

Vision-based survey method for extraordinary loads on buildings

Yang LI^{a,b}, Jun CHEN^{a,c*}, Pengcheng WANG^{a,d}

^a College of Civil Engineering, Tongji University, Shanghai 200092, China

^b China United Engineering Co. Ltd., Hangzhou 310052, China

^c State Key Laboratory of Disaster Reduction in Civil Engineering, Tongji University, Shanghai 200092, China

^d Shanghai Key Laboratory of Engineering Structure Safety, Shanghai Research Institute of Building Sciences Co. Ltd., Shanghai 200032, China

*Corresponding author. E-mail: cejchen@tongji.edu.cn

© The Author(s) 2024. This article is published with open access at link.springer.com and journal.hep.com.cn

ABSTRACT The statistical modeling of extraordinary loads on buildings has been stagnant for decades due to the laborious and error-prone nature of existing survey methods, such as questionnaires and verbal inquiries. This study proposes a new vision-based survey method for collecting extraordinary load data by automatically analyzing surveillance videos. For this purpose, a crowd head tracking framework is developed that integrates crowd head detection and reidentification models based on convolutional neural networks to obtain head trajectories of the crowd in the survey area. The crowd head trajectories are then analyzed to extract crowd quantity and velocities, which are the essential factors for extraordinary loads. For survey areas with frequent crowd movements during temporary events, the equivalent dynamic load factor can be further estimated using crowd velocity to consider dynamic effects. A crowd quantity investigation experiment and a crowd walking experiment are conducted to validate the proposed survey method. The experimental results prove that the proposed survey method is effective and accurate in collecting load data and reasonable in considering dynamic effects during extraordinary events. The proposed survey method is easy to deploy and has the potential to collect substantial and reliable extraordinary load data for determining design load on buildings.

KEYWORDS load survey, extraordinary load, live load, object detection, multiple-object tracking

1 Introduction

Designing safe and economical buildings requires reliable knowledge of live loads. Live loads have two components. One is the sustained load, including furniture, equipment, and normal working personnel, which exists near-continuously. The other is the extraordinary load caused by events such as crowding of people or furniture stacking, which occurs for a short duration and has a relatively high intensity. Owing to the spatial and temporal variability of the live load, sustained and extraordinary loads are modeled as a Poisson square wave and compound Poisson point processes, respectively [1–3]. Accordingly, probabilistic live load models can

provide a suitable design value based on the model parameters obtained from live load surveys.

Over the past several decades, a few live load surveys have been conducted in different countries [4–9]. Furthermore, continuous progress has been made in sustained load studies using field survey data. However, the understanding of extraordinary loads has been hindered by the scarcity of data. Unlike the data for sustained loads, which can be obtained by direct measurement at an arbitrary time, data for extraordinary loads are difficult to obtain owing to the transient nature of events. Thus, in earlier studies, the model parameters of the extraordinary load have been estimated using engineering judgment with little or no survey data [10–12]. However, it is essential to determine the number of people and weight of stacked furniture in areas where extraordinary events

occur. To provide design load based on observational data, only a few extraordinary load surveys have been conducted, and these have relied on questionnaires or questioning. Paloheimo and Ollila [13] investigated temporary floor loads due to personnel crowding and determined the number of persons present during transient load events. Choi [14] conducted a questionnaire survey of occupants to collect detailed information about extraordinary events, including locations, occurrences, area dimensions, number of people, and weight of stacked items. Out of the 1989 extraordinary events studied, only 54 were furniture stacking events, and the vast majority were crowding events [15]. Andam [16] obtained extraordinary load data in residences by questioning the tenants regarding the unusually large number of visitors to their homes in the last five years and the visitor frequency. However, questionnaire and questioning methods rely on the subjective understanding of occupants and are liable to errors [17], especially for areas with frequent crowd movements, such as corridors and stairs. Surveyors are unlikely to perform lengthy surveys, so this less accurate method has had to be accepted. In addition, the aforementioned extraordinary load surveys considered only the static load effects. For access areas, such as corridors and stairs, the effect of the crowd static load evidently differs from the actual crowd movement effect. In emergencies, such as during a fire or earthquake, extraordinary loads and a reduction in structural resistance often occurs simultaneously. It is necessary to consider the dynamic effects caused by crowds walking or running in access areas [18]. Thus, the dynamic effect of live load should be included in the design loads [19,20], and in field surveys as well.

