Title	Development of an occupancy measurement system for micro-zones within open office spaces based on multi-view multi-person 3D pose estimation
Author(s)	Chen, Sihua; Fukuda, Tomohiro; Yabuki, Nobuyoshi
Citation	Journal of Building Engineering. 2025, 111, p. 113037
Version Type	VoR
URL	https://hdl.handle.net/11094/102617
rights	This article is licensed under a Creative Commons Attribution 4.0 International License.
Note	

Contents lists available at ScienceDirect

Journal of Building Engineering

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

Development of an occupancy measurement system for micro-zones within open office spaces based on multi-view multi-person 3D pose estimation

Sihua Chen [a,*], Tomohiro Fukuda [a,**], Nobuyoshi Yabuki [b,c]

[a] Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, The University of Osaka, Osaka, 565-0871, Japan
[b] Advanced Research Laboratories, Tokyo City University, Tokyo, 158-8557, Japan
[c] The University of Osaka, Osaka, Japan

ARTICLE INFO

Keywords:
Occupancy measurement
Open office spaces
Functional zones
Multi-camera calibration
3D pose estimation

ABSTRACT

A micro-zone constitutes the fundamental unit of open spaces. Occupancy information of micro-zones with functional attributes is critical for layout optimization, space management, and energy control in open-plan workplaces. However, existing occupancy measurement methods predominantly focus on macro-scales and commonly suffer from reliance on additional devices, rigid behavioral constraints, and limited access to authentic human occupancy data. To address these limitations and bridge the gap in micro-scale occupancy analysis, this study proposes a design framework for occupancy measurement systems to investigate functional zone utilization in open office spaces. The framework takes multi-view RGB images as input, applies 3D pose estimation to infer spatial positions of human joints, and further determines the occupancy state and metrics of the entire space and functional zones by analyzing overlap between projected joint coordinates and zone boundaries. To evaluate the feasibility of the framework, a prototype system is developed and deployed in a small-scale open office space, achieving a Precision of 100 % and an F1 score of 89.19 % in occupancy state measurement. These results demonstrate that the framework effectively supports occupancy surveys of functional zones and holds potential for real-world application. This study presents a low-cost, multi-scale occupancy measurement approach that innovatively introduces pose estimation technology into building occupancy investigation. While enabling holistic spatial occupancy perception, it fills the gap in functional zone analysis. This work provides a theoretical foundation for the development of CCTV-based occupancy measurement methods and offers data support for decision-making in the sustainable design and operation of indoor open office environments.

1. Introduction

Over the past few decades, open-plan layouts have gradually become the mainstream choice for organizational offices, owing to their advantages in fostering creativity, enhancing communication efficiency, and reducing real estate costs [1,2]. These layouts typically lack enclosed physical boundaries such as walls or full-height partitions, and instead implicitly divide spaces into multiple

* Corresponding author.
** Corresponding author.
E-mail addresses: sihua.chen@it.see.eng.osaka-u.ac.jp (S. Chen), fukuda.tomohiro.see.eng@osaka-u.ac.jp (T. Fukuda).

https://doi.org/10.1016/j.jobe.2025.113037
Received 16 December 2024; Received in revised form 30 April 2025; Accepted 27 May 2025
Available online 12 June 2025

micro-zones through elements like furniture, plants, or ceiling designs [3,4]. Micro-zones are the basic constituent units of open-plan layouts, and occupancy data for micro-zones with specific functional attributes (i.e., functional zones) hold significant importance for layout optimization, space management, and energy control in open office spaces. Such data enable designers to evaluate the rationality of functional layouts based on empirical evidence; allow facility managers to allocate detailed spatial resources according to actual usage patterns; and support technical staff in optimizing energy control strategies based on fine-grained user demand [5–7]. However, effective methods for measuring occupancy in functional zones are currently lacking, and developing occupancy measurement techniques with high spatial resolution has become one of the critical challenges for achieving sustainable design and operation of indoor open office environments.

Nomenclature

PIR	Passive Infra-Red	CCTV	Closed-Circuit Television Video
WLC	Wireless Local Communication	RTMDet	Real-Time Multi-object Detector
RFID	Radio Frequency IDentification	MS COCO	Microsoft Common Objects in COntext
UWB	Ultra-Wide Band	OsNet	Omni-scale Network
3DPE	3D Pose Estimation	TP	True Positive
CVA	Cross-View Association	FP	False Positive
3DPR	3D Pose Reconstruction	TN	True Negative
SOTA	State-Of-The-Art	FN	False Negative
GT	Ground Truth	mIoU	mean Intersection over Union
BBox	Bounding Box	K	Intrinsic matrix
PnP	Perspective-n-Point	D	Distortion coefficients
RE	Reprojection Error	RVec	Rotation Vector
ReID	Re-IDentification	TVec	Translation Vector
PC	Personal Computer	RMS	Root Mean Square
WSL	Windows Subsystem for Linux	mRMS	mean Root Mean Square
FPS	Frames Per Second	Dim	Dimension

Traditional occupancy measurement methods are typically conducted at macro scales, taking rooms [8], floors [9], or entire buildings [10] as the basic units of surveys. In recent years, as building energy control systems have evolved toward demand-driven refinement, fine-grained occupancy data have gained increasing attention [11,12]. Such data can be described in multiple dimensions including time, space, and occupants. In the spatial dimension, existing research has achieved occupancy detection for furniture or micro-zones within the same space: The former leverages sensors such as temperature and pressure [13,14], which depend on physical contact between occupants and furniture, thus struggling to detect occupancy once occupants leave furniture; The latter employs passive infra-red (PIR) sensor-based methods [15], which cannot effectively detect stationary occupants. Moreover, PIR-based methods define micro-zones based on the fields of views of sensors rather than functional layout, thereby limiting the application value of occupancy data. A review of representative occupancy measurement studies in recent years (Table 1) reveals that existing technologies, whether targeting macro [16,9,10] or micro scales [13–15,17], struggle to obtain occupancy data for functional zones within the same space. Furthermore, they face several limitations during deployment and application, such as requiring additional devices and tags [18,19]; imposing specific constraints on occupants' states [15], behaviors [14], and carried items [20]; or failing to capture actual human occupancy data [9,21]. Therefore, there is an urgent need to develop a new measurement method that overcomes these limitations while enabling occupancy investigation across functional zones within open office spaces.

The key to occupancy investigation for functional zones lies in precise human positioning to classify occupied zones. Table 2 examines the methods with positioning capabilities from Table 1. It reveals that most employ indirect positioning strategies to obtain human locations, such as approximating positions through chairs [13], carried signal tags [20], or mobile devices [9], while only a few methods adopt direct positioning, such as capturing exact coordinates of human head centers or body centroids [16,24]. However, positioning strategies in existing occupancy measurement methods present two issues: First, methods based on indirect positioning cannot track human positions directly, thus failing to acquire genuine occupancy data; Second, these positioning strategies typically output single-point coordinates, which insufficiently represent the actual human occupancy range on a plane, resulting in their inability to perform accurate micro-zone classification. Errors in zone classification are particularly likely when human positions approach micro-zone boundaries. Therefore, it is necessary to explore a direct human positioning strategy that provides high-resolution location information to support precise identification of occupied functional zones in complex open office space scenarios.

More recently, advances in computer vision technology have provided viable pathways for acquiring high-resolution human position data. Among these, multi-view 3D pose estimation (3DPE) algorithms can infer spatial coordinates of multiple human joints (i.e., keypoints) and identify their identities through three stages: 2D pose estimation (2DPE), cross-view association (CVA), and 3D pose reconstruction (3DPR) [28]. Although these algorithms can obtain occupants' depth information based on multi-view RGB images, they still exhibit evident limitations in modular integration capability and system robustness: First, the three stages are highly coupled and often designed for specific configuration models, making it difficult to integrate state-of-the-art (SOTA) models; Second, commonly used 3D geometric structure models [29], epipolar constraint methods [30], or multi-path matching algorithms [28] in the CVA stage generally rely on geometric similarities of 2D poses, causing cross-view identity association accuracy to be susceptible to pose estimation errors. Given that closed-circuit television video (CCTV) systems have been widely deployed in open office spaces due

Table 1
Comparison between related work and the proposed solution.

	Sensor types	Sensor quantities	Other devices	Spatial resolution	User positioning	User requirements
[16]	vision	RGB camera × 1	×	room	✓	□
[22]	vision	CCTV camera × 6	×	room	✓	□
[23]	vision	thermal imaging camera × 1	√ (data processing)	room	□	□
[24]	vision	stereo camera × 2	√ (data processing)	room	✓	□
[18]	vision	RGB-D camera × 1	√ (data processing)	room	✓	□
[17]	motion	PIR × 1	√ (mobile platform)	desk	✓	□
[15]	motion	PIR × 7	×	micro-zones	✓	✓ (status)
[13]	motion, pressure	PIR × 768, pressure × 2288	×	chair	✓	✓ (status, behavior)
[14]	environment	temperature × 20	√ (data recording)	chair	✓	✓ (behavior)
[25]	environment	CO2 × 1, temperature × 1	×	room	□	□
[26]	environment, pressure	CO2, temperature, humidity, PIR (total of 66)	×	building	□	□
[27]	vision, environment	thermal imaging camera × 1、air quality × 1	√ (data processing)	room	□	□
[20]	WLC[a]	RFID[b]	√ (signal tags)	room	✓	✓ (carrying)
[19]	WLC	UWB[c]	√ (signal tags)	room	✓	✓ (carrying)
[9]	WLC	Wi-Fi	√ (mobile devices)	floor	✓	✓ (carrying)
[10]	WLC, vision, motion, environment	Wi-Fi, RGB camera, PIR, CO2 (multiple)	√ (mobile devices)	building	✓	✓ (carrying, behavior)
This work	**vision**	**RGB camera (multiple)**	×	**micro-zones**	✓	□

[a] WLC: Wireless Local communication.
[b] RFID: Radio frequency identification.
[c] UWB: Ultra-wide band.

Table 2
Comparison between the methods with positioning capabilities and the proposed solution.

	Positioning targets	Positioning outputs	Positioning resolution	Real human positioning
[16]	head centers	single-point coordinates	accurate	✓
[22]	human BBoxes[a]	single-point coordinates	fuzzy	□
[24]	human centroids	single-point coordinates	accurate	✓
[18]	human centroids	single-point coordinates	accurate	✓
[17]	desk-tops	occupied desk-tops	fuzzy	□
[15]	human bodies	occupied zones	fuzzy	✓
[13]	chairs	occupied chairs	fuzzy	□
[14]	chairs	occupied chairs	fuzzy	□
[20]	signal tags	single-point coordinates	accurate	□
[19]	signal tags	single-point coordinates	accurate	□
[9]	mobile devices	single-point coordinates	fuzzy	□
[10]	mobile devices	single-point coordinates	fuzzy	□
This work	**human joints**	**multi-point coordinates**	**accurate**	✓

[a] BBox: Bounding box.

to security and management requirements [22], studying RGB camera-based occupancy measurement methods has a solid implementation foundation and contributes to reducing hardware and technical costs. However, the limitations of existing 3DPE algorithms not only restrict the scalability of occupancy measurement methods but also affect the reliability of measurement results due to reduced indoor positioning precision. Therefore, further research is needed to enhance the modular capability and system robustness of 3DPE algorithms.

In summary, to address the limitations of existing occupancy measurement methods and bridge the gap in micro-scale occupancy investigation, this study aims to propose a measurement method supporting occupancy investigation for functional zones within open office spaces. The proposed method leverages image sequences captured by multiple RGB cameras, along with 3DPE technology, to infer the spatial coordinates of human joints. Based on predefined rules for determining occupancy states, the method analyzes the overlap between projected joint coordinates and functional zone boundaries to calculate occupancy states and related metrics for both the entire space and its functional zones. To validate the method’s feasibility, a prototype occupancy measurement system was developed and deployed in a real-world small-scale open office space. A ground truth (GT) data acquisition approach is also designed to evaluate system performance. Experimental results show that the system achieved a Precision of 100 % and an F1 score of 89.19 % in the occupancy state measurement task, indicating that the proposed method effectively supports occupancy surveys of functional zones and demonstrates good potential for real-world applications.

