Thermal value inside/outside the mask of the false negative prediction.

Fig. 18 presents a False Positive case involving object ID 9. This object is a thin, distant structural component. Due to the sparsity of the LiDAR point cloud at longer ranges, the captured geometry was incomplete and noisy. When compared to the baseline, the "missing" points caused by sensor resolution were misinterpreted by the algorithm as significant geometric deformation, triggering a false alarm. This indicates a limitation in processing objects with low point density.

For water leakage detection, the primary failure mode is thermal equilibrium. Fig. 19 illustrate an FN case where the system failed to identify a leak. As seen in the thermal distribution analysis (Fig. 20), the temperature difference between the wet surface (inside the mask) and the surrounding area (outside the mask) was minimal. This occurred because the leak was small and had cooled to the ambient temperature of the plenum space. Without a significant thermal gradient, the relative thermal difference metric remained low, causing the MLP to classify the state as normal. This highlights the challenge of using passive thermography for detecting "isothermal" leaks that have stabilized the environment.

4. Conclusion

This paper presented STF-Inspector, a segmentation–tracking–fusion framework for mobile-robot inspection of building plenum spaces. It combined open-set object segmentation (SAM3) with temporal association (AOT) to enable object-centric, history-aware fault detection by fusing RGB, depth, thermal, and environmental data. The pipeline used global-to-object point-cloud registration with a percentile-based distance metric, and incorporated relative thermal contrast with ambient context to improve robustness to occlusion, incomplete views, and sensor noise common in crowded plenums.

Experiments in an operational plenum showed that the percentile-based point-cloud difference was more robust than Hausdorff distance for deformation detection under partial observations, and that relative thermal contrast identified leakage better than absolute temperature differences. Ablation results indicated that the feature design remained stable even without temperature/humidity inputs, suggesting that the proposed physically grounded features captured the dominant fault cues. Overall, the study argued that decoupling foundation-model perception from physics-based feature extraction could yield reliable inspection in visually diverse real-world plenums, and highlighted that maintaining temporal object identity was as critical as instantaneous perception for safety-relevant inspection.

Failure analysis revealed limitations: the geometric module struggled with small in-plane displacements when the observed surface was nearly planar in a single view and was vulnerable to sparse or low-quality depth measurements for distant or light-absorbing materials; the leakage module could miss isothermal leaks when wet areas equilibrated with the environment and thermal gradients faded. These issues suggested that some fault modes were inherently ambiguous under passive single-view sensing.

Future work will focus on: (1) multi-view or active viewpoint planning to reduce planar ambiguity and improve in-plane sensitivity; (2) complementary sensing or active thermography for sparse point clouds and low-contrast leaks; (3) uncertainty-aware inference to separate occlusion/sensor degradation from real changes; (4) long-term deployments addressing seasonal/operational drift via baseline updating and asset-specific normalization; and (5) operational support for opening/removing ceiling tiles to reduce reliance on existing access points. STF-Inspector targets preventive maintenance in large infrastructures and can integrate with CMMS and digital-twin/BIM workflows for trend monitoring, prioritized work orders, and data-driven inspection scheduling.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.autcon.2026.106889.

CRediT authorship contribution statement

Gao Lou: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Shuai Liu:** Writing – review & editing, Software, Methodology. **Yi Shen:** Software, Data curation. **Wing Yin Wong:** Writing – review & editing, Validation. **Yanglong Lu:** Writing – review & editing, Supervision. **Molong Duan:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT and Grammarly AI to perform grammar checks. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the financial support from the GREAT Smart Cities Institute, the HKUST-MTR Joint Research Laboratory (Grant No. MTR22EG01-H), and the HKUST Frontier Technology Research for Joint Institutes with Industry Scheme (Grant No. OKT26EG04). Professors Xiaomeng Li and Dan Xu are gratefully acknowledged for their valuable suggestions and constructive feedback throughout the development of this work.

Data availability

Data will be made available on request.

