Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Smart building evacuation by tracking multi-camera network and explainable Re-identification model

Yifei Ding[a], Xinghao Chen[a], Yuxin Zhang[a,b,c,*], Xinyan Huang[a,**]

[a] Research Center for Smart Urban Resilience and Firefighting, Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong

[b] State Key Laboratory of Disaster Reduction in Civil Engineering, Tongji University, Shanghai, China

[c] Department of Geotechnical Engineering, Tongji University, Shanghai, China

ARTICLE INFO

Keywords:
Computer vision
Multi-camera tracking
Person re-identification
Evacuation monitoring
Emergency management
Building digital twin

ABSTRACT

Real-time crowd data from surveillance devices is essential for the emergency decision-making and management inside complex buildings. Traditional evacuation monitoring with single-camera tracking often leads to erratic information, so multi-camera tracking for building occupants is critical to enhance evacuation safety and emergency response. This research proposes a novel real-time multi-camera tracking framework for the detection, tracking and re-identification (Re-ID) of evacuees across multi-camera. The framework consists of (1) a multi-camera network, (2) human detection model, (3) tracking model, (4) an explainable attention-aided Re-ID (AAR) model, and (5) a module of feature matching and re-distribution algorithm. The attention-aided Re-ID model presents outstanding performance on both the standard big benchmarks and our custom dataset. Moreover, a simple evacuation drill is conducted to demonstrate real-time multi-camera tracking, showing good accuracy in Re-ID and personnel counting, where the overall Re-ID tracking accuracy exceeds 75% and the personnel counting accuracy is approaching 100%. Lastly, the class activation map (CAM) illustrates the model explainability and limitations. The proposed multi-camera tracking framework helps develop a more automated monitoring system and an intelligent digital twin for building emergency safety management.

Nomenclature

Symbols		**Abbreviations**	
F_i^j	Feature vector in FD	AI	Artificial Intelligence
FD	Feature dictionary	AAR	Attention-Aided Re-ID
D	Euclidean distance	AG	Aggregation Gate
x	Input feature of a CNN block	CAM	Class Activation Map
$\widetilde{x}$	Residual feature	CNN	Convolutional Neural Network
y	Output feature of a CNN block	CMC	Cumulated Matching Curve
f	Feature tensor	CCTV	Closed-Circuit Television
MLP	Multiple convolutional layer perceptron	CV	Computer Vision
AvgPool	Average pooling	DeepSORT	Deep Simple Online and Realtime Tracking
MaxPool	Maximum pooling	FC	Fully-connected
Conv	Convolutional computing	ID	Person Identity
Cat	Concat computing	Re-ID	Re-identification of Person Identity
Avg	Average value computing	YOLO	You Look Only Once
Max	Maximum computing	HA-CNN	Harmonious Attention CNN
q	True value	ResNet	Residual Neural Network
p	Prediction value	FD	Feature Dictionary
K	Number of person identity	OSNet	Omni-Scale Network
M	Feature map	RFZ	Receptive Field Size
w	Weight of feature map	R-k	Rank-k Metric
		mAP	Mean Average Precision
Greeks		*RTA*	Re-identification Tracking Accuracy

(continued on next column)

(continued)

(continued on next page)

* Corresponding author. Research Centre for Smart Urban Resilience and Firefighting, Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong.

** Corresponding author.

E-mail addresses: yx.zhang@polyu.edu.hk (Y. Zhang), xy.huang@polyu.edu.hk (X. Huang).

https://doi.org/10.1016/j.engappai.2025.110394
Received 4 April 2024; Received in revised form 21 January 2025; Accepted 21 February 2025
Available online 5 March 2025

(*continued*)

σ	Sigmoid function	*TRF*	True Re-identification Tracking Frames
ε	Hyper-parameter	*TTF*	Total Tracking Frames
$\mathscr{L}$	Cross-entropy loss		

1. Introduction

With the rapid development of urbanization, the number of public emergency incidents has increased heavily in the past few decades. For instance, the tragic London Grenfell Tower Fire in 2017 cause 72 deaths, exposing critical issues concerning building evacuation (McKenna et al., 2019). In contrast, timely and efficient evacuation in the early stage of the disasters could greatly reduce injuries and casualties. All 180+ occupants successfully evacuated from a residential fire in Hangzhou without injury in 2018 before the firefighters arrived, mainly attributed to good evacuation guidance at the early stage (Real Life Fire Cases, 2020). In a typical emergency disaster, the complex building structure and nervous atmosphere inhibit occupants from self-evacuating timely and efficiently. Therefore, the evacuation guidance based on the real-time information of indoor crowd flow indoor in different emergency scenarios is vitally important for the onsite evacuation and rescue.

Most of traditional surveillance system installed in public buildings includes dozens and hundreds of CCTV cameras, depending on the complexity of building, and they are limited to recording, storage, and playback of videos (Zhang et al., 2025a). However, it is extremely labour-intensive (if not impossible) for human eyes and brain to process massive historical and real-time video data, not mention to guide occupants to evacuate safely in an emergency. Therefore, the intelligent evacuation management system based on multi-sensors, CCTV cameras, and computing technologies is necessary for the development of contemporary smart buildings (Ibrahim et al., 2016). Such a system should be smart enough to automatically track every occupant across different cameras and provide adequate and timely feedback for safely guiding the evacuation and rescue processes. Specifically, it will monitor evacuees (Yogameena and Nagananthini, 2017; Vanem and Ellis, 2010), predict evacuation development and potential crowd disaster (Cheng et al., 2017; Wang et al., 2020), and provide egress path guidelines (Zhang et al., 2014; Zhou et al., 2019a).

Accurate locating and tracking of the evacuees are two essential tasks of an intelligent surveillance system. With the assistance of computer vision (CV), pattern recognition and deep learning, automatic filtering of irrelevant information and precise recognising of a person in one video stream is widely developed and deployed in smart building safety systems (Baduge et al., 2022). The relevant computer vision algorithm, especially convolutional neural network (CNN)-based object detection (Liu et al., 2016; Lin et al., 2016), tracking algorithms (Bewley et al., 2016), and derivative applications or systems (Li et al., 2023; Tesfaye et al., 2019) have emerged as the research hotspots. The accurate detection and tracking of multiple humans could provide rapid population flow statistics, population density calculation, and evacuees' positioning, which are vitally essential for the analysis and judgment of evacuation conditions and human behaviours in a fire scenario (Zhang et al., 2025b). In practical applications, one of the most crucial problems of intelligent evacuation monitoring is multi-camera tracking, which requires to re-identify people's identity among different cameras in different scenarios (i.e., Person Re-ID).

Person Re-ID is a well-explored problem focused on retrieving specific individuals across non-overlapping cameras (Ye et al., 2021). Specifically, when given a query person-of-interest, its primary objective is to determine whether the target person appears in other locations at a different time captured by the distinct cameras, or even the same camera at a different time instant (Gheissari et al., 2006). Therefore, Re-ID could establish a connection among distinct video streams and share the person's features at different times. In addition, multi-camera tracking using person Re-ID also contributes to the intelligent emergency digital twins. Combining computer vision technology, building information modelling, and multimodal data, emergency digital twin system enables the establishment of a connection between the virtual world and physical reality (Zhang et al., 2025). It is widely applied on global remote surveillance, forecast, and operation in safety engineering, i.e., smart firefighting (Zhang et al., 2022, 2024a), risk assessment (Li et al., 2022a) and emergency management (Ding et al., 2023; Zhang et al., 2024b). Moreover, explainability of AI is also significant for real-time adoptions and application research to solve *Black-Box* problem (Barredo et al., 2020). Explainable AI could enhance the understand about how AI learning and working and why the AI decisions, which benefits human's trust, confidence, and further model improvements (Ding et al., 2022). To this end, this work also discuss the explainability of the Re-ID model for error analysis.

According to the author's knowledge, few frameworks and demonstrations of an evacuation monitoring system based on Re-ID have been presented, requiring an in-depth exploration. In addition, how to avoid repeat personnel counting across multi-camera is another critical pending issue. The previous human tracking system based on YOLO (Chen et al., 2020) and DeepSORT (Bewley et al., 2016) performs good in single camera view, but it cannot build a connection among distinct cameras, resulting in huge errors in multi-camera tracking. Therefore, creating interoperability and connection on evacuation information among distinct viewpoints to pursue multi-camera tracking would promote a huge step forward for emergency safety. In our previous work (Ding et al., 2023), a primary digital twin framework was demonstrated to monitor the building evacuation process with a single camera and represent evacuees virtually, but it still expand to full-view tracking.

This paper improves the evacuation digital twin system by tracking evacuees across multi-camera network and continuously monitoring the movement of individual throughout the evacuation process. The proposed intelligent system includes (1) multi-camera network (2) human detection model, (3) tracking model, and (4) attention-aided re-identification (AAR) model, and (5) module of feature matching and ID redistribution. The remaining parts of the paper include a short review of the technical background (Section 2), the methodology developed for this study (Section 3), the evaluation of the improved AAR model (Section 4), the demonstration of the proposed framework (Section 5), and the explainability analysis and future perspectives (Section 6) before conclusions. The major contribution of this work concludes.

(1) We proposed a multi-camera tracking framework for evacuation safety.
(2) Our attention-aided Re-ID model outperformed both standard benchmarks and custom datasets.
(3) We developed a novel Re-ID dataset annotation tool for efficient model research and created a mini-scale testing dataset.
(4) Our method achieved satisfactory accuracy in Re-ID tracking and personnel counting in a real-time evacuation drill.
(5) We used class activation maps to demonstrate the explainability of our model, highlighting its limitations and areas for improvement.

2. Technical background

2.1. Computer vision-based people detection and tracking

Video analytics equipped with computer vision techniques have been widely applied in emergency management and evacuation analysis. Related research includes the public emergency crowd disaster avoidance and crowd behaviour analysis (Yogameena and Nagananthini, 2017; Dong et al., 2020), crowd density estimation (Huang et al., 2023), evacuation movement modelling (Li et al., 2022b), egress route planning (Liu et al., 2018; Deng et al., 2022), and virtual reality application (Zhao et al., 2022; Dang et al., 2024) et al. Among them, object detection models, i.e., Faster R-CNN (Girshick, 2015), and YOLO algorithms

(Redmon et al., 2016) are widely used to locate the observed targets. For example, Li et al. (2023) proposed a real-time detection method to extract the risk factors about important human factors in the evacuation processes. Huang et al. (2023) applied an object detection model to count the population and further estimate crowd density for evacuation simulations. Li et al. (2022b) utilized an object detection model to extract and model evacuees' movement features from the real seismic evacuation videos.

Moreover, video analytics have also been applied to monitoring evacuation processes and public crowd conditions including pedestrian dynamic tracking, locating evacuee distribution, and behaviour classification, which enable extracting spatial, temporal, and semantic information for onsite emergency safety assessment. For example, Cheng et al. (2021) combined the deep learning model and video tracking algorithm to tally the evacuee number and further predict the congestion area and the proposed method contributes to the real-time evacuation navigation. Khlevnoi et al. (2022) applied person-tracking algorithms and video processing techniques to estimate the evacuee's real-time moving speed and analysed the relationship between the population density and the crowd movement speed during the evacuation process. Wong et al. (2021) proposed a robust method of pedestrian tracking to recognize the person's attribute and trajectory for facilitating the real-time analysis of human behaviours. Yu et al. (2021) extracted the evacuee distribution from the monitoring device and presented the crowd information with 3D visualization for real-time autonomous evacuation management.

In our previous work (Ding et al., 2023), an intelligent digital twin tracking system was proposed to automatically monitor fire evacuation where computer vision and deep learning method were used to analyse the evacuee's movement characteristics. Then, a computer vision-based monitoring method is proposed to recognize and classify special pedestrian behaviours i.e., pregnancy, and disability, contributing to an intelligent underprivileged evacuation monitoring framework (Ding et al., 2024). So far, no tracking system can monitor evacuee movement beyond a confined sub-region or analyse the walking trajectories across multi-camera with a Re-ID function for the whole evacuation process.

2.2. *Person Re-ID technology*

In the computer vision field, person re-identification (Re-ID) aims to retrieve a person of interest across multiple non-overlapping cameras (Ye et al., 2021). The classical object detection (Liu et al., 2016; Lin et al., 2016; Redmon et al., 2016) and tracking algorithms (Bewley et al., 2016; Wojke et al., 2017) would process the raw video to generate the bounding boxes, which are mature in the previous single-camera tracking. In contrast, extracting personal features and making correlations across different cameras are still not widely applicable, while are treated as key research points of the multi-camera tracking. In the evacuation monitoring, person Re-ID model contributes to avoidance of repeat personnel counting across multiple cameras.