This study presents a low-cost, multi-scale occupancy measurement approach that innovatively applies 3DPE technology into building occupancy investigation. While enabling holistic spatial occupancy perception, it fills the gap in functional zone analysis. The acquired occupancy data demonstrate strong interpretability and practical value, providing effective data support for layout optimization, resource management, and energy control in open office environments. Compared to existing building occupancy measurement methods, the proposed approach: (1) relies solely on RGB cameras without additional equipment; (2) enables measurements under occupants’ natural behavioral patterns; (3) captures actual human occupancy data and addresses the limitations of single-point outputs in micro-zonal classification problems by describing human bodies using multiple points. Although the measurement performance of the system instance developed based on this method remains constrained by the upper bounds of the algorithms used in the 2DPE stage and the multi-viewpoint requirement of triangulation in the 3DPR stage, the findings of this study provides a theoretical foundation for the development of CCTV-based spatial occupancy measurement methods and offers data support for decision-making in the sustainable design and operation of indoor open office environments. The main contributions of this paper are as follows:

- **A design framework for occupancy measurement systems**: It obtains occupancy states and related metrics for the entire space and its functional zones within open office spaces based on multiple RGB cameras.
- **A multi-view multi-person 3DPE framework**: It infers spatial coordinates of human joints from multi-view RGB image sequences, achieving modular integration of 3DPE algorithms and avoiding error accumulation and robustness issues caused by reliance on 2D pose geometry.
- **A global calibration workflow for multi-camera systems**: It estimates and validates both intrinsic and global extrinsic parameters in indoor multi-camera systems, addressing cases where parameters are unknown or subject to deviation.
- **A prototype system for occupancy measurement**: It validates the feasibility and effectiveness of the proposed design framework by testing on a multi-person activity dataset in a small-scale open office space.

The remainder of this paper is organized as follows: Section 2 presents the methodology, proposing a design framework for occupancy measurement systems; Section 3 covers experiments and materials, developing a prototype system based on a real office scenario and proposing a method for obtaining GT data; Section 4 shows results and analysis, demonstrating the measurement results of the system and the evaluation of its performance; Section 5 discusses the advantages of the proposed method compared to existing occupancy measurement methods and clarifies its scope of application and limitations; Finally, Section 6 concludes the paper and outlines directions for future research.

2. Methodology

This section proposes a design framework for occupancy measurement systems that supports investigation for both entire space and functional zones within open office spaces. The framework takes multi-view RGB image sequences as input and applies 3DPE to extract spatial coordinates of human joints. Based on predefined occupancy rules, it then analyzes the overlap between projected joint coordinates and functional zone boundaries to output occupancy states and related metrics for both the entire space and functional zones. As shown in Fig. 1, the framework comprises four parts: preparation work (Section 2.1), data collection and preprocessing (Section 2.2), data processing (Section 2.3), and occupancy computation (Section 2.4).

2.1. Preparation work

The preparation phase includes the division of functional zones, the formulation of occupancy rules, and installation and global calibration of cameras. The zone boundaries defined in zone division are used to subsequently classify human joint positions into specific functional zones. Occupancy rules determine the occupancy states of both the entire space and its functional zones by setting thresholds on keypoint distribution and identity counts. The installation of multiple cameras should enable that the combined field of view fully covers and clearly captures the 3D spatial range of human activity. To associate coordinates of projected joints and zone vertices for interaction analysis, both must be aligned within a unified spatial coordinate system (i.e., the global coordinate system). This alignment depends on camera parameters—particularly the global extrinsic parameters that represent each camera's position and orientation in the global coordinate system—thereby necessitating global calibration of the multi-camera system.

To address scenarios where camera parameters are unknown or subject to deviation, this framework establishes a global calibration workflow for indoor multi-camera systems. This workflow utilizes Zhang's method [31] and the Perspective-n-Point (PnP) method [32] to obtain camera parameters and evaluates their accuracy through the reprojection error (RE). Zhang's method estimates the intrinsic matrix by detecting corner points across multiple images of the same checkerboard [31]. The PnP method estimates the position and orientation of multiple cameras in a unified coordinate system using known correspondences between 3D spatial points and their 2D image projections (i.e., feature points) [32]. Both methods offer advantages of operational simplicity, high precision, and strong robustness [33]. Additionally, they are based on geometric imaging principles and are not restricted by specific camera structures or lens types, thus demonstrating good universality [34]. In particular, the PnP method requires no specialized calibration tools and supports flexible feature point selection in unstructured environments [32], making it suitable for open office spaces. However, to ensure the robustness of the PnP solution, feature point selection must adhere to the following criteria: (1) no fewer than 6 points; (2) distinctive texture and clear edges; (3) clearly visible across all camera views; (4) distributed as uniformly as possible in space, avoiding coplanarity or excessive concentration [35]. Considering the significant impact of camera parameters on the precision

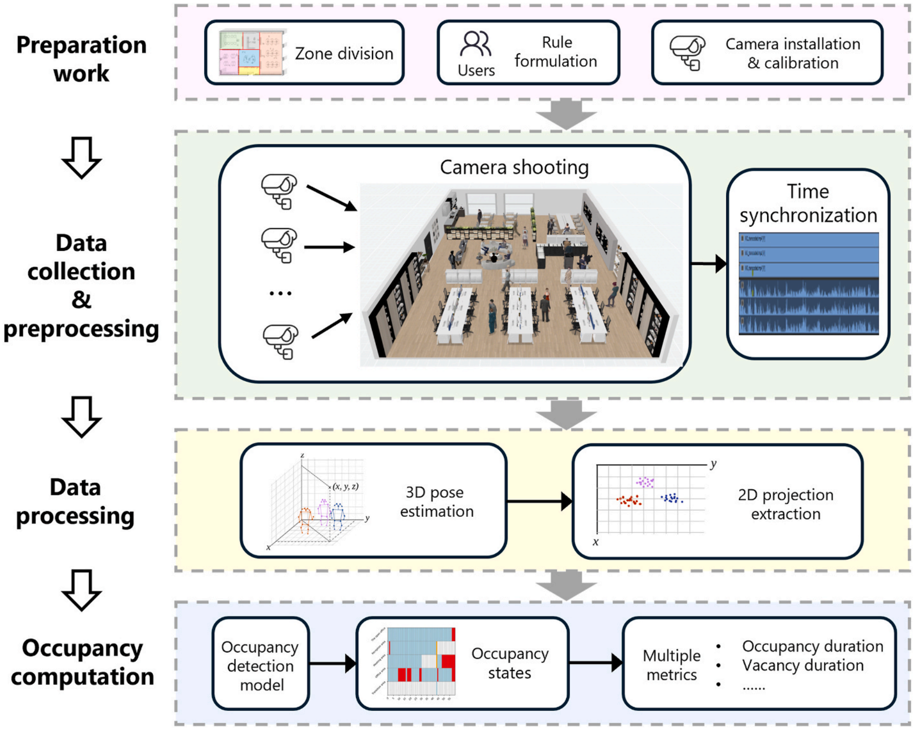


Fig. 1. The overview of the design framework for occupancy measurement systems.

of vision-based indoor positioning, rigorous assessment of parameter accuracy is indispensable. The RE refers to the distance between a 3D spatial point projected onto the image plane through a known camera model and its original 2D image point [36], thus its value can serve as an important indicator for evaluating the accuracy of camera parameters.

2.2. Data collection and reprocessing

To capture occupancy behavior data under natural patterns, the multi-camera system operates in a non-interventional environment. The acquired multi-view image sequences require time synchronization and key frame extraction to meet the processing requirements of downstream tasks. An audio signal-based method is employed for time synchronization, achieving frame-level alignment through pulse peak identification [37]. Compared to hardware-, timestamp-, and vision-based synchronization approaches, this method offers advantages of low cost, ease of deployment, and resistance to occlusion, while demonstrating strong robustness in complex environments [38]. This robustness primarily stems from the ability to artificially design synchronization signals with distinctive acoustic characteristics, which maintain high identifiability amid background noise interference and thereby enhance synchronization accuracy under non-ideal acoustic conditions [39]. Although the proposed design framework does not restrict audio signal acquisition methods, practical applications still impose requirements on both the triggering mechanism and the spatial location of the sound source. To ensure accurate multi-view video synchronization, the employed sound source must satisfy two basic conditions: (1) The propagation path difference (ΔPD) between the sound source and various cameras should be less than the propagation distance of sound waves corresponding to the maximum allowable time deviation, to avoid synchronization errors caused by differences in arrival times; (2) The sound source should possess sufficient loudness and instantaneity to generate clearly identifiable pulse peaks in the audio tracks of all cameras.

The following presents the calculation method for ΔPD referenced in the second condition. According to formula (1):

$$C = 331.6 + 0.6T \tag{1}$$

where the speed of sound in air (C) depends on temperature (T) and is independent of sound frequency. Therefore, sound signals can be generated in various ways that satisfy the first condition. To simplify calculations, open office spaces are assumed to be isothermal environments (i.e., constant T), so that C remains constant. Since differences in sound wave arrival times at the cameras (ΔPT) may

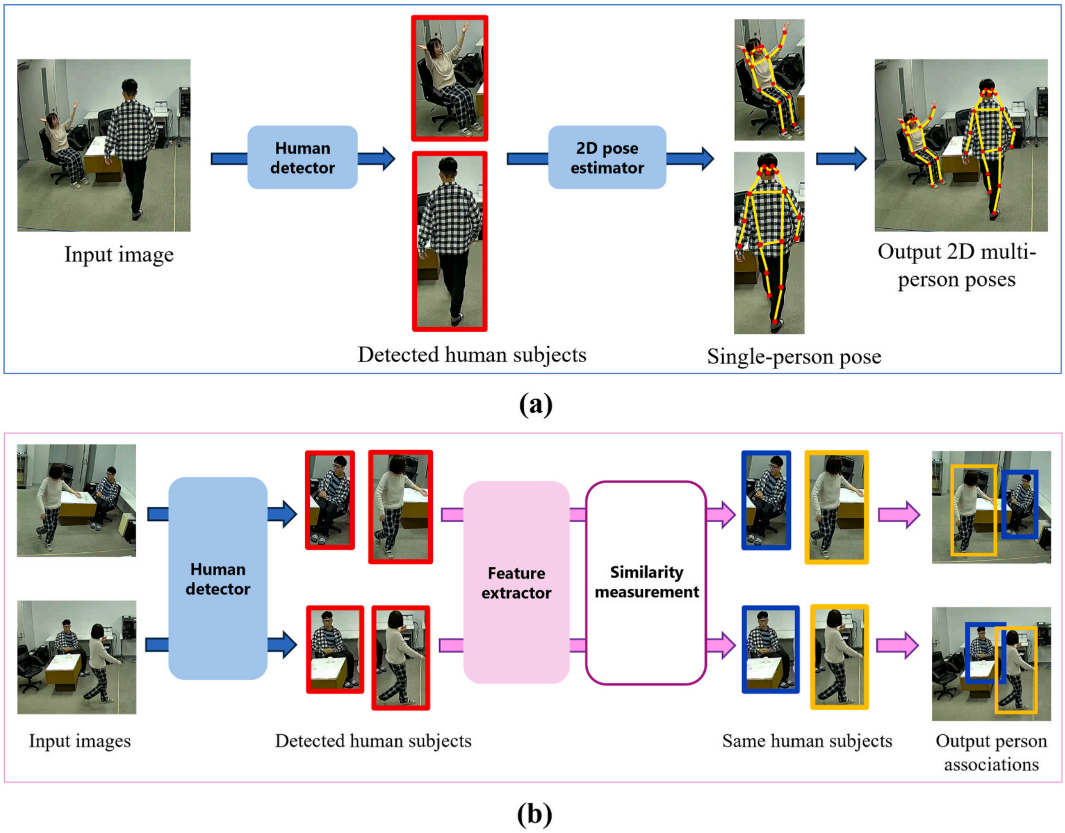


Fig. 2. The main steps in the 2DPE and CVA stages.
(a) The top-down approach in the 2DPE stage.
(b) Using ReID for cross-view matching during the CVA stage.

introduce synchronization errors, a maximum allowable ΔPT must be defined to ensure the accuracy of multi-view sequence synchronization. formula (2) for calculating *ΔPT* is as follows:

$$\Delta PT = \Delta PD / C \tag{2}$$

given that C is constant, *ΔPT* is directly proportional to *ΔPD*. The maximum value of *ΔPD* can therefore be derived by substituting the threshold value of *ΔPT* into formula (2).