References

[1] S.H. Cha, C. Koo, T.W. Kim, T. Hong, Spatial perception of ceiling height and type variation in immersive virtual environments, Build. Environ. 163 (2019) 106285, https://doi.org/10.1016/j.buildenv.2019.106285.
[2] Z.-Z. Hu, J.-P. Zhang, F.-Q. Yu, P.-L. Tian, X.-S. Xiang, Construction and facility management of large MEP projects using a multi-scale building information model, Adv. Eng. Softw. 100 (2016) 215–230, https://doi.org/10.1016/j.advengsoft.2016.07.006.
[3] Building and Construction Authority, Advisory on good practices for the inspection and maintenance of suspended ceilings (Circular ref. APPBCA-2021-12), 2021. https://www1.bca.gov.sg/docs/default-source/docs-corp-news-and-publications/circulars/good-practices-for-the-inspection-and-maintenance-of-suspended-ceilings.pdf.
[4] S. Selvakumaran, A.A. Hayat, K. Elangovan, K. Manivannan, M.R. Elara, Design of a self-reconfigurable robot with roll, crawl, and climb features for false ceiling inspection task, in: H. Muthusamy, J. Botzheim, R. Nayak (Eds.), Robotics, Control and Computer Vision, Springer Nature Singapore, Singapore, 2023, pp. 467–477, https://doi.org/10.1007/978-981-99-0236-1_37.
[5] Y. Nishimura, H. Mochiyama, T. Yamaguchi, Propeller-type wall-climbing robot for visual and hammering inspection of concrete surfaces, IEEE Access 12 (2024) 70963–70972, https://doi.org/10.1109/ACCESS.2024.3402218.
[6] R. Parween, T.Y. Wen, M.R. Elara, Modelling and analysis of a vertical wall robot for ceiling inspection, in: Proceedings of the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, Las Vegas, NV, USA, 2022, pp. 679–687, https://doi.org/10.1109/CCWC54503.2022.9720745.
[7] R.-J. Yan, E. Kayacan, I.-M. Chen, L.K. Tiong, J. Wu, QuicaBot: quality inspection and assessment robot, IEEE Trans. Autom. Sci. Eng. 16 (2) (2019) 506–517, https://doi.org/10.1109/TASE.2018.2829927.
[8] S. Liu, M. Duan, Cable-driven parallel robot trajectory generation with optimized orientation considering disturbance rejection, Mech. Mach. Theory 210 (2025) 106016, https://doi.org/10.1016/j.mechmachtheory.2025.106016.

[9] Y. Zhou, M. Duan, Vibration compensation of an extendable variable-stiffness boom-lift-mounted robot, IEEE/ASME Trans. Mechatron. 29 (4) (2024) 2812–2820, https://doi.org/10.1109/TMECH.2024.3402054.

[10] W. Cai, L. Huang, Z. Zou, Actively-exploring thermography-enabled autonomous robotic system for detecting and registering HVAC thermal leaks, Autom. Constr. 152 (2023) 104901, https://doi.org/10.1016/j.autcon.2023.104901.

[11] L. Ge, A. Sadhu, Deep learning-enhanced smart ground robotic system for automated structural damage inspection and mapping, Autom. Constr. 170 (2025) 105951, https://doi.org/10.1016/j.autcon.2024.105951.

[12] A. Semwal, R.E. Mohan, L.M.J. Melvin, P. Palanisamy, C. Baskar, L. Yi, S. Pookkuttath, B. Ramalingam, False ceiling deterioration detection and mapping using a deep learning framework and the teleoperated reconfigurable 'Falcon' robot, Sensors 22 (1) (2022) 262, https://doi.org/10.3390/s22010262.

[13] G.A. Atkinson, W. Zhang, M.F. Hansen, M.L. Holloway, A.A. Napier, Image segmentation of underfloor scenes using a mask regions convolutional neural network with two-stage transfer learning, Autom. Constr. 113 (2020) 103118, https://doi.org/10.1016/j.autcon.2020.103118.