With the improvement and increased use of surveillance facilities in buildings, it is possible to conduct extraordinary load surveys using surveillance videos instead of questionnaires or questions. However, it is difficult for investigators to view surveillance videos for a long period of time to capture extraordinary load data. In recent years, machine learning methods, especially deep learning methods, have been widely applied in the field of civil engineering [21–24]. Computer vision technologies based on convolutional neural networks (CNNs) [25–27] have attracted significant attention and have been used in various research tasks, such as crack detection [28], damage detection [29,30], and intelligent construction [31,32]. However, these methods focus on object recognition and localization in a single frame, and cannot track the state change of an object or multiple objects in an area. Luo et al. introduced vision-based object detection, multiple-object tracking (MOT), and action recognition to implement automatic detection and visualization of the dynamic workspaces of workers on foot [33]. Xiao et al. [34] proposed a vision-based method consisting of instance segmentation and Kalman filtering

to track workers in off-site construction, and similar frameworks have been adopted to automatically track construction machines in videos [35]. Wang et al. combined the masks R-CNN [36] and DeepSort [25] algorithms to collect precast wall locations and temporal information from surveillance videos to automatically monitor construction progress [37]. Hence, MOT-based computer vision methods can facilitate automated and sustainable inspection and monitoring. It is feasible to adopt MOT methods to monitor and record the number of people or stacked furniture in an area through surveillance videos, thereby enabling automated extraordinary load investigations.

In contrast to furniture stacking, the collection of load data caused by crowd gathering is more challenging owing to high crowd density and crowd movement. With increasing crowd density in a scene, individual people's visibility decreases with increasing mutual occlusions, leading to a significant decrease in the accuracy of obtained crowd quantities. To address this challenge, in this study, people in densely crowded environments are tracked efficiently with distinctly visible heads. Moreover, an equivalent dynamic load factor (EDLF) is introduced to amplify the crowd quantity to account for the dynamic effects caused by crowd movement in access areas. Therefore, this study examines the vision-based survey method with crowding events as the primary concern.

This study proposes a new vision-based survey method to automatically and sustainably collect extraordinary load data using surveillance videos, providing a realistic and accurate basis for determining design load on buildings. This method integrates object detection and MOT to track humans using visible heads. Based on stable head tracking, variations in crowd quantity during extraordinary events in the survey area are captured, and the velocity of the crowd from the trajectory analysis is used to consider dynamic effects owing to crowd movement. The remainder of this paper is organized as follows. First, the overall architecture of the vision-based survey method for extraordinary loads is explained. Second, head image data sets for constructing stable crowd head tracking and experiments for survey method validation are presented. Subsequently, the performance of the proposed survey method is evaluated. Moreover, the applications and limitations of the proposed method are discussed. Finally, the conclusions of this study are summarized.

2 Overall architecture

Two key issues must be addressed to realize extraordinary load surveys using surveillance videos. The first issue is determination of the number of people gathered

in the survey area during extraordinary events. The number of people entering the survey area should be assessed. Therefore, the survey area is calibrated to obtain intrusion lines (i.e., boundaries of the area) in the video frames. Each participant is tracked to judge whether they crossed the intrusion line and entered the survey area. The second issue is consideration of the dynamic effects of crowd movement in the access areas. The solution to this question is to amplify the crowd size using the EDLF to consider actual dynamic effects. To obtain the EDLF, the velocity of each person through the access area is calculated. By analyzing the crowd trajectory, the velocity can be estimated based on the actual distance between intrusion lines. Figure 1 presents an overview of the vision-based survey method for extraordinary loads.

Based on the previous analysis of the two core questions, the proposed method needs to realize stable crowd tracking to obtain the crowd quantity and EDLF. Therefore, a crowd head detector (CHD) and a crowd head tracker (CHT) are built to implement the vision-based survey method. The following sections elaborate the details of the proposed method.

2.1 Crowd head detector

This study adopts a tracking-by-detection strategy to track crowd heads. In this paradigm, the detector provides localization information of the targets in each frame for the subsequent MOT task. The accuracy and operating speed of the detector directly determine the tracking system performance. Considering that single-stage detectors balance well between accuracy and speed, the CHD is built based on fine-tuning the YOLOv5 [38]

detector. Figure 2 shows the overall architecture of the CHD, and a description of each block in the figure is presented in the legend.