Therefore, a sound signal satisfying the first condition should be triggered from a position that meets the second condition during data collection. And this signal should be ensured to be captured by all cameras. Subsequently, multi-view sequences containing this signal are exported and are synchronized based on the waveform characteristics of the audio. Finally, key frames are extracted to reduce the computational burden associated with processing similar frames.

2.3. Data processing

This section proposes an indoor real-time multi-point positioning strategy, which consists of two key steps: 3DPE and 2D projection extraction. The strategy utilizes 3DPE algorithms to recover the spatial coordinates of human joints in the global coordinate system from multi-view keyframe images, and projects them onto the ground plane, achieving high-resolution multi-point representation of individual positions. The implementation of the 3DPE algorithm relies on a constructed multi-view multi-person 3DPE framework which takes multi-view images as input and outputs 3D coordinates of all human joints based on camera parameters. The framework comprises three stages: 2DPE, CVA, and 3DPR. Specifically, the 2DPE stage estimates the 2D coordinates of human keypoints (i.e., 2D poses) in each image; The CVA stage judges the identities of 2D poses through appearance features, thereby establishing cross-view 2D pose correspondences; The 3DPR stage reconstructs the 3D pose of each person based on corresponding multi-view 2D poses. The first two stages are implemented using MMPose (for PE tasks) [40] and Torchreid (for pedestrian re-identification tasks) [41], respectively. These two open-source deep learning toolkits support a variety of SOTA algorithms and are designed with high modularity and scalability, thereby providing a foundational framework for the modular integration and system-level extension of the 3DPE algorithms. Additionally, this 3DPE framework completes cross-view identity matching through purely appearance-based cues, avoiding error accumulation and robustness issues caused by traditional matching algorithms' geometric dependence on 2D poses.

The implementation of the multi-view multi-person 3DPE framework is as follows: The 2DPE stage adopts a top-down approach, which can reduce background interference and narrow the search space while preserving image resolution. This approach first uses a human detector to human BBoxes from images, then applies a pose estimator within the BBox images to generate 2D poses (Fig. 2a); The CVA stage applies re-identification (ReID) algorithms to BBox images to extract and express individual appearance features, followed by measuring the similarity among these features (Fig. 2b). According to preset similarity thresholds, multi-view BBox images of the same individuals can be filtered out, thereby establishing identity associations for their corresponding 2D poses; The 3DPR stage uses calibration parameters (Section 2.1) to generate projection matrices for each camera, constructing mapping relationships from 3D physical space to 2D images. Finally, triangulation is applied to multi-view 2D pose sets of the same individuals to infer their 3D pose in the global coordinate system.

2.4. Occupancy computation

This section proposes a design concept for an occupancy measurement model. The model takes the ground projection coordinates of human joints and vertex coordinates of functional zones, both represented in the global coordinate system, as input. By analyzing their interaction, it outputs the occupancy states and related metrics of the entire space and its functional zones under continuous frames. The model consists of two parts: occupancy detection and occupancy computation. Occupancy detection is responsible for the output of occupancy states, with specific processes including: For keypoint data of each frame, the model first classifies all keypoints to their respective zones and counts the distribution of keypoints and number of identities in each functional zone; Second, decision conditions are designed according to occupancy rules established in Section 2.1 to determine the occupancy states of the entire space and its functional zones; Finally, by iteratively executing the above steps for all frame data, occupancy state time series for the entire space and various functional zones are generated. The occupancy computation part is responsible for the output of occupancy metrics. Occupancy metrics are derived from occupancy state data, with names and mathematical formulas that can be flexibly designed to meet individual requirements.

3. The experiment and materials

To validate the feasibility of the design framework proposed in Section 2, this section establishes an occupancy measurement prototype system in a real building environment: Section 3.1 demonstrates the development process of the system; Section 3.2 proposes a GT data acquisition method to evaluate system performance; Furthermore, the experiment also creates a test dataset containing human BBox annotations to evaluate the performance of the human detector.

3.1. Prototype system development

3.1.1. Zone division and rule formulation

Considering the requirements for environmental control, simulation costs, and debugging efficiency, the system is developed and tested in a small open office space (approximately 7 m in width, 6.5 m in depth, and 2.6 m in height). To simulate typical functional zones in open offices, the experimental space is divided into three functional zones: a reception zone, an office zone and a meeting zone (Fig. 3a). The boundaries between these zones are defined as the buffer zone (Fig. 3b), marked with colored tape to facilitate the acquisition of GT data (see Section 3.2). Table 3 formulates the determination rules for occupancy states of the entire space, functional zones, and the buffer zone. To simplify the collection of GT data, the thresholds for keypoint distribution and identity count are set at 100 % and 1, respectively, meaning that if all keypoints of at least one individual are located within a certain functional zone (or the entire space), that functional zone (or the entire space) is considered occupied. The buffer zone is introduced to account for situations in which the occupancy conditions of the entire space are satisfied, while those of the functional zones are not.

3.1.2. Hardware and software configuration

Table 4 summarizes the hardware and software configurations used for system development. Data collection employed three identical IP cameras (referred to as “main cameras”). IP cameras can access local area networks via Wi-Fi and enable remote access to image streams from computer terminals through standard network protocols [42]. The Eseecloud platform is used for multi-screen monitoring and data export of camera sources. Data processing is performed on a personal computer (PC) equipped with an NVIDIA GeForce RTX 3080 GPU. This GPU features powerful Tensor Cores and CUDA parallel computing capabilities, significantly accelerating AI inference and training processes. The operating system is Windows 10 Pro integrated with the Windows Subsystem for Linux (WSL), which supports the development of deep learning projects in a Linux-like environment while ensuring compatibility with the Windows ecosystem [43]. This system configuration enabled cross-operating system resource management and invocation, enhancing the efficiency and flexibility of occupancy measurement system’s development.

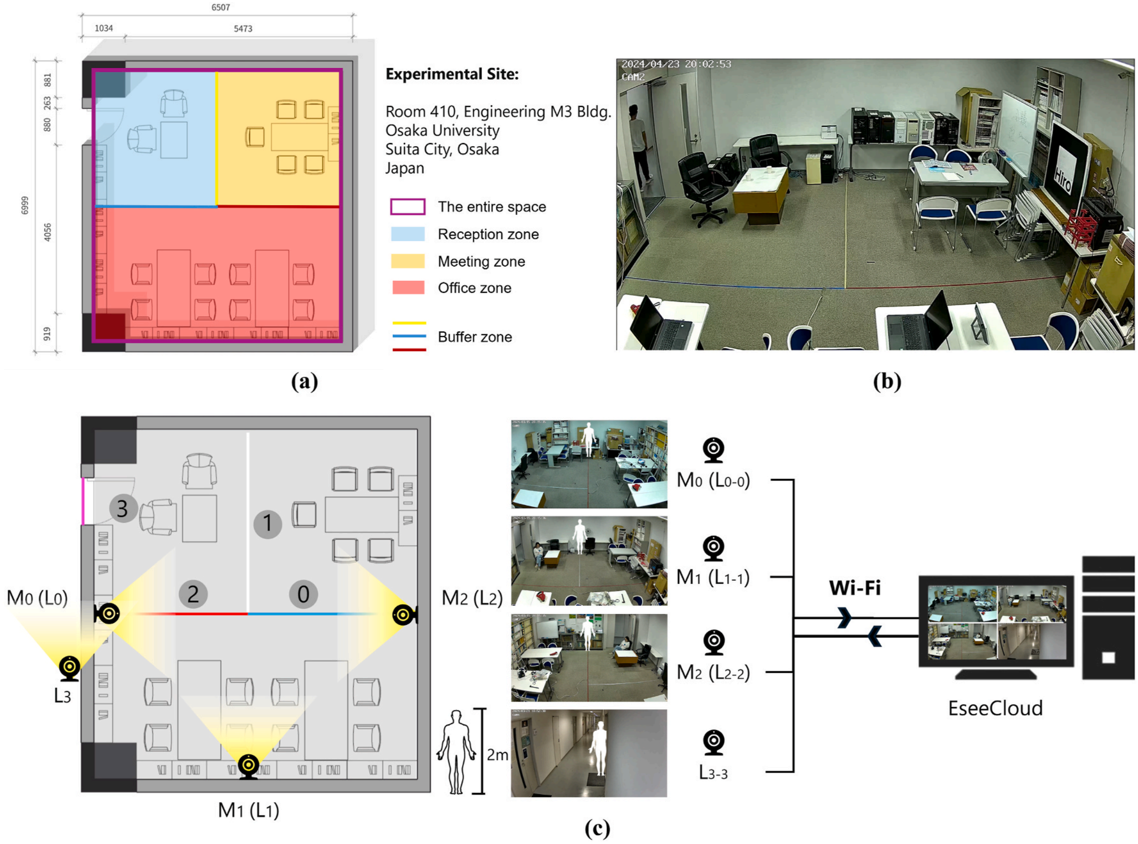


Fig. 3. Arrangement of experimental space and camera installation.
(a) Plan layout and zone division.
(b) Realistic layout and ground markings.
(c) Installation location of cameras and remote control.

Table 3
Occupation rules for the entire space, the buffer zone and functional zones.

Rule number	Detection target	Decision criteria for occupancy states
Rule_0	The entire space	All keypoints of at least one person are located within the open space.
Rule_1	The buffer zone	The keypoints of at least one person are distributed across multiple functional zones.
Rule_2	Functional zones	All keypoints of at least one person are entirely located within a specific functional zone.

Table 4
Hardware and software configuration.

Hardware	**Device Info.**		
	Home-built desktop PC	OS	Windows 10 64bit + WSL 2
		CPU	Inter(R) Core(TM) I9-9900 KF
		GPU	NVIDIA GeForce RTX 3080
		RAM	32.0 GB
	Hiseeu Security Camera Set	Camera	fisheye, wireless, five million pixels
Software	**Application Info.**		
	PyCharm	Pro 2023.3.3	
	Adobe Premiere	Pro 2024	
	Audacity	3.7.3	

