Engineering Applications of Artificial Intelligence 174 (2026) 114440

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Deep learning-driven digital twin system for pedestrian tracking and evacuation load assessment in public spaces

Huakai Sun [a,b], Yifei Ding [c], Ruiwen Fan [a,b], Yuxin Zhang [c], Tianhang Zhang [a,b,c,*], Xinyan Huang [c], Ke Wu [a,b,d,**]

[a] Key Laboratory of Disaster Control and Emergency Response in Civil Engineering, Ministry of Emergency Management, Zhejiang University, China
[b] Research Center for Urban Fire Safety Engineering, Zhejiang University, Hangzhou, China
[c] Research Centre for Smart Urban Resilience and Firefighting, The Hong Kong Polytechnic University, Hong Kong SAR, China
[d] Zhejiang Key Laboratory of Offshore Civil Engineering and Materials, Zhejiang University, China

ARTICLE INFO

Keywords:
Evacuation
Digital twin
Deep learning
Multi-object tracking
Pedestrian localization

ABSTRACT

Real-time pedestrian localization is essential for effective emergency evacuation in large indoor public spaces. This study presents an intelligent digital twin system for evacuation monitoring, integrating deep learning and computer vision. The system includes four components: (1) Internet of Things sensor network, (2) cloud computing server, (3) Artificial Intelligence processing engine, and (4) interactive user interface. The Artificial Intelligence engine introduces three innovations: automated detection and tracking of pedestrian coordinates using You Only Look Once-Pose (YOLO-Pose) and Deep Simple Online and Realtime Tracking (DeepSORT); transformation of multi-camera data into a unified world coordinate system; and the Multi-Object Matching Operation (MOMO) algorithm for identity association. These enable accurate detection, localization, and counting while minimizing identifiability. The system was validated in controlled experiments and a high-speed rail station waiting hall with dense, dynamic pedestrian flow. It achieves high localization precision, with a root mean square error of 5.3 cm, a mean absolute error of 4.8 cm, and a people counting accuracy of 92.34% while processing 30 frames per second video at 27.8 ms per frame. These results demonstrate the potential of the digital twin framework in intelligent evacuation management. The main contribution in Artificial Intelligence is the Multi-Object Matching Operation algorithm, and the engineering contribution is the realization of a real-time digital twin system in a large public facility.

1. Introduction

Driven by rapid urbanization, cities worldwide are witnessing substantial population growth and increasingly dense human activity. This trend has resulted in frequent crowd gatherings across diverse public spaces such as shopping malls, office buildings, transportation hubs, and sports venues (Chen et al., 2021; Sun et al., 2025b; Sun and Chen, 2023; Wu et al., 2025). In such densely populated indoor environments, the risk of large-scale casualties and property loss significantly increases in the event of sudden disasters, including fires, earthquakes, floods, toxic gas leaks, or even terrorist attacks (Sun et al., 2025a; T. Zhang et al., 2024; Zhang et al., 2022a; Wu et al., 2026). For instance, from January to October 2024, China experienced 745,000 fires, resulting in 1381 deaths and 2063 injuries. The direct property damage amounted to 6.15 billion Renminbi (RMB), approximately 861 million United States Dollar (USD). Notably, many of these incidents occurred in urban commercial and residential complexes with high population densities (Xie et al., 2025).

The increasing frequency and severity of public safety incidents underscore the urgent need for intelligent technologies to support the development of effective evacuation strategies (Y. Zhang et al., 2025a, 2025b). To address this need, growing attention has been directed toward the application of Artificial Intelligence (AI) in the field of safety and emergency management. As an advanced technological paradigm, AI offers new possibilities for enhancing emergency preparedness,

* Corresponding author. author. Key Laboratory of Disaster Control and Emergency Response in Civil Engineering, Ministry of Emergency Management, Zhejiang University, China.
** Corresponding author. author. Key Laboratory of Disaster Control and Emergency Response in Civil Engineering, Ministry of Emergency Management, Zhejiang University, China.
E-mail addresses: thz@zju.edu.cn (T. Zhang), wuke@zju.edu.cn (K. Wu).

https://doi.org/10.1016/j.engappai.2026.114440
Received 10 November 2025; Received in revised form 2 March 2026; Accepted 6 March 2026

Abbreviation

2D	Two Dimensional
3D	Three Dimensional
AI	Artificial Intelligence
BT	Bluetooth
CCTV	Closed-Circuit Television
CDF	Cumulative Distribution Function
COCO	Common Objects in Context
DeepSORT	Deep Simple Online and Realtime Tracking
FPS	Frames Per Second
GPU	Graphics Processing Unit
I/O	Input/Output
IDF1	ID F1 Score
IoT	Internet of Things
MAE	Mean Absolute Error
MOMO	Multi-Object Matching Operation
MOTA	Multiple Object Tracking Accuracy
MySQL	My Structured Query Language
PCA	People Counting Accuracy
PnP	Perspective-n-Point
RAM	Random Access Memory
Re-ID	Cross-camera Person Re-identification
ResNet	Residual Neural Network
RFID	Radio Frequency Identification
RMB	Renminbi
RMSE	Root Mean Square Error
SORT	Simple Online and Realtime Tracking
US	Ultrasound
USD	United States Dollar
UWB	Ultra Wide Band
ViT	Vision Transformer
Wi-Fi	Wireless Fidelity
YOLO	You Only Look Once

detection, and response by leveraging data analysis, pattern recognition, and automated decision-making. These capabilities have the potential to significantly improve the efficiency and effectiveness of safety interventions, particularly in complex and dynamic environments (Z. Li et al., 2025; M. Liu et al., 2025; Xiao et al., 2023; Zou et al., 2025). Recent research trends increasingly emphasize the transition from isolated algorithm development toward system-level intelligent perception and real-time decision support in complex indoor environments (Kreuzer et al., 2024).

Among the various applications of AI, situational awareness of personnel status serves as a fundamental component for improving evacuation response capabilities (Ouyang et al., 2025). Studies indicate that the majority of disaster casualties result from a lack of real-time, dynamic personnel information, such as the pedestrians' positions, their numbers, and rescuers' conditions, etc., at the escape and rescue phase in public areas (Guyo et al., 2023; Li et al., 2023). Notable disaster events that illustrate these challenges are shown in Fig. 1. Crucially, the precise location of personnel is one of the most vital pieces of information in disaster situations, playing a pivotal role for both evacuees and rescue teams (Li et al., 2014; Wong and Lee, 2023, 2025; Zhu et al., 2025). On one hand, dynamically tracking the positions of evacuees allows for the real-time planning of evacuation routes, helping to prevent individuals from being trapped in hazardous areas and enhancing evacuation efficiency. On the other hand, by knowing the location of rescue personnel, resources, and teams can be allocated effectively, preventing rescuers from becoming lost or searching blindly in the risky zone. However, the majority of existing evacuation simulation studies have relied on assumed personnel locations, which fail to accurately reflect the actual distribution of individuals in a real evacuation scenario (Han et al., 2024). This limitation reduces the practical applicability and effectiveness of the generated evacuation plans. In fact, existing research has shown that the initial location distribution of personnel significantly impacts the evacuation process (Yang et al., 2024). Therefore, the real-time acquisition of indoor personnel location data based on AI is of paramount importance for guiding and optimizing emergency evacuation strategies during public place disasters. In particular, the integration of AI with IoT infrastructure and digital twin technologies has recently emerged as a research hotspot for enabling global, real-time situational awareness in large-scale facilities (Sacoto-Cabrera et al., 2025).

Motivated by the emerging demand for global situational awareness and system-level integration in intelligent evacuation management, this study presents an AI-enabled Digital Twin framework designed for facility-wide evacuation monitoring. Distinct from algorithm-centric studies that focus on modifying deep learning network architectures, this work focuses on the engineering integration of computer vision with safety management, establishing a complete data pipeline from IoT sensors to a visualization interface for real-time situational awareness.

Fig. 1. Some disasters with serious consequences due to the lack of live information. (a) Türkiye–Syria earthquake (Turkey-Syria earthquake, 2023), (b) Grenfell Tower fire (Gregory, 2024), (c) Port of Santos toxic gas leak (Nuvem tóxica atinge quatro cidades no litoral de SP; vazamento continua, 2016).

Furthermore, unlike the single-camera dependent digital twin systems or traditional multi-camera monitoring systems, which often suffer from limited coverage and information fragmentation, the proposed system integrates data from multiple cameras to provide real-time, collaborative pedestrian information. The proposed system framework embodies the core architecture of a digital twin by establishing a closed-loop synchronization between the physical evacuation environment and its virtual counterpart. Instead of functioning merely as a monitoring tool, the system integrates four main components to construct a high-fidelity virtual replica: the IoT sensor network serves as the physical sensing layer to capture raw crowd dynamics; the cloud server and AI engine constitute the virtual modeling layer to transform fragmented video data into a unified, semantically rich digital model; and the user interface functions as the interaction layer, visualizing the indoor evacuation scenario in real time to support physical-world decision-making. To evaluate the system's performance, pedestrian experiments and real-world applications were conducted, demonstrating its potential in emergency evacuation research. This study aims to achieve the intelligent perception of pedestrian location information in public place evacuations and provide effective support for evacuation decision-making and emergency response.

The remaining part of the current work is organized as follows: Section 2 reviews the related work. Section 3 introduces the digital twin framework and the functionality of each component. Section 4 provides a detailed explanation of the multi-source video fusion method for pedestrian position monitoring, powered by the AI engine. Section 5 presents and discusses the demonstration of the digital twin system through real-time pedestrian monitoring tests. Section 6 discusses the research work with its application prospects and limitations. Finally, the main conclusions are summarized in Section 7.

2. Related work

To position the proposed system within the current research landscape, this section reviews recent advancements and emerging trends in three interconnected domains that have become central research hotspots in intelligent evacuation management: indoor pedestrian localization, digital twin applications in smart firefighting, and multi-camera tracking technologies. By synthesizing these directions, we identify existing limitations and clarify the research gap addressed in this study.

2.1. Indoor pedestrian localization

Common indoor pedestrian localization technologies include Wireless Fidelity (Wi-Fi), Bluetooth (BT), Radio Frequency Identification (RFID), Ultrasound (US), and Ultra Wide Band (UWB), all of which rely on radio signal transmission and reception to achieve location tracking (Sesyuk et al., 2022). These methods typically require the installation of devices in specific locations, leading to higher implementation costs. Additionally, they are prone to issues such as multipath effects and non-line-of-sight interference, which can compromise localization precision and hinder widespread adoption (Hayward et al., 2022; Zafari et al., 2019). In contrast, visual-based localization utilizes cameras as data acquisition devices, processing image data through algorithms to extract features for indoor localization. With the advancement of the Internet of Things (IoT), big data, and AI, real-time localization of individuals through monitoring systems has become feasible.

Compared to other technologies, visual-based localization offers superior stability and accuracy (Wang et al., 2024). The widespread deployment of surveillance cameras provides a solid infrastructure for mass adoption, eliminating the need for additional equipment or the requirement for individuals to carry devices, making it a cost-effective and scalable solution (Piasco et al., 2018). Chen et al. (2023) proposed a pedestrian location framework based on monocular cameras, which consists of three parts: coarse localization, auxiliary information generation, and information fusion. Niu et al. (2021) proposed a pedestrian three-dimensional (3D) localization method using monocular images combined with ground point clouds, calculating foot and head coordinates via collinearity equations under the assumption that pedestrians stand perpendicular to the ground. Sun et al. (2017) investigated human body localization with a single camera and introduced a device-free method based on panoramic cameras and indoor maps. Sato et al. (2020) developed a pedestrian localization method that estimates foot location using anthropometric features, such as face length, and maps it to the floor plane via perspective transformation. However, standalone localization algorithms are insufficient for comprehensive emergency management. In practical applications, visual-based pedestrian localization requires a mature system for data collection, transmission, processing, visualization, and interaction (Zhang et al., 2022b). Furthermore, the protection of personal identity of individuals during evacuation is a critical concern in video surveillance. Considering these factors, the integration of digital twin systems into smart firefighting applications is necessary to address both technical and ethical challenges (J. Chen et al., 2025b, 2025a).

2.2. Digital twins in smart firefighting

Digital Twin is a technology that creates a dynamic virtual replica of a physical object or system, enabling real-time monitoring, simulation, and predictive analysis through data synchronization (Y. Liu et al., 2025). Recently, several digital twin systems have already been applied to smart firefighting, reflecting a growing research focus on coupling physical sensing systems with virtual simulation environments for proactive risk management. For instance, Ding et al. (2023) introduced an intelligent emergency digital twin system based on computer vision and deep learning, which was evaluated in a stairwell. Zhang et al. (2024) developed a tunnel fire digital twin framework that considers vehicle classification and entry speed at tunnel entrances to estimate fire risks and evacuation safety in real time. To further enhance disaster resilience, Setijadi Prihatmanto et al. (2025) investigated the application of Digital Twin Cities for flood evacuation, highlighting how the integration of 3D city models with AI and IoT sensors can significantly improve predictive capabilities for identifying optimal evacuation routes in real time. Additionally, Jahangir et al. (2025) proposed the Building Simulation Identity Card framework to standardize and integrate specialized simulation models, such as earthquake propagation and occupant movement, thereby enabling more holistic and interoperable emergency evacuation simulations. These developments indicate a growing transition from static simulation models toward real-time, perception-driven digital twin systems for emergency management (Wang et al., 2026).

It is noteworthy that despite these advancements, most practical vision-based digital twin systems primarily focus on specific building areas and lack global monitoring capabilities for relying only on a single camera. This reliance on a single camera limits the monitoring coverage and makes the system susceptible to interference, occlusion, and blind spots, which may restrict the system's capacity to provide a holistic view of the evacuation process or other critical events. This limitation reveals a critical gap between current research prototypes and the emerging demand for facility-wide, collaborative evacuation perception systems.

2.3. Multi-camera pedestrian tracking

To overcome the constraints of single-view perception and achieve facility-wide monitoring, it is essential to employ multiple cameras to provide multi-view information, thereby expanding the monitoring field and enhancing the robustness of pedestrian localization through collaborative localization (Wang et al., 2025). Additionally, the world's increase in surveillance cameras in recent years has created favorable conditions for multi-camera pedestrian localization (Tran et al., 2022). Obviously, multi-camera collaborative perception and cross-view information fusion have become important research hotspots in intelligent surveillance and smart safety management in recent years (Kim et al.,

2026). Currently, to fully utilize these visual resources, research primarily focuses on multi-camera pedestrian tracking and follows a "tracking-by-detection" paradigm. In this framework, state-of-the-art detectors like the You Only Look Once (YOLO) series (Lu et al., 2026) are employed to detect pedestrians. Subsequently, algorithms such as Deep Simple Online and Realtime Tracking (DeepSORT) (Wojke et al., 2017) are widely adopted to track pedestrian trajectories within single camera views. For cross-camera identity association, cross-camera person Re-identification (Re-ID) techniques (Zheng et al., 2019) utilize deep convolutional neural networks to extract discriminative appearance features, achieving high matching accuracy on benchmark datasets.

However, directly applying these traditional multi-camera tracking systems to emergency evacuation monitoring faces significant limitations. First, high-accuracy Re-ID models often require pixel-level feature extraction for every target, leading to computational latency that hinders real-time processing in high-density crowds. Second, reliance on facial or appearance features raises privacy concerns. Third, and most critically, while these systems excel at logical linking (identifying "who is who"), they often lack the capability to integrate multi-view data into a unified global coordinate system for precise spatial localization. Consequently, information between cameras is not effectively integrated, resulting in information isolation. The integration of visual data from multiple cameras to overcome information fragmentation and enable real-time collaborative localization of individuals during evacuation remains a significant challenge (Abbas et al., 2025; D. Li et al., 2025; Qiu et al., 2025). Therefore, further research is needed to develop and demonstrate a comprehensive digital twin framework for global personnel evacuation monitoring in public places.

3. Digital twin system framework

The framework of the proposed intelligent monitoring system mainly consists of four components (see Fig. 2): (1) IoT sensor network, (2) Cloud server, (3) AI engine, and (4) User interface. The IoT sensor network is installed in the building before disaster incidents to collect the on-site data, with closed-circuit television (CCTV) gathering critical information about the personnel during an evacuation. This data is then transmitted to a cloud server, which stores and organizes it in a standardized format for easy access. The cloud server also facilitates communication between the IoT network and the AI engine. The AI engine processes the data to identify pedestrians' positions and provide real-time alerts, helping to guide evacuation efforts and prioritize safety measures. The user interface serves as the platform for visualizing the pedestrian movement information, allowing for cyber-physical interaction and enabling decision-makers to respond promptly. The following subsections detail these four components.