[14] N. Techasarntikul, T. Mashita, Instance segmentation of images above the ceiling using Mask R-CNN, in: Proceedings of the 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), IEEE, Cape Town, South Africa, 2021, pp. 1–6, https://doi.org/10.1109/ICECET52533.2021.9698438.

[15] T. Czerniawski, F. Leite, Automated segmentation of RGB-D images into a comprehensive set of building components using deep learning, Adv. Eng. Inform. 45 (2020) 101131, https://doi.org/10.1016/j.aei.2020.101131.

[16] E. Menendez, J.G. Victores, R. Montero, S. Martínez, C. Balaguer, Tunnel structural inspection and assessment using an autonomous robotic system, Autom. Constr. 87 (2018) 117–126, https://doi.org/10.1016/j.autcon.2017.12.001.

[17] P. Jiao, X. Ye, C. Zhang, W. Li, H. Wang, Vision-based real-time marine and offshore structural health monitoring system using underwater robots, Comput. Aided Civ. Inf. Eng. 39 (2) (2024) 281–299, https://doi.org/10.1111/mice.12993.

[18] M. Previtali, L. Barazzetti, R. Brumana, F. Roncoroni, Thermographic analysis from UAV platforms for energy efficiency retrofit applications, J. Mobile Multimedia 9 (1–2) (2013) 66–82. https://ieeexplore.ieee.org/abstract/document/11079922.

[19] E. McLaughlin, N. Charron, S. Narasimhan, Automated defect quantification in concrete bridges using robotics and deep learning, J. Comput. Civ. Eng. 34 (5) (2020) 04020029, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000915.

[20] G. Lee, D. Kim, D. Oh, S. Lee, S. Cho, S.-H. Sim, Zero-shot framework for automated inspection of bridge expansion joints using road scan images, Autom. Constr. 179 (2025) 106444, https://doi.org/10.1016/j.autcon.2025.106444.

[21] P. Palanisamy, R.E. Mohan, A. Semwal, L.M. Jun Melivin, B. Félix Gómez, S. Balakrishnan, K. Elangovan, B. Ramalingam, D.N. Terntzer, Drain structural defect detection and mapping using AI-enabled reconfigurable robot raptor and IoRT framework, Sensors 21 (21) (2021) 7287, https://doi.org/10.3390/s21217287.

[22] X. Han, S. Chen, Z. Fu, Z. Feng, L. Fan, D. An, C. Wang, L. Guo, W. Meng, X. Zhang, R. Xu, S. Xu, Multimodal fusion and vision–language models: a survey for robot vision, Inf. Fusion 126 (2026) 103652, https://doi.org/10.1016/j.inffus.2025.103652.

[23] R. Xu, J. Zhang, J. Sun, C. Wang, Y. Wu, S. Xu, W. Meng, X. Zhang, MRFTrans: multimodal representation fusion transformer for monocular 3D semantic scene completion, Inf. Fusion 111 (2024) 102493, https://doi.org/10.1016/j.inffus.2024.102493.

[24] Y. Zhang, D. Sidibé, O. Morel, F. Mériaudeau, Deep multimodal fusion for semantic image segmentation: A survey, Image Vis. Comput. 105 (2021) 104042, https://doi.org/10.1016/j.imavis.2020.104042.

[25] A.B. Badiru, F.W. Ciarallo, E.G. Mbonimpa, Handbook of Scholarly Publications from the Air Force Institute of Technology (AFIT), Volume 1, 2000–2020, 1st ed., CRC Press, 2022 https://doi.org/10.1201/9781003220978.

[26] L. Han, J. Chen, H. Li, G. Liu, B. Leng, A. Ahmed, Z. Zhang, Multispectral water leakage detection based on a one-stage anchor-free modality fusion network for metro tunnels, Autom. Constr. 140 (2022) 104345, https://doi.org/10.1016/j.autcon.2022.104345.

[27] B. Jalil, G.R. Leone, M. Martinelli, D. Moroni, M.A. Pascali, A. Berton, Fault detection in power equipment via an unmanned aerial system using multi modal data, Sensors 19 (13) (2019) 3014, https://doi.org/10.3390/s19133014.