3.1.3. Camera installation and calibration

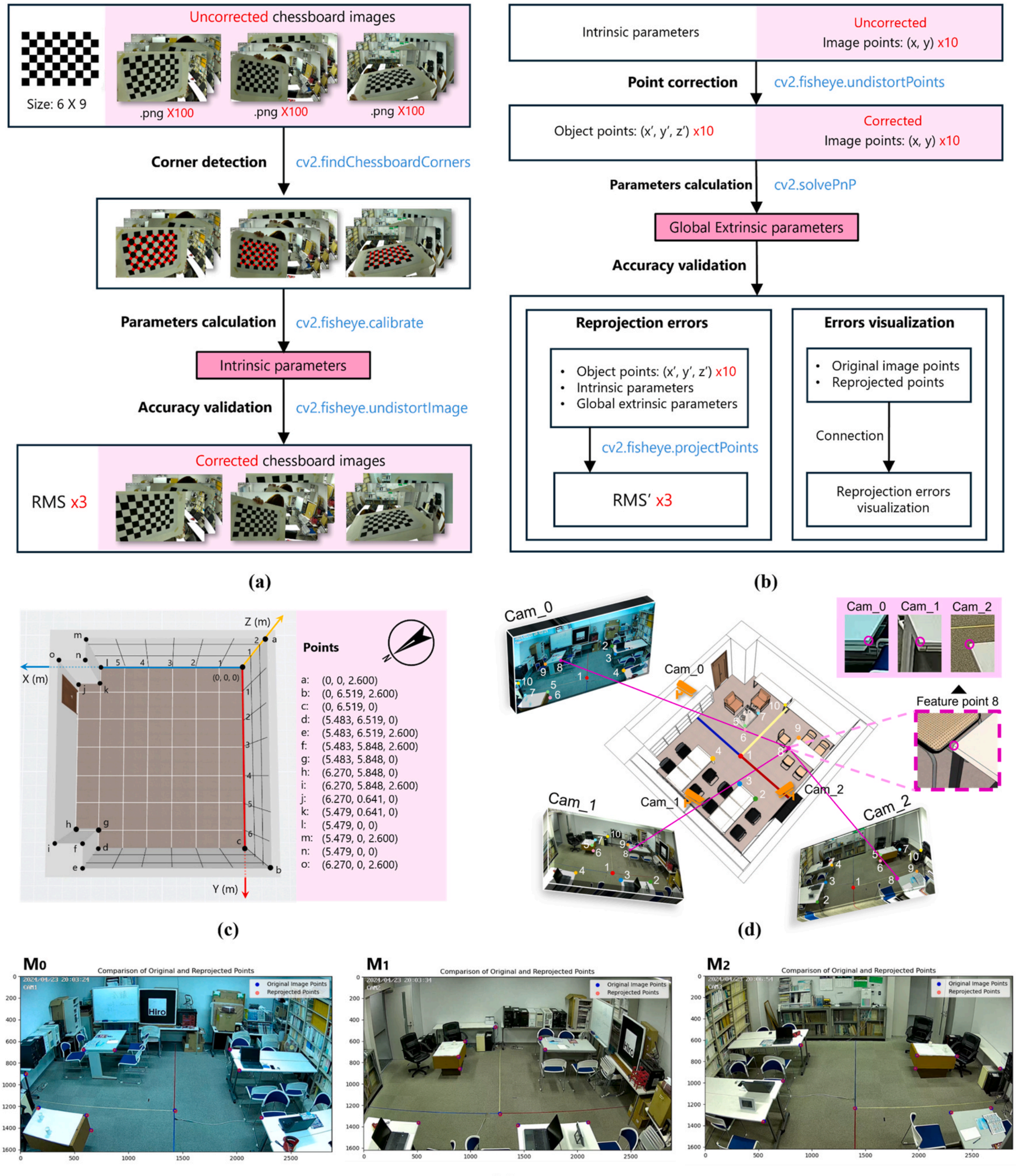
Fig. 3c illustrates the installation positions of the cameras and their field of view ranges. M0, M1, and M2 represent the main cameras mounted on three walls. The images collected by the main cameras are used for both system development and performance evaluation; thus their installation must accommodate both requirements: the planar positions and horizontal rotation angles of the cameras are set according to system evaluation requirements (detailed in Section 3.2), while the installation heights and vertical rotation angles are adjusted based on system development needs. To minimize occlusion within the field of view, all cameras are installed at the top of the walls. Additionally, considering that the average height range of Asian populations is 1.5 m–1.8 m [44], the camera views are required to cover the main height space of human activities—that is, to ensure visibility of space within 2 m above the ground on the opposite wall.

According to the global calibration workflow established in Section 2.1, the multi-camera system is calibrated. Since all cameras used fisheye lenses, distortion correction is necessary during the calibration process. Fig. 4a describes the process of obtaining intrinsic parameters: First, multiple images of the same checkerboard in different positions and orientations are captured and corner detection is performed; Subsequently, the camera's intrinsic parameters are estimated using OpenCV's fisheye calibration module based on the detected corners. Fig. 4b outlines the inference of global extrinsic parameters based on a predefined physical coordinate system (Fig. 4c): Static scene images are first captured from three viewpoints, and spatial feature points are selected in overlapping view regions. (Fig. 4d shows the correspondence between the 10 selected spatial feature points and their multi-view 2D projections. To illustrate that these feature points meet the selection criteria, the spatial position of point 8 and its projection area in each viewpoint image are magnified); Next, the 3D physical coordinates and corrected 2D image coordinates of these feature points are measured, and the cv2.solvePnP function is used to estimate the extrinsic parameters; Finally, RE is employed to evaluate the accuracy of parameters. The evaluation for intrinsic parameters is implemented by visually correcting the original distorted images based on the obtained intrinsic parameters and calculating REs of checkerboard corners in the corrected images. The evaluation for extrinsic parameters is conducted by calculating the REs of all feature points in each camera view (Fig. 4e).

3.1.4. Video collection and preprocessing

This experiment designs three typical work scenarios: customer reception, multi-party meetings, and desktop work. Table 5 provides detailed descriptions of the execution location, number of participants, and primary behavioral patterns associated with each scenario. A total of three participants—two males and one female—are recruited. They simulate scenario tasks under varying personnel configurations through free combinations, thereby generating diverse interaction relationships and occlusion conditions. During the simulation, participants are only required to perform the core behavioral contents of each scenario, without strict standardization of action details. This manner preserves individual variability and behavioral naturalness, enabling the system to be tested under conditions that closely resemble real-world working environments. Moreover, no intervention is introduced to control occlusion conditions. All person-to-person (intra-class) and person-to-item (inter-class) occlusions occur naturally, allowing for the assessment of the system's robustness under occlusion conditions. This experimental design facilitates the collection of representative sample data and effectively meets the initial testing requirements of the system in terms of occlusion, occupant quantity, and behavioral diversity.

In this experiment, the ΔPT threshold is set to 1 s, consistent with the key frame extraction frequency of one frame per second (1 FPS). The experimental space is considered an isothermal environment, maintained at a constant temperature of 24 °C by the air conditioning system. According to formulas (1) and (2), the maximum value of ΔPD is 346 m, far exceeding the dimensions of the space. Therefore, triggering audio signals for synchronization from any position indoors could satisfy the requirements of the first condition regarding sound source setting (Section 2.2). For the second condition, this experiment employed standardized hand



a:	(0, 0, 2.600)
b:	(0, 6.519, 2.600)
c:	(0, 6.519, 0)
d:	(5.483, 6.519, 0)
e:	(5.483, 6.519, 2.600)
f:	(5.483, 5.848, 2.600)
g:	(5.483, 5.848, 0)
h:	(6.270, 5.848, 0)
i:	(6.270, 5.848, 2.600)
j:	(6.270, 0.641, 0)
k:	(5.479, 0.641, 0)
l:	(5.479, 0, 0)
m:	(5.479, 0, 2.600)
n:	(5.479, 0, 0)
o:	(6.270, 0, 2.600)

Fig. 4. Global calibration workflow for multi-camera systems.
(a) Acquisition of intrinsic parameters.
(b) Acquisition of global extrinsic parameters.
(c) The global coordinate system.
(d) Ten feature points in 3D space and their projections in three image views.
(e) Visualization of reprojection errors during the validation of extrinsics.

clapping as the audio signal source, with signal consistency and reproducibility ensured by standardizing the posture, position, and force of clapping [45]. This method is operationally simple and requires no additional equipment. However, acoustic calibration is still necessary to ensure that the signals collected by the multi-camera system meet synchronization requirements: First, the experimental environment is maintained in the same acoustic state as during formal collection; Subsequently, the built-in microphones of each camera collect hand clapping signals from a fixed position, located at the intersection of multiple viewpoints to facilitate visual

Table 5
The type and corresponding behavior pattern of work scenarios.

Scenario type	Execution location	Participants	Behavioral Patterns
Client reception	Reception zone	2	● **Entry** (seating, waiting) ● **Reception service** (water and material provision by staff) ● **Face-to-face communication** (verbal interaction) ● **Information registration and brief task handling** (computer-based data entry by staff) ● **Conclusion** (farewell, handshake, visitor departure)
Multi-party Meeting	Meeting zone	≥2	● **Entry and pre-meeting preparation** (seating, waiting, distribution of materials) ● **Speech and presentation** (speaking by presenter, use of whiteboard) ● **Information recording** (note-taking, typing by participants) ● **Collaborative discussion** (turn-taking verbal exchanges among participants) ● **Conclusion** (individual departure from the meeting zone)
Desktop work	Office zone	≥1 (independent)	● **Entry and task preparation** (seating, moving furniture, using device, retrieving and placing items) ● **Desk work execution** (e.g., typing, reviewing documents, phone operations, handwriting) ● **Postural adjustment** (e.g., stretching, bending, head turning, leaning on desk) ● **Low-frequency peer interaction** (brief verbal, bodily, and gestural communication) ● **Conclusion** (individual departure from the office zone)

synchronization. Calibration is then conducted using Audacity. The results show that the signal presents stable acoustic characteristics across all camera positions: sound pressure level 85 ± 2 dB, rise time 4.2 ± 0.3 m s, duration 95 ± 5 m s, signal-to-noise ratio 42 ± 3 dB, meeting ISO 3382-3 requirements for impulse sound sources. Therefore, standardized hand clapping is suitable for the synchronization needs of multi-camera systems in this experimental environment.

Fig. 5 illustrates the data collection and preprocessing process: First, video recording is initiated, and all participants begin performing the target scenario tasks. During this process, a trained participant continuously performed standardized hand clapping at the aforementioned calibration position to generate synchronization signals (Fig. 5a); After task completion, three channels of multi-view image sequences containing these synchronization signals are exported through Eseecloud, and ffmpeg is used to convert the original video encoding format (H.265) to the more compatible H.264 format before importing into Adobe Premiere Pro; Next, the three sequences are synchronized based on pulse peaks and video images, and 120s shared segments (14.268 FPS) are extracted from each sequence; Finally, key frames are extracted at a frequency of 1 FPS, generating 120 images from each sequence for subsequent processing, as shown in Fig. 5b.

3.1.5. *Indoor real-time multi-point localization*

This section establishes an implementation scheme based on the positioning strategy proposed in Section 2.3. To obtain 3D spatial coordinates of human keypoints, an algorithm instance is constructed based on the proposed 3DPE framework. Multi-point positioning

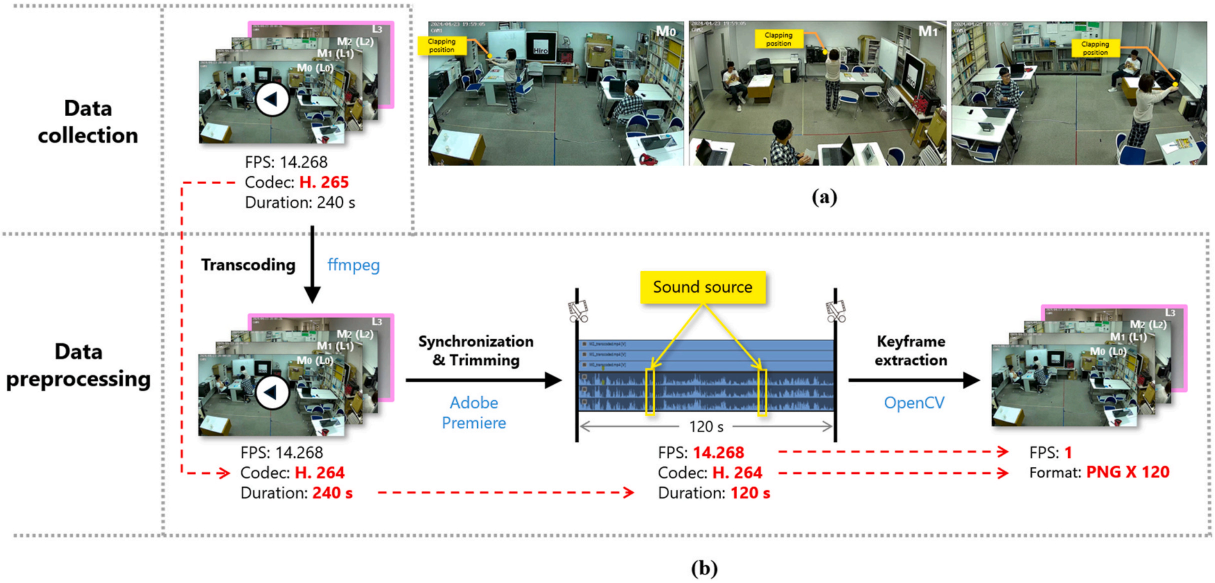


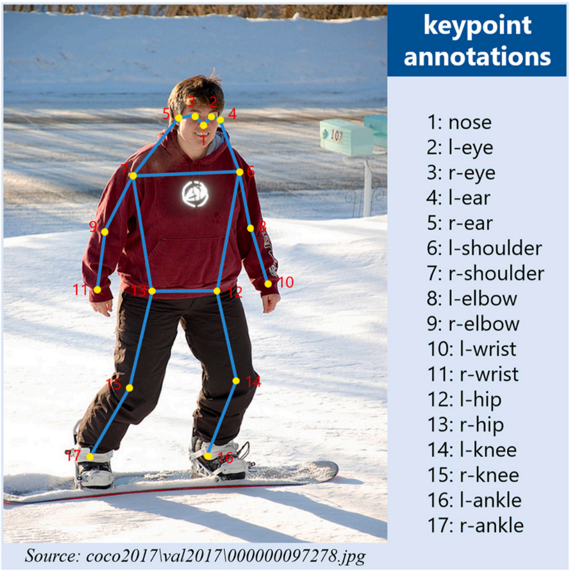
Fig. 5. Data collection and preprocessing.
(a) The position of the sound source (clapping hands).
(b) The dataflow during the data collection and preprocessing.

of individuals on a plane is achieved through further 2D projection extraction. In the 3DPE algorithm, the 2DPE stage employs the Real-Time Object Detector RTMDet [46] as the human detector and Deeppose [47] as the pose estimator to conduct top-down 2D pose inference. Both selected models are pre-trained on the Microsoft Common Objects in Context (MSCOCO) dataset [48]. The MSCOCO dataset provides annotations for human BBoxes and seventeen keypoints in images (Fig. 6), covering various poses, occlusions, and scenarios. The scale and diversity of the data set enable models trained on it to possess good generalization capabilities, accurately detecting human regions and locating keypoints in complex environments; The CVA stage adopts the Omni-Scale Network (OsNet) [49] as the feature extractor. OsNet selects a model pre-trained on the Market-1501 dataset [50]. The Market-1501 dataset contains numerous multi-view pedestrian images in complex scenarios, with each image annotated with pedestrian identity IDs and camera IDs. Therefore, models trained on this dataset can effectively extract cross-view pedestrian features for matching BBox images of the same person across different viewpoints; The 3DPR stage utilizes triangulation functions to fuse multi-view 2D poses of the same person into 3D poses in the global coordinate system.