A CHD consists of a backbone network and a path aggregation network [39]. The backbone network is the primary feature extractor, comprising continuous convolution and merge operations. Each convolution operation has a convolution layer, batch normalization [40] layer, and SiLU [41] activation layer, termed the CBS block. To promote the learning capability of the entire network, a modified cross-stage partial structure [42] with bottleneck layers, called the BCSP block, is introduced in the backbone network. The detailed design of the BCSP block is shown by the dotted rectangular box in Fig. 2. As the spatial pyramid pooling [43] operation can increase the receptive field and separate useful context features, a modified version with a faster operating speed, denoted as the SPPF block, is added to the backbone network. The structure of the SPPF block is shown in the dot-dashed rectangular box in Fig. 2. The path aggregation network collects features from the backbone network and provides merged features with richer and proper representations to promote detection accuracy. A top-down path with lateral connections, similar to the feature pyramid network [44], is used to propagate semantically strong features for high performance. Next, to enhance the entire feature pyramid with accurate localization signals existing in the lower layers, a bottom-up path augmentation is adopted. Finally, the CHD provides three output feature maps to detect and locate crowd heads at different scales.

Five CHD models with different parameters and layers are investigated for different accuracies and operating speed requirements. The five models differ primarily in

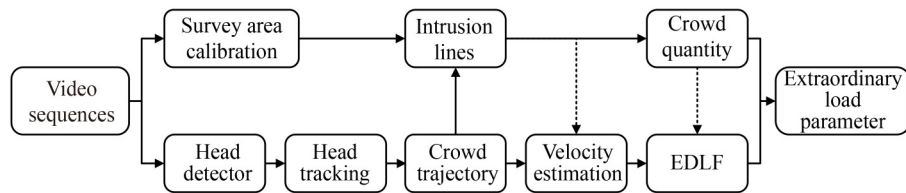


Fig. 1 Overview of vision-based extraordinary load survey method.

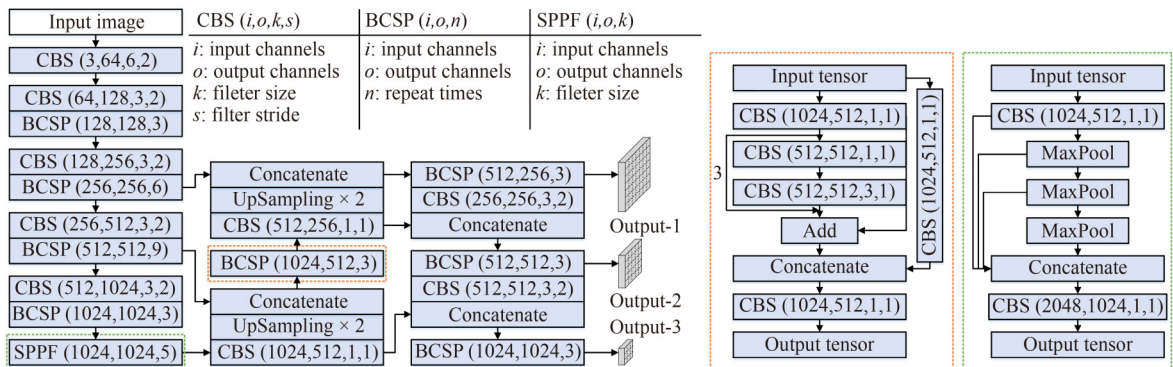


Fig. 2 Architecture of crowd head detector.

tracks are split into matched and unmatched tracks. The tentative tracks, unmatched tracks with an age of 1, and unmatched head detections are collected to perform the second association using the intersection over union (IoU) matching procedure. Unmatched head detections, unmatched tracks, and matched tracks are then processed in the data update process. Unmatched head detections are initiated as new head tracks with tentative status, and each new head track is assigned a unique identity (ID). The confirmed tracks over the max-age and the tentative tracks in the unmatched tracks are deleted. Furthermore, the matched tracks in the first and second associations are updated using the location information and feature embeddings of the head detections. Finally, the new head tracks, confirmed tracks remaining in the unmatched tracks, as well as updated tracks, constitute the head tracks for the next frame. The ID and pixel position of the updated tracks with confirmed status are visualized in the current frame and used to conduct an extraordinary load survey.