As shown in Fig. 1(a), the evacuee in the same video domain has the same identity but it changes across camera views. Assuming that Re-ID model is introduced in the monitoring system (see Fig. 1(b)), the evacuee would retain constant ID in both single camera field and non-overlapping multi-camera. Therefore, building a discriminative and robust Re-ID model is the core step for developing a multi-camera tracking framework. Person Re-ID faces two major challenges. First, the intra-class (instance/identity) variations are typically big due to the changes in camera viewing conditions. Second, there are also small inter-class variations, for example, people in public spaces often wear similar clothes. To increase the performance of Re-ID technology in various application scenarios, both the corresponding dataset and the innovative Re-ID model vitally matter.

2.2.1. *Person Re-ID dataset*

Re-ID dataset typically consists of images or videos captured from multiple cameras in various environments, such as shopping malls, airports, or streets. Each dataset usually contains many individuals, and each individual is captured from different camera angles and under different lighting conditions. The dataset is annotated to provide ground truth labels for person identities. Each person in the dataset is assigned a unique identifier, and these annotations are used to train and evaluate person Re-ID algorithms. The annotations in the dataset can vary with the specific dataset and its purpose.

Typically, the Re-ID dataset includes a bounding box train subset, a bounding box test subset (gallery subset), and a query subset. As a retrieval task, the query data can be represented by an image or a video sequence, so the commonly used datasets also be classified as image datasets (Zheng et al., 2015, 2017; Wei-Shi et al., 2009; Loy et al., 2013; Gray and Tao, 2008; Li et al., 2013, 2014; Li and Wang, 2013) and video datasets (Hirzer et al., 2011; Wang et al., 2014; Zheng et al., 2016; Wu et al., 2018; Li et al., 2018a). Market1501 (Zheng et al., 2015), CUHK03 (Li et al., 2014), DukeMTMC (Zheng et al., 2017) are generally used as the big benchmark of person Re-ID. Market1501 dataset consists of a

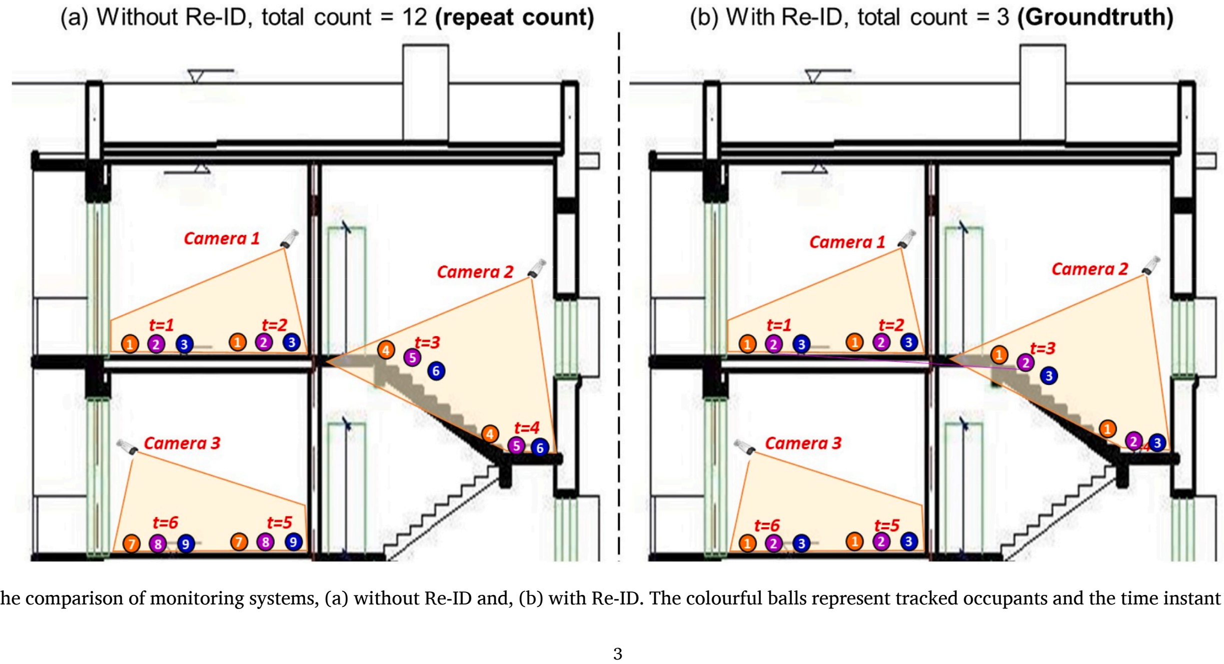

(a) Without Re-ID, total count = 12 (repeat count)
(b) With Re-ID, total count = 3 (Groundtruth)
Camera 1
Camera 2
Camera 3
t=1
t=2
t=3
t=4
t=5
t=6
1
2
3
4
5
6
7
8
9

Fig. 1. The comparison of monitoring systems, (a) without Re-ID and, (b) with Re-ID. The colourful balls represent tracked occupants and the time instant notes as *t*.

train set, a test set and a query set with a total of 1501 identifies, 6 cameras per identify, and 32,668 boxes, and all images are normalized to 128 × 64 pixel size. Market1501 is annotated with the Deformable Part Model (DPM) and uses cumulated matching curve (CMC) as the evaluation metric, while some datasets are annotated by hand and use mean average precision (mAP) to assess the Re-ID performance. Some famous big Re-ID image datasets are summarized in Table .1.

The existing popular datasets could be regarded as the benchmark to evaluate the person feature representation performance of the novel Re-ID models. However, due to the significant impact of the environmental factors on the Re-ID model performance, we need also to test the applied model using the custom dataset to fit every new scenario. To this end, we propose a new dataset from the evacuation scenario. In addition, the existing image label annotation tools such as LabelImage (Tzutalin, 2018) and DarkLabel are either not suitable for Re-ID annotation tasks or complicated to operate. To overcome this issue, a new lightweight Re-ID annotation tool is developed in this work for faster and easier generation of our own Re-ID dataset.

2.2.2. Person Re-ID model

The prevailing architecture of person Re-ID model for feature representation learning is based on convolutional neural networks (CNNs) i.e., Densenet (Huang et al., 2017), Inception (Szegedy et al., 2017), and ResNet (He et al., 2016) because deep CNNs performs strong ability to feature extraction and image classification. There are four major aspects of feature representation learning routes, which are categorized as global feature, local feature, auxiliary feature, and video feature. Among them, global feature learning is the most mature method and facilitates numerous state-of-art models (Szegedy et al., 2017; Hou et al., 2019; Zhou et al., 2019b; Li et al., 2018b; Zhang et al., 2018). For example, the Interaction-and-Aggregation Network (IANet) proposed by Hou et al. (2019) introduces spatial and channel interaction-and-aggregation into CNNs, which performed superior on three Re-ID benchmark datasets. OSNet proposed by Zhou et al. (2019b) aggregates multiple convolutional streams with diverse receptive fields in the residual block to learn multi-scale features and corporates the point-wise and depth-wise convolutional method to achieve a light-weight network.

To enhance the representation learning ability of CNNs, attention schemes are widely studied in image classification, object detection, and absolutely, Re-ID tasks. The most popular strategies are pixel-level attention and channel-wise feature response re-weighting. For instance, Li et al. (2018b) proposed Harmonious Attention CNN (HA-CNN) model to combine soft pixel attention learning and hard regional attention learning to optimise feature representations simultaneously, which contributes to improving the person Re-ID in misaligned images. Moreover, channel attention (Hu et al., 2018) and spatial attention (Woo et al., 2018) are widely introduced in convolutional blocks and corresponding combining structures achieve significant optimization on benchmark computer vision datasets. Based on this point, this paper inserts channel attention block and spatial attention block into the OSNet architecture to improve the feature representation performance of cross-camera evacuees.

Table 1
Statistic of some popular Re-ID image datasets.

Dataset	#ID	#Camera	#Image	Annotation	Evaluation
VIPeR (Gray and Tao, 2008)	632	2	1264	hand	CMC
iLIDS (Wei-Shi et al., 2009)	119	2	476	hand	CMC
GRID (Loy et al., 2013)	250	8	1275	hand	CMC
CUHK01 (Li et al., 2013)	971	2	3884	hand	CMC
CUHK02 (Li and Wang, 2013)	1816	10	7264	hand	CMC
CUHK03 (Li et al., 2014)	1467	2	13164	hand	CMC
Market1501(Zheng et al., 2015)	1501	6	32668	Hand + DPM	CMC + mAP
DukeMTMC (Zheng et al., 2017)	1404	8	36411	Hand + DPM	CMC + mAP

3. Methodology

3.1. Overall framework of the proposed method

The proposed framework is composed of five critical components: (1) multi-camera network to record raw videos (2) detection model (YOLOv7) to locate person, (3) tracking model to distribute ID, (4) attention-aided Re-ID model for feature extraction, and (5) The module of feature matching and ID re-distribution for person re-identification. Fig. 2 illustrates the overall framework of the proposed multi-camera evacuation Re-ID tracking methodology (see Algorithm I in Appendix).

In Step-1, the framework processes the first input video from a single camera. Firstly, the detection model draws a bounding box for each evacuee object and calculates the 2D position coordinates. The tracking model is then used to assign a unique ID number for objects and the ID remains constant throughout the entire video. After that, the attention-aided Re-ID model extracts features from each image in the bounding box and the screenshot image is automatically processed to a size of 128 × 256 pixel for Re-ID model inference. The extraction result of each person's image is a feature vector with 512 units that denotes $\overrightarrow{\boldsymbol{F}_1^j}$, where **1** represents the first camera and ***j*** represents the person ID. The feature vector with its corresponding ID forming as the feature pair is stored in a dictionary for subsequent feature matching, where the ID is the key, and the feature vector is the value of the dictionary.

In Step-2, the temporal videos from the other distinct camera view are sequentially processed as the Step-1. Each obtained feature vector that denotes $\mathbf{F}_i^j$, where ***i*** represents the camera ID, is automatically stored in the feature dictionary (***FD***), which is formed as,

$$FD = \begin{bmatrix} F_1^1 & F_1^2 & \cdots & F_1^{j-1} & F_1^j \\ F_2^1 & F_2^2 & \cdots & F_2^{j-1} & F_2^{j-1} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ F_i^1 & F_i^2 & \cdots & F_i^{j-1} & F_i^j \end{bmatrix} \tag{1}$$

The module of feature matching and ID re-distribution would be responsible for the comparison of the new feature vector and the previously stored vectors in the dictionary, so it can determine whether to inherit the existing ID or assign a new ID. The Euclidean distance is used as the metric to evaluate the similarity between the new and previous features. We denote the new feature vector as $\overrightarrow{\boldsymbol{F}}_{\boldsymbol{new}}$, i.e.,

$$\overrightarrow{F}_{new} = (q_1, q_2, q_3, \dots, q_{512}) \tag{2}$$

The previous feature vector in the dictionary denotes $\overrightarrow{\boldsymbol{F}}_{\boldsymbol{previous}}$, i.e.,

$$\overrightarrow{F}_{previous} = (p_1, p_2, p_3, \dots, p_{512}) \tag{3}$$

The Euclidean distance of two vectors is formulated by,

$$\mathscr{D}(q, p) = \sqrt{\sum_{i=1}^{512} (q_i - p_i)^2} \tag{4}$$

After a loop of the above procedures, the framework would output the real-time multi-camera tracking results, which contains the bounding boxes of evacuees, the local counting in one single camera view, the global counting in overall camera network, and the continued ID of each evacuee. The above information is useful for the provision of the location data and continued trajectory of each individuals, as well as the