3.1. IoT sensor network

The IoT sensor network plays a crucial role in gathering real-time data from the environment. In this system, cameras serve as the primary sensors, capturing visual information of the building's interior. These cameras are strategically deployed throughout the site to cover critical evacuation areas and provide continuous monitoring. Specifically, the deployment follows three empirical principles: (1) Occlusion & Blind Spot Minimization: Cameras are mounted at elevated positions (typically >2.0 m) with a depression angle to minimize pedestrian-to-pedestrian occlusion and reduce monitoring blind spots; (2) Field-of-View Overlap: Adjacent cameras maintain an overlapping region (typically 10%-30%) to ensure stable identity handover for cross-camera association; and (3) Effective Resolution: The camera spacing is constrained to ensure that the pixel density of pedestrians at the farthest edge of the monitoring range satisfies the minimum input requirements for the detection and tracking model. The sensor network enables real-time monitoring of pedestrian movements by processing the visual data captured by these cameras. The integration of multiple cameras ensures comprehensive coverage, reducing blind spots and improving the robustness of the system. Furthermore, the sensor network is

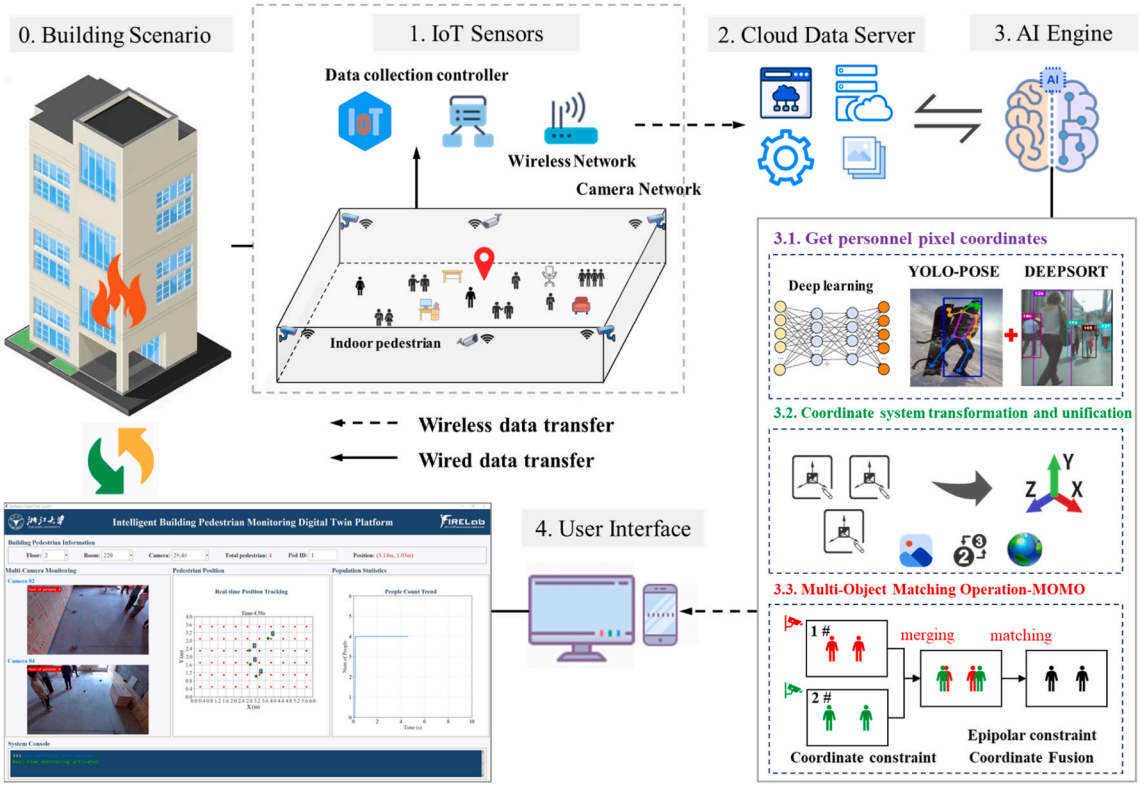

Fig. 2. Framework of the intelligent monitoring digital twin system.

engineered for scalability, thus allowing it to accommodate the dynamic nature of indoor evacuation scenarios, where the number of people and their locations can rapidly change.

3.2. Cloud server

The cloud server functions as the central hub for data storage, processing, and management. All the data collected by the IoT sensor network is transmitted to the cloud, where it is stored and processed in real time. The cloud server facilitates the synchronization of the data from various sensors, ensuring that the information is up-to-date and readily available for analysis. It also supports the system's scalability by providing the requisite computational resources to handle large volumes of data, particularly in the context of densely populated buildings. Furthermore, the cloud server ensures data redundancy and security, safeguarding the information against potential failures or cyber threats. The data interaction uses the My Structured Query Language (MySQL) Connector.

3.3. AI engine

The AI engine is the core component of the system, responsible for processing and analyzing the data gathered by the IoT sensor network. It employs deep learning and computer vision techniques to accurately identify and track pedestrians within the building. The AI engine uses multi-source video fusion to combine data from different cameras, enabling the system to provide a comprehensive and accurate representation of the emergency evacuation scenario. It performs real-time pedestrian localization, detecting individual positions and movements, and predicting the total number of people. By leveraging AI, the engine enhances the system's ability to adapt to dynamic changes in the evacuation environment. More detailed information, including the pedestrian coordinate acquisition, coordinate transformation and unification, and multi-camera pedestrian matching, is given in Section 4.

3.4. User interface

The user interface serves as the front end of the digital twin system, providing emergency responders, safety managers, and other stakeholders with a clear and intuitive view of the evacuation scenario. The interface visualizes real-time data collected from the IoT sensor network, such as the location of pedestrians and the number of pedestrians. The user interface is designed to be user-friendly, with easy-to-read charts that highlight key evacuation data. It allows responders to monitor the situation timely and make informed decisions, improving the overall efficiency and safety of the evacuation process.

4. Multi-source video fusion monitoring pedestrian method (AI engine)

4.1. Principle of pedestrian visual-based localization with monocular camera

Monocular cameras are currently the primary imaging devices used in public place scene monitoring (Kim et al., 2023). This study focuses on collaborative pedestrian localization monitoring using multiple monocular cameras. To achieve this, a fundamental understanding of monocular camera-based visual localization is required, which involves key concepts such as the camera imaging model, camera calibration, and monocular camera-based pedestrian coordinate transformation theory.

4.1.1. Camera imaging model

4.1.1.1. Pinhole camera model. The process of camera imaging essentially involves the projection of three-dimensional spatial information onto a two-dimensional image plane. The most commonly used model in computer vision for describing this process is the pinhole camera model (Devernay and Faugeras, 2001). It represents the camera imaging process using pixel coordinates $(O_{uv} - uv)$, image coordinates $(O_{xy} - xy)$, camera coordinates $(O_c - X_cY_cZ_c)$, and world coordinates $(O_w - X_wY_wZ_w)$, as shown in Fig. 3.

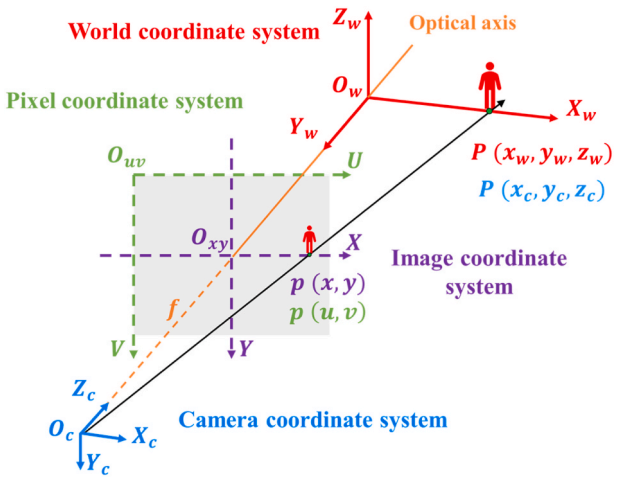

Fig. 3. Pinhole camera model.

The transformation relationship of the four coordinate systems is shown in Fig. 4. The transformation from the world coordinate (x_w, y_w, z_w) to the camera coordinate (x_c, y_c, z_c) is achieved through a rigid body transformation, which involves both rotation and translation. The transformation is typically represented by a rotation matrix and a translation vector. The conversion from the camera coordinate to the image coordinate (x, y) is based on perspective projection. This projection depends on the camera's intrinsic parameters, such as focal length and principal point. The image coordinate system typically has its origin at the center of the image. Next, the transformation from the image coordinate to the pixel coordinate (u, v) is an affine transformation, which involves scaling and shifting the coordinates. In the pixel coordinate system, the origin is located at the upper-left corner of the image, and the coordinates are measured in pixels.

Finally, the complete coordinate transformation from the world coordinate system to the pixel coordinate system involves combining all these steps. The total transformation can be represented by Eq. (1):

$$\begin{cases} z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_x} & 0 & u_0 \\ 0 & \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R}|\mathbf{T}] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \\ \mathbf{K} = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, f_x = \frac{f}{d_x}, f_y = \frac{f}{d_y} \end{cases} \tag{1}$$

where z_c represents the point coordinate on the Z-axis in the camera coordinate system; (u, v) represents the point coordinates in the pixel coordinate system; d_x and d_y represent the width and length of a pixel, respectively; (u_0, v_0) represents the coordinates of the origin of the image coordinate system in the pixel coordinate system; f represents the camera focal length; $\mathbf{R}$ and $\mathbf{T}$ represent the camera rotation matrix and translation vector, respectively; $[\mathbf{R}|\mathbf{T}]$ represents the camera extrinsic

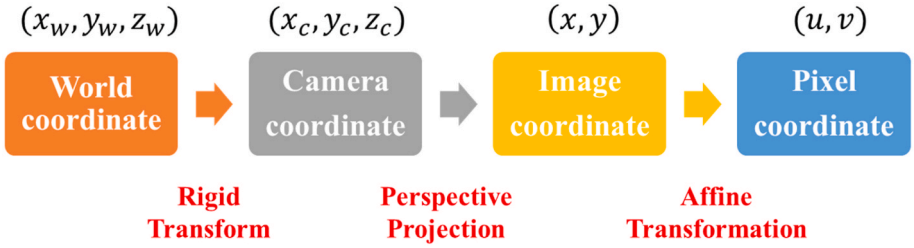

Fig. 4. Transformation relationship of the four coordinate systems.

matrix, which will change whenever the camera moves or changes its viewpoint; **K** represents the camera intrinsic matrix, which is fixed and characterizes the internal properties of the camera; f_x and f_y represent the equivalent focal length expressed in pixel width and height, respectively; (x_w, y_w, z_w) represents the point coordinates in the world coordinate system.

4.1.1.2. Camera distortion model. In practical camera imaging, lens distortion is an important factor that must be accounted for to achieve accurate image measurements. There are two primary types of lens distortion: radial distortion and tangential distortion.

Radial distortion occurs due to the lens's shape, particularly at the edges of the image, which can be characterized in Eq. (2):

$$\begin{cases} x_{distorted} = x\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) \\ y_{distorted} = y\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) \end{cases} \tag{2}$$

where (x, y) represents the undistorted image coordinates; $(x_{distorted}, y_{distorted})$ represents the distorted image coordinates; k_1, k_2 and k_3 are the radial distortion coefficients; $r^2 = x^2 + y^2$ is the radial distance from the image center.

Tangential distortion arises when the lens and the image plane are not perfectly parallel, leading to an asymmetrical distortion. Tangential distortion is typically modeled in Eq. (3):

$$\begin{cases} x_{distorted} = x + \left[2p_1 xy + p_2\left(r^2 + 2x^2\right)\right] \\ y_{distorted} = y + \left[p_1\left(r^2 + 2y^2\right) + 2p_2 xy\right] \end{cases} \tag{3}$$

where p_1 and p_2 are the tangential distortion coefficients.

In sum, by combining Eqs. (2) and (3), five camera distortion coefficients are used and the full camera distortion model can be written as Eq. (4):

$$\begin{cases} x_{distorted} = x\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) + 2p_1 xy + p_2\left(r^2 + 2x^2\right) \\ y_{distorted} = y\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) + 2p_2 xy + p_1\left(r^2 + 2y^2\right) \end{cases} \tag{4}$$

4.1.2. Camera calibration

As previously analyzed, to achieve the conversion from pixel coordinates to world coordinates, it is essential to know the camera's intrinsic parameters, extrinsic parameters, and distortion coefficients. Camera calibration estimates the parameters necessary for accurate localization. Zhang's method (Zhang, 2000), a widely used approach, employs a planar pattern (e.g., checkerboard) to compute intrinsic parameters (focal length, principal point), extrinsic parameters (rotation, translation), and distortion coefficients. Due to its simplicity and accuracy, it is adopted in this study for camera calibration.

4.1.3. Pedestrian coordinate transformation based on monocular camera

According to Eq. (1), it can be observed that to convert from pixel coordinates to world coordinates, additional information z_c is required. As shown in Fig. 3, for any given point P (x_c, y_c, z_c) in 3D world, its projection onto the image plane through the camera C results in the image point p (x, y). Since the transformation from the camera coordinate system to the image coordinate system involves perspective projection, which is a one-to-many relationship, any point along the perspective ray O_cP will correspond to the same image point p. Therefore, if the pixel location of point p in the image is known, it can only determine that the real-world point P lies somewhere along the ray O_cP, but the 3D coordinates of the point P in the real world cannot be directly obtained.

However, once point P lies on a specific plane Z in the 3D world is known, the intersection of the perspective ray O_cP with the plane Z can be used to solve for the unique coordinates of point P. In pedestrian localization scenarios, as the pedestrian is assumed to be on the ground, so the ground plane is often selected as the reference plane Z, where the coordinate z_w of the point P in the world coordinate system is typically set to 0. This assumption allows for the accurate determination of the 3D location of the pedestrian.

Since the rotation matrix **R** in the extrinsic parameters is a 3 × 3 matrix and the translation vector **T** is a 3-dimensional vector, the extrinsic parameters can be expressed in Eq. (5):

$$\begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5}$$

Where r_{ij} denotes the element in the i-th row and j-th column of the rotation matrix **R**, and t_i denotes the i-th component of the translation vector **T**.

First, substituting the extrinsic parameter form from Eq. (5) into Eq. (1), the following Eq. (6) is obtained.

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x r_{11} + u_0 r_{31} & f_x r_{12} + u_0 r_{32} & f_x r_{13} + u_0 r_{33} & f_x t_1 + u_0 t_3 \\ f_y r_{21} + v_0 r_{31} & f_y r_{22} + v_0 r_{32} & f_y r_{23} + v_0 r_{33} & f_y t_2 + v_0 t_3 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{6}$$

Then, substituting $z_w = 0$ into Eq. (6) and $z_c = r_{31}x_w + r_{32}y_w + t_3$ can be obtained.

Next, substituting $z_c = r_{31}x_w + r_{32}y_w + t_3$ into Eq. (6) yields Eq. (7).

$$\begin{cases} (r_{31}u - f_x r_{11} - u_0 r_{31})x_w + (r_{32}u - f_x r_{12} - u_0 r_{32})y_w = f_x t_1 + (u_0 - u)t_3 \\ \left(r_{31}v - f_y r_{21} - v_0 r_{31}\right)x_w + \left(r_{31}v - f_y r_{22} - v_0 r_{32}\right)y_w = f_y t_2 + (v_0 - v)t_3 \end{cases} \tag{7}$$

To simplify the calculation, make $A = r_{31}u - f_x r_{11} - u_0 r_{31}$, $B = r_{32}u - f_x r_{12} - u_0 r_{32}$, $C = f_x t_1 + (u_0 - u)t_3$, $D = r_{31}v - f_y r_{21} - v_0 r_{31}$, $E = r_{31}v - f_y r_{22} - v_0 r_{32}$, $F = f_y t_2 + (v_0 - v)t_3$, and Eq. (7) can be converted to Eq. (8).

$$\begin{cases} Ax_w + By_w = C \\ Dx_w + Ey_w = F \end{cases} \tag{8}$$

Finally, we can establish a relationship between the pedestrian pixel coordinates and the world coordinates, as shown in Eq. (9):

$$\begin{cases} x_w = \dfrac{CE - BF}{AE - BD} \\ y_w = \dfrac{AF - CD}{AE - BD} \end{cases} \tag{9}$$

According to Eq. (9), once the pixel coordinates (u, v) of the pedestrian on the ground are obtained after distortion correction, along with the camera intrinsic matrix $\mathbf{K}$ and camera extrinsic matrix $[\mathbf{R}|\mathbf{T}]$, the world coordinates (x_w, y_w, 0) of the pedestrian can be computed.

4.2. Pedestrian detection and tracking

After understanding the conversion relationship between pixel coordinates and world coordinates, to intelligently acquire the pedestrian's pixel coordinates, it is necessary to detect the pedestrian in the image and obtain the coordinates of their foot position on the ground. In this study, we use YOLOv8-Pose for this task.

YOLOv8-Pose is a cutting-edge pose estimation model built on the YOLOv8 architecture, which is widely known for its real-time object detection capabilities (Maji and Mathew, 2022). YOLOv8-Pose extends this framework by adding a dedicated human pose estimation module, enabling it to accurately detect and localize key body keypoints in real time. The overall network architecture of YOLOv8-Pose model can be seen in Fig. 5a. For human pose estimation, the model works by detecting 17 body keypoints, as shown in Fig. 5b.

In this study, we focus specifically on the left and right ankle points as the key body landmarks to represent the pedestrian's foot position on the ground. More precisely, we calculate the midpoint between the left and right ankle keypoints, which provides a robust and accurate estimate of the pedestrian's foot position in the image. This midpoint is then used as the pedestrian's foot pixel coordinate on the ground, which will later be transformed into world coordinate for further localization and tracking. Notably, the ankle keypoints physically remain in proximity to the floor level, so the validity of this ground-plane mapping ($z_w = 0$) holds robustly across common postures (e.g., standing, walking, sitting, or squatting). Although minor vertical deviations (e.g., shoe sole height) exist, they are negligible relative to the high deployment altitude of the cameras (common >2 m). To ensure the reproducibility of the proposed method, key implementation details are specified as follows. In this study, we use the YOLOv8m-pose model for this task. The model is pretrained on the Common Objects in Context (COCO) dataset, and we specifically filter the output to detect only the "person" class, utilizing a confidence threshold of 0.5 to ensure detection reliability. To handle potential keypoint occlusions in complex crowd scenarios, an adaptive extraction strategy is implemented: when both ankle keypoints are visible with high confidence, the midpoint between the left and right ankles is calculated as the primary ground contact proxy. Conversely, if one or both ankles are occluded, the system automatically falls back to using the bottom center of the detected bounding box.