The workflow of the 3DPE algorithm is shown in Fig. 7, comprising three stages: The first stage is top-down 2DPE. After installing and configuring MMpose, the human detector is initially applied to key frame images, to obtain human BBoxes. Subsequently, the pose estimator is applied to BBox images to estimate 2D human poses; The second stage is CVA. The BBox images are resized to 256 × 128 pix to meet the input requirements of the feature extractor. After installing and configuring Torchreid, the feature extractor is invoked to extract human features, represented in the form of 512-dimensional feature vectors. Considering the scale differences across multi-view images, cosine similarity is used to measure the similarity between feature vectors, enabling identity matching of cross-view BBox images and the association of corresponding 2D poses; The third stage is 3DPR. Camera parameters are utilized to correct the 2D poses and to construct the projection matrices for each camera. Subsequently, 3D poses in the global coordinate system can be reconstructed based on their corrected multi-view 2D pose sets of the same persons in the same frame by OpenCV's triangulation function.

3.1.6. Occupancy state and metric calculation

This subsection constructs an occupancy measurement model based on the occupancy rules defined in Table 3 and the design concept proposed in Section 2.4. This model outputs occupancy states and related metrics for the entire space, the buffer zone and functional zones. Table 6 lists the main variables involved in occupancy detection, including the occupancy state of the entire space (O_e), the buffer zone (O_b) and functional zones (O_f), as well as the current processing frame (i). Fig. 8 illustrates the steps of the occupancy measurement model's detection process: First, the vertex coordinates of functional zones are input, and path objects are generated accordingly; Then, the testing period is specified to import the corresponding keypoint projection data, followed by the initialization of O_e, O_b, and O_f; During frame-by-frame processing, the model sequentially checks the decision conditions defined by Rule_0, Rule_1, and Rule_2 based on the projected keypoint data to update the target variables; After processing all frames, the final values and visualization results of O_e, O_b, and O_f represent the occupancy state time series of the entire space, various functional zones, and the buffer zone during the test period. Regarding occupancy computation, the model designs three categories of occupancy metrics: occupancy and vacancy duration, occupancy frequency, and the ratios of occupied functional zones by count and area. Occupancy and vacancy duration reflect the overall accumulation of occupancy states, while occupancy frequency—defined as the ratio of occupancy duration to total duration—captures trends in occupancy dynamics. The ratios of occupied functional zones, calculated by either count or area relative to the total, provide a micro-zonal perspective on spatial occupancy analysis.



Source: coco2017\val2017\000000097278.jpg

Fig. 6. MS COCO convention: seventeen human keypoints.

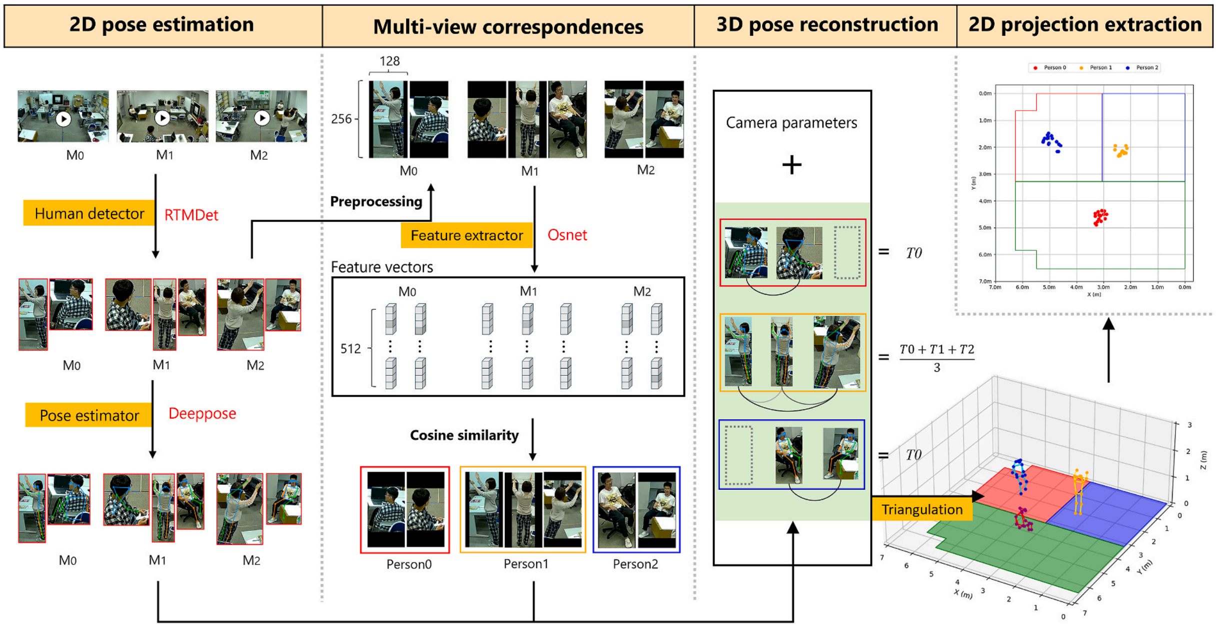


Fig. 7. The overview of 3DPE and 2D projection extraction.

Table 6
Core variables in occupancy detection.

Variable	Definitions
i	Current iteration frame
O_e	The occupancy states of the entire space
O_b	The occupancy states of the buffer zone
O_f	The occupancy states of functional zones

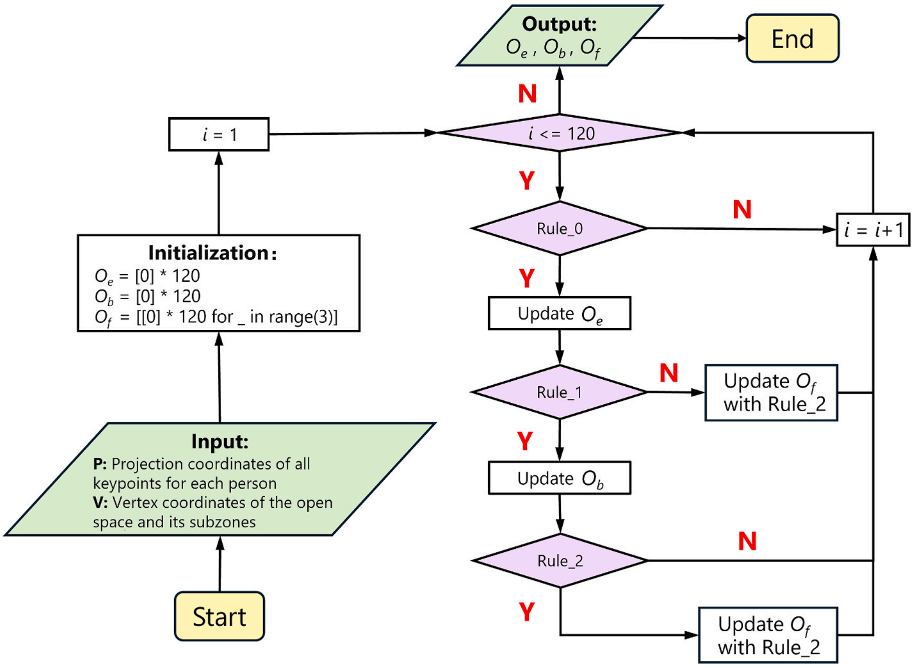


Fig. 8. Decision process for occupancy detection.

3.2. System performance evaluation

As occupancy metrics are derived from occupancy states, the evaluation of the system is based solely on occupancy state data. GT data for occupancy states are obtained through manual observation. As shown in Fig. 3b, ground marking lines are fixed at the space entrance and the buffer zone. Cameras (referred to as “line-monitoring cameras”) are installed in positions parallel to these lines to observe whether personnel completely crossed the marking lines. A total of four line-monitoring cameras (L0–L3) are deployed in this evaluation, with L0, L1, and L2 simultaneously serving as main cameras to reduce hardware costs. Since the indoor perspective parallel to line 3 is obstructed by the door frame, L3 is mounted on the ceiling of the outdoor corridor for observation. To ensure consistency of the image data used for system measurement and evaluation, video clips from the same time period are extracted from videos captured by M0 (L0), M1 (L1), M2 (L2), and L3, and they are subjected to unified data preprocessing. Furthermore, the system’s measurement and evaluation follow the same occupancy determination rules, where the keypoint distribution threshold is set to 100 % (Table 3) in order to facilitate the manual acquisition of GT occupancy states.

Given that occupancy states are binary data, the evaluation process employs a confusion matrix [51] for result analysis. It uses metrics such as Precision, Recall, and F1 Score to assess the system’s occupancy detection performance. Treating the occupancy state as the positive class and the non-occupancy state as the negative class, the key metrics are calculated as follows:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \tag{3}$$

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \tag{4}$$

$$F1\ Score = 2*\frac{Precision*Recall}{Precision + Recall} \tag{5}$$

Where *TP* represents the number of samples correctly classified as positive class, *FP* represents the number of samples incorrectly classified as positive class, and *FN* represents the number of samples incorrectly classified as negative class. Furthermore, this evaluation creates a small office activity dataset with BBox annotations during the test period to evaluate the detector’s performance. To analyze the overlap degree between predicted BBoxes and their GT data, the mean Intersection over Union (mIoU) metric [52] is introduced. This metric refers to the average IoU value for multiple objects in a single category. For category *j*, the calculation formula for mIoU is given as follows:

$$mIoU = \frac{1}{N}\sum_{j=0}^{N}\frac{|A_j \cap B_j|}{|A_j \cup B_j|} \tag{6}$$

where N represents the number of objects in category j, Aj denotes the GT data of the object’s BBox, and Bj refers the predicted data of the object’s BBox.

4. Results and analysis

This section analyzes the data acquired during system development and operation: Section 4.1 presents camera parameters and their accuracy assessment; Section 4.2 demonstrates the system outputs and the evaluation of its occupancy measurement performance, followed by a further analysis of the causes of measurement failures.

Table 7
The intrinsic parameters and global extrinsic parameters of main cameras.

Camera	M0	M1	M2
Dim[a] (pixels)	(2880, 1620)	(2880, 1620)	(2880, 1620)
K (pixels)	[[2137.188, 0.0, 1474.343], [0.0, 2146.274, 786.619], [0.0, 0.0, 1.0]]	[[2124.271, 0.0, 1420.908], [0.0, 2136.550, 799.075], [0.0, 0.0, 1.0]]	[[2093.615, 0.0, 1461.155], [0.0, 2098.679, 803.407], [0.0, 0.0, 1.0]]
D	[[0.00138], [-0.41285], [0.71809], [-0.54802]]	[[-0.0296], [-0.35519], [0.75338], [-0.71428]]	[[-0.02398], [-0.26747], [0.29080], [-0.11620]]
RMS (pixels)	2.7381	2.3239	2.5982
RVec (radians)	[1.52841736, 1.53305531, −0.91159104]	[0.05347739, 2.63492917, −1.64313241]	[1.57404595, −1.58060185, 0.95472509]
TVec (m)	[-3.40788845, −0.57982493, 6.55640845]	[2.60494037, −0.56635505, 6.99195665]	[3.21711789, 2.28999862, 1.15726483]
mRMS (pixels)	5.543236255645752	4.301755905151367	4.417686462402344

[a] Dim: Dimension.