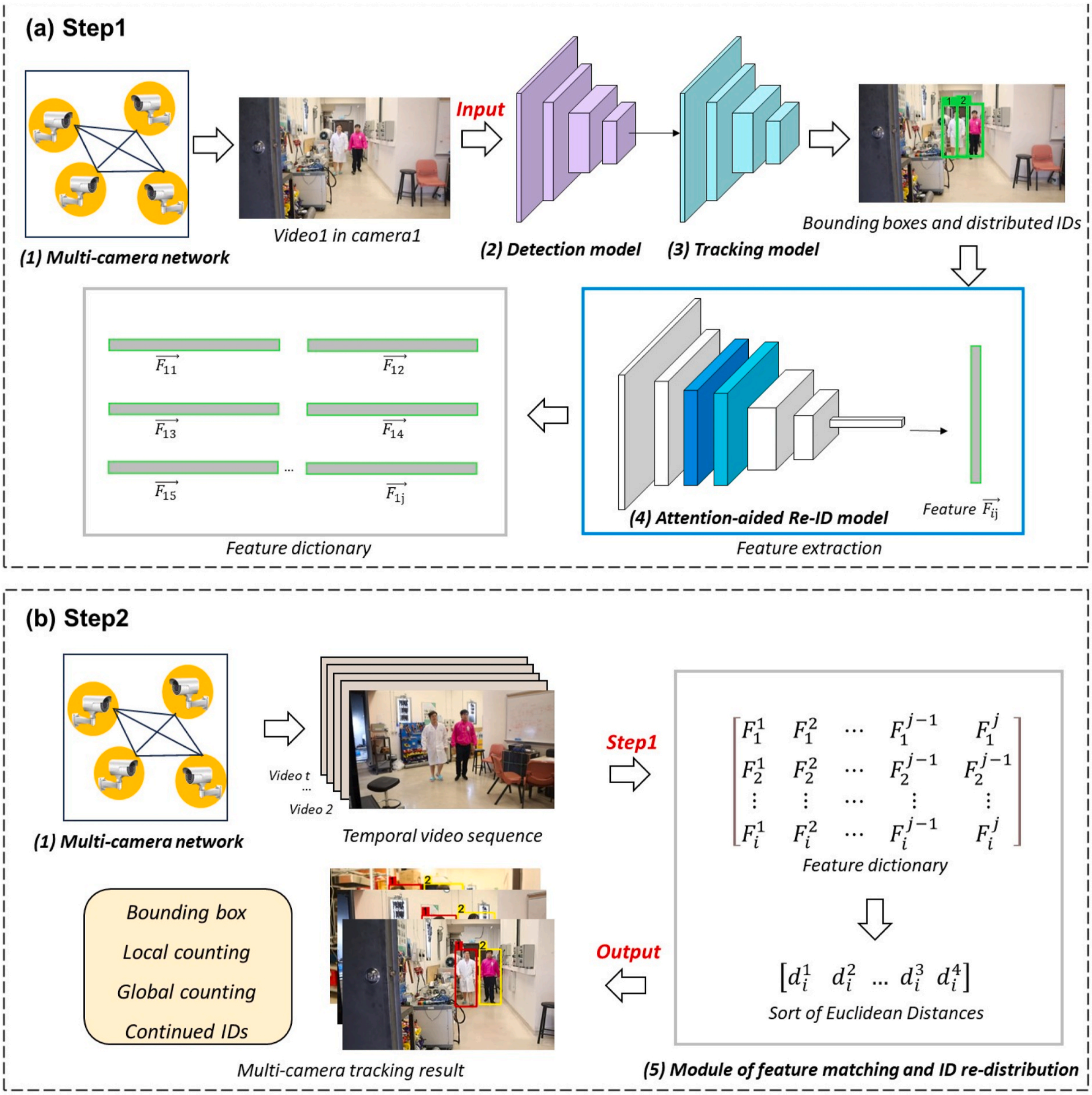

(a) Step1
(1) Multi-camera network
Video1 in camera1
Input
(2) Detection model
(3) Tracking model
Bounding boxes and distributed IDs
1
2
$\overrightarrow{F_{11}}$
$\overrightarrow{F_{12}}$
$\overrightarrow{F_{13}}$
$\overrightarrow{F_{14}}$
$\overrightarrow{F_{15}}$
...
$\overrightarrow{F_{1j}}$
Feature dictionary
(4) Attention-aided Re-ID model
Feature $\overrightarrow{F_{ij}}$
Feature extraction
(b) Step2
(1) Multi-camera network
Video t
...
Video 2
Temporal video sequence
Step1
$\begin{bmatrix} F_1^1 & F_1^2 & \cdots & F_1^{j-1} & F_1^j \\ F_2^1 & F_2^2 & \cdots & F_2^{j-1} & F_2^{j-1} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ F_i^1 & F_i^2 & \cdots & F_i^{j-1} & F_i^j \end{bmatrix}$
Feature dictionary
$[d_i^1 \ \ d_i^2 \ \ ... \ \ d_i^3 \ \ d_i^4]$
Sort of Euclidean Distances
Output
Bounding box
Local counting
Global counting
Continued IDs
2
Multi-camera tracking result
(5) Module of feature matching and ID re-distribution

Fig. 2. The flowchart of the overall methodology for multi-camera person *Re-ID* tracking. (a) Step-1: single video processing and feature extraction; (b) Step-2: multi-video processing, feature matching and Re-ID tracking.

personnel dynamic statistics. The evacuation director or rescue commander could make more comprehensive judgements and decisions of further operation benefiting from the dynamic outputs of the proposed framework.

3.2. Attention-aided Re-ID (AAR) model

In this paper, we proposed improved feature extraction network named Attention-aided Re-ID (AAR), which utilize the Osnet (Zhou et al., 2019b) as backbone and combine the spatial and channel attention mechanism to extract personnel specific features. As shown in Fig. 3, the AAR model extracts 512-unit feature vectors from a processing image with a size of 128 × 256. The structure of the proposed network is composed of five convolutional layers, two transition layers, one channel attention block, one spatial attention block, and one fully-connected layer. The conv2, conv3, and conv5 are composed of two omni-scale residual bottlenecks.

OSNet adopts an omni-scale residual block based on the residual bottleneck to realise the multi-scale feature learning. The baseline residual block aims to learn a residual feature $\widetilde{\boldsymbol{x}}$ via mapping function F when given the input $\boldsymbol{x}$, i.e.,

$$y = x + \widetilde{x} \quad with \quad \widetilde{x} = F(x) \tag{5}$$

Where $\boldsymbol{F}$ denotes a lite $\mathbf{3} \times \mathbf{3}$ convolutional layer to learn single-scale features with receptive field size (RFZ) for 3. In the OSNet block, multi-scale feature learning is used by fusing features with diverse RFZs controlled by multiple convolution layers with different depths as.

$$y' = x + \widetilde{x}' \quad with \quad \widetilde{x}' = \sum_{RFZ} F'(x), RFZ = 3,5,7,9 \tag{6}$$

In addition, the OSNet proposed a dynamic and unified aggregation gate (AG) to combine the different output streams in a dynamic way for learning effective omni-scale features. The aggregation gate is designed as a learnable neural network block to achieve dynamic scale feature fusion. Now, the omni-scale residual $\widetilde{\boldsymbol{x}}$ is formulated as

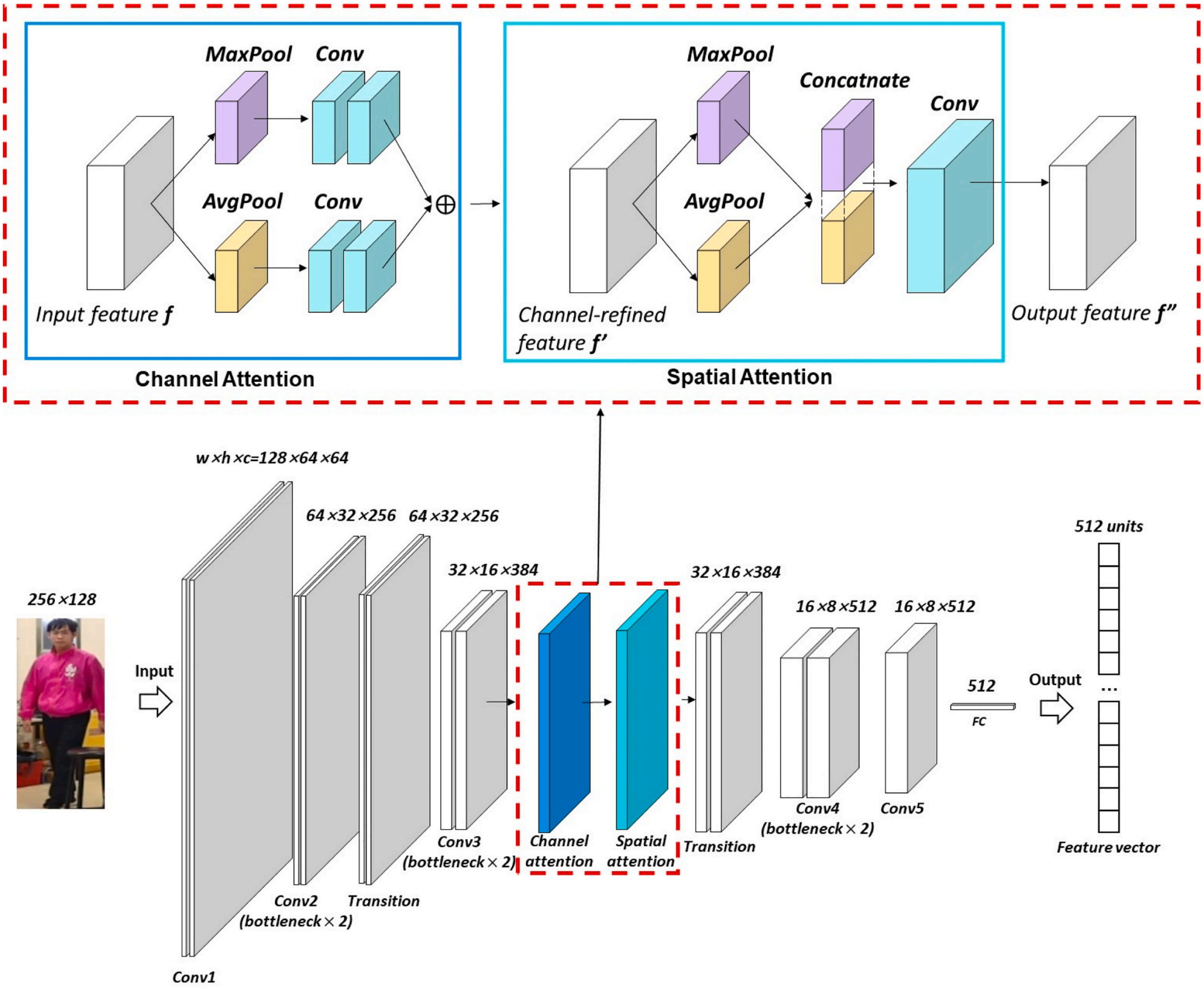


Fig. 3. The architecture of the proposed attention-aided Re-ID (AAR) model.

$$\widetilde{x}'' = \sum_{RFZ} G(F'(x)) \odot F'(x), RFZ = 3, 5, 7, 9 \tag{7}$$

$$y'' = x + \widetilde{x}'' \tag{8}$$

Where $\boldsymbol{G}(\boldsymbol{F}'(\boldsymbol{x}))$ is a data-conditioned vector with a length spanning the entire channel dimension of the input $\boldsymbol{F}'(\boldsymbol{x})$ and $\odot$ denotes the Hadamard product. $\boldsymbol{G}(\boldsymbol{x})$ is implemented as a mini-network composed of a non-parametric global average pooling layer.