Additionally, to continuously track the pedestrian's position across multiple frames, DeepSORT is utilized. DeepSORT extends the Simple Online and Realtime Tracking (SORT) algorithm by incorporating deep learning-based appearance features, enhancing tracking accuracy through combined motion and appearance information. It uses Kalman filtering and the Hungarian algorithm for motion tracking, while deep features help distinguish pedestrians and manage occlusions. The flowchart of DeepSORT is shown in Fig. 6. Note that in this framework, DeepSORT (initialized with standard ckpt.t7 weights) is strictly used for intra-camera temporal tracking. The identity maintenance across different camera views is subsequently handled by the MOMO algorithm (See Section 4.4), thereby mitigating reliance on appearance embeddings, which are often unreliable under significant viewpoint changes.

By employing YOLOv8-Pose for pedestrian detection and DeepSORT for continuous tracking, we can accurately and efficiently obtain the position of pedestrians in the camera's field of view, enabling real-time localization on the ground for subsequent conversion into world coordinates. Crucially, this keypoint-based detection paradigm constitutes the technical path for the system's "non-intrusive representation," offering reduced identifiability compared to traditional face-based surveillance. Unlike methods that rely on facial recognition, the proposed system utilizes the YOLO-Pose model to extract only the geometric coordinates of body keypoints, thereby filtering out biometric facial features during the inference phase. Regarding system architecture, a strict separation between the frontend and backend is implemented. The backend processes raw video feeds temporarily in memory to extract coordinates (x, y, t, ID) without permanently storing video files. Subsequently, only these anonymous data streams are transmitted to the frontend. Consequently, in practical real-world deployments, the user interface visualizes pedestrians solely as abstract data points or generic avatars rather than displaying live video feeds (note: the video feeds presented in this paper are exclusively for the purpose of experimental validation and algorithm demonstration), effectively decoupling evacuation monitoring from personal identity intrusion.

4.3. Multi-camera system calibration

This study develops a digital twin system for global pedestrian evacuation monitoring based on a multi-camera system. A multi-camera system typically consists of several cameras placed at strategic locations within the facility. Each camera captures real-time video footage and

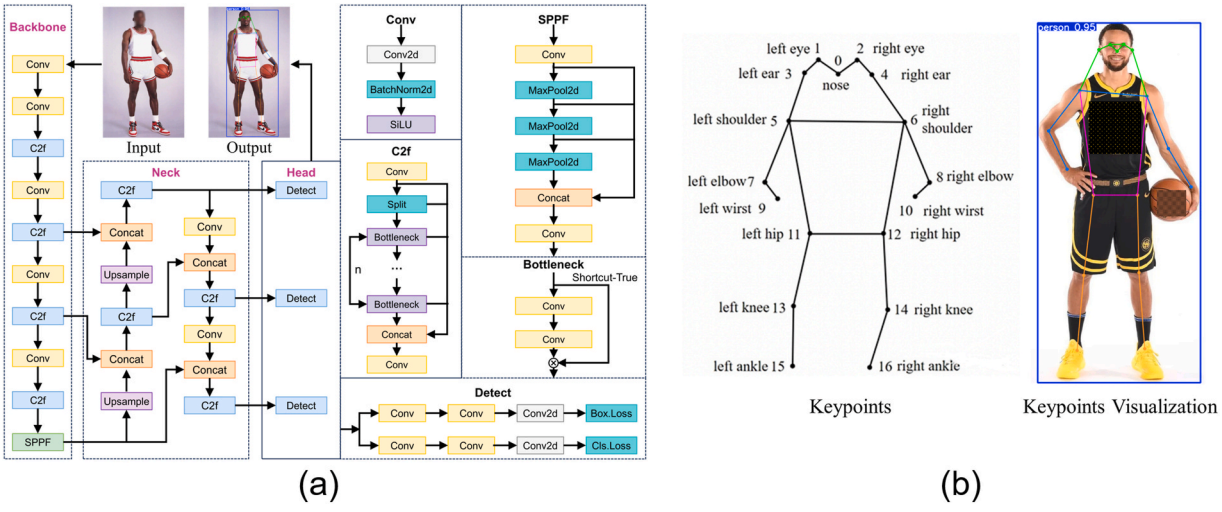

Fig. 5. YOLOv8-Pose model. (a) overall network architecture, (b) 17 Body keypoints and their visualization.

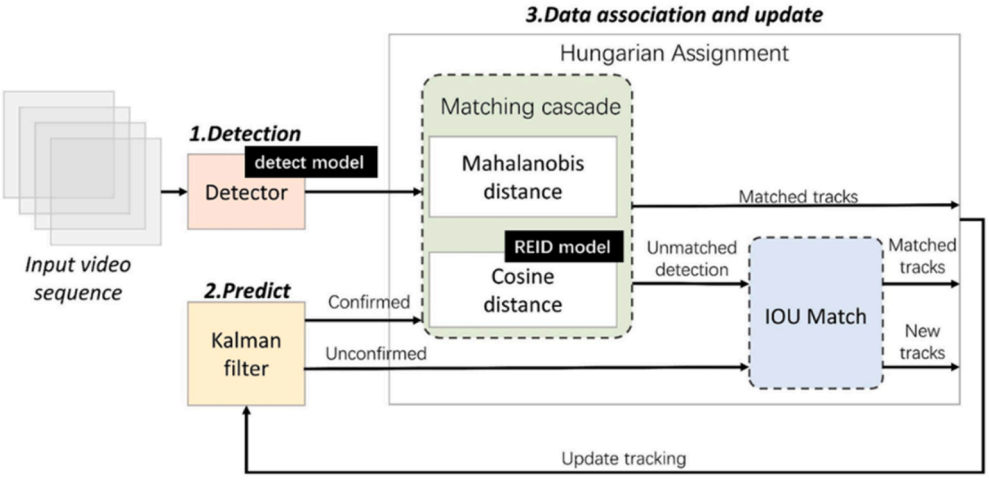

Fig. 6. Flowchart of DeepSORT algorithm.

locates pedestrians using visual techniques. However, the coordinates provided by each camera are specific to its own coordinate system, which is generally defined for the camera's position and orientation within the building.

To integrate localization data from multiple cameras, it is necessary to transform each camera's coordinate system to a unified world coordinate system, typically based on a fixed reference frame for the indoor environment. Due to the high camera deployment density within the facility, a pairwise camera combination method can be used to solve the positional relationships between multiple camera systems and achieve a unified world coordinate system. Therefore, a multi-camera system calibration method based on cascade transformation is proposed to solve the coordinate system unification problem. Unlike methods that require measuring the absolute physical locations of every camera or establishing world coordinates for all reference points, our approach relies on pairwise relative pose estimation. In this framework, one of the cameras is designated as the global reference. For adjacent cameras, the calibration prerequisites are limited to their intrinsic parameters and a set of common feature points (e.g., Chessboard grid corner points or floor tile intersections) visible in their overlapping fields of view. Crucially, the absolute world coordinates of these common points are not required. By establishing feature correspondences between adjacent views, the precise relative spatial relationships (rotation matrix and translation vector) are mathematically derived. This process allows the pose of each camera to be logically derived in a chain-like manner, effectively unifying the entire network into a single world coordinate system without extensive manual surveying.

Fig. 7 presents the calibration-based multi-camera cascade relationship. In Fig. 7, there are n camera coordinate systems, namely O_{c_1}, O_{c_2}, O_{c_3}, O_{c_4}, …, O_{c_n}. For the sake of illustration, the O_{c_1} camera coordinate system is designated as the global reference camera coordinate system. The transformation relationships between the O_{c_1} camera coordinate system and the other camera systems (O_{c_2} to O_{c_n}) are then calculated. Finally, the conversion between the O_{c_1} camera coordinate system and the unified world coordinate system is established to obtain the world coordinates of all the cameras.

The multi-camera pose transfer concept operates by calculating the relative positional relationship of the n-th camera with respect to the (n - 1)-th camera. Similarly, the relative position between the (n - 1)-th and (n - 2)-th cameras is determined, and this process continues in this manner. Eventually, the relative position of the n-th camera with respect to the first camera (O_{c_1}) is obtained. This allows us to derive the absolute pose information of the n-th camera relative to the reference camera coordinate system defined by O_{c_1}.

Let the matrix $\mathbf{H}_i^j$ represent the transformation matrix from the O_{c_i} camera coordinate system to the O_{c_j} camera coordinate system, which is shown in Eq. (10):

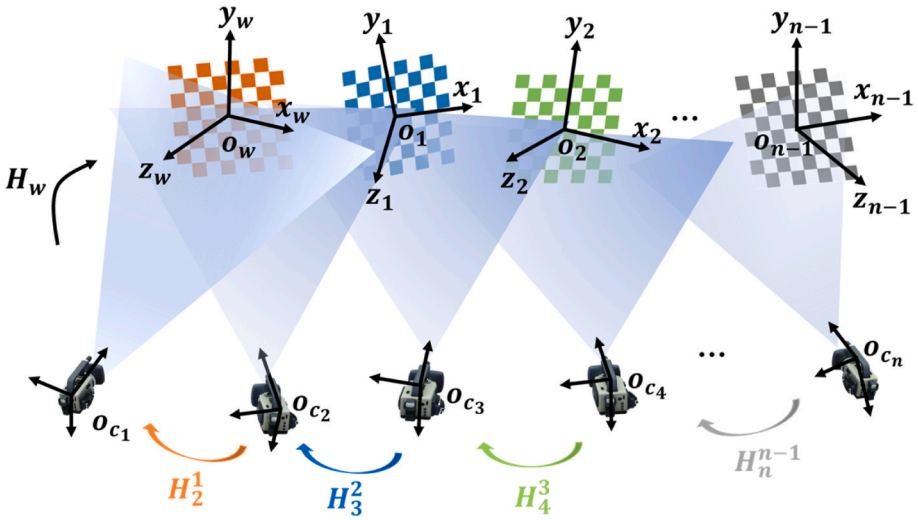

Fig. 7. Calibration-based multi-camera cascade relationship.

$$\mathbf{H}_i^j = \begin{bmatrix} \mathbf{R}_i^j & \mathbf{T}_i^j \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \tag{10}$$

The relative pose transformation matrix between the n-th camera and the (n - 1)-th camera is expressed as $\mathbf{H}_n^{n-1}$. Therefore, the transformation matrix from the n-th camera coordinate system to the O_{c_1} camera coordinate system is given by Eq. (11):

$$\mathbf{H}_n^1 = \prod_{i=1}^{n-1} \mathbf{H}_{n-i+1}^{n-i} = \mathbf{H}_n^{n-1} \cdot \mathbf{H}_{n-1}^{n-2} \cdot \cdots \cdot \mathbf{H}_3^2 \cdot \mathbf{H}_4^3 \cdot \mathbf{H}_2^1 \tag{11}$$

Next, we describe the multi-camera pose transformation relationship using rotation matrices and translation vectors. Let $\mathbf{P}^n$ represent the coordinates of a spatial point p in the O_{c_n} camera coordinate system. The transformation of this point's coordinates to the O_{c_1} coordinate system can be expressed as Eq. (12):

$$\begin{cases} \mathbf{P}^1 = \mathbf{R}_n^1 \cdot \mathbf{P}^n + \mathbf{T}_n^1 \\ \mathbf{R}_n^1 = \prod_{i=1}^{n-1} \mathbf{R}_{n-i+1}^{n-i} \\ \mathbf{T}_n^1 = \sum_{i=1}^{n-1} \left(\prod_{j=i+1}^{n-1} \mathbf{R}_j^{j-1} \right) \mathbf{T}_{n-i+1}^{n-i} \end{cases} \tag{12}$$

where $\mathbf{P}^1$ represent the coordinates of a spatial point p in the O_{c_1} camera coordinate system; $\mathbf{R}_n^1$ is the rotation matrix and $\mathbf{T}_n^1$ is the translation vector that transforms the coordinates from O_{c_n} to O_{c_1}.

Once the coordinates in the O_{c_1} camera coordinate system are obtained, we can associate them with the transformation matrix $\mathbf{H}_w$ that relates the O_{c_1} camera coordinate system to the unified world coordinate system. This gives the world coordinate $\mathbf{P}_w$ of the point p, as expressed in Eq. (13):

$$\mathbf{P}_w = \mathbf{H}_w \cdot \mathbf{P}_1 \tag{13}$$

By utilizing this series of transformations, the world coordinates of the point p can be accurately determined, even if it is initially detected by any camera in the system. This cascade calibration strategy not only extends the monitoring coverage but also explicitly establishes the topological relationships between cameras, which is essential for the subsequent cross-view identity matching. It is important to note that while this cascade strategy may carry the risk of error accumulation, this potential issue is effectively mitigated by the use of global static reference points (e.g., the standardized floor tile grid) in this study. These physical references act as absolute geometric anchors ("ground truth") during the pairwise calibration process. Since each camera's extrinsic parameters can be refined by aligning with these fixed visible references rather than relying solely on the relative extrapolation from the previous camera, the propagation of calibration error is constrained within a stable budget, ensuring the overall precision required for crowd localization.

4.4. Multi-object matching operation-MOMO

When a pedestrian is monitored by multiple cameras simultaneously, performing target matching across different camera views becomes imperative to prevent the duplicate recognition of the same individual. This step is one of the critical aspects of multi-camera information fusion. Currently, several methods are available for pedestrian matching across multiple cameras including image stitching (Brown and Lowe, 2007), Re-ID, and location-based pedestrian matching. The detailed overview and comparison of these approaches can be found in Appendix A. Based on this comparative analysis, the location-based method is selected for this study due to its computational efficiency and robustness in privacy-sensitive environments.

However, in practical applications, simply relying on coordinate mapping is insufficient. Due to various factors (such as discrepancies in camera calibration, camera angles, or pedestrian foot position coordinate detection error), the coordinates of the same pedestrian in a unified coordinate system may not align perfectly, as shown in Fig. 8, leading to decreased matching accuracy. Considering the monitoring ranges of cameras, the most common scenario involves a pedestrian being captured simultaneously by two cameras. Therefore, the key focus is on solving the matching problem for the same pedestrian observed by two cameras. To address this challenge, we propose a deterministic matching-and-fusion heuristic termed the Multi-Object Matching Operation (MOMO). Unlike iterative optimization solvers, MOMO is designed as a streamlined, rule-based pipeline that executes a sequential consistency check to merge pedestrian data from the unified world coordinate system. Specifically, two constraints are introduced: the distance constraint and the epipolar constraint.

The distance constraint assumes that the same pedestrian in a unified world coordinate system should have identical or nearly identical coordinates across different camera views. This principle enables the matching of pedestrian targets by evaluating the spatial proximity of their coordinates. Let $\mathbf{P}_1$ and $\mathbf{P}_2$ represent the coordinates of a pedestrian captured in the world coordinate system by camera 1 and camera 2, respectively. A spatial distance threshold ϕ is introduced to assess whether these points represent the same pedestrian. If the Euclidean distance between these two points $d = \|\mathbf{P}_1 - \mathbf{P}_2\| \leq \phi$, it is determined that $\mathbf{P}_1$ and $\mathbf{P}_2$ likely represent the same pedestrian. Conversely, if $d > \phi$, then $\mathbf{P}_1$ and $\mathbf{P}_2$ are deemed to represent different pedestrians. To determine an appropriate threshold ϕ, we adopt Hall's four zones of social distance theory (Lipman and Hall, 1970), which categorizes interpersonal distances as follows: intimate distance (0 - 0.45 m), personal distance (0.45 - 1.2 m), social distance (1.2 - 3.6 m), and public distance (3.6 - 7.6 m) in social contexts. For public spaces, the intimate space distance of 0.45 m is selected as the threshold. This reflects the assumption that two different individuals in public areas are unlikely to have inter-personal distances smaller than 0.45 m.

To further enhance the robustness of pedestrian matching, the epipolar constraint is introduced while satisfying the distance constraint.

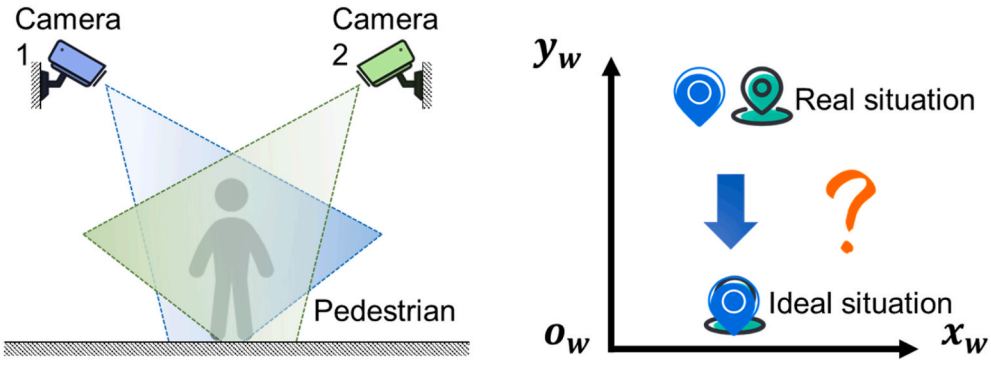

Fig. 8. Example of location-based pedestrian matching method.

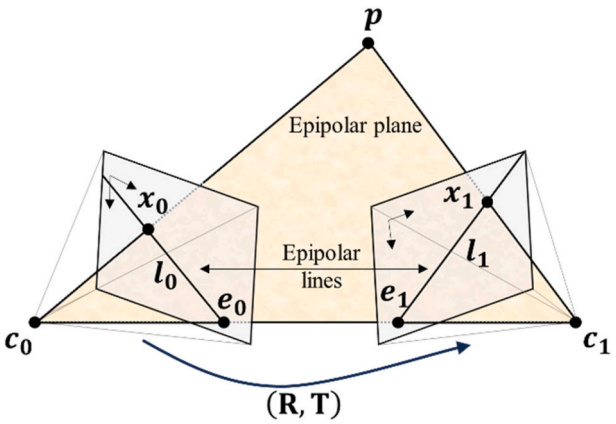

Fig. 9. Schematic diagram of the epipolar constraint.