[28] X. Li, S. Yu, Y. Lei, N. Li, B. Yang, Dynamic vision-based machinery fault diagnosis with cross-modality feature alignment, IEEE/CAA J. Automatica Sinica 11 (10) (2024) 2068–2081, https://doi.org/10.1109/JAS.2024.124470.

[29] D. Mussina, A. Irmanova, P.K. Jamwal, M. Bagheri, Multi-modal data fusion using deep neural network for condition monitoring of high voltage insulator, IEEE Access 8 (2020) 184486–184496, https://doi.org/10.1109/ACCESS.2020.3027825.

[30] P. Mouzenidis, A. Louros, D. Konstantinidis, K. Dimitropoulos, P. Daras, T. Mastos, Multi-modal variational Faster R-CNN for improved visual object detection in manufacturing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, IEEE/CVF, 2021, pp. 2587–2594. https://openaccess.thecvf.com/content/ICCV2021W/CVinHRC/html/Mouzenidis_Multi-Modal_Variational_Faster_R-CNN_for_Improved_Visual_Object_Detection_in_ICCVW_2021_paper.html.

[31] C. Zhang, Z. Yin, R. Qin, Attention-enhanced co-interactive fusion network (AECIF-net) for automated structural condition assessment in visual inspection, Autom. Constr. 159 (2024) 105292, https://doi.org/10.1016/j.autcon.2024.105292.

[32] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, Z. Luo, Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE/CVF, 2022, pp. 5802–5811, https://doi.org/10.48550/arXiv.2203.16220.

[33] J. Grandio, B. Riveiro, D. Lamas, P. Arias, Multimodal deep learning for point cloud panoptic segmentation of railway environments, Autom. Constr. 150 (2023) 104854, https://doi.org/10.1016/j.autcon.2023.104854.

[34] C. Liu, M. Xie, C. Yuan, F. Liang, Z. Dong, B. Yang, Training-free open-set 3D inventory of transportation infrastructure by combining point clouds and images, Autom. Constr. 178 (2025) 106377, https://doi.org/10.1016/j.autcon.2025.106377.

[35] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, et al., Grounded SAM: assembling open-world models for diverse visual tasks, arXiv preprint (2024), https://doi.org/10.48550/arXiv.2401.14159 arXiv:2401.14159.

[36] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K.V. Alwala, H. Khedr, A. Huang, et al., SAM 3: segment anything with concepts, arXiv preprint (2025), https://doi.org/10.48550/arXiv.2511.16719 arXiv:2511.16719.

[37] Z. Yang, Y. Wei, Y. Yang, Associating objects with transformers for video object segmentation, in: Advances in Neural Information Processing Systems vol. 34, Curran Associates, Inc., 2021, pp. 2491–2502, in: https://proceedings.neurips.cc/paper_files/paper/2021/file/147702db07145348245dc5a2f2fe5683-Paper.pdf.

[38] B. Jafari, A. Khaloo, D. Lattanzi, Deformation tracking in 3D point clouds via statistical sampling of direct cloud-to-cloud distances, J. Nondestruct. Eval. 36 (2017) 65, https://doi.org/10.1007/s10921-017-0444-2.

[39] W. Graves, K. Aminfar, D. Lattanzi, Full-scale highway bridge deformation tracking via photogrammetry and remote sensing, Remote Sens. 14 (12) (2022) 2767, https://doi.org/10.3390/rs14122767.

[40] V.S. Kalasapudi, P. Tang, W. Xiong, Y. Shi, A multi-level 3D data registration approach for supporting reliable spatial change classification of single-pier bridges, Adv. Eng. Inform. 38 (2018) 187–202, https://doi.org/10.1016/j.aei.2018.06.010.

[41] H. Guan, T. Xiao, W. Luo, J. Gu, R. He, P. Xu, Automatic fault diagnosis algorithm for hot water pipes based on infrared thermal images, Build. Environ. 218 (2022) 109111, https://doi.org/10.1016/j.buildenv.2022.109111.