To improve its robustness and performance, we introduce the scheme of channel attention and spatial attention into the OSNet structure for feature fusion, shown in Fig. 3. We denote the output feature of the OSNet block $\boldsymbol{y}''$ as the following input feature of the attention block $\boldsymbol{f}$, and $\boldsymbol{f}'$ denotes the channel-refined feature. $\boldsymbol{f} \bullet \boldsymbol{f}'$ represents inner product $\boldsymbol{f}$ and $\boldsymbol{f}'$, and this product is the input tensor of spatial attention block. $\boldsymbol{f}''$ denotes the spatial-refined feature. The convolutional attention scheme is computed as

$$f' = \sigma[MLP(AvgPool(f)) + MLP(MaxPool(f))] \tag{9}$$

$$f'' = \sigma\{Conv[Cat(Avg(f' \bullet f), Max(f' \bullet f))]\} \tag{10}$$

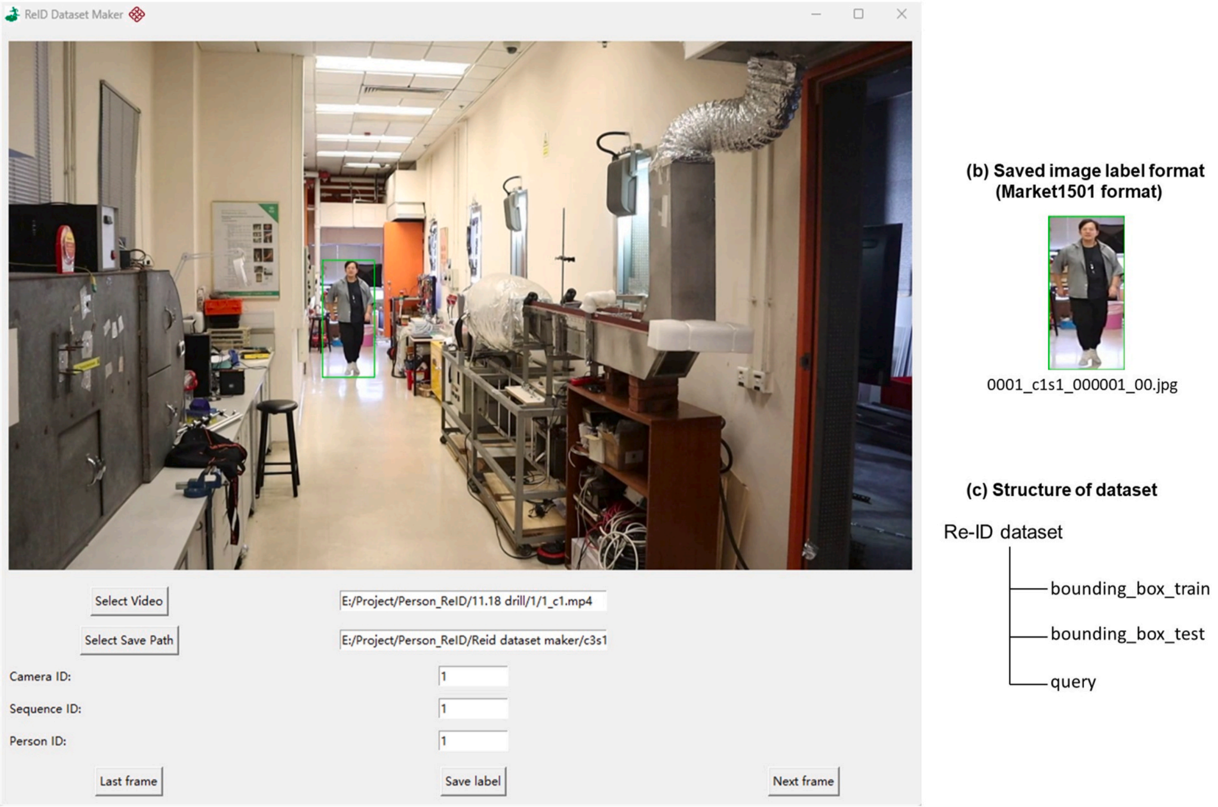
$$out = f'' \bullet (f' \bullet f) \tag{11}$$

Where σ denotes the sigmoid function, *MLP* denotes the multiple convolutional layer perceptron with one hidden layer. *AvgPool* and *MaxPool* denotes average pooling and maximum pooling to reduce the learnable parameter. *Conv* represents 2D convolutional computation that $Conv \in \mathbb{R}^{1\times7\times7}$ in this block. *Cat* denotes the *concatnate* computation meaning that joints two tensors together. *Avg* and *Max* denote computing the mean value and maximum value across columns of a tensor. *out* represents the final output of the attention block, which is an inner product of channel attention output $(f' \bullet f)$ and spatial attention output f''.

3.3. Custom Re-ID dataset

There is few dedicated Re-ID dataset for emergency evacuation scenarios, therefore, we collected some evacuation tests and fire drill videos and annotated a novel customised test dataset to further verify the feasibility of our proposed model. To make the dataset more convenient, we designed a novel annotation tool specifically for Re-ID dataset making and generating a standard data label format that is the same as Market1501. The interface of the proposed novel annotation tool is shown in Fig. 4(a), and its principle is illustrated in Algorithm III in Appendix.

Our dataset annotation tool automatically saves the image label with the standard Market1501 format, shown in Fig. 4(b). The label naming format contains person ID, camera ID, video sequence number, frame number, and code of annotation method ("00" represents manual annotation). After labelling all images, the annotation data is stored as the standard Re-ID dataset structure composed of three subfolders named "bounding_box_train", "bouding_box_test" and "query" and the former two subsets act as the training and test gallery set. The traditional Re-ID task is to retrieve the same identity from the gallery as the query



ReID Dataset Maker

Select Video: E:/Project/Person_ReID/11.18 drill/1/1_c1.mp4

Select Save Path: E:/Project/Person_ReID/Reid dataset maker/c3s1

Camera ID: 1

Sequence ID: 1

Person ID: 1

Fig. 4. The user interface of the developed dataset making tool and annotation example.

set. We use this software to make a novel Re-ID dataset, where the videos and images are all collected from the emergency drill tests. Totally 15 videos from 5 cameras and 487 images with 12 identities are annotated, some image examples are shown in F ig. A1. This mini-scale dataset is annotated as the test set to further evaluate the proposed Re-ID model and to perform the feasibility of the annotation software.

3.4. Module of feature matching and re-distribution

In this section, the proposed innovative module of feature matching and re-distribution is introduced, shown in Fig. 5 and Algorithm III in Appendix. D_{min} represents the minimum value of computed Euclidean distance and T represents the similarity threshold.

After the generation of new person feature from the Re-ID model, the proposed module would calculate the Euclidean distances of the new feature vector with all previous vectors and then sort the distances, and the minimum one is compared to a preset threshold value, which equals to 0.5 in this work. If the minimum distance is lower than the threshold, the module judges the corresponding feature is from the same person and it succeeds the previous ID. Otherwise, the module would assign a new ID for this person. This module realise the person Re-ID during multi-camera tracking that the same person remains a constant ID and the new person obtains a new ID, which not only eliminates the repeat counting of total evacuees, but also builds connection among distinct cameras.

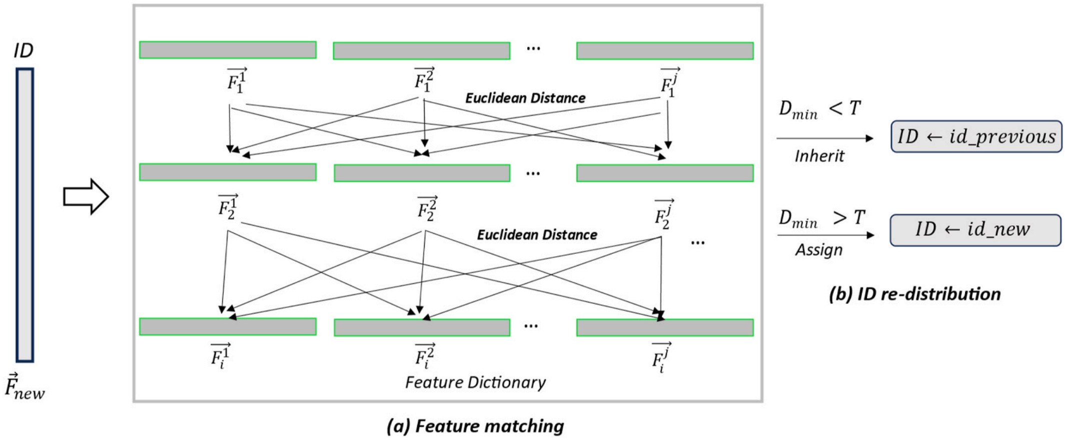


Fig. 5. The illustration of feature matching and ID re-distribution.

4. Improved attention-aided Re-ID model

4.1. Implementation details

The OSNet model and our proposed model use the linear fully-connected combing the softmax as the classification layer on the top layer. Each person's identity is regarded as a specific category to follow the standard image classification paradigm during the training process. As the supervision learning task, the training uses the cross-entropy loss with label smoothing as the loss function as

$$q_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for } i = 1, \ldots, K \text{ and } z = (z_1, \ldots, z_K) \in \mathbb{R}^K \tag{12}$$

$$p_i = \begin{cases} 1 - \varepsilon, \text{if i is true} \\ \frac{\varepsilon}{K-1}, \text{if i is flase} \end{cases} \tag{13}$$

$$\mathscr{L}(p, q) = -\sum_{i=1}^{C} p_i \log(q_i) \tag{14}$$

where $\boldsymbol{q}$ represents True, $\boldsymbol{p}$ represents Prediction and $\boldsymbol{\varepsilon}$ is a hyper-parameter. $\boldsymbol{K}$ is the number of person identities (e.g., $\boldsymbol{K} = \mathbf{1501}$ in Market1501 and $\boldsymbol{K} = \mathbf{1467}$ in CUHK03).

In addition, person matching during the training and test is based on the cosine distance using 512-unit feature vectors extracted from the last fully-connected layer. The training batch size and the weight decay are set to 32 and 5e-4 respectively. The learning rate is set to 0.0003 and the Adam optimiser is used. The data augmentation methods during the training select random flip and random crop. The total training epoch is set to 100. Table 2 summarize hyper-parameters for model training and testing. The models were trained in a server, and its hardware and software configurations are listed in Table 3.

In this work, we conduct two experiments to evaluate the proposed Re-ID model. One is the same-domain person Re-Identification experiment, where the models are trained and tested on the same dataset. In experiment 1, three classical datasets are used. The experiment 2, the model is pre-trained by benchmarks and tested by the customised evacuation Re-ID dataset.

4.2. Evaluation results

This section performs the evaluations of the proposed AAR model (OSNet + attention) in the conventional person Re-ID datasets and our custom evacuation dataset. To demonstrate the superiority of the proposed model, we conduct a series of experiments to compare it with other classical Re-ID models. For evaluation metrics, cumulative matching characteristics (CMC) rank accuracy, specifically Rank-k (R-k) and mean average precision (mAP) is used. R-k means the percentage of predictions where the top k prediction matches the ground truth label. mAP is a popular metric in evaluating the accuracy of object detection and classification as

$$AP = \int_0^1 p(r)dr \tag{15}$$

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{16}$$

Table 2
The summary of hyper-parameters for model training and test.

Max Epoch	Training batch size	Test batch size	Loss function
100	32	100	Cross-entropy loss
Optimiser	Initial learning rate	Pre-trained (Y/N)	Data augmentation
Adam	0.0003	Y	Flip & Crop

Table 3
The configuration of hardware and software for model experiments.

Item	Configuration
Hardware	CPU: Inter(R) Xeon(R) Gold 6252 CPU @ 2.10GHZ; GPU: NVIDIA Tesla V100 16 GB
Software	Linux user-ProLiant-DL380-Gen10 5.4.0 Ubuntu UTC x86_64 CUDA 11.6; Python 3.7.16; Torch 1.13.1; Torchreid 1.4.0; Opencv-python 4.8.1

Where, $\boldsymbol{AP}$ represents the average precision of each person's identity, defined as the area under the precision-recall curve above. K is still the number of person identities, same as Eq. (12).

In Experiment A, three popular Re-ID benchmarks are used to evaluate the proposed model, including Market1501, CUHK03, and DukeMTMC, where mAP and R-1 are used as metrics. The overall dataset statistic is shown in Table 1. The evaluation results of Experiment A are shown in Table 4, which illustrates that OSNet achieves the best overall performance compared with other published models (Densenet-20, HA-CNN, ResNet-50, and Shufflenet) and almost reaches the saturation in the Market1501 and DukeMTMC. After combining the spatial attention and channel attention, our model achieves significant improvements in three benchmarks and obviously surpasses the original OSNet in both mAP and R1 metrics. Our model proves that the convolutional attention block also contributes to better Re-ID feature fusions to optimise the feature extraction in the Re-ID task.

In Experiment B, we use our test dataset to test the pre-trained classical models (Densenet-20, HA-CNN, ResNet-50, Shufflenet, and OSNet) and our proposed model trained by Market1501. The results are shown in Table 5, which shows that our model also achieves the best performance among all tested models. Besides mAP and Rank-1, Rank-5, Rank-10, and Rank-20 are used in this experiment as the evaluation metrics that are widely used in the person Re-ID research. Our model achieves superiority in most metrics when tested by our customised dataset, which also illustrates the feasibility and robustness of our dataset in the person Re-ID task.

4.3. Ablation study

Fig. 6 provides an illustrative depiction of the diverse position choices of spatial attention and channel attention block. The primary model is model 1 incorporating an attention block positioned behind Conv3, which also serves as the second Bottleneck of the OSNet framework. Model 2, model 3, model 4, and model 5 individually explore the insertion of an attention module at various positions of the original structure, while ensuring that the overall parameters of the neural network remain unaltered. As for model 6, we attempted to combine the attention module and the OSBlock to create a new bottleneck, which makes the total number of trainable parameters for characteristic learning increase.