This constraint is rooted in photogrammetry and leverages the geometric relationship between two cameras and their captured images. In photogrammetry, an epipolar plane is formed by a 3D object point and the optical centers of two cameras. The intersection of the epipolar plane with the two image planes creates epipolar lines. For any pixel in the left image, its corresponding point in the right image must lie on the epipolar line. The schematic diagram of the epipolar constraint can be seen in Fig. 9. c_0 and c_1 represent the optical centers of two cameras. A point p in 3D space has projections onto the image planes of these two cameras, denoted as x_0 and x_1 respectively. The line connecting c_0 and c_1 intersects the image planes at two points, e_0 and e_1, known as the epipoles. The lines l_0 and l_1, which pass through the corresponding epipoles, are called epipolar lines. The plane formed by the points c_0, c_1 and p is referred to as the epipolar plane.

For pedestrians, the pixel coordinates of their foot positions in the two images are checked to determine whether they are corresponding points. If the points lie on each other's epipolar lines, they are deemed the same pedestrian.

The relationship between the two points is expressed by the epipolar constraint equation, as calculated in Eq. (14):

$$\mathbf{p}_2^T \mathbf{F} \mathbf{p}_1 = 0 \tag{14}$$

where are $\mathbf{p}_1$ and $\mathbf{p}_2$ are the 2D pixel coordinates of the pedestrian's midpoint of the both ankles in camera 1 and camera 2 images respectively, and $\mathbf{F}$ is the fundamental matrix derived from the intrinsic and extrinsic parameters of the two cameras.

In a word, pedestrians whose coordinates meet the distance constraint are treated as preliminary matches. Among these candidates, the epipolar constraint is used for further verification, reducing false positives caused by positional inaccuracies or noise. Therefore, by combining these constraints, it is possible to achieve higher accuracy in pedestrian matching for multi-camera systems, where quick and accurate pedestrian matching is essential for decision-making.

After matching the same pedestrian across dual-camera views, it is essential to compute the fused world coordinates. As shown in Section 5.1.2, the estimation error decreases as the target approaches the camera. Therefore, we first obtain the pedestrian's coordinates from each camera and use their midpoint as an initial estimate. This is then refined by assigning greater weight to the closer camera, with weights inversely proportional to the distance. The detailed calculation method is as follows:

Step 1: Midpoint Initialization

The geometric midpoint $\mathbf{P}_{\text{mid}}$ serves as an initial spatial reference:

$$\mathbf{P}_{\text{mid}} = \frac{\mathbf{P}_1 + \mathbf{P}_2}{2} \tag{15}$$

Step 2: Adaptive Weight Assignment

Assign normalized inverse distance weights to prioritize contributions from closer cameras:

$$w_i = \frac{\frac{1}{d_i}}{\frac{1}{d_1} + \frac{1}{d_2}} \tag{16}$$

$$d_i = \|\mathbf{P}_{\text{mid}} - \mathbf{C}_i\|_2 \tag{17}$$

where w_i is the weight of Camera i; d_i is the Euclidean distance between $\mathbf{P}_{\text{mid}}$ and the optical center of Camera i; $\mathbf{C}_i$ is the 3D position of Camera i.

Step 3: Final Coordinate Fusion

The final world coordinate $\mathbf{P}_{\text{world}}$ is calculated as the weighted average of $\mathbf{P}_1$ and $\mathbf{P}_2$:

$$\mathbf{P}_{\text{world}} = w_1 \cdot \mathbf{P}_1 + w_2 \cdot \mathbf{P}_2 \tag{18}$$

5. Demonstration of digital twin for real-time monitoring pedestrian tests

To demonstrate and evaluate the performance of the proposed intelligent digital twin system, three tests were conducted, including two in a controlled laboratory environment (one static and one dynamic) and one in a real-world scenario. **Test 1**, the static test, involved four stationary pedestrians positioned within a multi-camera monitoring area. This test is designed to assess the fundamental capabilities of the system, including pedestrian detection, tracking accuracy, cross-camera matching, and localization precision; **Test 2** followed a similar setup, but the pedestrians were allowed to walk randomly within the monitored area. This dynamic setting enabled the evaluation of the system's head count accuracy under motion; **Test 3** utilized multi-view video footage captured from the waiting hall of a high-speed railway station. This scenario presented a more realistic and complex environment, with varying crowd densities and diverse pedestrian activities. As such, test 3 serves to validate the system's real-time performance and robustness in monitoring multiple pedestrians in large-scale public infrastructure. Detailed descriptions of the test setups and configurations are provided in Table 1.

5.1. Demonstration of experimental scenario test

5.1.1. Experimental setup

The experimental area of the laboratory test (tests 1 & 2) consisted of a 4 × 6 m^2 indoor plane, within which 60 calibration points were systematically arranged on the floor to facilitate both camera calibration and the evaluation of localization precision. Four participants were involved in the experiment, with two staff members providing on-site guidance to ensure smooth and safe execution. Additionally, four cameras were strategically mounted on the walls to cover the entire experimental area, capturing the complete process. The detailed layout of the

Table 1
Detailed descriptions of pedestrian tests.

No.	No. of people	Scenario	Description
Test 1	4	Static test	Pedestrians stand still in a multi-camera monitoring area.
Test 2	4	Dynamic test	Pedestrians walk randomly in the experimental area.
Test 3	Dynamic change	Real scenario	Scene of people's activities in the waiting hall of a high-speed railway station.

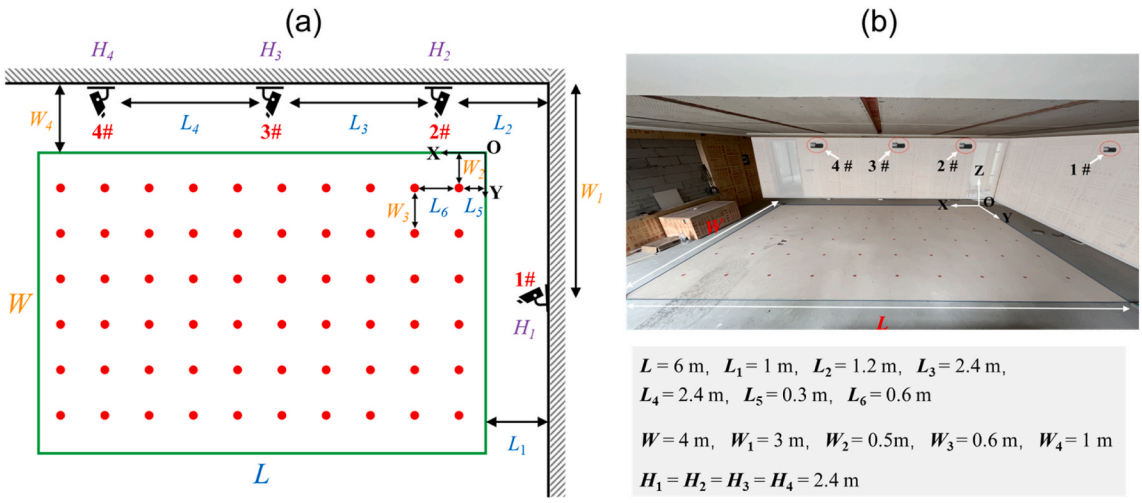

Fig. 10. Layout of the experimental area. (a) a diagram, and (b) on-site photo of the test scene.

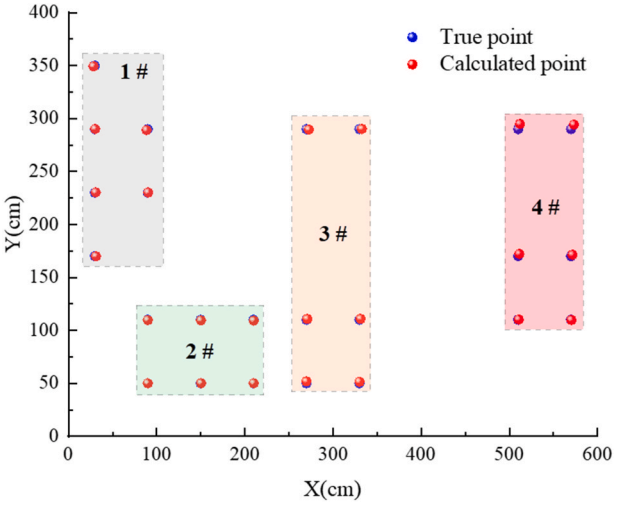

Fig. 11. Comparative results of true and calculated points.

experimental scene is illustrated in Fig. 10, and the camera calibration procedure is thoroughly documented in Appendix B. The experimental setup not only served to assess the accuracy and precision of the digital twin system in localizing pedestrians but also simulated realistic conditions pertinent to emergency evacuation scenarios to some extent. The rigorous calibration process and strategic camera placement further contributed to the system's reliability.

5.1.2. Precision evaluation of coordinate mapping model

To verify the reliability of both the camera calibration and the coordinate mapping model, six calibration points were selected for each camera within the experimental scene, and their world coordinates were measured. The pixel coordinates of these calibration points were extracted from the surveillance images and then transformed into world coordinates using the mapping model. The discrepancies between the true points and the computed ones were analyzed, with the comparative results illustrated in Fig. 11.

This study utilized two error metrics, named the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE), to assess the precision of the coordinate mapping model. RMSE quantifies the overall deviation between the computed values and the true values, where lower values indicate a closer approximation to the ground truth, whereas higher values reflect larger discrepancies. In contrast, MAE, calculated as the average of the absolute errors, offers a more intuitive understanding of the average magnitude of the errors. It is important to note that in the context of emergency response, these metrics serve as critical safety indicators rather than mere statistical values. High localization precision (indicated by low RMSE and MAE) is essential for ensuring the system's reliability in facilitating precise evacuation and rescue operations, where large positional errors may lead to erroneous decision-making in complex environments. The formulas for RMSE and MAE are provided in Eqs. (19) and (20).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_{true,i} - X_{model,i}\right)^2} \tag{19}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|X_{true,i} - X_{model,i}\right| \tag{20}$$

where N is the total number of samples, $X_{true,i}$ represents the true value for sample i, $X_{model,i}$ represents the calculated value for sample i.

The error metrics calculated for each camera are summarized in Table 2. This level of precision is crucial for ensuring reliable localization within the digital twin system. Notably, Camera 2# shows significantly lower error compared to the other cameras, which suggests that the precision of world coordinate estimation improves when the calibration points are near the camera. Based on these findings, the target matching process is designed to assign a greater weight to the camera that is closer to the target, thereby further enhancing the system's overall localization precision.

Besides, given that the accuracy of foot position detection and tracking plays a critical role in determining the system's overall localization performance, we further evaluated the effectiveness of our proposed method by conducting a comparative analysis against two state-of-the-art baseline approaches widely used in pedestrian localization: the Bottom Midpoint method and the Geometric Center method. The results demonstrate that using ankle midpoints significantly reduces projection errors compared to these baselines, providing a robust foundation for the high-precision localization results reported in Tests 1,

Table 2
Error metrics calculated for each camera.

Camera number	RMSE (cm)	MAE (cm)
1#	1.2	1.1
2#	0.4	0.4
3#	2.1	2.0
4#	3.4	2.8
overall	2.1	1.5

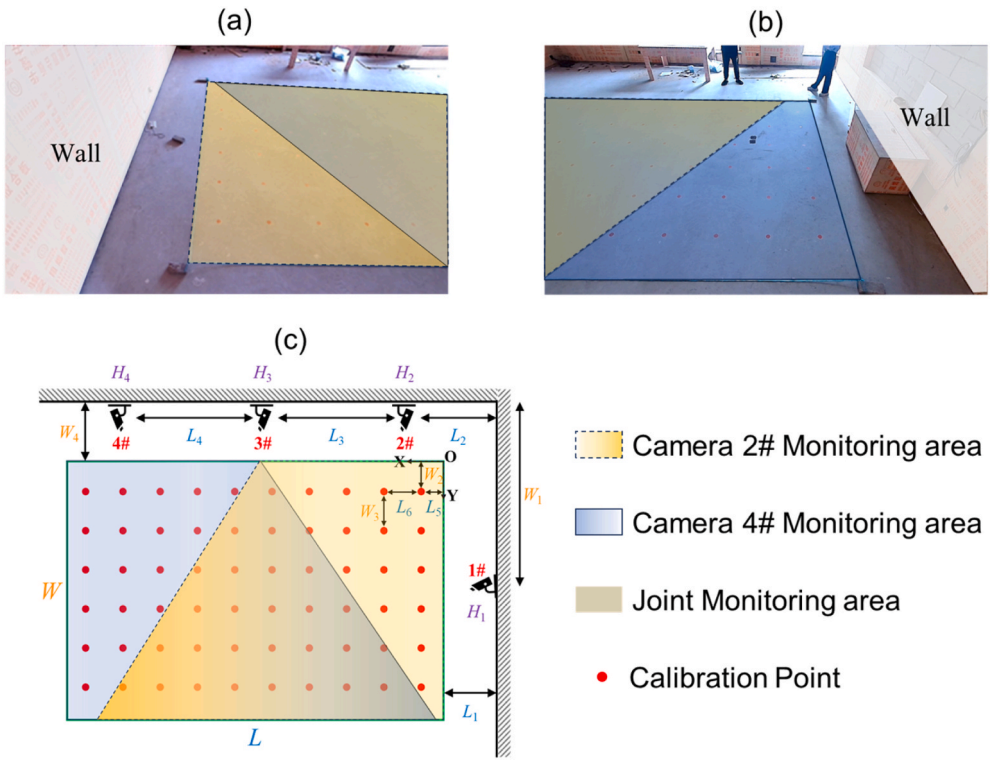

Fig. 12. Coverage regions of Camera 2# and 4#. (a) View of Camera 2# monitoring area after distortion correction, (b)View of Camera 4# monitoring area after distortion correction, (c)Schematic of Camera 2# and 4# monitoring area.

2, and 3. Detailed comparison procedures and evaluation metrics can be found in Appendix C.

5.1.3. Demonstration of the static test

To validate the system's performance in multi-person scenarios, Test 1 was conducted to assess the global monitoring capability for multiple static pedestrians. In this experiment, four participants were positioned within the monitoring areas of Cameras 2# and 4#. Although four cameras were deployed in the experimental environment, this study specifically selected the representative pair of Cameras 2# and 4# as a minimal validation unit. This dual-camera setup is sufficient to rigorously verify the core mechanisms of the proposed multi-camera fusion

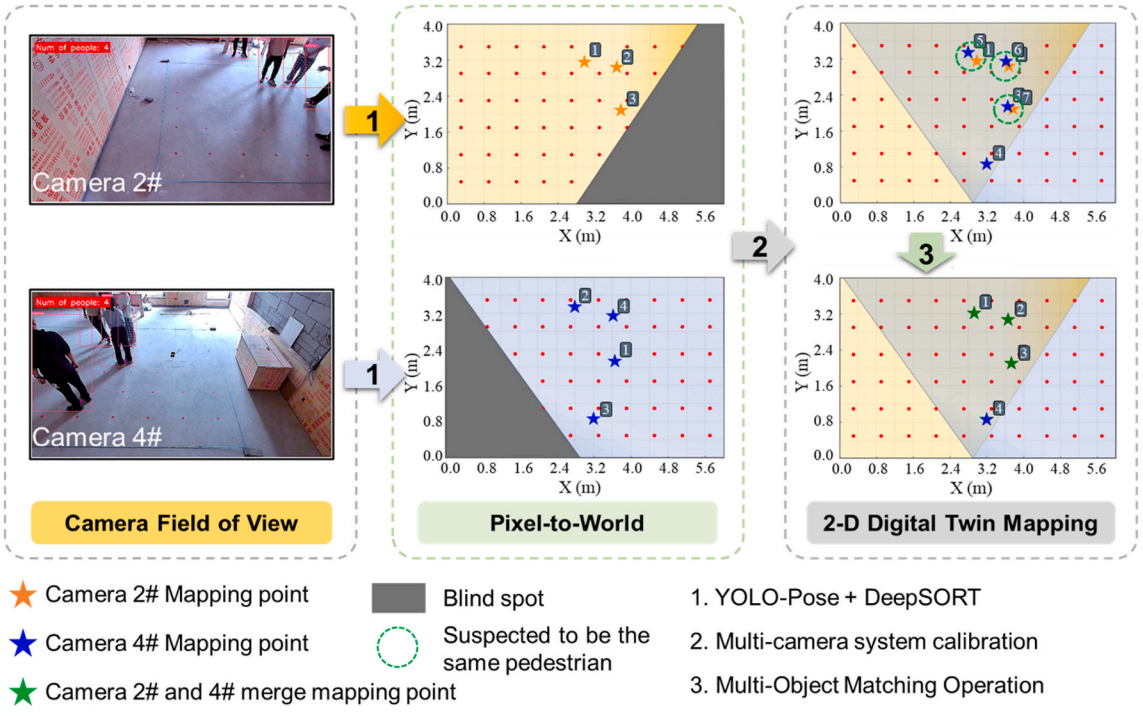

Fig. 13. Demonstration of the operational mechanism of the digital twin system.

algorithm—including global coordinate unification and cross-view identity matching—while providing a clearer demonstration of the fusion logic that is mathematically scalable to larger camera networks. The coverage regions of Cameras 2# and 4# are illustrated in Fig. 12. A demonstration video of Test 1 (see Video S1) shows that the system supports user-defined monitoring zones and camera selection. The video clearly demonstrates that the system can automatically detect, track, and localize multiple pedestrians in real time. Moreover, it provides an aggregated count of individuals in the monitored area along with the corresponding coordinates and unique IDs, which is essential for analyzing multi-dimensional movement information during building evacuation scenarios.