Head feature embeddings play an important role in the data association process. To obtain well-discriminating feature embedding, a small and fast HReID model is built using a structural re-parameterization method [47]. The training and inference architectures of the HReID model are decoupled, as shown in Table 2. The HReID model contains five stages with different numbers of convolutional layers. In the training phase, each layer has an ID branch, a 3×3 branch, and a 1×1 branch. In the inference phase, the three branches are transformed into a single branch with a kernel size of three. The global feature of dimensionality 256 is computed in the pooling layer and normalized to be compatible with the cosine distance metric. The HReID model is trained on a HReID data set and is discussed in the following results section.

2.3 Extraordinary load survey

With the development of robust crowd tracking, an extraordinary load survey can be conducted automatically

and sustainably through surveillance cameras. From the previous analysis, it is clear that the number of people in the survey area at the time of an extraordinary event needs to be determined during the investigations. For access areas, to consider the dynamic effects caused by crowd movement, it is necessary to further obtain pedestrian velocity to calculate the EDLF.

2.3.1 Survey area calibration

To achieve extraordinary load surveys, the survey area needs to be calibrated first. In the calibration process, the horizontal and vertical intrusion lines and characteristic data of the survey area, such as occupant type, area use, and area size, should be obtained. As shown in Fig. 4, vertical intrusion lines are responsible for crowd counting to conduct an extraordinary load survey of the rooms. Horizontal intrusion lines are used to count people and calculate the velocity to achieve an extraordinary load survey of the access areas. The vertical intrusion lines can be obtained by selecting two points with suitable pixel spacing at the room entrance using an arbitrary frame of the surveillance video. The procedure to obtain horizontal intrusion lines is simple and requires only a person and an object of fixed height, such as a leveling rod. In this study, the leveling rod is set at the mean height of the people for crowd head tracking. The surveyor places the leveling rod on both ends of the horizontal intrusion line for a few seconds to obtain its pixel position in the surveillance images. In addition, the actual distance between two adjacent horizontal intrusion lines needs to be measured for subsequent velocity estimation. The above calibration procedure for the intrusion lines needs to be carried out only once, when the camera field of view is fixed.

2.3.2 Trajectory analysis and velocity estimation

When a person enters the monitored area, the CHT model tracks the head and assigns a unique tracking ID. When a

Table 2 Architecture of the HReID model

Name	Layer number	Training kernel/stride	Inference kernel/stride	Output size
Stage 1	1	$3 \times 3/2, 1 \times 1/2$	$3 \times 3/2$	$24 \times 128 \times 128$
Stage 2	1	$3 \times 3/2, 1 \times 1/2$	$3 \times 3/2$	$48 \times 64 \times 64$
	1	$3 \times 3/1, 1 \times 1/1$	$3 \times 3/1$	$48 \times 64 \times 64$
Stage 3	1	$3 \times 3/2, 1 \times 1/2$	$3 \times 3/2$	$96 \times 32 \times 32$
	3	$3 \times 3/1, 1 \times 1/1$	$3 \times 3/1$	$96 \times 32 \times 32$
Stage 4	1	$3 \times 3/2, 1 \times 1/2$	$3 \times 3/2$	$192 \times 16 \times 16$
	3	$3 \times 3/1, 1 \times 1/1$	$3 \times 3/1$	$192 \times 16 \times 16$
Stage 5	1	$3 \times 3/2, 1 \times 1/2$	$3 \times 3/2$	$256 \times 8 \times 8$
AdaptiveAvgPool	1	—	—	256

person's head trajectory crosses the boundary intrusion line of the survey area, it can be determined that he or she enters or exits the survey area using the movement direction. Whether the head trajectory intersects the intrusion line in a plane can be determined as follows:

$$(\overrightarrow{P_1Q_1} \times \overrightarrow{P_1Q_2}) \cdot (\overrightarrow{P_2Q_1} \times \overrightarrow{P_2Q_2}) \leq 0, \quad (2)$$

$$(\overrightarrow{Q_1P_1} \times \overrightarrow{Q_1P_2}) \cdot (\overrightarrow{Q_2P_1} \times \overrightarrow{Q_2P_2}) \leq 0, \quad (3)$$

where Q_1Q_2 represents the two endpoints of the intrusion line and P_1P_2 represents the head trajectory of the adjacent frame, as shown in Fig. 4. $\overrightarrow{P_1Q_1}$ denotes a vector calculated using the pixel coordinates. The direction of movement is confirmed using the changes in the pixels. The frame numbers of the same person crossing the adjacent intrusion lines can be obtained using Eqs. (2) and (3). Thus, the average velocity of the person crossing the intrusion region can be calculated as follows:

$$v = \frac{\Delta d}{f_2 - f_1}, \quad (4)$$

where Δd represents the measured distance between adjacent intrusion lines. f_1 and f_2 indicate the frames in which the person crosses the first and second intrusion lines, respectively. The above process is performed for each person entering the monitored area in each frame. Thus, the counting change at each boundary intrusion line and the velocity of each person in the access area are recorded. The number of crowds can be calculated per frame by combining the data from all boundary intrusion lines of the survey area. For survey areas with multiple

entrances and exits, such as access areas, data from multiple surveillance cameras can be used for investigations after time synchronization.