Table 6 evaluates the above distinct architecture designs of our Re-ID model with an attention scheme in the benchmarks. The results show that the primary model achieves the best performance among the tested models in both mAP and Rank-1. However, positioning the attention module behind the first two layers or the last two layers leads to unsatisfactory results. The reason could be placing attention too forward may result in insufficient feature fusion, i.e., underfitting, while placing it in the last few layers may result in too much feature fusion, i.e., overfitting. Therefore, placing it in the middle position would achieve the most optimal optimization results. As for model 6 which combines the attention and the OSBlock, its performance is close to the primary model, but it significantly increases the overall parameters which is detrimental to inference and training speed. Above all, model 1 is selected as the primary Re-ID model for the overall tracking framework.

Table 4
Experiment A- Results (%) comparison of our model on classical big Re-ID datasets.

Model	Publication	Market1501		CUHK03		DukeMTMC	
		mAP	R-1	mAP	R-1	mAP	R-1
Densenet-201	CVPR	61.0	80.8	39.6	40.5	44.4	66.4
HA-CNN	CVPR	67.0	85.5	32.6	31.0	58.3	74.8
ResNet-50	CVPR	67.3	82.6	36.7	37.3	56.1	74.4
Shufflenet	CVPR	42.2	66.2	13.3	12.6	32.9	52.4
OSNet	ICCV	76.7	92.2	46.0	47.4	65.6	83.3
AAR (ours)	/	**76.8 ↑**	**91.9 ↓**	**46.1 ↑**	**47.6 ↑**	**66.1 ↑**	**83.6 ↑**

Table 5
Experiment B- Results (%) comparison of our model on our custom small test dataset.

Model	Publication	Test dataset in this work				
		mAP	R-1	R-5	R-10	R-20
Densenet-201 (Huang et al., 2017)	CVPR	13.9	19.2	36.5	50.6	64.7
HA-CNN (Li et al., 2018b)	CVPR	7.8	3.8	12.2	21.2	42.9
ResNet-50 (He et al., 2016)	CVPR	12.2	20.5	38.5	44.2	59.6
Shufflenet (Zhang et al., 2018)	CVPR	7.9	1.3	11.5	21.8	53.2
OSNet (Zhou et al., 2019b)	ICCV	37.2	51.3	**70.5**	**80.8**	92.9
AAR (ours)	/	**38.5 ↑**	**51.9 ↑**	**62.2 ↓**	**76.3 ↓**	**92.9**

5. Demonstration of real-time multi-camera tracking

5.1. Test setup

In this section, we conduct a simplified simulated evacuation scenario to demonstrate the real-time performance of our proposed multi-camera tracking framework. The demonstration is conducted in a university complex and a laboratory is assumed to occur in an emergency accident, i.e., building fire, seismic, or chemical explosion. A total of 5 participants played the roles of the evacuees and evacuated immediately along the safety route from Foor N to Floor N-1 (assumed as Refuge Floor). There are three cameras located in the critical components of the evacuation route. The detailed layout of camera locations and route illustration are shown in Fig. 7. Specifically, Camera 1 is located at the laboratory gate to monitor the evacuees departing the first site of the emergency site. After leaving the most dangerous location, the participants would enter the view of Camera 2 located at the entrance of the staircase, which is the critical midway of the evacuation route. Camera 3 is in the staircase platform between Floor N and Floor N-1. After the participants successfully pass Camera 3, they are determined to complete the evacuation safely.

The videos from the three camera views are used to test the evacuee Re-ID monitoring performance of our proposed model. The first content is testing the person re-identification tracking accuracy across three cameras. The ideal target is that participants go across distinct camera views retaining a constant individual ID. The second content is testing the performance of dynamic statistics of personnel counting, which contains local counting in single camera view and global counting across all cameras.

5.2. Performance of person Re-ID

In this section, the performance of person re-identification during tracking process is tested. The proposed AAR model that combines the OSNet backbone and the attention block is used to extract the distinct individual features. The model is re-trained by a united dataset that combines Market1501, CUHK03, and DukeMTMC for stronger robustness in new testing data. The training process is shown in F ig. A2, which illustrates that the training accuracy achieves over 99.8% and the loss decreases to 1.17 after 100 training epochs. After that, the weights of the well-trained model are embedded in the framework to replace the pre-trained weights.

The recorded videos from three camera views are input into the Re-ID tracking system. The system output is the real-time tracking result, which consists of the bounding boxes with pixel coordinates of each person, the Re-ID tracking results that are the distributed ID and re-

Table 6
Ablation study on position choices of attention block.

Model	Position of attention block	mAP	R-1
1	**Conv3+attention (primary)**	**76.8**	**91.9**
2	Conv1+attention	75.5	90.9
3	Conv2+attention	76.0	91.3
4	Conv4+attention	75.5	90.9
5	Conv5+attention	75.2	90.5
6	Attention in OSBlock	76.5	91.7

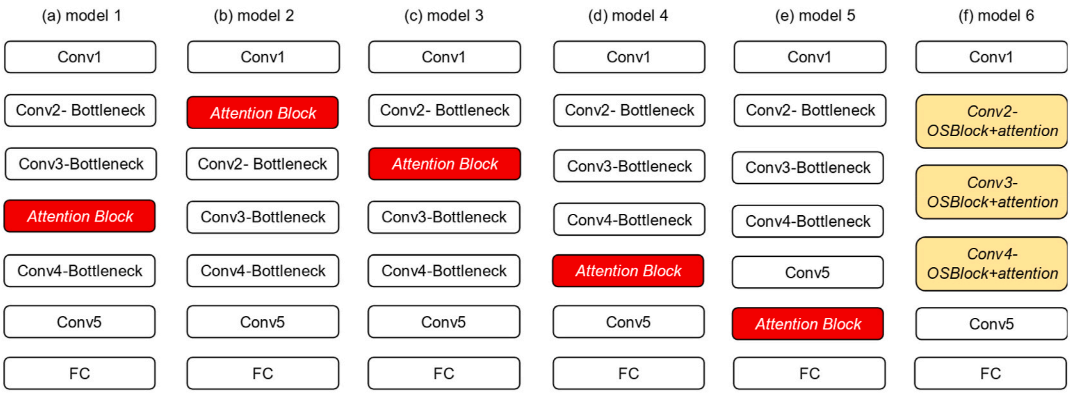


Fig. 6. Different position choices of attention block.

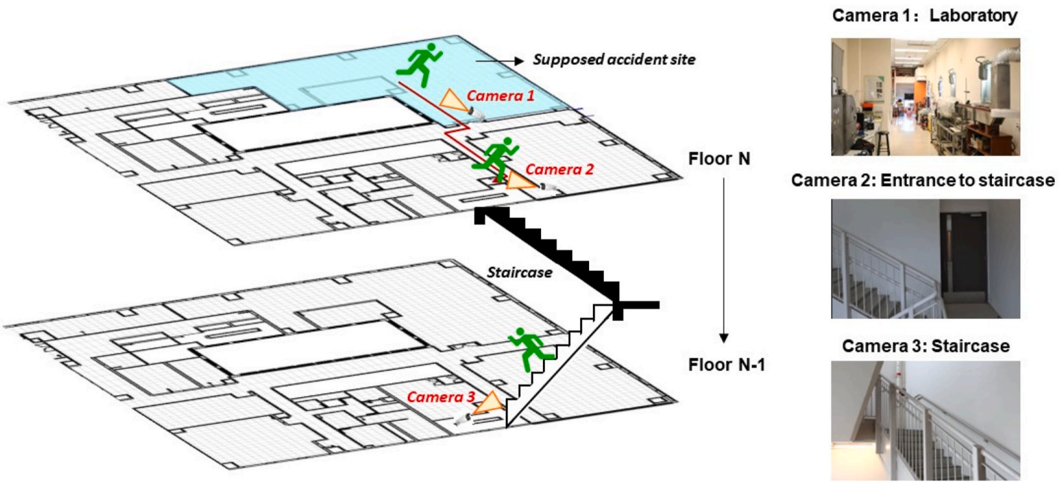


Fig. 7. Camera layout and experimental egress route.

distributed ID, as well as the person counting. Some examples of true Re-ID tracking results are shown in Fig. 8, and the false re-identification examples are shown in Fig. 9. In Fig. 8(b), the initial IDs of five participants distributed by Camera 1 are 1, 2, 4, 5, 7, and each ID number corresponds to one constant colour. In Camera 2 and Camera 3, the true re-identification instances from Fig. 8(c and d) show that five participants are re-distributed with the same ID number and corresponding box colour as the initial ID in Camera 1.

However, the participants in some tracking video frames also be erroneously identified as different ID numbers with Camera 1, which is regarded as false Re-ID tracking. For instance (Fig. 9), the participant with initial ID 2 is recognised as ID 7 during Camera 2, and ID 1 is mistakenly recognised as ID 4 in Camera 3. Commonly, the tracked participants retain the true ID in most video frames, but sometimes several frames perform false tracking results. To evaluate the video Re-ID tracking results more scientifically, we proposed a new metric named Re-ID Tracking Accuracy (RTA) for each person, shown as Eqs. 17 and 18. *RTA* is calculated as the number of correctly Re-ID frames (*TRF*) divided by the total number of tracking frames (*TTF*).

$$RTA_i = \frac{TRF_i}{TTF_i} \times 100\% \tag{17}$$

$$Overall\ RTA = \frac{\sum TRF_i}{\sum TTF_i} \times 100\% \tag{18}$$

where, RTA_i represents Re-ID tracking accuracy in view camera i; TRF_i represents the frames of true Re-ID tracking in camera i; TTF_i represents the total tracking frame of one person in camera i.

The quantitative Re-ID results are shown in Table 7, which illustrates the Re-ID accuracy of each participant across distinct cameras. It shows that the overall RTA of each tracking object exceeds 53%, and the best *RTA* score achieves 100%. ID 4 and ID 5 achieve outstanding results both in Camera 2 and Camera 3, and the overall *RTA* exceeds 92% as well. Some tracking individual performs well in one camera but is identified worse in the other camera. For example, the Re-ID system always confused ID 2 and ID 7, and ID 1 was erroneously determined as ID 4 in many frames of Camera 3. The detailed reasons for the error will be discussed in Section 6.1. Overall, the testing Re-ID tracking performance is satisfactory where the average overall RTA exceeds 75%, and the error is acceptable.

5.3. Performance of personnel counting

In this section, we illustrate the performance of another significant function of the proposed monitoring framework which is dynamic

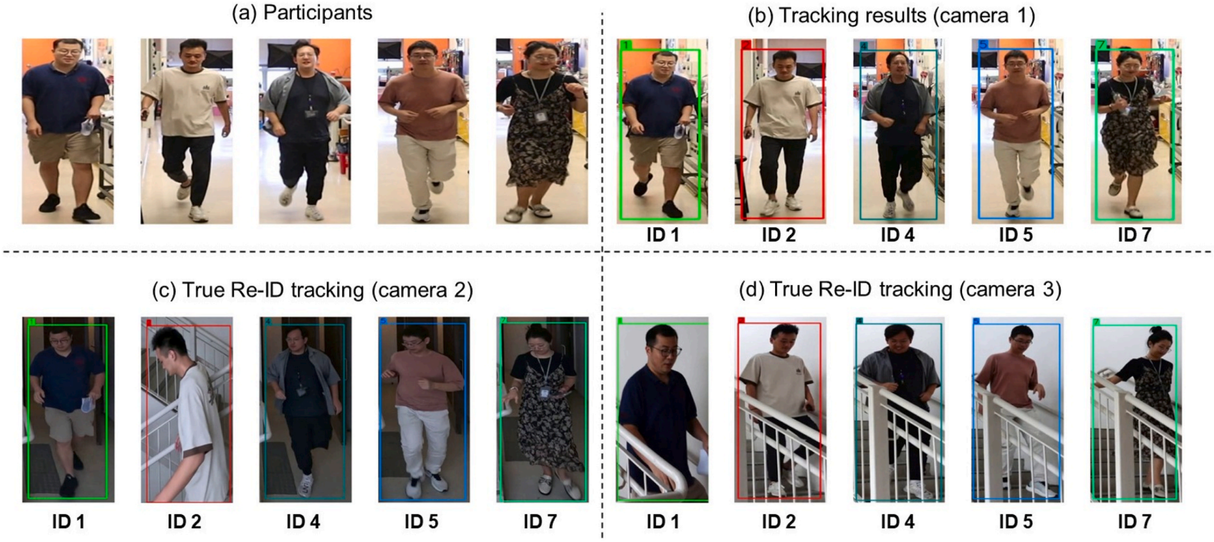

(a) Participants
(b) Tracking results (camera 1)
ID 1
ID 2
ID 4
ID 5
ID 7
(c) True Re-ID tracking (camera 2)
(d) True Re-ID tracking (camera 3)
ID 1
ID 2
ID 4
ID 5
ID 7
ID 1
ID 2
ID 4
ID 5
ID 7

Fig. 8. Examples of true Re-ID tracking results.