Supplementary video related to this article can be found at https://doi.org/10.1016/j.engappai.2026.114440

Fig. 13 illustrates the operational mechanism of the digital twin system by presenting a snapshot of the live surveillance feed and its corresponding digital twin visualization captured at 4.00 s in the demonstration video. First, using the YOLO-POSE and DeepSORT algorithms, pedestrians are detected and tracked in the original video frames from Cameras 2# and 4#, and the pixel coordinates of their ankle midpoints are extracted. Next, these pixel coordinates of each respective camera are transformed into world coordinates through the coordinate mapping model. Subsequently, the multi-camera system calibration technique is employed to unify the pedestrian coordinates from different cameras into a unified world coordinate system. In this instance, Camera 2# detects and tracks three pedestrian ankle midpoints, while Camera 4# detects four. By employing the proposed MOMO algorithm for the same object matching, data from both cameras are effectively fused, resulting in the digital twin interface accurately displaying the positions of four pedestrians. This outcome validates the robustness of the multi-target matching algorithm and the precision of the multi-camera calibration.

As illustrated in Table 3, a quantitative comparison is presented between the real positions, measured with a tape measure (with a minimum scale of 0.1 cm) once pedestrians were stationary, and the calculated positions from the system. The results show a maximum localization error of 3.6 cm among the four subjects, with an overall RMSE of 2.6 cm and an MAE of 2.3 cm, underscoring the high localization performance achieved in static conditions.

Besides, accurately determining the number of individuals within a monitored area is crucial for formulating effective evacuation strategies and allocating resources efficiently. The system fuses information from multiple camera feeds to determine the total number of people within the monitored zone. To assess the counting accuracy, a metric called **People Counting Accuracy (PCA)** is introduced. PCA is defined as the ratio of the number of frames in which the predicted count matches the true count to the total number of frames observed during the statistical period, with its formula provided in Eqs. (21) and (22). This metric is particularly suitable for evacuation load assessment, as accurate real-time headcounts are essential for detecting overcrowding risks, managing dynamic exit flow, and supporting timely emergency response actions in public spaces.

Table 3
Comparative results of real and calculated positions for Test 1.

ID	Time/s	Real position (X, Y)/cm	Calculated position (X, Y)/cm	Localization error/cm	RMSE/cm	MAE/cm
1	0.17	(290, 320)	(292, 321)	(2, 1) = 2.2	2.6	2.3
1	7.70	(287, 220)	(290, 222)	(3, 2) = 3.6		
2	8.60	(362, 323)	(364, 326)	(2, 3) = 3.6		
2	16.20	(285, 235)	(284, 235)	(-1, 0) = 1.0		
3	18.17	(285, 162)	(286, 163)	(1, 1) = 1.4		
3	24.60	(383, 145)	(380, 147)	(-3, 2) = 3.6		
4	21.07	(373, 290)	(374, 290)	(1, 0) = 1.0		
4	25.27	(315, 350)	(313, 350)	(-2, 0) = 2.0		

$$PCA = \frac{1}{N}\sum_{i=1}^{N}\delta(y_i, \hat{y}_i) \tag{21}$$

$$\delta(y_i, \hat{y}_i) = \begin{cases} 1 & if\ y_i = \hat{y}_i \\ 0 & otherwise \end{cases} \tag{22}$$

where N represents the total number of frames; $\delta(\cdot)$ is the indicator function which equals 1 if the condition is met and 0 otherwise; y_i represents the actual number of people in frame i; and $\hat{y}_i$ represents the predicted number of people in frame i. To ensure the rigorousness of the evaluation, the ground truth y_i for all test scenarios was established through a strict manual annotation protocol. Specifically, the number of pedestrians was manually counted frame-by-frame by a primary researcher. To ensure data reliability and high inter-annotator agreement, the counts were subsequently cross-verified by a second independent annotator. This dual-verification process eliminates potential counting errors, providing a precise and reliable baseline for calculating the frame-level accuracy.

It is important to note that, unlike fundamental computer vision research that prioritizes trajectory consistency using metrics like Multiple Object Tracking Accuracy (MOTA) or ID F1 Score (IDF1), this study focuses on the engineering application of evacuation load assessment. In emergency scenarios, the precise physical location of individuals (reflected by RMSE/MAE) and the macroscopic distribution of the crowd (reflected by PCA) are more critical for risk assessment and decision-making than long-term identity maintenance. Since this system employs the mature DeepSORT algorithm for temporal tracking, the "identity association" in this work specifically emphasizes the capability of the MOMO algorithm to successfully match and fuse targets across spatially overlapping camera views to prevent duplicate counting, rather than optimizing the ID switch rate of single-view tracking.

The temporal evolution of the predicted and actual people counts can be seen in Video S1. In the Test 1 scenario, the system achieves a PCA of 96.67%, with an RMSE and MAE of 0.034 persons, respectively. Discrepancies in the people counting results are primarily due to occasional missed or false detections when pedestrians moved to the next stand positions, as shown in Fig. 14. For example, at 5.93 s, one same pedestrian is obscured in both camera views, resulting in a missed detection (see Fig. 14a); at 14.37 s, an error in detecting the ankle midpoint in Camera 2# leads to a failure in matching the same pedestrian across the cameras, causing a false detection (see Fig. 14b); and at 25.63 s, the pedestrian's ankle midpoint is not detected in Camera 2# while being obscured in Camera 4#, leading again to a missed detection (see Fig. 14c). In cases where the pedestrian count is accurate, it can be observed that the pedestrian is detected by at least one camera with a correct identification of the ankle midpoint, thereby preventing under-reporting. To prevent false reporting, it is necessary for the same pedestrian's accurate ankle midpoint to be simultaneously detected by multiple cameras and subsequently merged through the MOMO algorithm.

To further dissect the mechanism of the proposed framework and compare it against standard geometric baselines, an ablation study was conducted on the Test 1 scenario to evaluate the specific contributions of the constraint modules and the fusion module in the MOMO algorithm. The "Distance Only" and "Epipolar Only" configurations serve as baseline methods representing traditional geometry-based matching strategies. The results, detailed in Table 4, highlight a distinct functional separation. The "Distance Constraint" and "Epipolar Constraint" primarily serve as association filters. Using either constraint individually results in suboptimal People Counting Accuracy (PCA of 85.33% and 78.67%, respectively) due to the inability to fully exclude false positives. However, their combination creates a robust filter that eliminates mismatches, significantly elevating the PCA to 96.67%.

On the other hand, the "Weighted Fusion" strategy is specifically designed to enhance localization precision rather than association logic. As

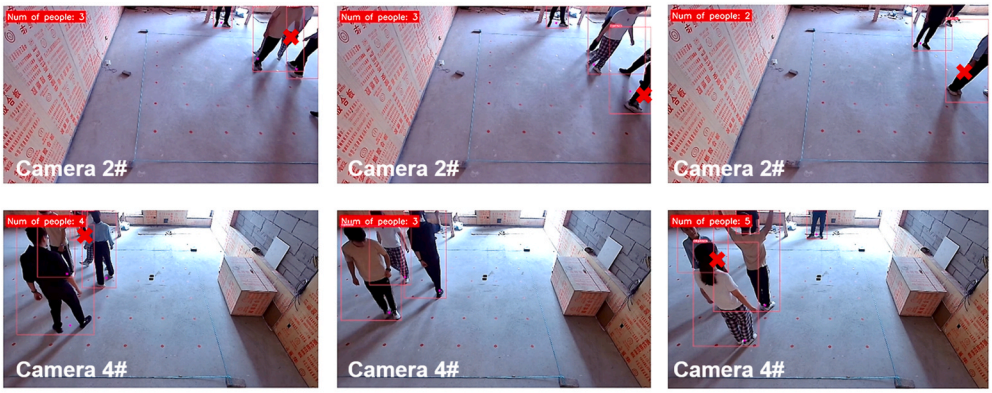

Fig. 14. Causes of underreporting and false reporting in the predicted number of people for Test 1. (a) Same Pedestrian Obscured in Both Views, (b) Error in Ankle Midpoint Detection in Camera 2#, (c) Ankle Midpoint Missing in Camera 2# and Occluded in Camera 4#.

Table 4
Ablation study of MOMO algorithm components in Test 1.

Method	Components	PCA	RMSE (cm)	MAE (cm)
Distance Only	Distance Constraint	85.33%	/	/
Epipolar Only	Epipolar Constraint	78.67%	/	/
Dist. + Epi.	Distance + Epipolar	96.67%	3.8	3.4
MOMO	**Dist. + Epi. + Fusion**	**96.67%**	**2.6**	**2.3**

shown in the comparison between "Dist. + Epi." (utilizing arithmetic mean fusion) and "MOMO" (utilizing adaptive weighted fusion), the introduction of the weighting mechanism does not alter the PCA but substantially improves the spatial metrics. It reduces the RMSE from 3.8 cm to 2.6 cm and the MAE from 3.4 cm to 2.3 cm. This quantitative evidence confirms that while geometric constraints ensure the correct identity ("Who is who"), the adaptive fusion is indispensable for the high-precision localization ("Where is exactly") required for digital twin fidelity.

5.1.4. *Demonstration of the dynamic test*

Subsequently, the system's performance was evaluated under dynamic conditions in Test 2, where multiple pedestrians were instructed to walk randomly within the experimental area without following any predefined routes. This setup was utilized to simulate the unpredictable nature of crowd movement and interaction, captured by Cameras 2# and 4# (see Video S2). Compared to static scenarios where pedestrians remain stationary, dynamic movement conditions present significantly greater challenges to the localization performance of the system due to motion blur, occlusion, and rapid position changes.

Supplementary video related to this article can be found at https://doi.org/10.1016/j.engappai.2026.114440

Fig. 15a shows a snapshot from the video at 22.37s, Camera 2# detects three pedestrians via ankle midpoint localization while Camera 4# captures four targets. Through the novel proposed MOMO algorithm, the system successfully matches and fuses these observations, accurately displaying all four pedestrians in the digital twin interface, validating the reliability of multi-target localization in dynamic environments. A critical demonstration of the system's robustness occurs at 15.50s, as shown in Fig. 15b, where Camera 2# identifies pedestrian ID 2's ankle midpoint, while Camera 4# fails due to occlusion. By integrating multi-camera data, the system reconstructed the global information, effectively expanding the monitoring Field of View and mitigating single-view limitations. This integration of local surveillance data broadens the monitoring field, ensuring comprehensive situational awareness,

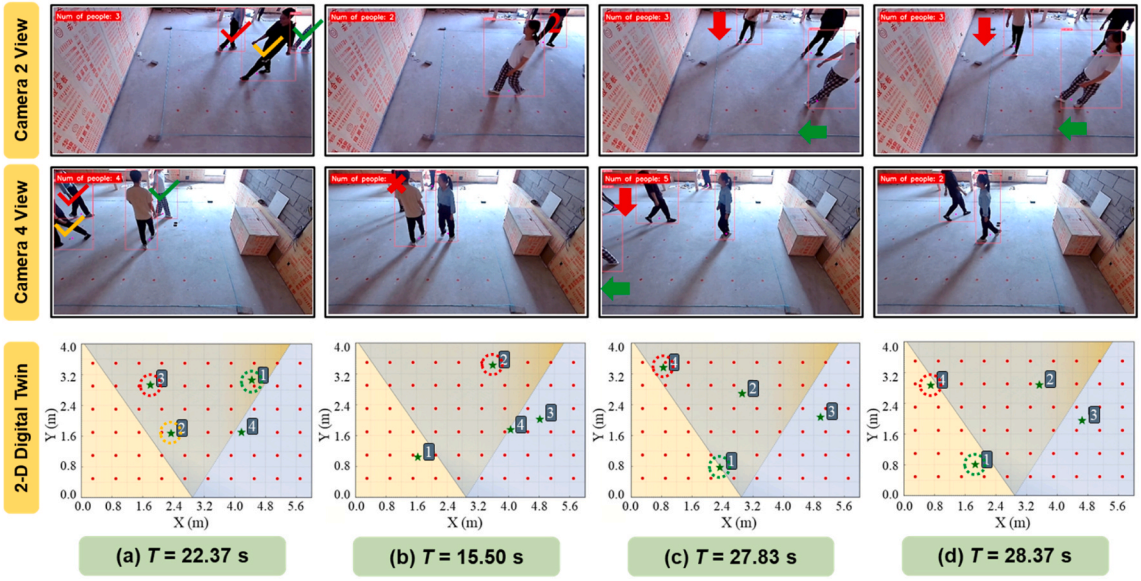

Fig. 15. Demonstration of the Digital Twin system for Test 2. (a) Multi-Target Matching and Fusion via MOMO Algorithm in Dynamic Environments, (b) Occlusion Recovery and Field-of-View Expansion via Multi-Camera Fusion, (c) and (d) Seamless Cross-Camera Trajectory Stitching and ID Retention.

which is a critical factor in effective evacuation management. Further evidence of the system's advanced capabilities is illustrated in Fig. 15c and d, which showcases smooth cross-camera handover of pedestrians ID 1 and 4 as they transition between camera coverage zones. The arrows indicate pedestrian movement directions, demonstrating the system's ability to maintain continuous trajectory tracking through robust multi-target matching and multi-camera calibration methods. During cross-camera transitions, the system accurately associates the same pedestrian's identity across junction of the monitoring area by leveraging spatiotemporal consistency and ankle midpoint trajectory features. This includes automatic trajectory stitching and identity retention even under partial occlusion or brief visual loss. Such seamless transition capability is particularly critical for evacuation scenarios where uninterrupted monitoring of pedestrian flow patterns is essential for safety management.

Table 5 presents a comparison between the actual world coordinates, which were recorded when pedestrians passed certain calibration points, and the calculated positions. The maximum localization error among the four pedestrians is 8.5 cm, with an RMSE of 4.8 cm and an MAE of 4.4 cm, demonstrating that the system maintains good performance even under dynamic conditions. Furthermore, in Video S2, the performance metrics of Test 2 reveal a PCA of 93.44%, with RMSE and MAE at 0.07 and 0.067 persons, respectively. The slight decrease in counting accuracy observed during dynamic test, as compared to the static test, is primarily attributed to two factors: the increased frequency of occlusion caused by pedestrian intersecting walking paths, which raises the likelihood of missed detections, and the greater motion amplitude of pedestrians, which can lead to inaccuracies in ankle midpoint localization. Overall, the static and dynamic test results demonstrate that the intelligent digital twin system exhibits robust performance in monitoring and analyzing multi-person scenarios within a public space environment. The system's high-precision localization, reliable multi-camera fusion, and comprehensive digital twin visualization not only support detailed analysis of pedestrian movement but also play a critical role in enhancing global situational awareness during emergency evacuations and public safety management.

5.2. Demonstration of real scenario test

5.2.1. Real scenario introduction

Beyond testing the system under experimental conditions, assessing its performance in real-world environments is of paramount importance. Real-world environments are inherently more variable and complex, thereby presenting a greater challenge to the digital twin system's reliability and robustness.

In this section, the system was deployed to monitor pedestrian movement within a designated area in the waiting hall of a high-speed rail station. The complex environment of the high-speed rail station, characterized by variable lighting, dynamic crowd movements, and intricate architectural features, served as a rigorous testbed for the digital twin system. The selected monitoring zone measured 6.8 m by 3.2 m, and two cameras, labeled C1 and C2, were installed to cover this area. These cameras are the same models described in Section 5.1, and their calibration results are provided in Appendix D. Fig. 16 displays the distortion-corrected images from these cameras, along with a schematic diagram of the monitored area. To calibrate the camera and verify the localization precision, the intersections between floor tiles (each floor tile measuring 0.4 m by 0.4 m) were utilized as calibration points.

Table 5
Comparative results of real and calculated positions for Test 2.

ID	Time/s	Real position/cm	Calculated position/cm	Localization error/cm	RMSE/cm	MAE/cm
1	0.70	(150, 170)	(153, 162)	(3, −8) = 8.5	4.8	4.4
1	7.13	(210, 170)	(210, 172)	(0, 2) = 2.0		
2	12.37	(90, 110)	(88, 113)	(-2, 3) = 3.6		
2	13.43	(150, 170)	(155, 169)	(5, −1) = 5.1		
3	16.13	(450, 170)	(455, 169)	(5, −1) = 5.1		
3	21.90	(150, 290)	(149, 293)	(-1, 3) = 3.2		
4	23.77	(450, 290)	(446, 290)	(-4, 0) = 4.0		
4	27.70	(90, 350)	(94, 349)	(4, −1) = 4.1		

5.2.2. Real scenario test results

The system demonstration video (see Video S3) clearly illustrates how the digital twin interface dynamically visualizes pedestrian positions and population counts within the real-world monitored area, providing critical support for multi-camera surveillance in large public infrastructure scenarios. Fig. 17a captures a key moment at 1.50s, showing both the real-time camera feed and corresponding digital twin representation. Here, two seated pedestrians are simultaneously detected in the overlapping coverage area of two cameras. Through the proposed MOMO algorithm, the system successfully identifies and displays both individuals with precise localization in the digital twin interface. Notably, the system also accurately locates small pedestrian groups, which is a common but challenging scenario in evacuations where people move in a group and are in close proximity. This shows the system's capability to reliably output location information for clustered individuals.

Supplementary video related to this article can be found at https://doi.org/10.1016/j.engappai.2026.114440

Subsequent frames further illustrate the system's robustness in a realistic setting. For example, in Fig. 17b and c, multiple cameras complement each other's fields of view, so that even when pedestrian ID 4 is occluded in one camera's view, another camera can capture and display that individual's information on the digital twin interface. Moreover, these frames validate the system's ability to perform continuous cross-camera localization, as evidenced by the seamless handover of tracking for pedestrians with IDs 5 and 6 as they move from the monitoring region of Camera C2 to that of Camera C1. This capability underscores the system's strong global surveillance potential. In Fig. 17d, the system demonstrates strong perception capabilities even in complex human-object interaction scenarios, effectively handling occlusions caused not only by other pedestrians but also by items such as suitcases. This ensures comprehensive situational awareness under challenging and dynamic conditions.