4.1. Camera parameters

Table 7 presents the camera parameters and evaluation results obtained based on the proposed global calibration workflow for indoor multi-camera systems. The intrinsic parameters include the intrinsic matrix (K) and distortion coefficients (D), while the extrinsic parameters consist of rotation vectors (RVec) and translation vectors (TVec). The root mean square (RMS) of REs for the checkerboard corners and the mean reprojection error (mRMS) of feature points in each viewpoint image quantify the calibration accuracy. The results indicate that despite the three cameras being identical models, differences exist in their intrinsic parameters, K and D, primarily stemming from camera manufacturing tolerances and variations in calibration environments. Therefore, to obtain reliable parameters, each camera in the multi-camera system requires independent calibration. The RMS values for all three viewpoints are ~0.1 % of the image width, indicating that the obtained intrinsic parameters accurately characterize the optical properties of fisheye cameras, thereby enabling precise image distortion correction. The mRMS values are ~0.17 % of the image width, reflecting that the positions and orientations of cameras in the global coordinate system have been accurately estimated. These intrinsic parameters will be used to convert 2D keypoints from image coordinate systems to camera coordinate systems, while extrinsic parameters will be used to align the coordinate systems of different cameras to a unified 3D physical space, with the accuracy of both jointly determining the accuracy of reconstructing multi-view 2D poses into 3D poses.

4.2. Occupancy measurement

4.2.1. Output results

The system is tested from 19:58:40 to 20:00:39 on April 23, 2024. Fig. 9 displays the comparison results between the system-output occupancy states and GT values for the entire space, various functional zones and the buffer zone with a keypoint confidence threshold of 0.3. Overall, the system demonstrates good spatiotemporal consistency in most zones. However, false negatives (FNs) occur in certain time segments, indicating the presence of missed detections. These issues primarily arise in the meeting and office zones. Table 8 presents a quantitative evaluation of the system's performance across two time windows and the entire testing period. The results show that the system achieves 100 % Precision and ~88 % Accuracy across all tested periods, demonstrating a high level of stability and reliability in its measurement capability. However, the Recall varies noticeably, dropping from 83.81 % in the first time window to 76.10 % in the second. This suggests a higher proportion of missed detections during the second period, which negatively impacts the overall F1 score. A detailed analysis of the causes of missed detections is provided in Section 4.2.2.

Fig. 10 show the differences between three categories of occupancy metrics output by the system and their GT data. Since the occupancy metrics are derived from occupancy state data, the discrepancies from GT data primarily originate from errors in occupancy state detection. Fig. 10a displays the occupancy and vacancy durations of the entire space, various functional zones and the buffer zone during the two time windows. Despite the occupancy duration of the overall space approaching full capacity levels, significant differences still exist among functional zones. This metric provides crucial data support for optimizing occupancy duration in each functional zone, contributing to balanced allocation of spatial resources; Fig. 10b depicts occupancy proportions of functional zones by count and area at different moments, offering a micro-zonal perspective for the evaluation of spatial occupancy density; Fig. 10c1 and 10c2 show real-time variation trends in occupancy frequency. The frequency curves of various functional zones reveal differentiated usage patterns across different time periods, enabling the identification of special events and providing a deeper understanding of usage demands across functional zones. By continuously monitoring occupancy frequency and setting threshold criteria, vacancy alerts

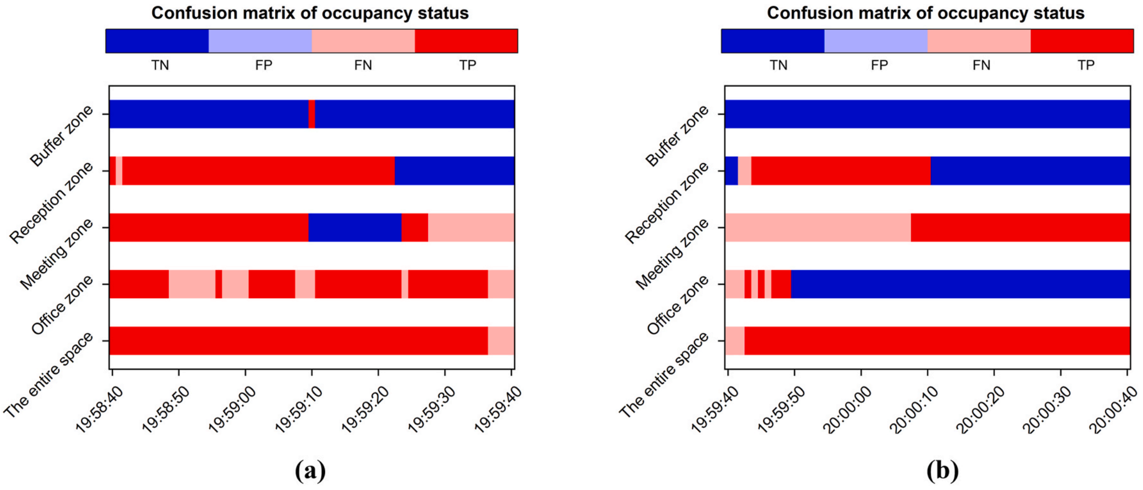


Fig. 9. Occupancy states for the entire space and its functional zones.
(a) Occupancy states (19:58:40–19:59:39).
(b) Occupancy states (19:59:40–20:00:39).

Table 8
Performance metrics of occupancy state detection.

Time Range	TP	TN[a]	FP	FN	Accuracy	Precision	Recall	F1	Specificity
19:58:40–19:59:39	176	90	0	34	88.67 %	100 %	83.81 %	91.19 %	100 %
19:59:40–20:00:39	121	141	0	38	87.33 %	100 %	76.10 %	86.43 %	100 %
19:58:40–20:00:39	297	231	0	72	88.00 %	100 %	80.49 %	89.19 %	100 %

[a] TN: True Negative.

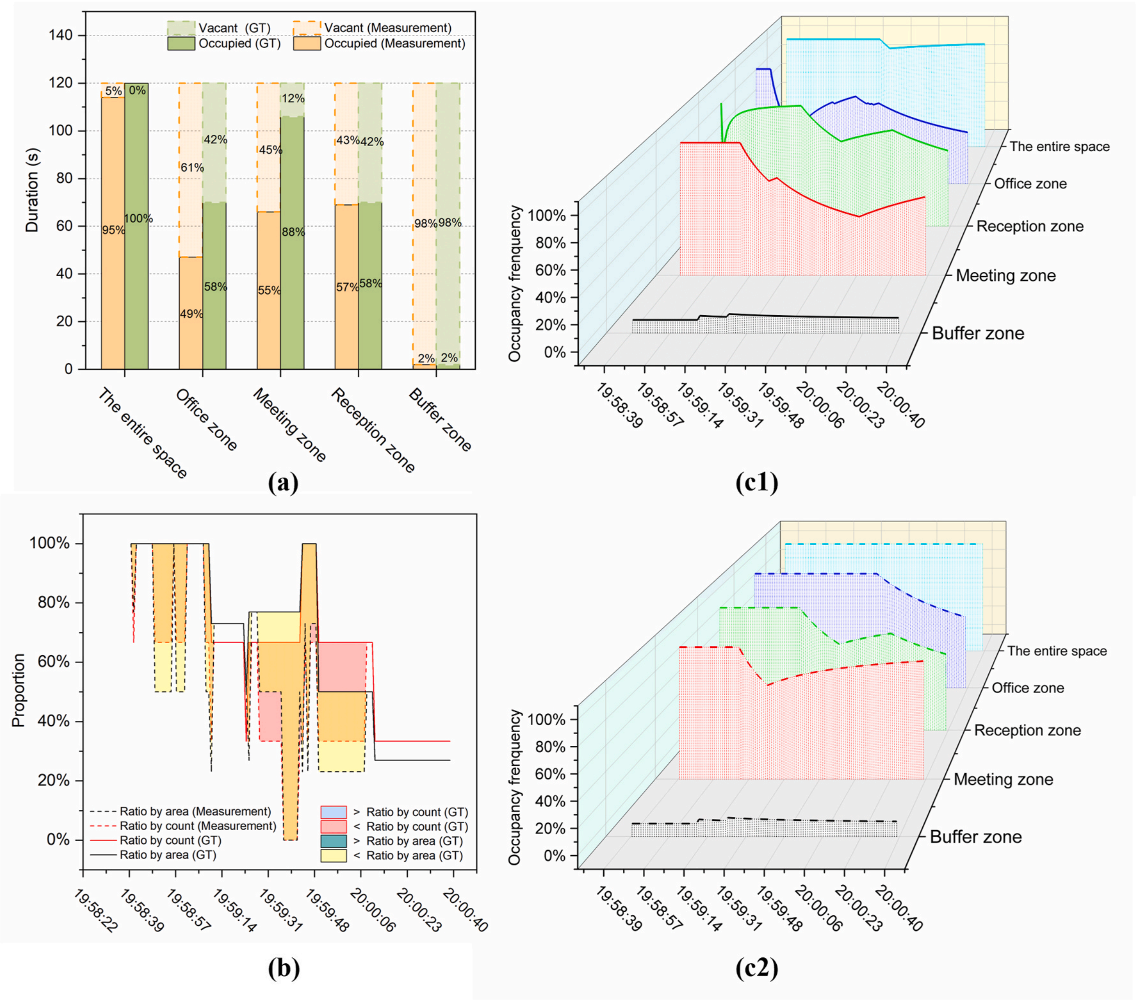


Fig. 10. Occupancy metrics for the entire space and its functional zones (19:58:40–20:00:39).
(a) Occupied and vacant duration.
(b) Variations in the ratio of occupied zones by count and area.
(c1) Variations of occupancy frequency (Measurement).
(c2) Variations of occupancy frequency (GT).

can be triggered at multiple scales of the space. Moreover, modeling these frequency curves as time series also allows future occupancy trends to be predicted.

4.2.2. Performance analysis

The occurrence of missed detections in the system's occupancy state measurement is primarily attributed to the failure of 3D human pose reconstruction, which leads to the absence of 2D keypoint projection data required for occupancy computation. To further investigate the specific mechanism behind these missed detections, Fig. 11 compares the outputs of the human detection, 2D pose estimation, and triangulation modules with their GT data. The corresponding statistical results are presented in Table 9. Figs. 11a and b illustrate the number of detected BBoxes and estimated 2D poses across three camera views, with black rectangles highlighting

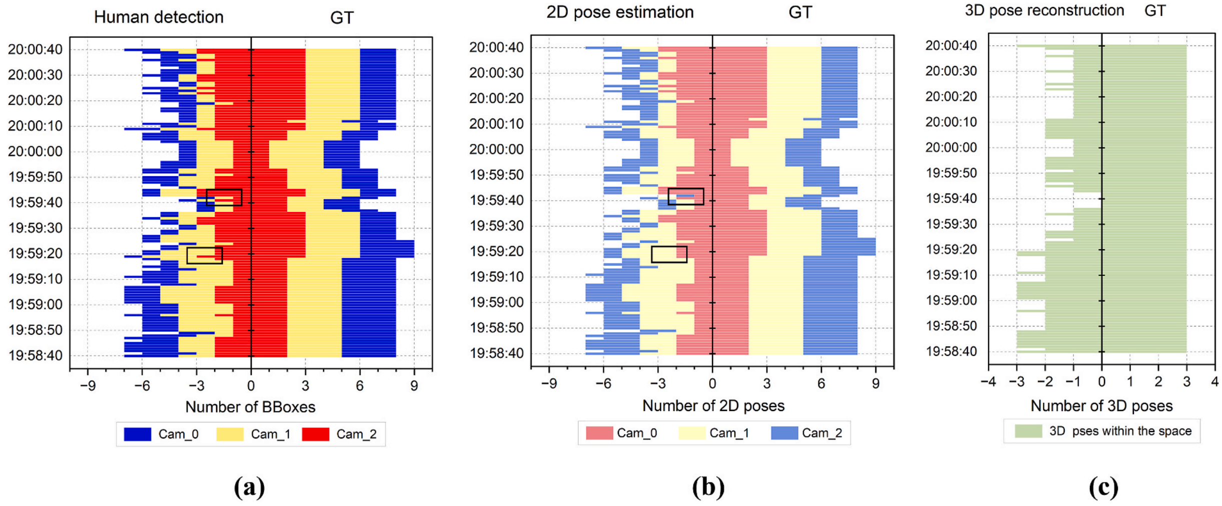


Fig. 11. Outputs of the detector, the estimator and triangulation (19:58:40–20:00:39).
(a) The total number of BBoxes from three view images.
(b) The total number of 2D poses from three view images.
(c) The total number of 3D poses within the space.