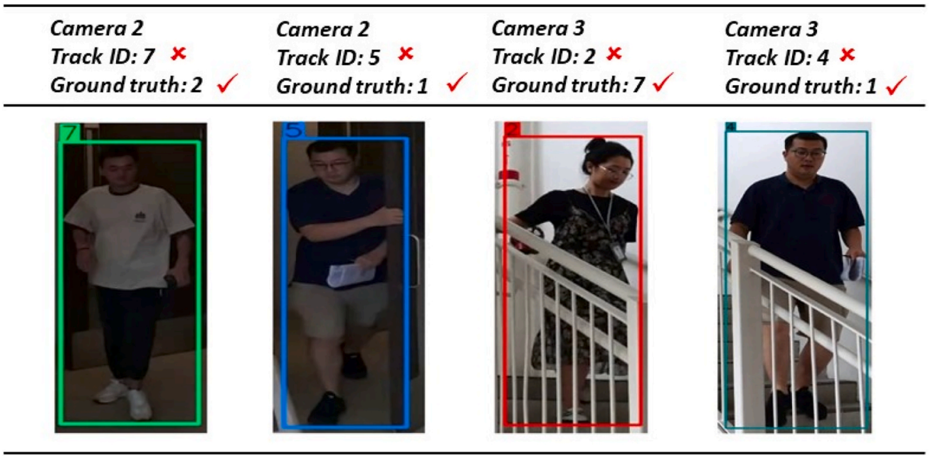


Fig. 9. Examples of false Re-ID tracking results.

Table 7
The quantitative statistic of Re-ID tracking results.

ID	Camera 2			Camera 3			Overall RTA (%)
	TRF	TTF	RTA (%)	TRF	TTF	RTA (%)	
1	110	130	84.6	53	174	30.5	53.6
2	41	122	33.6	160	164	97.6	70.3
4	62	67	92.5	171	185	92.4	92.5
5	93	93	100.0	165	183	90.2	93.5
7	115	126	91.3	72	159	45.3	65.6

personnel counting. The statistical result is shown in Fig. 10, where the entire test time of the evacuation process is divided into three periods, when the tracked crowd passes Camera 1, Camera 2, and Camera 3, respectively. The red curve represents the global counting to track the total number of being tracked evacuees during the test process. It obviously shows that the global counting increases gradually with the participants across Camera 1 and maintains the stable value of five during the following time. The dash curves represent the local counting that how many people pass each camera at every moment. The local counting statistic fluctuates with the dynamic variation of the number of people detected by each camera.

The personnel counting is automatically calculated by our proposed algorithm, and its accuracy has been proved by comparing the statistical results with the real video observations. The global counting of the entire evacuation process and the local counting in each critical component along the egress route are vitally significant for evacuation monitoring. For example, the local counting could provide how many people have left the danger area or have passed the safety exit. The global counting could provide how many people undergo evacuation and the overall progress of the evacuation.

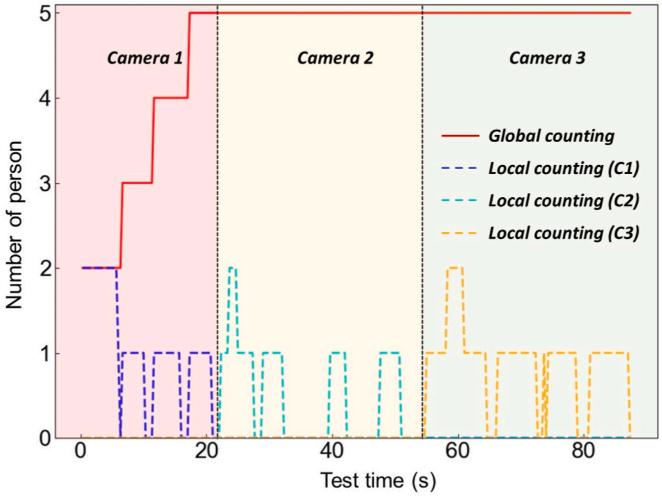


Fig. 10. Dynamic statistic of automatic personnel counting.

6. Model explainability discussion and perspective

6.1. Explainable error analysis by class activation maps

In this section, the class activation map (CAM) (Zhou et al., 2016) method is applied to explain the reason for the re-identification errors and the challenges of the CNN-based framework. CAM displays a visual heatmap to illustrate the most notable features that the CNN model pays the most attention to. Therefore, the CAM could be used to interpret the prediction results of a CNN model. The last convolutional layer contains the richest spatial and semantic information which therefore could be taken full advantage of by CAM. Specifically, the algorithm of CAM computes the weighted mean value of all feature maps as the output. We denote the feature map of a class output by each convolution layer as M_k^c, and its corresponding weight is w_k^c, so the CAM of each class is formed as

$$CAM^c = \sum_k w_k^c \times M_k^c \tag{19}$$

where c represents a class that each person's identity is regarded as a specific class in the Re-ID task; k represents the number of the convolutional layers of the model.

In the demonstration result shown in Table 7 in Section 5.2.1, the Re-ID model always confused the ID 2 and ID 7 in Camera 2, as well as ID 1 and ID 4 in Camera 3. Therefore, we use the class activation map of these participants to analyse the error reason. In Fig. 11(a), the heatmap gathers on the black collar area and the black logo of the white T-shirt, where the white-black feature is extracted. In Fig. 11(b), the heatmap gathers on the white-black dress, so the most significant feature is also the white-black piece. Therefore, the white-black piece was determined as the same characteristic between the two people and mistakenly recognised as the same identity class. Similarly, the belongings taken by the person could also become the primary focus of feature extraction. For example, a roll of white paper taken by ID 1 in Fig. 11(c) is extracted by the Re-ID model in Camera 3, and this feature is coincidently determined as the same as the white ornament in ID 4 of Fig. 11(c), which results in the false Re-ID results of these two identities.

Apart from the problem of misidentification caused by clothing and ornament, another limitation is that the model is likely to focus on some conspicuous but irrelevant objects in the image background. For

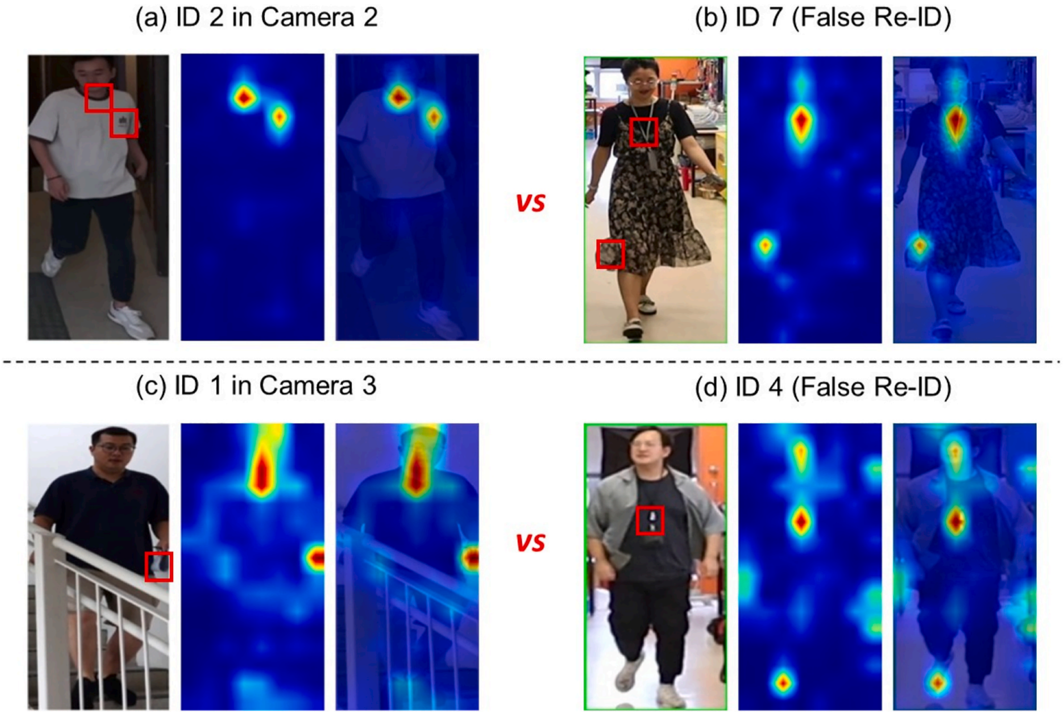


Fig. 11. Illustration of the error analysis by class activation maps. The red box represents the focused positions. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

example, a chair and a red object are extracted in Fig. 12(a), and the obvious white object is extracted in Fig. 12(b), as well as the red footing is focused on Fig. 12(c). Therefore, the background factors have a significant side effect on the feature extraction and identification of the CNN model. In addition, some variations of other factors, i.e., lighting conditions, camera angle and focal length, gesture, and posture of the person, significantly impact the Re-ID performance.

6.2. Perspective of the application in a digital twin system

Even though the existence of the above limitations and problems, our proposed Re-ID tracking framework is a creative method for smart evacuation monitoring and contributes to the application of visualised digital twin system. To better introduce our related perspective, we design a user interface demonstration of an ideal emergency digital twin system shown in Fig. 13. The core of this digital twin system is driven by the person Re-ID tracking framework in this paper. The computer vision algorithm and building modelling contribute to the visualization of data mapping. The multi-modal sensors, i.e., temperature and humidity sensors, CCTV cameras, broadcast, fire alarm, and corresponding wireless and wired communication networks comprise the IoT module to strengthen our system.

As for the specific functions of the emergency digital twin system, firstly, it could more intuitively display the evacuation process, automatically monitor the evacuation safety, and integrate the evacuation information. Fig. 13 shows the intelligent real-time monitoring interface based on our proposed Re-ID tracking framework, which can automatically detect, track, and re-identify occupants and compute the number of evacuee populations (see a demonstration in Video S1). The real surveillance and tracking views could also be selected and switched conveniently. In the same, the tracking results containing trajectories and corresponding IDs of each occupant would be synchronously mapped on the virtual interface of the internal building. Through the virtual building view, the dynamic evacuation conditions on every floor could be intuitively monitored by the evacuation and rescue directors outside the building. Apart from that, the system could also monitor emergency risk through the sensor data and surveillance interface, and promptly activate alarm and issue evacuation response by call and broadcast once an emergency accident occurs.