In real-world evacuation scenarios, pedestrian postures can vary significantly, encompassing individuals seated in wheelchairs, those moving in an upright position, and others who are standing still. The video captures these diverse postures, and a detailed analysis of the positional data confirms the system's high localization precision across varied postures. Table 6 illustrates that the maximum localization error observed is 8.3 cm, with an RMSE of 5.3 cm and an MAE of 4.8 cm, which is well within the practical requirement of maintaining localization errors below 30 cm (Liu et al., 2019; Zhu et al., 2020). Additionally, the temporal evolution of the predicted versus actual headcounts in the monitoring area can be seen in Video S3. In this real-world public infrastructure scenario, the system achieves a PCA of 92.34%. Such a high frame-level matching accuracy, along with low RMSE (0.0783 persons) and MAE (0.0772 persons), indicates the system's capability to deliver precise real-time occupancy data. These results are of particular significance for emergency evacuation management, as accurate and timely estimation of crowd density and movement is essential for dynamic risk assessment and load-based decision-making.

5.2.3. Real-time performance evaluation of the digital twin system

Real-time performance is a critical metric in the design of digital twin systems. To evaluate this aspect, a test was conducted using a video with a resolution of 2304 × 1296, consisting of 2000 frames at a frame rate of 30 frames per second (FPS). The experimental platform is equipped with Windows 10 (64-bit), an Intel Core i7-10875H processor, an NVIDIA GeForce RTX 2060 Graphics Processing Unit (GPU), and 16 GB of

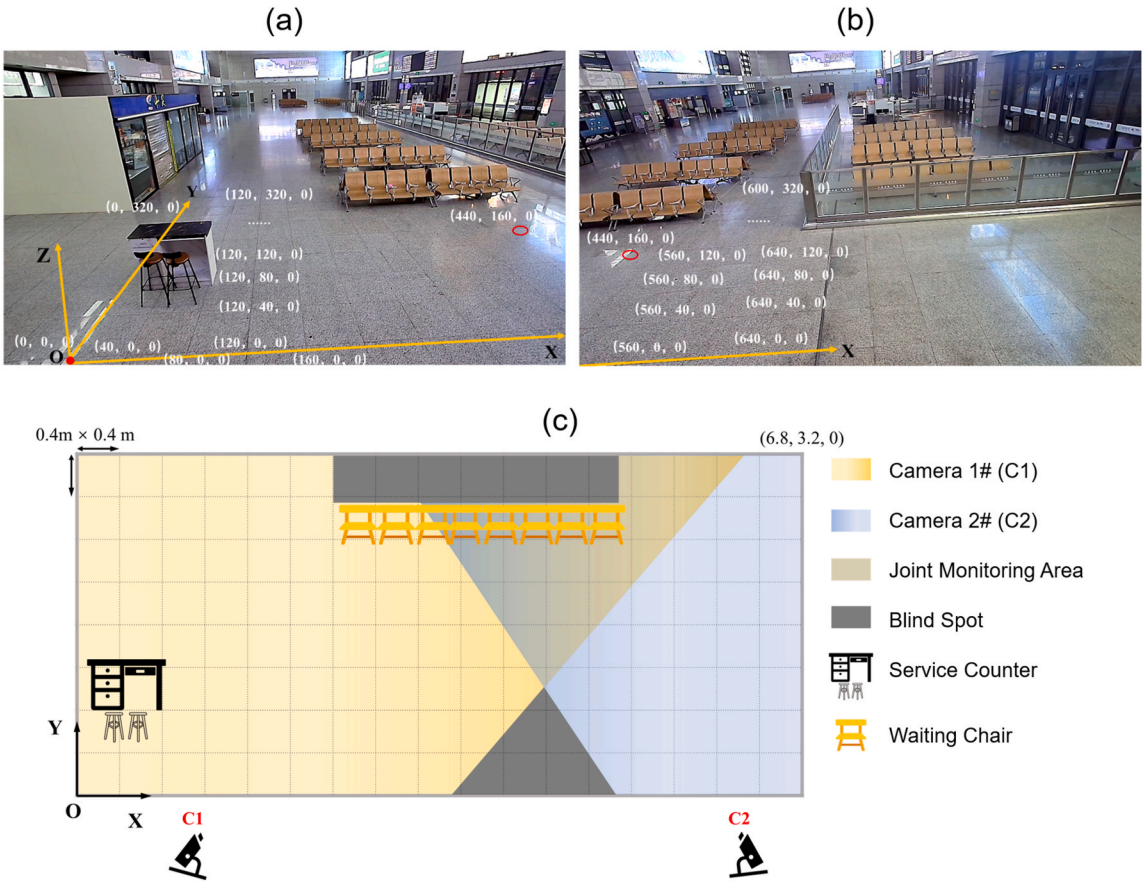

Fig. 16. Coverage regions of Camera C1 and C2. (a) View of Camera C1 monitoring area after distortion correction, (b)View of Camera C2 monitoring area after distortion correction, (c)Schematic of Camera C1 and C2 monitoring area.

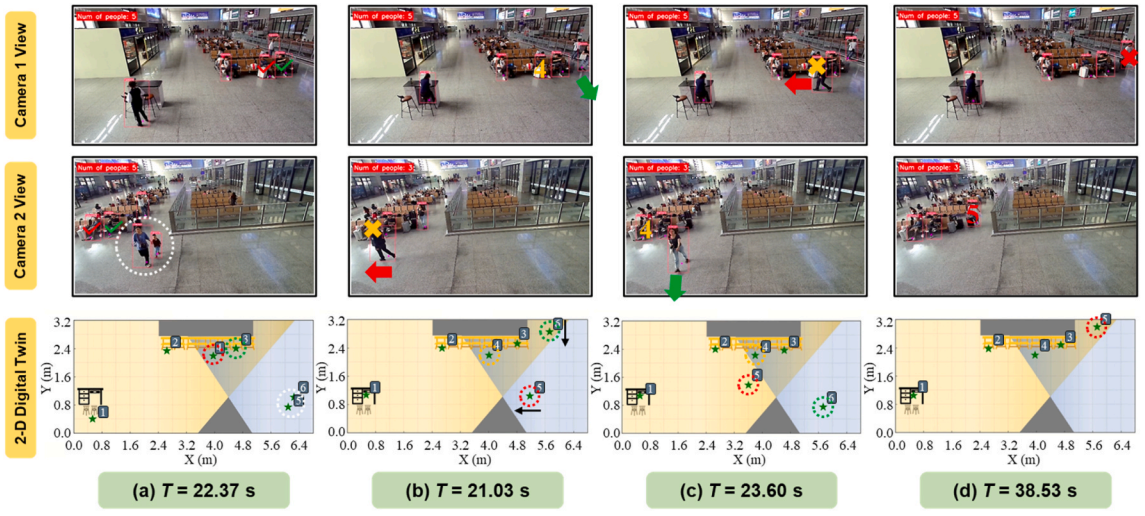

Fig. 17. Demonstration of the Digital Twin system for Test 3. (a) Precise Multi-Target Matching and Small Group Identification, (b) and (c) Occlusion Handling and Seamless Cross-Camera Tracking, (d) Robust perception in complex human-object interaction scenarios with occlusions.

Random Access Memory (RAM). Under these conditions, the system achieves an average processing time of 27.8 ms per frame for the core AI pipeline, with a total runtime of 55.6 s for the entire video. To provide a rigorous end-to-end latency analysis, this 27.8 ms latency consists of approximately 16.0 ms for pedestrian detection (YOLOv8-Pose), 9.0 ms for pedestrian tracking (DeepSORT), and 2.8 ms for the calibration and MOMO Association module. This computational efficiency presents a significant advantage over traditional appearance-based Re-ID methods in emergency scenarios. As noted in the comprehensive survey (Ye et al., 2022), state-of-the-art Re-ID systems typically rely on deep convolutional networks (e.g., Residual Neural Network (ResNet) or Vision Transformer (ViT) backbones) to extract discriminative feature embeddings. This process is computationally intensive, often requiring inference times exceeding 100 ms per frame on standard hardware,

Table 6
Comparative results of real and calculated positions for Test 3.

ID	Time/s	Posture	Real position/cm	Calculated position/cm	Localization error/cm	RMSE/cm	MAE/cm
1	0.70	Stand	(80, 40)	(76, 46)	(-4, 6) = 7.2	5.3	4.8
2	17.37	Sit	(270, 250)	(270, 245)	(0, −5) = 5.0		
3	20.50	Sit	(460, 240)	(463, 243)	(3, 3) = 4.2		
4	34.80	Sit	(400, 230)	(398, 222)	(-2, −8) = 8.3		
5	8.07	Squat	(320, 80)	(319, 80)	(-1, 0) = 1.0		
5	19.27	Walk	(640, 80)	(640, 85)	(0, 5) = 5.0		
6	22.00	Walk	(560, 200)	(557, 201)	(-3, 1) = 3.2		

which introduces latency bottlenecks for real-time tracking. In contrast, the proposed MOMO algorithm leverages lightweight geometric constraints, completing the association and fusion process in about 2 ms. This ensures that the system maintains high-frame-rate monitoring capabilities essential for capturing rapid crowd movements during evacuations. Given that maintaining a processing time below 33 ms per frame is essential to prevent lag in a 30 FPS environment, the measured performance satisfies the real-time requirements.

Regarding system-level integration, the database Input/Output (I/O) and user interface visualization are implemented using an asynchronous multi-threading architecture. This ensures that these peripheral tasks execute in parallel without blocking the AI inference loop. Furthermore, while the current test on a single RTX 2060 validates the algorithm's efficiency, the system architecture supports horizontal scalability; for large-scale multi-camera deployments, utilizing server-grade hardware (e.g., multi-GPU clusters) would allow simultaneous processing of multiple high-resolution streams while maintaining the validated 30 FPS real-time standard. This efficiency ensures that the digital twin system can respond instantly and provide accurate positional feedback, thereby enhancing its overall operational reliability in real-time evacuation applications.

6. Discussion

6.1. Application prospects

The digital twin system in this study holds significant promise for applications in facility evacuation and crowd management. By integrating multi-camera data and employing robust calibration techniques, the system can provide real-time visualization and tracking of individual positions and headcounts in complex indoor environments. In real-world scenarios such as a waiting hall at high-speed rail station, the system has proven to seamlessly merge data from overlapping camera views, accurately track pedestrian movement even across camera boundaries, and reliably identify small groups and individual poses. These capabilities are crucial for generating a global comprehensive situational awareness during emergency evacuations, where understanding crowd dynamics in real-time can greatly enhance decision-making for resource allocation and safe egress. Unlike traditional surveillance systems that rely on discrete, single-view feeds, the proposed framework realizes a “Global Digital Twin” capability. The core value of a digital twin lies in its holistic perception and continuous mirroring of physical entities. By fusing multi-camera data into a unified world coordinate system, our approach eliminates visual blind spots and data silos, ensuring that the virtual model maintains spatial continuity and completeness. This globally consistent digital mapping is a prerequisite for advanced digital twin functions, such as high-fidelity evacuation simulation and future state prediction, distinguishing it from passive monitoring technologies.

The application prospects of this digital twin system are extensive. In the context of indoor evacuation, the system accurately captures real-time pedestrian locations, providing strong support for evacuation path planning to ensure the most efficient and congestion-free escape routes. Furthermore, its precise localization capabilities facilitate targeted emergency response and rescue operations by guiding first responders to individuals in distress, ensuring a rapid and coordinated intervention (Long et al., 2026). For crowd management, it enables the generation of dynamic density heatmaps, allowing for continuous monitoring of crowd distribution and movement patterns to prevent overcrowding and enhance situational awareness. Seamlessly integrating with emergency management and disaster response systems, this digital twin framework enhances the overall efficiency and effectiveness of evacuation strategies and crisis management.

6.2. Limitations

Despite the system's promising application potential, several practical considerations must be addressed for reliable deployment in real-world scenarios. First, as indicated by the missed and false detections observed in the preceding PCA analysis, missed detections often occur when pedestrians happen to be located in blind spots not covered by any camera, or are captured by only a single camera in which ankle midpoint detection fails, resulting in undercounting. This issue is particularly pronounced in scenarios with high pedestrian flow density, where mutual occlusion significantly reduces the visibility of lower-body keypoints. Crucially, it is important to acknowledge that the validation in this study—conducted in limited monitored zones (4 × 6 m in the lab and 6.8 × 3.2 m in the station)—serves primarily as a “proof-of-concept” for the system architecture. While the proposed framework is theoretically scalable to large public spaces, its reliability in facility-scale deployments with genuinely high-density flows has not yet been fully stress-tested. Specifically, in extreme crowding where interpersonal distances consistently fall below the 0.45 m social distance threshold assumed by the MOMO algorithm, the matching accuracy may degrade. On the other hand, false positives may arise when incorrect ankle midpoint localization in a single camera view causes the same pedestrian to be counted multiple times due to failure in merging by the MOMO algorithm. To address these challenges, future work will go beyond simple node detection enhancement by introducing temporal logic and trajectory prediction algorithms. By leveraging historical movement data to infer obscured ankle positions, the system can maintain tracking continuity even in dense crowds. Additionally, investigating the quantitative relationship between crowd density levels and detection accuracy will be a critical direction to develop density-adaptive algorithms. Second, the computational complexity associated with real-time data processing and cross-camera matching necessitates further optimization to ensure scalability, particularly in densely populated scenarios where the volume of data can be substantial. Third, environmental factors characteristic of emergency scenarios, such as the accumulation of smoke or reduced lighting due to power failures, may impact the visibility of standard optical cameras. While the system's multi-camera fusion strategy offers a degree of robustness by leveraging redundant views to overcome localized visual obstructions, relying solely on the visible spectrum remains a limitation in extremely low-visibility conditions. The extensibility of the proposed digital twin framework allows for the future integration of multi-modal sensors; therefore, incorporating infrared thermal imaging devices alongside optical cameras will be a key focus of future work to ensure reliable pedestrian detection and tracking in smoke-filled or dark environments. Fourth, the current system relies on a flat ground plane assumption (z_w

= 0) and does not yet account for complex terrains such as stairs, ramps, or uneven floors. Furthermore, while the proposed cascade calibration utilizes geometric anchors to constrain drift, a rigorous systematic uncertainty propagation analysis has not yet been conducted. Future work will focus on integrating 3D terrain modeling to handle multi-level environments and conducting a comprehensive sensitivity analysis to quantify how detection noise and calibration errors propagate through the digital twin system.

In summary, the digital twin system developed in this work represents a significant advancement in the domain of building evacuation and crowd management, offering detailed, real-time insights that can inform emergency response and facility design. Its ability to offer real-time, high-precision monitoring and data fusion across multiple camera views has the potential to significantly improve evacuation route planning, density analysis, and targeted rescue operations. Continued research is needed to address the challenges related to occlusion, environmental variability, ankle keypoints detection, computational efficiency, low-visibility conditions, and complex terrains, which will be critical for achieving widespread adoption in large-scale, real-world applications.

7. Conclusions

This study addresses the pressing need for real-time pedestrian localization in facility-scale emergency evacuations by proposing an intelligent digital twin system that integrates advanced visual localization technologies. To overcome the limitations of existing disaster safety digital twin frameworks—particularly their deficiencies in global, multi-perspective monitoring—we developed a system architecture comprising four core components: IoT sensors, a cloud server, an AI engine, and a user interface. This architecture enables continuous data acquisition, real-time processing, and intuitive visualization to support dynamic evacuation management.

At the core of the system lies the AI engine, which integrates three key technical innovations to enable robust pedestrian monitoring in complex environments. First, pedestrian detection and tracking are performed using YOLO-Pose and DeepSORT, facilitating the automated extraction of ankle midpoints in pixel coordinates to improve localization accuracy. Second, a novel multi-camera calibration approach is developed to transform and unify local camera coordinate systems into a unified world coordinate framework, ensuring consistent spatial alignment across multiple views. Third, a MOMO algorithm is proposed to address the challenge of identity association across overlapping camera fields. By bypassing traditional image stitching and cross-camera re-identification techniques, MOMO delivers a computationally efficient and scalable solution highly tailored to the dynamic conditions of emergency evacuation scenarios.

The system's performance is validated through a series of controlled laboratory tests and a real-world deployment in a high-speed rail station waiting hall—representing a large-scale and complex public infrastructure setting. Results demonstrate accurate pedestrian tracking, effective multi-camera fusion, and reliable identity matching. In the real-world scenario, the system achieved a localization RMSE of 5.3 cm and an MAE of 4.8 cm, well within the practical threshold of 30 cm. Furthermore, the People Counting Accuracy (PCA) reached 92.34%, exceeding industry benchmarks, while the average core AI processing latency of 27.8 ms per frame ensures real-time responsiveness at 30 FPS.

These findings substantiate the feasibility and effectiveness of the proposed system for real-time monitoring and dynamic evacuation assessment within built environments. The approach holds significant potential for seamless integration into intelligent emergency response and crowd management strategies across large-scale infrastructure facilities.

CRediT authorship contribution statement

Huakai Sun: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Yifei Ding:** Writing – review & editing, Investigation. **Ruiwen Fan:** Data curation. **Yuxin Zhang:** Resources. **Tianhang Zhang:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Xinyan Huang:** Resources, Funding acquisition. **Ke Wu:** Writing – review & editing, Resources, Funding acquisition, Formal analysis.

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

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LMS26E080005, the National Natural Science Foundation of China (52478422), MTR Funding Scheme (PTU-23005), and HK Research Grants Council Theme-based Research Scheme (T22-505/19-N). The authors would like to thank Jiangdong Li and Xinting Zheng (ZJU) for their great help in the lab-scale experiment. TZ thanks the support from PolyU Joint Postdoc Scheme with Non-local Institutions.

Appendix A

Current literature offers several approaches for cross-camera pedestrian matching paradigms. To identify the most suitable method for our real-time digital twin system, we evaluated the following three primary techniques.

Table A
Advantages and disadvantages of the three methods.