Table 9
Comparison between the outputs of the detector, the estimator and triangulation and their corresponding GT values (19:58:40–20:00:39).

Time ranges	BBoxes	GT	Ratio	mIoU	2D poses	GT	Ratio	3D poses	GT	Ratio
19:58:40–19:59:39	347	482	72.00 %	85.75 %	320	482	66.39 %	124	180	68.89 %
19:59:40–20:00:39	296	440	67.27 %	91.36 %	285	440	64.77 %	80	180	44.44 %
19:58:40–20:00:39	643	922	69.74 %	88.54 %	605	922	65.61 %	204	360	56.67 %

examples where the number of valid detections changes within the same keyframe. Fig. 11c shows the number of 3D poses reconstructed from the multi-view 2D poses. According to Table 9, during the entire testing period, the human detection module captures only 69.74 % of the GT BBoxes, although the mean Intersection over Union (mIoU) of these BBoxes reaches 88.54 %. Based on the detected BBoxes, the subsequent 2D pose estimation yields only 65.61 % of GT 2D poses (The GT numbers for BBoxes and 2D poses are the same). Ultimately, the 3D poses reconstructed via triangulation from these multi-view 2D poses account for only 56.65 % of the corresponding GT values. These results indicate a progressive loss of data throughout the multi-stage processing pipeline. This cumulative loss further leads to missed detections in occupancy computation based on 2D keypoint projections derived from 3D poses. As a result, such a cascading loss mechanism across the “detection–estimation–reconstruction” pipeline is the primary cause of FNs in the system’s occupancy state measurement.

By conducting a comparative analysis between the input data in which missed detections occur at each module and those that are successfully processed, three fundamental causes are further identified:

1. **Insufficient robustness of the detector to occlusion:** In meeting and office zones, the combination of high occupant density, dense furniture and frequent interactions creates complex occlusion scenarios, which can easily lead to the failure of human BBox detection.
2. **Limited adaptability of the pose estimator to incomplete human bodies:** Even when the BBoxes are correctly detected, pose estimation may fail due to low keypoint confidence, which often occurs when the human body is incomplete because of occlusion or partial departure from the camera’s field of view.
3. **Inadequate viewpoint information leading to triangulation failures:** When occupants are captured by only a single camera viewpoint, despite obtaining their 2D poses, the system cannot reconstruct their 3D poses due to not meeting the minimum number of viewpoints required for triangulation ($\geq$2).

In summary, the system achieves 100 % precision and an F1 score of 89.19 %, demonstrating its effectiveness in conducting both macro- and micro-scale occupancy investigations within the open office space. However, the system still exhibits certain performance limitations, as its measurement accuracy is limited by the upper bounds of the employed algorithms’ capacity and the inherent constraints of the system design methodology. The algorithmic limitations are reflected in the 2DPE stage, which applies RTMDet for human detection and DeepPose for pose estimation. Consequently, the system’s ability to handle variations in occupant quantity, behavioral patterns, and occlusion conditions largely depends on the generalization and transferability of these two models in real-world environments. The limitation of the system design methodology is in the 3DPR stage. This stage relies solely on triangulation, which cannot process occupants that appear in only a single camera view—making it impossible to reconstruct their 3D poses. As a result, the system fails to detect the occupancy states of such occupants, leading to missed detections.

5. Discussion

This section compares the proposed method with existing SOTA methods (Table 1) based on the following questions and discusses its limitations:

- What levels of spatial resolution does the method support for building occupancy measurement? How does the employed indoor positioning strategy contribute to occupancy measurement?
- What types of sensors are used in this method? Does it require any additional equipment?
- Does the method impose specific requirements on occupants? Is it capable of recognizing all natural human behaviors?
- What are the advantages of this method compared to current vision-based approaches? Is it constrained by the type of camera used?
- What is the applicable scope of this method?
- What are the limitations and corresponding improvement strategies of the method?

Previous occupancy measurement methods primarily focus on macro levels, such as occupancy of rooms [8], floors [9], or entire buildings [10]. Although some studies attempt to enhance spatial resolution of occupancy measurement by detecting furniture or micro-zone occupancy states[13–15], they still struggle to investigate the occupancy of functional zones within open spaces. In contrast, the proposed solution combines both macro and micro scales, supporting occupancy measurements for the room and micro-zones, capable of obtaining occupancy states and related metrics for customized functional zones. Furthermore, considering the gap between indirect positioning strategies in Table 2 and actual human occupancy data, this solution introduces a multi-point positioning strategy capable of directly tracking joints. This strategy overcomes the limitations of traditional single-point positioning strategies for individual position classification in complex scenarios by describing the planar projection range of the human body with multiple keypoints.

Sensor-based occupancy measurement methods typically require additional hardware for data recording and processing, or signal transmission and reception. For example, methods based on thermal cameras [23] or depth cameras [18] usually demand computing devices to meet high computational demands; Methods based on environmental sensors rely on specific equipment for long-term data collection [14,]; Similarly, WLC-based methods require the deployment of dedicated equipment and tags for signal transmission and reception [20,19]. However, the proposed solution is cost-effective, as it relies solely on RGB cameras and processes relatively simple data structures, resulting in low computing resource demands. Therefore, it can be seamlessly integrated into the existing CCTV infrastructure, reducing both hardware and technical costs.

Methods summarized in Table 1 typically impose specific requirements on occupants, such as motion states, behavior patterns, or carried items. For example, PIR sensor-based methods [15,13] require occupants to maintain motion states; WLC-based methods require occupants to carry tags or mobile devices [9,10,19,20]; And desk- or chair-based occupancy detection methods rely on direct contact between occupants and furniture [17,13,14]. In contrast, the proposed solution imposes no restrictions on occupants. Moreover, by invoking models trained on standard or specialized datasets, the solution can recognize diverse natural behaviors.

The image sensor-based methods listed in Table 1 present several limitations. For instance, depth cameras [18] have a restricted measurement range, making them unsuitable for long-distance monitoring; Stereo cameras require substantial computational resources for disparity calculations [24], imposing high hardware demands; And thermal imaging cameras [23] are highly sensitive to ambient temperature fluctuations, which can easily introduce measurement errors. Moreover, these vision sensors are typically expensive, which limits their feasibility for large-scale deployment. In contrast, the proposed solution relies solely on RGB cameras, avoiding the range limitations of depth cameras while remaining cost-effective. This makes it feasible to use multi-view configurations to mitigate occlusion issues, thereby offering potential for large-scale deployment. Additionally, the image sensor-based methods listed in Table 1 [16,22] typically assume known camera parameters, which limits their applicability in scenarios where parameters are unknown or subject to deviation. To address this issue, the proposed solution establishes global calibration workflow for indoor multi-camera systems, enabling precise estimation of both intrinsic and global extrinsic parameters for each camera. As the adopted calibration techniques are based on geometric principles and do not rely on specific camera structures or lens types, the workflow possesses strong generalizability and cross-device compatibility. Furthermore, the structured characteristics of the checkerboard pattern used for intrinsic calibration and the flexibility of the PnP method for extrinsic calibration enable this workflow to adapt effectively to complex environments.

The theoretical applicability of the proposed system design methodology can be articulated from the following three perspectives:

1. **Generality in type and flexibility in quantity for hardware:** It only requires that the combined field of view of the cameras fully cover and clearly capture the primary 3D spatial scope where human activities occur. There are no restrictions on camera type, and the number of cameras can be flexibly adjusted based on the spatial scale and ceiling height of the application environment.
2. **Non-interference with behavior and adaptability to quantity for occupants:** It supports occupancy measurement under natural behavioral patterns; however, its ability to recognize poses and count occupants is influenced by the performance of the applied algorithms and the requirement of viewpoint number for triangulation.

3. **Compatibility with cameras and algorithms for spatial scale:** The building spaces it can accommodate depend on the distribution and the resolution of cameras, as well as the performance of the applied algorithms. The captured images must be of sufficient clarity to ensure the accurate operation of the 2DPE stage.

In summary, the proposed system design method offers several advantages over existing occupancy measurement approaches and demonstrates strong theoretical applicability in terms of hardware utilization, the number and behavior of occupants, and spatial scales. However, the measurement performance of the system instance developed based on this method remains constrained by two key factors: first, the performance ceiling of the algorithms adopted in the 2DPE stage during system development; and second, a limitation in the system design methodology, whereby the 3DPR stage cannot handle targets that appear in only a single camera view. To address the aforementioned issues, the following strategies are considered in future optimizations of the method and system development: introducing object detection models with better occlusion awareness to enhance the robustness of human detection; integrating graph-based pose completion methods to recover missing keypoints caused by occlusion or limited fields of view; and incorporating monocular 3DPE techniques to improve the system's robustness and completeness under constrained viewpoint conditions.

6. Conclusion

This study proposes a low-cost, multi-scale occupancy measurement method for open office spaces, which innovatively introduces 3DPE technology into building occupancy investigation. The proposed method takes multi-view RGB image sequences as input and outputs the occupancy states and related metrics of both the entire space and its functional zones, enabling occupancy analysis at both the macro and micro scales. Moreover, compared with existing methods, the proposed approach exhibits three advantages: (1) Relying solely on RGB cameras without the need for additional devices; (2) Enabling the measurement of occupancy data under natural behavioral patterns; (3) Being capable of capturing actual human occupancy data. In addition, the localization strategy employed by this method overcomes the limitations of conventional single-point localization approaches in classifying human positions under complex scenarios by representing the planar projection of the body using multiple keypoints. The proposed method bridges the gap of association between occupancy data in open spaces and functional layouts. The resulting high-spatial-resolution occupancy data has strong interpretability and practical value, providing reliable data support for layout optimization, resource management, and energy control in open-plan office environments.

The main contributions of this study are as follows: (1) proposing a design framework for occupancy measurement systems; (2) constructing a 3DPE framework; (3) establishing a global calibration workflow for indoor multi-camera systems; (4) developing an occupancy measurement prototype system. The system achieves a Precision of 100 % and an F1 score of 89.19 % in occupancy state measurement, demonstrating that the framework effectively supports occupancy surveys for functional zones and holds potential for real-world applications. This research outcome not only provides a theoretical foundation for the development of CCTV-based occupancy measurement methods, but also offers data support for decision-making in the sustainable design and operation of indoor open office environments. To mitigate the limitations in system measurement performance caused by the applied algorithmic capacity ceiling in the 2DPE stage and the methodological design constraints in the 3DPR stage, future work focuses on enhancing the robustness of human detection and pose estimation models under occlusion conditions and exploring single-view-based 3DPE techniques.

CRediT authorship contribution statement

Sihua Chen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tomohiro Fukuda:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Conceptualization. **Nobuyoshi Yabuki:** Writing – review & editing, Supervision, Resources, Methodology.

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.

Acknowledgments

We would like to express our sincere gratitude to all the participants who contributed to this study by taking part in the experiments.

Data availability

Data will be made available on request.

References

[1] H.K. Abdullah, H.Z. Alibaba, Open-plan office design for improved natural ventilation and reduced mixed mode supplementary loads, Indoor Built Environ. 31 (2022) 2145–2167, https://doi.org/10.1177/1420326X20953458.

[2] J.Y. Jeon, H.I. Jo, B.B. Santika, H. Lee, Crossed effects of audio-visual environment on indoor soundscape perception for pleasant open-plan office environments, Build. Environ. 207 (2022) 108512, https://doi.org/10.1016/j.buildenv.2021.108512.

[3] O. James, P. Delfabbro, D.L. King, A comparison of psychological and work outcomes in open-plan and cellular office designs: a systematic review, Sage Open 11 (2021) 2158244020988869, https://doi.org/10.1177/2158244020988869.