7. Conclusions

The real-time evacuation tracking by surveillances has vital significance for monitoring safety and emergency management inside complex

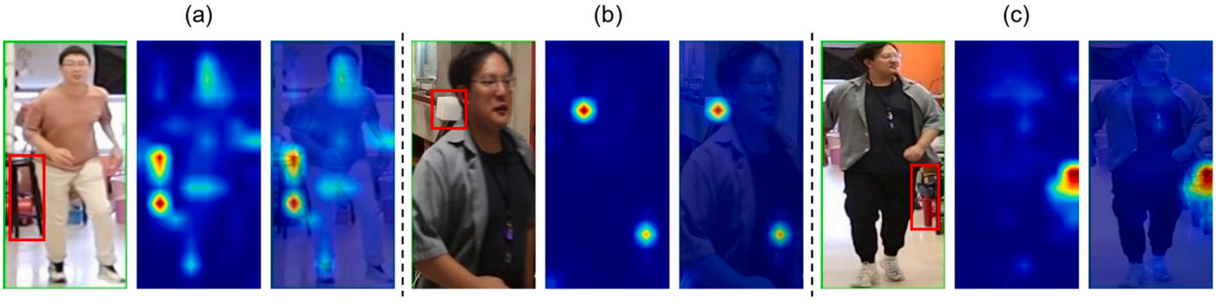
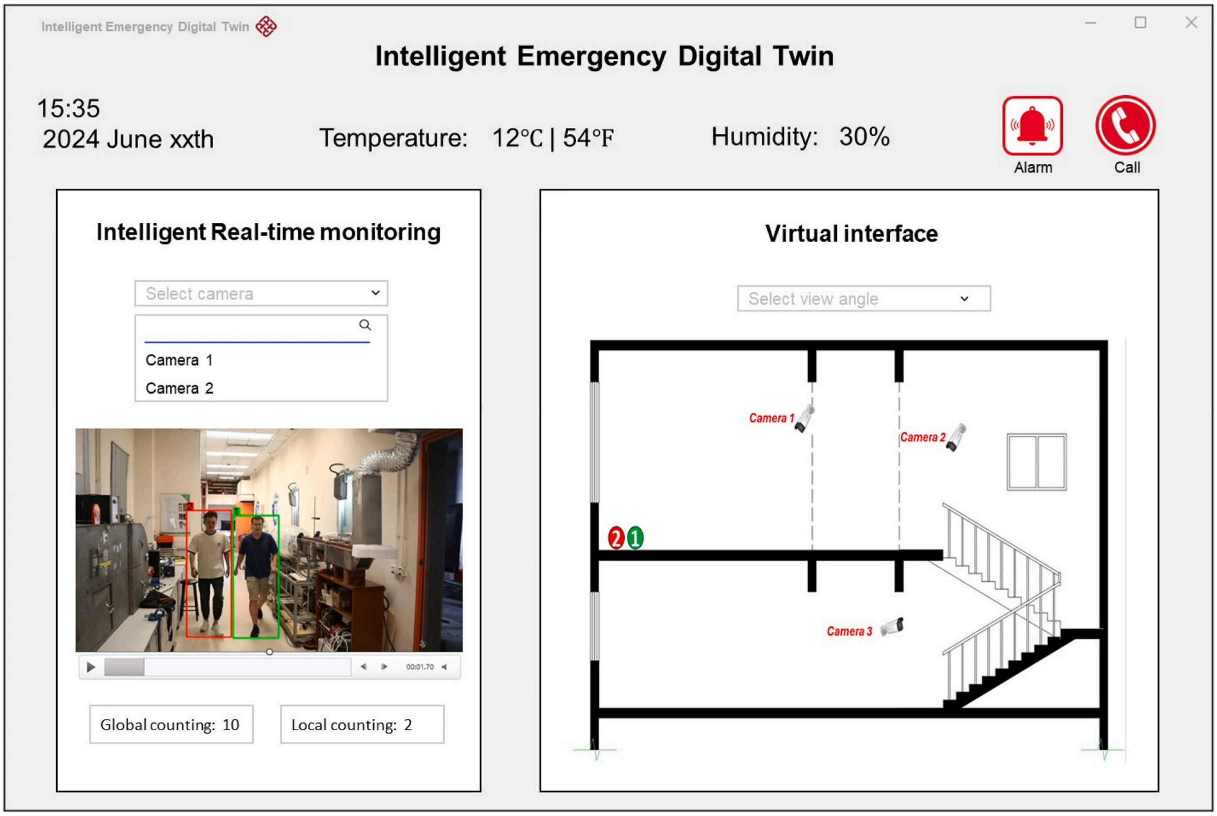


Fig. 12. Examples of background feature impact by using class activation maps with red box representing focused objects. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)



15:35
2024 June xxth
Temperature: 12°C | 54°F
Humidity: 30%

Global counting: 10
Local counting: 2

Fig. 13. User interface illustration of the proposed digital twin monitoring system (see Video S1).

buildings. However, traditional vision-based evacuation monitoring methods using single-camera tracking have difficulty in integrating information from different cameras. In this study, we propose a novel real-time multi-camera tracking framework for evacuation safety monitoring based on an improved explainable re-identification (Re-ID) model. Some main innovations and contributions are as followings.

(1) We proposed an innovative tracking evacuation safety framework consisting of multi-camera network, a detection model, a tracking model, an explainable Re-ID model, and a module of feature matching and ID re-distribution algorithm.
(2) To enhance the feature extraction ability of the Re-ID model, we proposed an attention-aided network, and the proposed model demonstrated outstanding performance on both standard large-scale benchmarks and our custom testing dataset.
(3) Additionally, we developed a novel Re-ID dataset annotation tool for more efficiently recognition model research and created a mini-scale testing dataset using this tool.
(4) To showcase the real-time multi-camera tracking capability of the proposed methodology, we conducted a simple evacuation drill, and the results indicated that our method achieved satisfactory accuracy in Re-ID and personnel counting.
(5) Furthermore, we employed class activation maps to demonstrate the explainability of our proposed model and highlight its limitations and potential improvement areas.

Overall, our approach enables the tracking and re-identification of individuals across different camera views which significantly enhances surveillance capabilities. The proposed multi-camera Re-ID tracking framework contributes to the development of automated monitoring system and intelligent digital twin for building emergency safety management. However, the proposed framework is limited by computational ability for processing multi-videos in parallel, and Re-ID accuracy is also affected by video angles, similar feature disturbance, and environmental factors. Moreover, the current attention scheme is still self-driven instead of manually controllable that is challenging to explain why certain features are deemed important by the model and the complex interactions between attention weights and input data. Additionally, the thick smoke condition also significantly hinders the system's effectiveness, which is a huge challenge in fire emergency application. In our future work, this Re-ID tracking framework could work with more diverse and advanced perception mechanisms, e.g., large language model and multimodal fusion of vision, radar, and infrared to enhance the feature recognition abilities in complex and challenging application scenarios.

CRediT authorship contribution statement

Yifei Ding: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Xinghao Chen:** Writing – original draft, Investigation, Formal analysis. **Yuxin Zhang:** Writing – review & editing, Supervision, Funding acquisition. **Xinyan Huang:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

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 work is funded by the Hong Kong Research Grants Council Theme-based Research Scheme (T22-505/19-N), the National Natural Science Foundation of China (52204232), and MTR Research Fund

(PTU-23005). The authors also thank all student volunteers from PolyU Fire Lab for supporting the demonstration. The research has obtained the ethical approval from the Hong Kong Polytechnic University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.engappai.2025.110394.

Appendix

Algorithm I. The procedure of the proposed Re-ID tracking framework (in Python)

```
Input: To-be-processed videos
Output: The Re-ID tracking results
Step1:
input video1 do
    # processed by detection model (YOLOv7)
    for person in video1 do
        # processed by tracking model (DeepSORT)
        # processed by Re-ID model
        $result \leftarrow [x1, x2, y1, y2, id, f]$ # bounding box coordinate, person id, feature vector
        # build a dictionary with id and feature
        $feature\_dictionary \leftarrow \{id : f\}$ # id is the key and feature is the value of this dictionary
    end
Step2:
for video in other camera view do:
    # processed by detection model
    local_count ← 0 # total evacuee number of this video
    for person in video do
        local_count + ← 1
        # processed by tracking model
        # processed by Re-ID model
        $result \leftarrow [x1, x2, y1, y2, id\_new, f\_new]$
        # compute Euclidean distance of each feature pair
        # features matching with the feature dictionary
        $Global\_count \leftarrow len(feature\_dictionary)$ # total evacuee number of whole process
        $final\_results \leftarrow [bounding\ box, local\ count, Global\ count, IDs]$
return final_results
```

Algorithm II. The scheme of channel attention and spatial attention (in Python)

```
def ca(x): # channel attention
    # processed by pooling layers, convolutional layers, and fully-connected layers
    return ca(x)
def spatial_attention(x): # spatial attention
    # processed by pooling layers, convolutional layers, and fully-connected layers
    return sa(x)
def channel_spatial_attention(x):
    out1 = x*ca(x)
    out2 = out1*sa(out1)
    return out2
```

Algorithm III. The Re-ID dataset annotation tool (in Python)

```
Input: To-be-annotated video; label save path
Output: The annotated images and labels
cam_id← input (“the camera ID”)
sequence_id← input (“the sequence ID of the annotating video”)
while true do:
    for person in video frame do: # annotate every person in the video frame
        # Draw bounding box for each person by tkinter
        cv2.rectangle (frame, start_point, end_point)
        # Save the image in the bounding box
        image ← frame[start_point[1] : end_point[1], start_point[0] : end_point[0]]
        person_id← input (“the person ID”)
        # rename the label as the standard format, like 0001_01_01_000025_00.jpg
        name ← f“{person_id : 04d}_{cam_id}_{sequence_id}_{int(frame) : 06d}_00.jpg”
end
```

Algorithm IV. The pseudo code of feature matching and ID re-distribution

```
Input: feature dictionary formed by Step 1
# features matching with the feature dictionary
DIS = [ ] # Create a list to save all compared distance among new feature and dictionary
for id_previous in feature_dictionary do:
    distance←eucidean_distance (f, f_new)
    DIS.append(distance) # add each distance in the list
sort (DIS)
if DIS[0] < threshod do # compare the minimum distance with the threshold
```

(continued on next page)

(*continued*)

Algorithm IV. The pseudo code of feature matching and ID re-distribution

```
ID←id_previous # inherit previous ID
if DIS[0] > threshod do # compare the minimum distance with the threshold
ID←id_new # assign a new ID
return ID
```

F ig.A1. shows the examples of our custom Re-ID dataset which is generated by our novel annotation tool, and the pixel size of each image is 128 × 256.

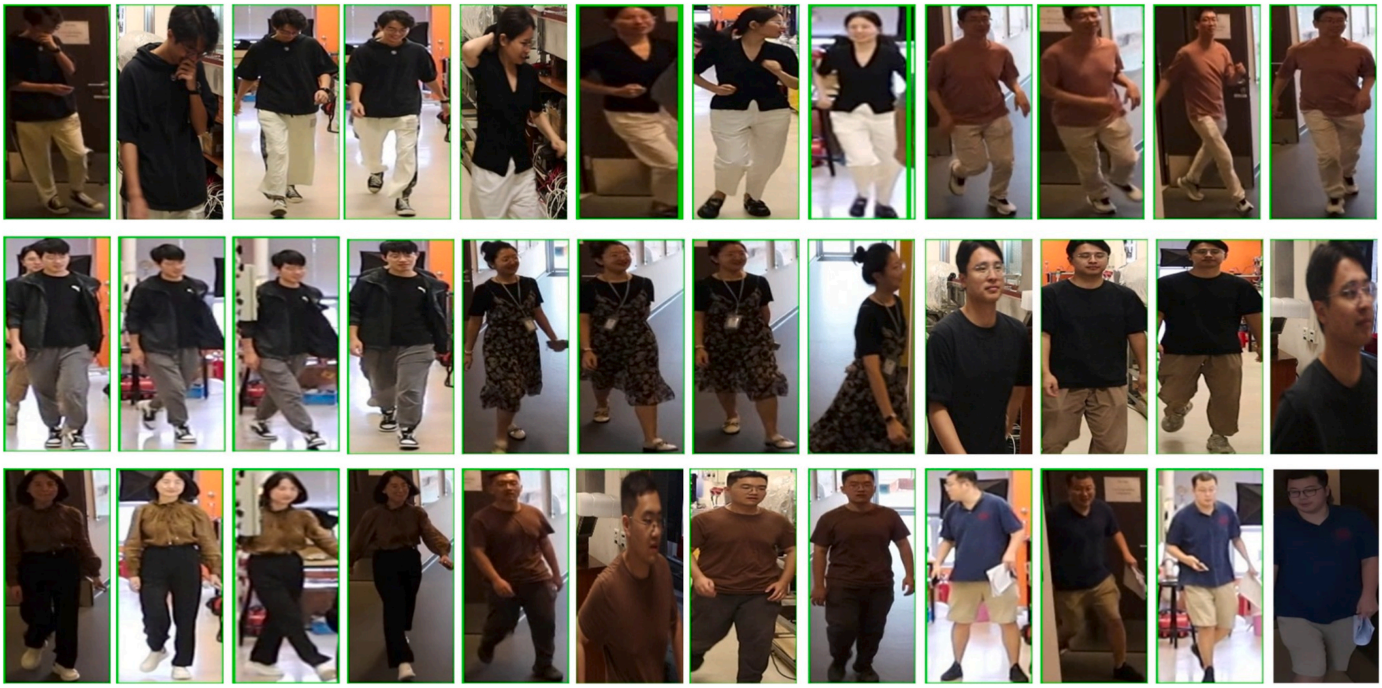


Fig. A1. The examples of custom Re-ID test dataset.