Method	Advantages	Disadvantages
Image Stitching	1. **Expands the field of view**. 2. Provides a **unified image** for analysis.	1.**Complex** stitching algorithms relying on camera calibration and **overlapping fields of view**. 2.Susceptible to **errors in dynamic scenes** and **distortion at stitching boundaries**. 3. High computational cost and **poor real-time performance**.
Re-ID	1. Does not rely on overlapping fields of view or camera calibration, **suitable for distributed monitoring scenarios**. 2.**Mature deep learning models** support large-scale environments.	1.Relies on **appearance-based features**, prone to **occlusion**, **low resolution**, and **lighting variations**. 2.High algorithmic complexity, **unsuitable for scenarios with strict real-time requirements**.

(*continued on next page*)

Table A (*continued*)

Method	Advantages	Disadvantages
Location-based pedestrian matching	1.Simple implementation and computationally efficient, **ideal for real-time scenarios**. 2.Avoids reliance on external appearance features, making it **robust to changes** in pedestrian appearance and lighting.	1.Relies on **accurate camera calibration and spatial mapping**, with potential errors affecting matching performance. 2.Primarily suitable for surveillance systems with **overlapping camera fields of view**.

Image stitching method involves combining images from multiple cameras into a single panoramic view, allowing pedestrians to be identified in a unified frame. Re-ID method leverages deep learning models to extract discriminative features from a pedestrian's appearance and match these features across camera views. Location-based pedestrian matching method uses spatial relationships between camera views to match pedestrian positions. By mapping coordinates from one camera's frame to another or a unified world coordinate system, pedestrians can be identified based on their positions. The advantages and limitations of these methods are summarized in Table A. Among these, location-based pedestrian matching is more suitable for indoor scenarios where cameras have overlapping fields of view and real-time performance is crucial. Its computational simplicity and robustness to changes in pedestrian appearance and environmental lighting make it a practical choice. Besides, the previously discussed multi-camera coordinate system calibration method in Section 4.3 provides a solid foundation for this approach. Therefore, the Location-based pedestrian matching method is chosen in this study.

Appendix B

This section presents the camera calibration results obtained from experimental scenarios using Zhang's calibration method. The calibration procedure was implemented through the Camera Calibration Toolbox in MATLAB software, which provides a systematic framework for determining intrinsic camera parameters with enhanced operational efficiency (Fetić et al., 2012). The camera used in the experiment is depicted in Fig B1, and its key parameter specifications are summarized in Table B1.

Fig. B1. HIKVISION D1+ surveillance camera physical picture.

Table B1
Key parameter specifications of camera.

Parameter category	Parameter information
Resolution	2304*1296
Frame rate	30FPS
Maximum lens angle	123.4°(D), 105°(D), 56.4°(D)
Aperture	F1.8
Size	77*32*35 mm

During camera calibration, a 9 × 12 chessboard pattern was used, which corresponds to an array of 8 × 11 internal corners with each square measuring 20 mm, as illustrated in Fig B2.

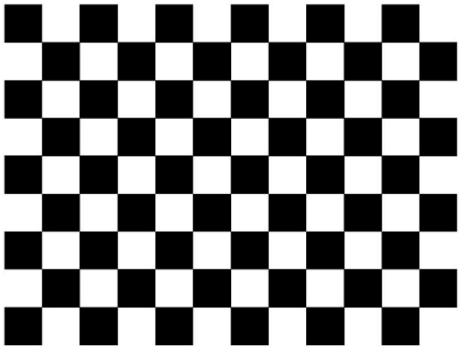

Fig. B2. 9 × 12 chessboard pattern.

Following established photogrammetric protocols, we captured 16 high-resolution images of the planar target at systematically varied orientations—a quantity within the recommended 15-20 image range for robust parameter estimation (Zhang, 2000), as shown in Fig B3. This multi-view acquisition strategy ensures sufficient geometric constraints for accurate intrinsic parameters and distortion coefficients determination.

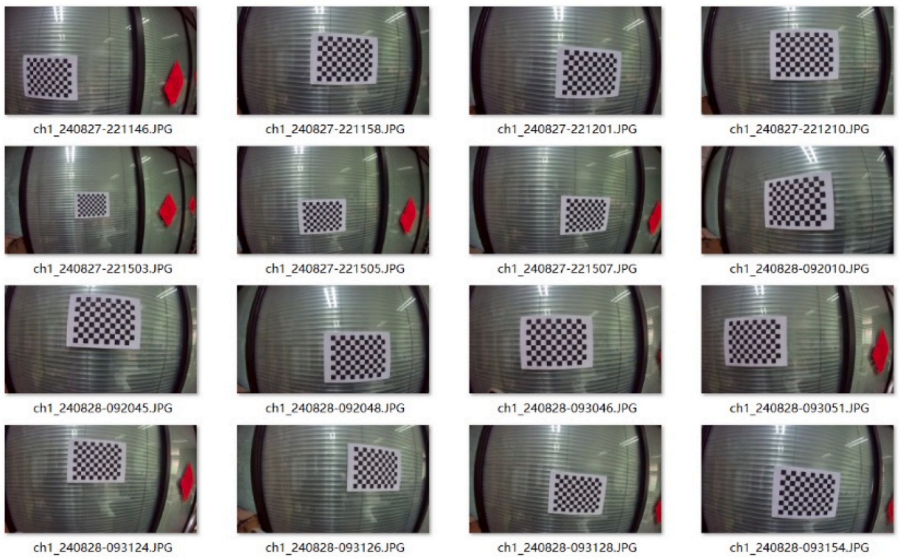

Fig. B3. Calibration images taken at different angles and distances.

The mean reprojection error is widely recognized as a key metric for evaluating the accuracy of camera calibration. A lower mean reprojection error signifies higher calibration precision, with values typically expected to be below 0.5. In this study, the calculated mean reprojection error is 0.18, as shown in Fig B4, which indicates that the calibration process achieved high accuracy.

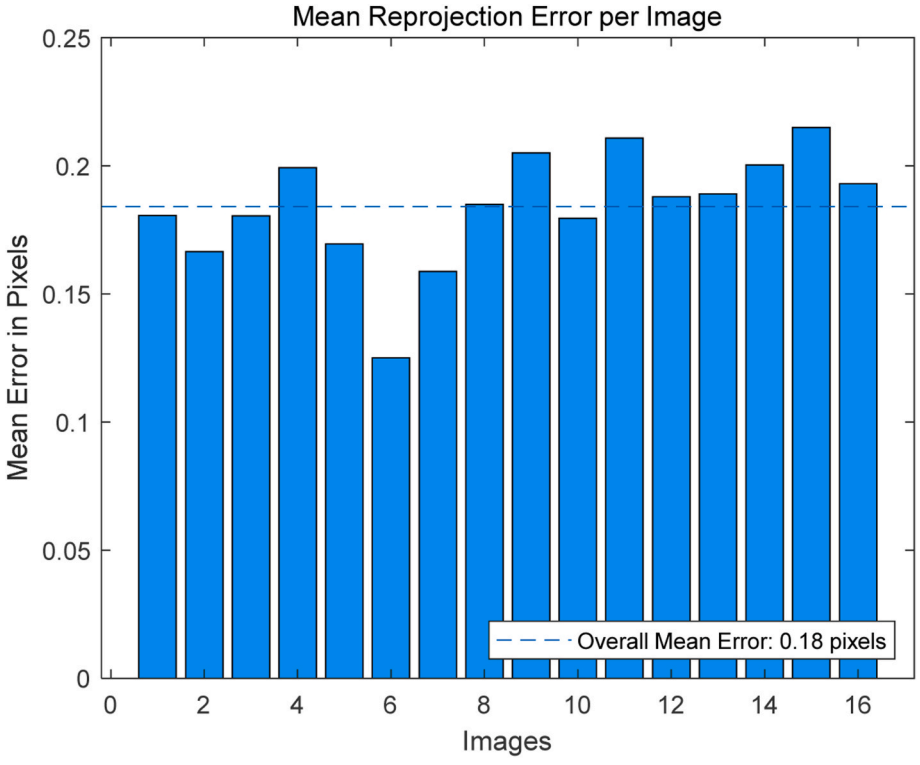

Fig. B4. Camera calibration mean reprojection error.

After obtaining the intrinsic parameters and distortion coefficients of the camera, the PnP algorithm was applied based on the correspondence between 3D world coordinates and their 2D image projections (Zhuang et al., 2023). The calibration points, visible on the ground in Fig. 10, were used as reference markers to solve for the extrinsic parameters. By combining these 2D-3D correspondences with the previously determined intrinsic parameters, a geometric constraint equation in Eq. (1) was used to accurately compute the extrinsic parameters. The camera's intrinsic parameters and distortion coefficients are detailed in Table B2, and the corresponding extrinsic parameters are summarized in Table B3.

Table B2
Intrinsic parameters and distortion coefficients of the camera.

Camera number	Parameter		Results
1#-4#	Intrinsic matrix **K**	$\begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 1225.7954 & 0 & 1153.8328 & 0 \\ 0 & 1228.4795 & 633.9847 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
	Distortion coefficients	$[k_1, k_2, p_1, p_2, k_3]$	$[-0.4130, 0.2335, 0.0020, -0.0002, -0.0759]$

Table B3
Extrinsic parameters in the experimental scenario.

Camera number	Rotation matrix **R**	Translation vector **T**
1#	$\begin{bmatrix} -0.0245 & -0.9953 & 0.0936 \\ -0.7277 & -0.0465 & -0.6843 \\ 0.6854 & -0.0849 & -0.7232 \end{bmatrix}$	$[182.2908 \quad 110.1149 \quad 255.0368]$
2#	$\begin{bmatrix} 0.9984 & -0.0103 & 0.0560 \\ 0.0315 & -0.7196 & -0.6937 \\ 0.0475 & 0.6943 & -0.7181 \end{bmatrix}$	$[-39.1665 \quad 98.4499 \quad 242.2851]$
3#	$\begin{bmatrix} 0.9987 & -0.0251 & 0.0450 \\ 0.0153 & -0.6894 & -0.7242 \\ 0.0492 & 0.7239 & -0.6881 \end{bmatrix}$	$[-271.9898 \quad 103.8886 \quad 224.9244]$
4#	$\begin{bmatrix} 0.9999 & 0.0056 & 0.0104 \\ 0.0116 & -0.6472 & -0.7622 \\ 0.0024 & 0.7623 & -0.6472 \end{bmatrix}$	$[-508.4192 \quad 105.2202 \quad 231.7118]$

Appendix C

To validate the accuracy and effectiveness of the proposed localization method in Section 3.2 and highlight its advantages, a comparative analysis was conducted against two commonly used methods from existing literature. Specifically, this appendix details the performance comparison among three approaches: (1) the proposed method using the midpoint between the left and right ankles, (2) the midpoint at bottom of detection frame (as in Huang et al. (2023), referred to as the **Bottom Midpoint method**), and (3) the geometric center of the detection frame (as in Ding et al. (2023), referred to as the **Geometric Center method**).

Tests C-1 and C-2 were conducted wherein a pedestrian was instructed to traverse a predetermined path marked by calibration points. To ensure a rigorous statistical comparison, the experiments utilized a self-recorded video dataset with strictly controlled variables. Each test consisted of a continuous 20-s sequence at 30 FPS (totaling 600 frames per test). The three comparison methods (Ankle Midpoint, Bottom Midpoint, and Geometric Center) were applied to the same video footage using identical camera calibration parameters and pose estimation settings, with the only variable being the pixel coordinate extraction logic. The pedestrian's pixel coordinates, captured from the surveillance images, were converted into world coordinates and used to reconstruct the motion trajectory. This computed trajectory was then compared against the designed trajectory to assess any deviations. Moreover, as the pedestrian passed each calibration point, the corresponding world coordinate was computed and compared with the actual world coordinate of calibration point, with the localization error evaluated using the Cumulative Distribution Function (CDF).

The actual trajectories from Tests C-1 and C-2, along with those obtained by the different localization methods, are illustrated in Figs. C1 and C2. Visual inspection obviously reveals that the trajectory produced by our method aligns more closely with the calibration points, exhibiting a higher degree of overlap with the intended path and thus more accurately representing the true motion trajectory.

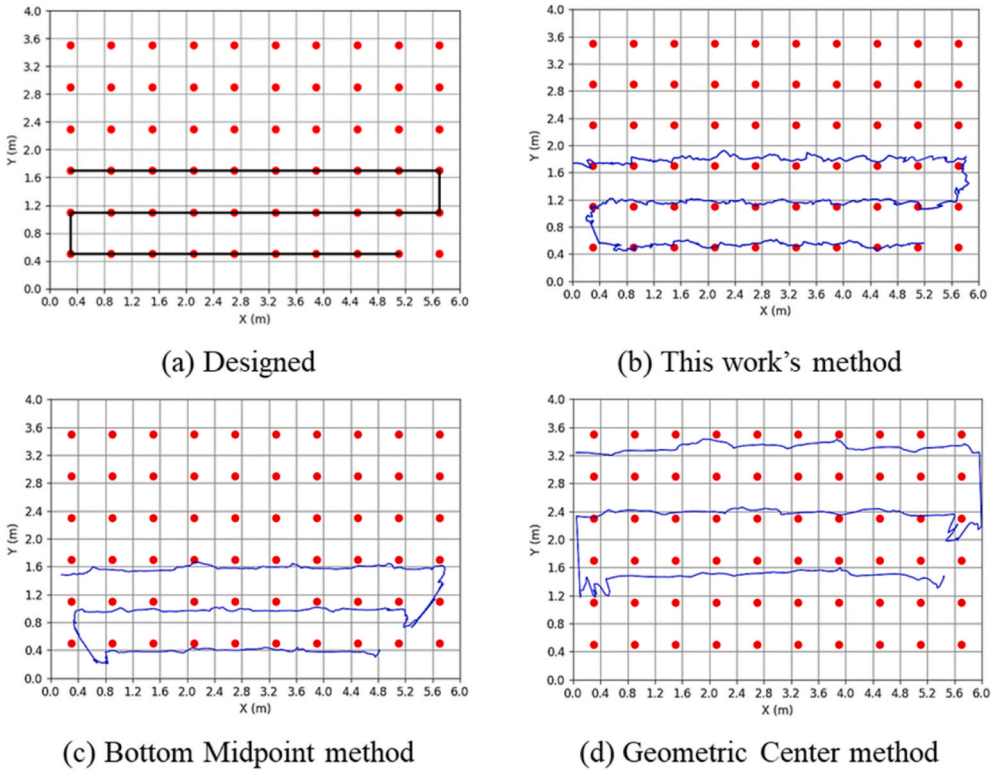

Fig. C1. Comparison of designed trajectory and calculated trajectory results by different methods for Test C-1. (Taken by Camera 1#).

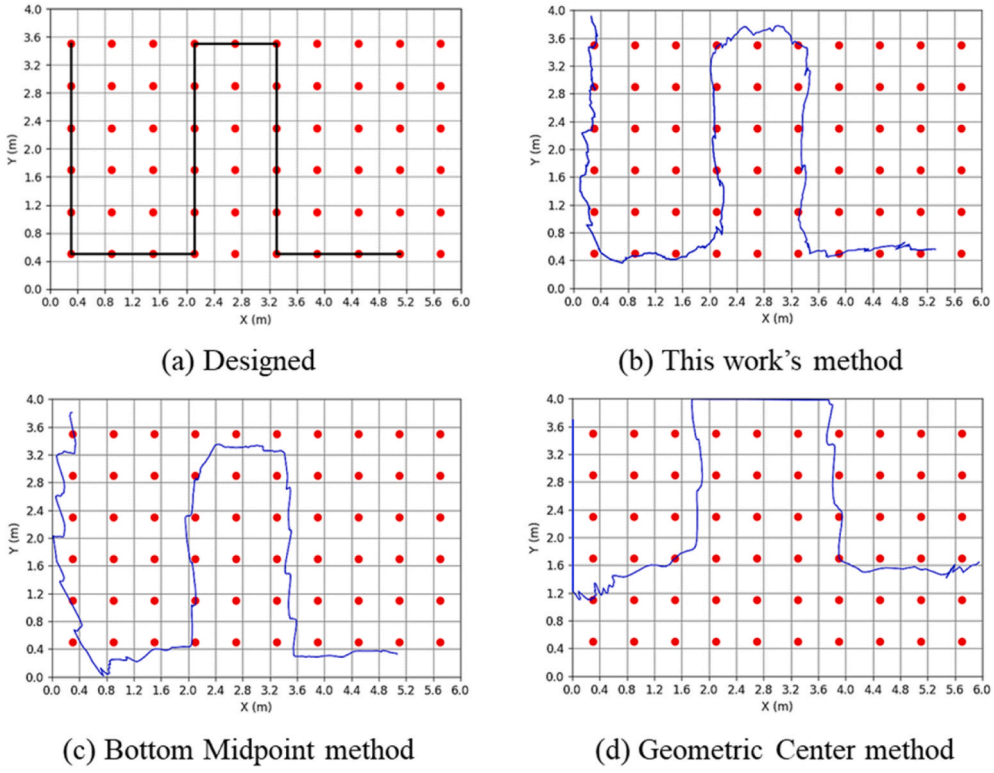

Fig. C2. Comparison of designed trajectory and calculated trajectory results for Test C-2. (Taken by Camera 3#).

Furthermore, the cumulative distribution functions of the localization errors, as shown in Figs. C3 and C4 for Tests C-1 and C-2, provide quantitative evidence of the superior performance of our method. In Test C-1, 79.2% of localized points achieved errors ≤10 cm (vs. 37.5% for Bottom Midpoint method and 4.2% for Geometric Center method), with 100% of points within 20 cm (vs. 75% and 8.3%), and a maximum error of 17.2 cm (vs. 33.6 cm and 132.4 cm). For Test C-2, 86.2% of points exhibited ≤10 cm errors (vs. 24.1% and 3.4%), 100% within 15 cm (vs. 86.2% and 10.3%), and a maximum error of 11.8 cm (vs. 24.1 cm and 103.7 cm). These results clearly demonstrate that using the midpoint between the left and right ankles for localization yields not only higher precision but also greater stability.