[4] F.D. Molli, D.D. Paoli, Middle managers in open-plan offices: feeling free and frustrated, Int. J. Work Organ. Emot. 11 (2020) 231–246. https://www.inderscienceonline.com/doi/10.1504/IJWOE.2020.111317.

[5] T. Labeodan, K. Aduda, W. Zeiler, F. Hoving, Experimental evaluation of the performance of chair sensors in an office space for occupancy detection and occupancy-driven control, Energy Build. 111 (2016) 195–206, https://doi.org/10.1016/j.enbuild.2015.11.054.

[6] Jones Lang LaSalle (JLL), Global occupancy planning benchmarking report. https://www.us.jll.com/en/trendsand-insights/research/occupancy-benchmarking-report, 2024, 2 December.

[7] C. de Bakker, M. Aries, H. Kort, A. Rosemann, Occupancy-based lighting control in open-plan office spaces: a state-of-the-art review, Build. Environ. 112 (2017) 308–321, https://doi.org/10.1016/j.buildenv.2016.11.042.

[8] W. Zhang, J. Calautit, P.W. Tien, Y. Wu, S. Wei, Deep learning models for vision-based occupancy detection in high occupancy buildings, J. Build. Eng. 98 (2024) 111355, https://doi.org/10.1016/j.jobe.2024.111355.

[9] Y. Zhou, J.K.W. Yeoh, W. Solihin, Studying the impact of building morphology on occupants' movement using a rule mining approach, Build. Environ. 249 (2024) 111116. https://doi.org/10.1016/j.buildenv.2023.111116.

[10] J. Gu, P. Xu, Y. Ji, A fast method for calculating the impact of occupancy on commercial building energy consumption, Buildings 13 (2023) 567, https://doi.org/10.3390/buildings13020567.

[11] T. Yang, A. Bandyopadhyay, Z. O'Neill, J. Wen, B. Dong, From occupants to occupants: a review of the occupant information understanding for building HVAC occupant-centric control, Build. Simulat. 15 (2022) 913–932, https://doi.org/10.1007/s12273-021-0861-0.

[12] X. Liang, J. Shim, O. Anderton, D. Song, Low-cost data-driven estimation of indoor occupancy based on carbon dioxide (CO2) concentration: a multi-scenario case study, J. Build. Eng. 82 (2024) 108180, https://doi.org/10.1016/j.jobe.2023.108180.

[13] X. Zhang, J. Fan, T. Peng, P. Zheng, C.K.M. Lee, R. Tang, A privacy-preserving and unobtrusive sitting posture recognition system via pressure array sensor and infrared array sensor for office workers, Adv. Eng. Inform. 53 (2022) 101690, https://doi.org/10.1016/j.aei.2022.101690.

[14] D.N. Wagner, A. Mathur, B.E. Boor, Spatial seated occupancy detection in offices with a chair-based temperature sensor array, Build. Environ. 187 (2021) 107360, https://doi.org/10.1016/j.buildenv.2020.107360.s.

[15] Enhanced multiplex binary PIR localization using the transferable belief model, in: A. Hadj Henni, R. Ben Bachouch, O. Bennis, N. Ramdani (Eds.), IEEE Sens. J. 19 (2019) 8146–8159, https://doi.org/10.1109/JSEN.2019.2918844.

[16] X. Zhao, S. Li, Z. Zhao, H. Li, A Cost-Effective System for Indoor Three-Dimensional Occupant Positioning and Trajectory Reconstruction, Buildings 13 (2023) 2832. https://doi.org/10.3390/buildings13112832.

[17] J. Andrews, M. Kowsika, A. Vakil, J. Li. A motion induced passive infrared (PIR) sensor for stationary human occupancy detection, 2020 IEEEION Position Locat. Navig. Symp, IEEE, Portland, OR, USA, 2020, pp. 1295–1304, https://doi.org/10.1109/PLANS46316.2020.9109909.

[18] R.C. Navarro, Indoor occupancy estimation for smart utilities: a novel approach based on depth sensors, Build, Environ 222 (2022) 109406, https://doi.org/10.1016/j.buildenv.2022.109406.

[19] S. Djosic, I. Stojanovic, M. Jovanovic, T. Nikolic, G.Lj Djordjevic, Fingerprinting-assisted UWB-based localization technique for complex indoor environments, Expert Syst. Appl. 167 (2021) 114188, https://doi.org/10.1016/j.eswa.2020.114188.

[20] Q. Ma, X. Li, G. Li, B. Ning, M. Bai, X. Wang, MRLIHT: mobile RFID-based localization for indoor human tracking, Sensors 20 (2020) 1711, https://doi.org/10.3390/s20061711.

[21] M. Yuan, Y. Wang, Z. Zhu, R. Zhang, H. Fan, Y. Sun, A user-centric temperature sensor deployment method under digital twin leveraging occupancy information, J. Build. Eng. 99 (2025) 111540, https://doi.org/10.1016/j.jobe.2024.111540.

[22] A. Ravi, K. Galmath, S. Hu, A. Misra, CS-light: camera sensing based occupancy-aware robust smart building lighting control, in: Proc. 8th ACM Int. Conf. Syst. Energy-Effic. Build. Cities Transp., ACM, Coimbra Portugal, 2021, pp. 61–70, https://doi.org/10.1145/3486611.3486657.

[23] M. Kraft, P. Aszkowski, D. Pieczyński, M. Fularz, Low-Cost Thermal Camera-Based Counting Occupancy Meter Facilitating Energy Saving in Smart Buildings, Energies 14 (2021) 4542. https://doi.org/10.3390/en14154542.

[24] H. Wang, C. Liang, G. Wang, X. Li, Energy-saving potential of fresh air management using camera-based indoor occupancy positioning system in public open space, Appl. Energy 356 (2024) 122358, https://doi.org/10.1016/j.apenergy.2023.122358.

[25] A.R. Mena, H.G. Ceballos, J. Alvarado-Uribe, Measuring Indoor Occupancy through Environmental Sensors: A Systematic Review on Sensor Deployment, Sensors 22 (2022) 3770. https://doi.org/10.3390/s22103770.

[26] Y. Ji, K. Ok, W.S. Choi. Occupancy detection technology in the building based on IoT environment sensors, Proc. 8th Int. Conf. Internet Things, ACM, Santa Barbara California USA, 2018, pp. 1–4, https://doi.org/10.1145/3277593.3277633.

[27] D.S. Barros, A.M.R. Da Cruz, S.I. Lopes, Hybrid building occupancy estimation using thermal imaging and environmental sensing, 2023, in: IEEE Int. Conf. Ind. Technol., ICIT, IEEE, Orlando, FL, USA, 2023, pp. 1–6, https://doi.org/10.1109/ICIT58465.2023.10143114.

[28] J. Dong, Q. Fang, W. Jiang, Y. Yang, Q. Huang, H. Bao, X. Zhou, Fast and robust multi-person 3D pose estimation and tracking from multiple views, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 6981–6992, https://doi.org/10.1109/TPAMI.2021.3098052.

[29] S. Ershadi-Nasab, E. Noury, S. Kasaei, E. Sanaei, Multiple human 3D pose estimation from multiview images, Multimed, Tool. Appl. 77 (2018) 15573–15601, https://doi.org/10.1007/s11042-017-5133-8.

[30] H. Qin, Y. Dai, Y. Jiang, D. Li, H. Liu, Y. Zhang, J. Li, T. Yang, Inside-out multiperson 3-D pose estimation using the panoramic camera capture system, IEEE Trans. Instrum. Meas. 73 (2024) 1–17, https://doi.org/10.1109/TIM.2023.3346490.

[31] Z. Zhang, A flexible new technique for camera calibration, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 1330–1334, https://doi.org/10.1109/34.888718.

[32] Q. Sun, T. Zhang, G. Zhang, K. Wang, D. Zhu, J. Li, X. Zhang, Efficient solution to PnP problem based on vision geometry, IEEE Rob. Autom. Lett. 9 (2024) 3100–3107, https://doi.org/10.1109/LRA.2023.3333740.

[33] X. Liang, Y. Du, D. Wei, An integrated camera parameters calibration approach for robotic monocular vision guidance, 34rd Youth Academic, Annual Conf. of Chinese Association of Autom. (YAC) (2019) 455–459. https://ieeexplore.ieee.org/abstract/document/8787642.

[34] C. Wang, J. Wang, C. Li, et al., Safe and Robust Mobile Robot Navigation in Uneven Indoor Environments, Sensors 19 (2019) 2993, https://doi.org/10.3390/s19132993.

[35] V. Lepetit, F. Moreno-Noguer, P. Fua, EPnP: an accurate O(n) solution to the PnP problem, Int. J. Comput. Vis. 81 (2009) 155–166, https://doi.org/10.1007/s11263-008-0152-6.

[36] H. Ju, L. Yunhui, Y. Ming, Multi-camera calibration method based on minimizing the difference of reprojection error vectors, J. Syst. Eng. Electron. 29 (2018) 844–853, https://doi.org/10.21629/JSEE.2018.04.19.

[37] S. Dagher, S. Ishiyama, Protocol for precise signal synchronization of electrophysiology, videography, and audio recordings using a custom-made pulse generator, STAR Protoc 4 (2023) 102306, https://doi.org/10.1016/j.xpro.2023.102306.

[38] J. Six, M. Leman, Synchronizing multimodal recordings using audio-to-audio alignment, J Multimodal User (2015) 223–229. https://link.springer.com/article/10.1007/s12193-015-0196-1.

[39] X. Zhou, Y. Dai, H. Qin, S. Qiu, X. Liu, Y. Dai, J. Li, T. Yang, Subframe-level synchronization in multi-camera system using time-calibrated video, Sensors 24 (2024) 6975, https://doi.org/10.3390/s24216975.

[40] MMPose Contributors, OpenMMLab pose estimation toolbox and benchmark, GitHub (3.2) (2024) v1 https://github.com/open-mmlab/mmpose.

[41] K. Zhou, T. Xiang, Torchreid: a library for deep learning person Re-identification in pytorch, GitHub v1 (0.6). https://github.com/KaiyangZhou/deep-person-reid.

[42] C. Stolojescu-Crisan, C. Crisan, B.-P. Butunoi, Access control and surveillance in a smart home, High-Confid, Comput. Times 2 (2022) 100036, https://doi.org/10.1016/j.hcc.2021.100036.

[43] P. Singh, Linux development on WSL, in: P. Singh (Ed.), Learn Window Subsystem Linux Pract. Guide Dev, IT Prof., Apress, Berkeley, CA, 2020, pp. 131–168, https://doi.org/10.1007/978-1-4842-6038-8_8.

[44] Fixtheheight. Average Height in Asia - FixTheHeight, 2024. https://www.fixtheheight.com/average-height-in-asia/ (accessed 2 December).

[45] J. Six, M. Leman, Synchronizing multimodal recordings using audio-to-audio alignment, J. Multimodal User Interfaces 9 (2015) 223–229, https://doi.org/10.1007/s12193-015-0196-1.

[46] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, K. Chen, RTMDet: an empirical study of designing real-time object detectors, arXiv (2022), https://doi.org/10.48550/arXiv.2212.07784 preprint arXiv:2212.07784.

[47] A. Toshev, C. Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks, IEEE, Columbus, OH, USA, 2014, pp. 1653–1660. https://doi.org/10.1109/CVPR.2014.214.

[48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, COCO - Common objects in Context , ECCV, Springer, vol. 1, 2020. https://cocodataset.org/#home.

[49] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person Re-identification, 2019, IEEECVF Int. Conf. Comput. Vis. ICCV (2019) 3701–3711, https://doi.org/10.1109/ICCV.2019.00380.

[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, market_1501. https://www.kaggle.com/datasets/pengcw1/market-1501.

[51] Z. Zhou, G. Li, H. Chen, H. Zhong, Fault diagnosis method for building VRF system based on convolutional neural network: Considering system defrosting process and sensor fault coupling, Build. Environ. 195 (2021). https://doi.org/10.1016/j.buildenv.2021.107775.

[52] M.-Y. Cheng, M.N. Sholeh, A. Kwek, Computer vision-based post-earthquake inspections for building safety assessment, J. Build. Eng. 94 (2024) 109909. https://doi.org/10.1016/j.jobe.2024.109909.