F ig.A2. shows the training results of our AAR network, and the used train dataset was combined with Market 1501, CUHK03 and DukeMTMC.

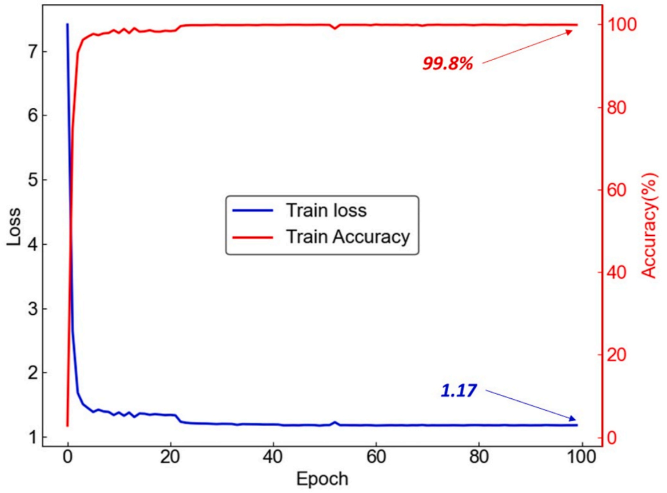


Fig. A2. Training process of our proposed model on the combined dataset.

Data availability

Data will be made available on request.

References

Baduge, S.K., Thilakarathna, S., Perera, J.S., Arashpour, M., Sharafi, P., Teodosio, B., et al., 2022. Artificial intelligence and smart vision for building and construction 4.0: machine and deep learning methods and applications. Autom. ConStruct. 141, 104440.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al., 2020. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking. 2016 IEEE Int. Conf. Image Process., pp. 3464–3468.

Chen, Y., Hu, S., Mao, H., Deng, W., Gao, X., 2020. Application of the best evacuation model of deep learning in the design of public structures. Image Vis Comput. 102, 103975.

Cheng, M.Y., Chiu, K.C., Hsieh, Y.M., Yang, I.T., Chou, J.S., Wu, Y.W., 2017. BIM integrated smart monitoring technique for building fire prevention and disaster relief. Autom. ConStruct. 84, 14–30.

Cheng, J.C.P., Chen, K., Wong, P.K.-Y., Chen, W., Li, C.T., 2021. Graph-based network generation and CCTV processing techniques for fire evacuation. Build. Res. Inf. 49, 179–196.

Dang, P., Zhu, J., Cao, Y., Wu, J., Li, W., Hu, Y., et al., 2024. A method for multi-person mobile virtual reality fire evacuation drills based on pose estimation: consistency of vision and perception. Saf. Sci. 170, 106334.

Deng, H., Tian, M., Ou, Z., Deng, Y., 2022. A semantic framework for on-site evacuation routing based on awareness of obstacle accessibility. Autom. ConStruct. 136, 104154.

Ding, W., Abdel-Basset, M., Hawash, H., Ali, A.M., 2022. Explainability of artificial intelligence methods, applications and challenges: a comprehensive survey. Inf. Sci. 615, 238–292.

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

Ding, Y., Chen, X., Wang, Z., Zhang, Y., Huang, X., 2024. Human behaviour detection dataset (HBDset) using computer vision for evacuation safety and emergency management. J. Saf. Sci. Resil. https://doi.org/10.1016/j.jnlssr.2024.04.002.

Dong, H., Zhou, M., Wang, Q., Yang, X., Wang, F.-Y., 2020. State-of-the-Art pedestrian and evacuation dynamics. IEEE Trans. Intell. Transport. Syst. 21, 1849–1866.

Gheissari, N., Sebastian, T.B., Hartley, R., 2006. Person reidentification using spatiotemporal appearance. 2006 IEEE Comput. Soc. Conf. Comput. Vis. pattern Recognit 2, 1528–1535. IEEE.

Girshick, R., 2015. Fast r-cnn. Proc. IEEE Int. Conf. Comput. Vis., pp. 1440–1448.

Gray, D., Tao, H., 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. Comput. Vision–ECCV 2008 10th Eur. Conf. Comput. Vision 262–275. Marseille, Fr. Oct. 12-18, 2008, Proceedings, Part I 10, Springer.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. pattern Recognit. 770–778.

Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H., 2011. Person re-identification by descriptive and discriminative classification. Image Anal. 17th Scand. Conf. SCIA 2011, Ystad, Sweden, May 2011. Proc. 17. Springer, pp. 91–102.

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X., 2019. Interaction-and-aggregation network for person re-identification. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 9317–9326.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. Proc. IEEE Conf. Comput. Vis. pattern Recognit. 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proc. IEEE Conf. Comput. Vis. pattern Recognit. 4700–4708.

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, 106285.

Ibrahim, A.M., Venkat, I., Subramanian, K.G., Khader, A.T., Wilde, P De, 2016. Intelligent Evacuation Management Systems: A Review, vol. 7. ACM Trans Intell Syst Technol.

Khlevnoi, O., Burak, N., Borzov, Y., Raita, D., 2022. Neural network analysis of evacuation flows according to video surveillance cameras. Int. Sci. Conf. "Intellectual Syst. Decis. Mak. Probl. Comput. Intell. Springer, pp. 639–650.

Li, W., Wang, X., 2013. Locally aligned feature transforms across views. Proc. IEEE Conf. Comput. Vis. pattern Recognit. 3594–3601.

Li, W., Zhao, R., Wang, X., 2013. Human reidentification with transferred metric learning. Comput. Vision–ACCV 2012 11th Asian Conf. Comput. Vision, Daejeon, Korea, Novemb. 5-9, 2012, Revis. Sel. Pap. Part I 11. Springer, pp. 31–44.

Li, W., Zhao, R., Xiao, T., Wang, X., 2014. DeepReID: deep filter pairing neural network for person Re-identification. 2014 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 152–159.

Li, M., Zhu, X., Gong, S., 2018a. Unsupervised Person Re-identification by Deep Learning Tracklet Association. Proc. Eur. Conf. Comput. Vis., pp. 737–753

Li, W., Zhu, X., Gong, S., 2018b. Harmonious attention network for person re-identification. Proc. IEEE Conf. Comput. Vis. pattern Recognit. 2285–2294.

Li, M., Feng, X., Han, Y., 2022a. Brillouin fiber optic sensors and mobile augmented reality-based digital twins for quantitative safety assessment of underground pipelines. Autom. ConStruct. 144, 104617.

Li, S., Tong, L., Zhai, C., 2022b. Extraction and modelling application of evacuation movement characteristic parameters in real earthquake evacuation video based on deep learning. Int. J. Disaster Risk Reduct. 80, 103213.

Li, J., Hu, Y., Zou, W., 2023. Dynamic risk assessment of emergency evacuation in large public buildings: a case study. Int. J. Disaster Risk Reduct. 91, 103659.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2016. Feature Pyramid Networks for Object Detection.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al., 2016. Ssd: Single Shot Multibox Detector. Eur. Conf. Comput. Vis., Springer, pp. 21–37.

Liu, H., Xu, B., Lu, D., Zhang, G., 2018. A path planning approach for crowd evacuation in buildings based on improved artificial bee colony algorithm. Appl. Soft Comput. 68, 360–376.

Loy, C.C., Liu, C., Gong, S., 2013. Person re-identification by manifold ranking. 2013 IEEE Int. Conf. Image Process. IEEE, pp. 3567–3571.

9 Real Life Fire Cases! Some People Lost Their Lives, Others Escaped! what Was the Cause? 2020.

McKenna, S.T., Jones, N., Peck, G., Dickens, K., Pawelec, W., Oradei, S., Harris, S., Stec, A.A., Hull, T.R., 2019. Fire behaviour of modern façade materials – Understanding the Grenfell Tower fire. J. Hazard. Mater. 368, 115–123. https://doi.org/10.1016/j.jhazmat.2018.12.077.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look once: Unified, Real-Time Object Detection.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. Proc. AAAI Conf. Artif. Intell. 31.

Tesfaye, Y.T., Zemene, E., Prati, A., Pelillo, M., Shah, M., 2019. Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets. Int. J. Comput. Vis. 127, 1303–1320.

Tzutalin, D., 2018. Tzutalin/labelimg: Labelimg Is a Graphical Image Annotation Tool and Label Object Bounding Boxes in Images.

Vanem, E., Ellis, J., 2010. Evaluating the cost-effectiveness of a monitoring system for improved evacuation from passenger ships. Saf. Sci. 48, 788–802.

Wang, T., Gong, S., Zhu, X., Wang, S., 2014. Person re-identification by video ranking. Comput. Vision–ECCV 2014 13th Eur. Conf. Zurich, Switzerland, Sept. 6-12, 2014, Proceedings, Part IV 13. Springer, pp. 688–703.

Wang, X., Jiang, F., Zhong, L., Ji, Y., Yamada, S., Takano, K., et al., 2020. Intelligent post-disaster networking by exploiting crowd big data. IEEE Netw 34, 49–55.

Wei-Shi, Z., Shaogang, G., Tao, X., 2009. Associating Groups of People. Proc. Br. Mach. Vis. . Conf., pp. 21–23

Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric. 2017 IEEE Int. Conf. Image Process., pp. 3645–3649.

Wong, P.K.-Y., Luo, H., Wang, M., Leung, P.H., Cheng, J.C.P., 2021. Recognition of pedestrian trajectories and attributes with computer vision and deep learning techniques. Adv. Eng. Inform. 49, 101356.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional Block Attention Module. Proc. Eur. Conf. Comput. Vis., pp. 3–19

Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., Yang, Y., 2018. Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning. Proc. IEEE Conf. Comput. Vis. pattern Recognit. 5177–5186.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H., 2021. Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. 44, 2872–2893.

Yogameena, B., Naganandhini, C., 2017. Computer vision based crowd disaster avoidance system: a survey. Int. J. Disaster Risk Reduct. 22, 95–129.

Yu, Q., Hu, L., Alzahrani, B., Baranawi, A., Alhindi, A., Chen, M., 2021. Intelligent visual-IoT-enabled real-time 3D visualization for autonomous crowd management. IEEE Wireless Commun. 28, 34–41.

Zhang, Q., Chen, T., Lv, X., 2014. New framework of intelligent evacuation system of buildings. Procedia Eng. 71, 397–402.

Zhang, X., Zhou, X., Lin, M., Sun, J., 2018. Shufflenet: an extremely efficient convolutional neural network for mobile devices. Proc. IEEE Conf. Comput. Vis. pattern Recognit. 6848–6856.

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

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. 100626. https://doi.org/10.1016/j.dibe.2025.100626.

Zhang, X., Jiang, Y., Wu, X., Nan, Z., Jiang, Y., Shi, J., et al., 2024a. AIoT-enabled digital twin system for smart tunnel fire safety management. Dev Built Environ, 100381.

Zhang, X., Zhang, T., Ding, Y., Huang, X., 2024b. In: Huang, X., Tam, W.C. (Eds.), Internet of Things and Digital Twin in Fire Safety Management BT - Intelligent Building Fire Safety and Smart Firefighting. Springer Nature Switzerland, Cham, pp. 335–361.

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, Y., Kinateder, M., Warren, W., Templeton, A., 2025. A virtual reality experiment on visual and auditory guidance for egress in road tunnel fires. Fire Technol.

Zhao, H., Schwabe, A., Schläfli, F., Thrash, T., Aguilar, L., Dubey, R.K., et al., 2022. Fire evacuation supported by centralized and decentralized visual guidance systems. Saf. Sci. 145, 105451.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q., 2015. Scalable person re-identification: a benchmark. Proc. IEEE Int. Conf. Comput. Vis. 1116–1124.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., et al., 2016. Mars: a video benchmark for large-scale person re-identification. Comput. Vision–ECCV 2016 14th Eur. Conf. Amsterdam, Netherlands, Oct. 11-14, 2016, Proceedings, Part VI 14. Springer, pp. 868–884.

Zheng, Z., Zheng, L., Yang, Y., 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. Proc. IEEE Int. Conf. Comput. Vis. 3754–3762.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. Proc. IEEE Conf. Comput. Vis. pattern Recognit. 2921–2929.

Zhou, M., Dong, H., Ioannou, P.A., Zhao, Y., Wang, F.-Y., 2019a. Guided crowd evacuation: approaches and challenges. IEEE/CAA J Autom Sin 6, 1081–1094.

Zhou, K., Yang, Y., Cavallaro, A., Xiang, T., 2019b. Omni-scale feature learning for person re-identification. Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 3702–3712.