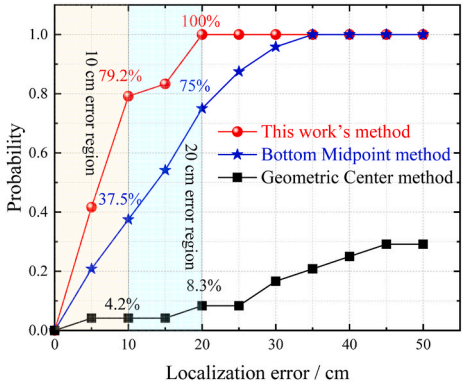

Fig. C3. Cumulative distribution function of localization error for Test C-1.

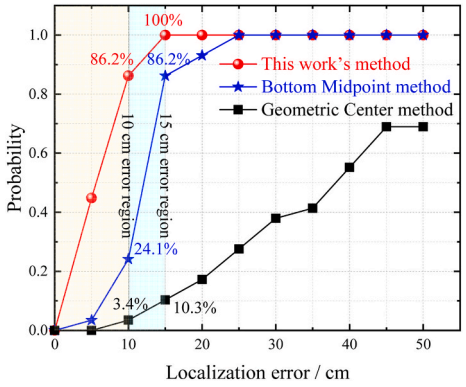

Fig. C4. Cumulative distribution function of localization error for Test C-2.

Appendix D

This section presents the camera calibration results in the real-world scenario. The equipment deployed at the high-speed rail station waiting hall also utilized the HIKVISION D1^{+}, consistent with that described in Appendix B. As intrinsic parameters and distortion coefficients remain invariant to camera pose or spatial configuration, their values are consistent with those previously reported in Appendix B. For extrinsic calibration, the intersections of floor tiles in Fig. 16 were utilized as reference markers, providing geometrically stable 3D-2D correspondences. By applying the same Perspective-n-Point algorithm described in Appendix B, the extrinsic parameters for the two cameras C1 and C2 were determined, with the results summarized in Table D1.

Table D1
Extrinsic parameters in the real-world scenario.

Camera number	Rotation matrix **R**	Translation vector **T**
C1	$\begin{bmatrix} 0.9987 & -0.0393 & -0.0314 \\ -0.0435 & -0.3603 & -0.9318 \\ 0.0253 & 0.9320 & -0.3615 \end{bmatrix}$	$[-145.6700 \quad 103.4267 \quad 199.1842]$
C2	$\begin{bmatrix} 0.9973 & -0.0318 & -0.0662 \\ -0.0734 & -0.3864 & -0.9194 \\ 0.0036 & 0.9218 & -0.3877 \end{bmatrix}$	$[-709.2039 \quad 152.7747 \quad 207.8757]$

Data availability

Data will be made available on request.

References

Abbas, Y., Alarfaj, A., Alabdulqader, E., Algarni, A., Ahmad, J., Liu, H., 2025. Drone-based public surveillance using 3D point clouds and neuro-fuzzy classifier. Comput. Mater. Contin. 82, 4759.

Brown, M., Lowe, D.G., 2007. Automatic panoramic image stitching using invariant features. In: International Journal of Computer Vision. https://doi.org/10.1007/s11263-006-0002-3.

Chen, C., Sun, H., Lei, P., Zhao, D., Shi, C., 2021. An extended model for crowd evacuation considering pedestrian panic in artificial attack. Phys. A Stat. Mech. its Appl. 571. https://doi.org/10.1016/j.physa.2021.125833.

Chen, H., Sun, J., Zhang, S., Yuan, H., Zhang, H., Zhang, J., 2023. 3D pedestrian localization fusing via monocular camera. J. Vis. Commun. Image Represent. 95. https://doi.org/10.1016/j.jvcir.2023.103871.

Chen, J., Ye, H., Ying, Z., Sun, Y., Xu, W., 2025a. Dynamic trend fusion module for traffic flow prediction. Appl. Soft Comput. 174, 112979.

Chen, J., Zhang, S., Xu, W., 2025b. Scalable prediction of heterogeneous traffic flow with enhanced non-periodic feature modeling. Expert Syst. Appl. 294, 128847.

Devernay, F., Faugeras, O., 2001. Straight lines have to be straight. Mach. Vis. Appl. 13. https://doi.org/10.1007/PL00013269.

Ding, Y., Zhang, Y., Huang, X., 2023. Intelligent emergency digital twin system for monitoring building fire evacuation. J. Build. Eng. 77, 107416.

Fetić, A., Jurić, D., Osmanković, D., 2012. The procedure of a camera calibration using camera calibration toolbox for MATLAB. In: MIPRO 2012 - 35th International Convention on Information and Communication Technology, Electronics and Microelectronics - Proceedings.

Gregory, A., 2024. No Title [WWW Document]. independent.

Guyo, E.D., Hartmann, T., Snyders, S., 2023. An ontology to represent firefighters data requirements during building fire emergencies. Adv. Eng. Inform. 56. https://doi.org/10.1016/j.aei.2023.101992.

Han, L., Feng, H., Liu, G., Zhang, A., Han, T., 2024. A real-time intelligent monitoring method for indoor evacuee distribution based on deep learning and spatial division. J. Build. Eng. 92, 109764. https://doi.org/10.1016/j.jobe.2024.109764.

Hayward, S.J., van Lopik, K., Hinde, C., West, A.A., 2022. A survey of indoor location technologies, techniques and applications in industry. Internet Things. https://doi.org/10.1016/j.iot.2022.100608.

Huang, S., Ji, J., Wang, Y., Li, W., Zheng, Y., 2023. A machine vision-based method for crowd density estimation and evacuation simulation. Saf. Sci. 167. https://doi.org/10.1016/j.ssci.2023.106285.

Jahangir, M.F., Kamari, A., Schultz, C.P.L., 2025. Unified and interoperable digital twin models for emergency evacuation using building simulation identity cards. Autom. ConStruct. 175. https://doi.org/10.1016/j.autcon.2025.106225.

Kim, D., Lee, J., Lee, D., 2026. Dual-stage graph-based association framework for cross-view person Re-Identification in construction worker monitoring. Buildings.

Kim, Junghoon, Chi, S., Kim, Jinwoo, 2023. 3D pose estimation and localization of construction equipment from single camera images by virtual model integration. Adv. Eng. Inform. 57. https://doi.org/10.1016/j.aei.2023.102092.

Li, D., Zhang, F., Rong, W., Yue, C., Zhang, Y., Liang, Y., Ren, J., 2025. Robust localization algorithm for micromanipulation targets under complex interference conditions. IEEE Trans. Autom. Sci. Eng. 22, 23959–23969.

Kreuzer, T., Papapetrou, P., Zdravkovic, J., 2024. Artificial intelligence in digital twins—A systematic literature review. Data Knowl. Eng. 151. https://doi.org/10.1016/j.datak.2024.102304.

Li, N., Becerik-Gerber, B., Krishnamachari, B., Soibelman, L., 2014. A BIM centered indoor localization algorithm to support building fire emergency response operations. Autom. ConStruct. 42, 78–89. https://doi.org/10.1016/j.autcon.2014.02.019.

Li, N., Huang, G., Jiang, H., Gao, X., Zhou, L., 2023. Fire propagation-driven dynamic intelligent evacuation model in multifloor hybrid buildings. Adv. Eng. Inform. 57. https://doi.org/10.1016/j.aei.2023.102097.

Li, Z., Cai, J., Chen, Q., Chen, L., Qing, M., Yang, S.X., 2025. An LSTM network with neural plasticity for driver fatigue recognition on real roads. IEEE Trans. Ind. Electron.

Lipman, A., Hall, E.T., 1970. The hidden dimension. Br. J. Sociol. 21. https://doi.org/10.2307/589150.

Liu, F., Wang, J., Zhang, J., Han, H., 2019. An indoor localization method for pedestrians base on combined UWB/PDR/floor map. Sensors (Switzerland) 19. https://doi.org/10.3390/s19112578.

Liu, M., Ma, W., Wang, C., Wang, P., Yang, M., 2025. Machine learning-enhanced rapid design of hydrodynamic shape for underwater vehicles pedigrees. IEEE/ASME Trans. Mechatronics.

Long, X., Chen, J., Yang, L., Huang, H., 2026. An emergency scheduling method based on AutoML for space maneuver objective tracking. Expert Syst. Appl. 298, 129759.

Lu, Z., Zhang, X., Cao, X., Hou, J., Yuan, X., 2026. SFEP-YOLO: a track obstacle detection model for autonomous electric locomotives in underground mine. IEEE Trans. Veh. Technol.

Maji, D., Mathew, M., 2022. YOLO-Pose:Enhancing YOLO for multi person pose estimation using object. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, pp. 2637–2646.

Niu, Y., Xu, Z., Xu, E., Li, G., Huo, Y., Sun, W., 2021. Monocular pedestrian 3d localization for social distance monitoring. Sensors 21. https://doi.org/10.3390/s21175908.

Nuvem tóxica atinge quatro cidades no litoral de SP, vazamento continua, 2016. [WWW Document]. alagoas24horas.

Ouyang, K., Fu, S., Chen, Y., Cai, Q., Heidari, A.A., Chen, H., 2025. Escape: an optimization method based on crowd evacuation behaviors. Artif. Intell. Rev. 58. https://doi.org/10.1007/s10462-024-11008-6.

Piasco, N., Sidibé, D., Demonceaux, C., Gouet-Brunet, V., 2018. A survey on visual-based localization: on the benefit of heterogeneous data. Pattern Recogn. 74. https://doi.org/10.1016/j.patcog.2017.09.013.

Qiu, X., Liao, S., Yang, D., Li, Y., Wang, S., 2025. Visual geo-localization and attitude estimation using satellite imagery and topographical elevation for unmanned aerial vehicles. Eng. Appl. Artif. Intell. 153, 110759.

Sacoto-Cabrera, E.J., Perez-Torres, A., Tello-Oquendo, L., Cerrada, M., 2025. IoT, AI, and digital twins in smart cities: a systematic review for a thematic mapping and research agenda. Smart Cities. https://doi.org/10.3390/smartcities8050175.

Sato, Toshio, Qi, X., Yu, K., Wen, Z., Myint, S.H., Katsuyama, Y., Tokuda, K., Sato, Takuro, 2020. Pedestrian positioning in surveillance video using anthropometric properties for effective communication. In: International Symposium on Wireless Personal Multimedia Communications. https://doi.org/10.1109/WPMC50192.2020.9309520. WPMC.

Sesyuk, A., Ioannou, S., Raspopoulos, M., 2022. A survey of 3D indoor localization systems and technologies. Sensors. https://doi.org/10.3390/s22239380.

Setijadi Prihatmanto, A., Prasetyadi, A., Yoganingrum, A., Sutriadi, R., Hadiana, A., 2025. The digital twin city in enhancing flood evacuation systems: future opportunities and challenges. IEEE Access 13. https://doi.org/10.1109/ACCESS.2025.3539346.

Sun, H., Zhu, K., Wang, G., Hu, H., Guo, P., Wu, K., Zhang, T., 2025a. Experimental study on pedestrian evacuation characteristics through building bottleneck group structure. Dev. Built Environ. 23. https://doi.org/10.1016/j.dibe.2025.100734.

Sun, H., Zhu, K., Zhang, W., Ke, Z., Hu, H., Wu, K., Zhang, T., 2025b. Emergency path planning based on improved ant colony algorithm. J. Build. Eng. 100, 111725.

Sun, H.K., Chen, C.K., 2023. Model considering panic emotion and personality traits for crowd evacuation. Chin. Phys. B 32. https://doi.org/10.1088/1674-1056/ac9e94.

Sun, Y., Zhao, K., Wang, J., Li, W., Bai, G., Zhang, N., 2017. Device-free human localization using panoramic camera and indoor map. In: 2016 IEEE International Conference on Consumer Electronics-China, ICCE-china 2016. https://doi.org/10.1109/ICCE-China.2016.7849743.

Tran, S.V.T., Nguyen, T.L., Chi, H.L., Lee, D., Park, C., 2022. Generative planning for construction safety surveillance camera installation in 4D BIM environment. Autom. ConStruct. 134. https://doi.org/10.1016/j.autcon.2021.104103.

Turkey-Syria earthquake, 2023. [WWW Document], 2023 actionaid.

Wang, Q., Chen, J., Song, Y., Li, X., Xu, W., 2024. Fusing visual quantified features for heterogeneous traffic flow prediction. Promet - Traffic & Transp. 36, 1068–1077.

Wang, S., Wang, Y., Chen, S., Zhou, Z., Liu, X., Li, Z., 2025. Interactive Siamese network-based roadside perception for multi-vehicle tracking. IEEE Trans. Intell. Transport. Syst.

Wang, Z., Yang, M., Wang, G., Lian, Y., Wang, Y., 2026. Digital twin-driven shape-performance-control-application integrated design for unmanned underwater vehicles. Sci. China Technol. Sci. 69, 1380301.

Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric. In: Proceedings - International Conference on Image Processing. ICIP. https://doi.org/10.1109/ICIP.2017.8296962.

Wong, M.O., Lee, S., 2025. Quantitative analysis of the effectiveness of real-time indoor navigation and information sharing for collaborative fire response. Dev. Built Environ. 23. https://doi.org/10.1016/j.dibe.2025.100690.

Wong, M.O., Lee, S., 2023. Indoor navigation and information sharing for collaborative fire emergency response with BIM and multi-user networking. Autom. ConStruct. 148. https://doi.org/10.1016/j.autcon.2023.104781.

Wu, K, Hu, H, Sun, H, et al., 2026. Experimental study on human visual response to safety signage under emergency lighting conditions. Dev. Built Environ., 100852

Wu, K., Sun, H., Zhu, Z., Hu, H., Xu, J., Zhu, K., Zhang, X., Zhang, T., 2025. Experimental study on the emergency evacuation behavior in building with bottleneck group. J. Build. Eng. 106. https://doi.org/10.1016/j.jobe.2025.112576.

Xiao, Z., Li, H., Jiang, H., Li, Y., Alazab, M., Zhu, Y., Dustdar, S., 2023. Predicting urban region heat via learning arrive-stay-leave behaviors of private cars. IEEE Trans. Intell. Transport. Syst. 24, 10843–10856.

Xie, W., Zeng, Y., Zhang, X., Wong, H.Y., Zhang, T., Wang, Z., Wu, X., Shi, J., Huang, X., Xiao, F., Usmani, A., 2025. AIoT-powered building digital twin for smart firefighting and super real-time fire forecast. Adv. Eng. Inform. 65, 103117. https://doi.org/10.1016/j.aei.2025.103117.

Yang, J., Zang, X., Chen, W., Luo, Q., Wang, R., Liu, Y., 2024. Improved social force model based on pedestrian collision avoidance behavior in counterflow. Phys. A Stat. Mech. its Appl. 642, 129762.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H., 2022. Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. 44. https://doi.org/10.1109/TPAMI.2021.3054775.

Zafari, F., Gkelias, A., Leung, K.K., 2019. A survey of indoor localization systems and technologies. IEEE Commun. Surv. Tutor. 21. https://doi.org/10.1109/COMST.2019.2911558.

Zhang, T., Ding, F., Wang, Z., Xiao, F., Lu, C.X., Huang, X., 2024. Forecasting backdraft with multimodal method: fusion of fire image and sensor data. Eng. Appl. Artif. Intell. 132, 107939.

Zhang, T., Wang, Z., Wong, H.Y., Tam, W.C., Huang, X., Xiao, F., 2022a. Real-time forecast of compartment fire and flashover based on deep learning. Fire Saf. J. 130, 103579.

Zhang, T., Wang, Z., Zeng, Y., Wu, X., Huang, X., Xiao, F., 2022b. Building artificial-intelligence digital fire (AID-Fire) system: a real-scale demonstration. J. Build. Eng. 62, 105363.

Zhang, X., Chen, X., Ding, Y., Zhang, Y., Wang, Z., Shi, J., Johansson, N., Huang, X., 2024. Smart real-time evaluation of tunnel fire risk and evacuation safety via computer vision. Saf. Sci. 177, 106563. https://doi.org/10.1016/j.ssci.2024.106563.

Zhang, Y., Ding, Y., Chraibi, M., Huang, X., 2025a. Multi-scale analysis of fire and evacuation drill in a multi-functional university high-rise building. Dev. Built Environ. 21, 100626. https://doi.org/10.1016/j.dibe.2025.100626.

Zhang, Y., Kinateder, M., Huang, X., Warren, W.H., 2025b. Modeling competing guidance on evacuation choices under time pressure using virtual reality and machine learning. Expert Syst. Appl. 262, 125582. https://doi.org/10.1016/j.eswa.2024.125582.

Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. 22. https://doi.org/10.1109/34.888718.

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J., 2019. Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2019.00224.

Zhu, J., Wong, M.O., Nisbet, N., Xu, J., Kelly, T., Zlatanova, S., Brilakis, I., 2025. Semantics-based connectivity graph for indoor pathfinding powered by IFC-graph. Autom. ConStruct. 171. https://doi.org/10.1016/j.autcon.2025.106019.

Zhu, R., Wang, Y., Cao, H., Yu, B., Gan, X., Huang, L., Zhang, H., Li, S., Jia, H., Chen, J., 2020. RTK/pseudolite/LAHDE/IMU-PDR integrated pedestrian navigation system for urban and indoor environments. Sensors (Switzerland) 20. https://doi.org/10.3390/s20061791.

Zhuang, S., Zhao, Z., Cao, L., Wang, D., Fu, C., Du, K., 2023. A robust and fast method to the Perspective-n-Point problem for camera pose estimation. IEEE Sens. J. 23. https://doi.org/10.1109/JSEN.2023.3266392.

Zou, B., Xia, K., Ma, J., Long, X., 2025. Deep learning driven prediction of dynamic stress-strain response in limestone: insights into transient mechanical behavior under complex loadings for shield tunneling. Eng. Appl. Artif. Intell. 162, 112554.