2.3.3 Calculation of equivalent dynamic load factor

Based on the obtained pedestrian velocity, the dynamic effect due to crowd movement is considered by amplifying the crowd size using the EDLF. Human-induced loads are generally treated as periodic excitations and expressed as a Fourier series model [48]. The load induced by the crowd, considering only the contribution of the first harmonic, can be described as follows:

$$F_c(t) = \sum_{i=1}^N G_i [1 + \alpha_i \sin(2\pi f_i t - \varphi_i)], \quad (5)$$

where $F_c(t)$ is the load induced by the crowd, N is the quantity of crowds, and G_i , f_i , α_i , and φ_i represent the weight, movement frequency, first-order dynamic load factor (DLF), and phase angle of i th pedestrian, respectively. Because the variation in the weight of pedestrians has a negligible effect on the statistical distribution of the dynamic response [49], the crowd load using the mean weight of pedestrians is expressed as follows:

$$F_c(t) = \bar{G} \left[N + \sum_{i=1}^N \alpha_i \sin(2\pi f_i t - \varphi_i) \right], \quad (6)$$

where $\bar{G}$ denotes the mean weight of the crowd. The relationships among the velocity v , movement frequency f and first-order DLF α of pedestrians are as follows [50,51]:

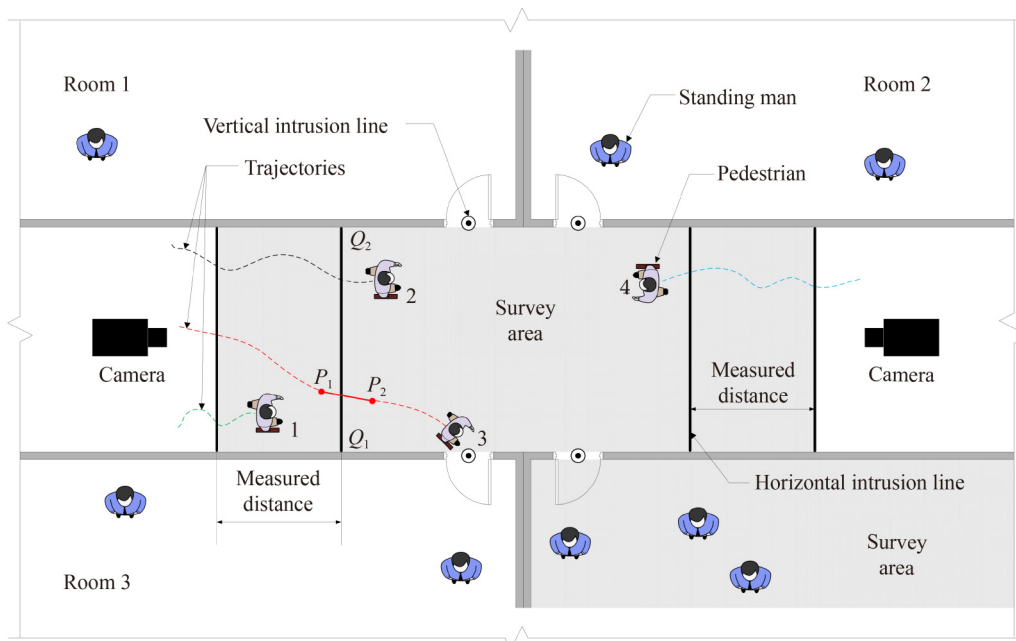


Fig. 4 Extraordinary load survey process using surveillance cameras.

$$f = 1.7345v^{0.4153}, \quad (7)$$

$$\alpha = 0.2358f - 0.201. \quad (8)$$

Considering that the pedestrian movement frequency tends to be the same when the crowd density is high, the mean velocity of the crowd is utilized to replace the velocity of each person. The crowd load in a frame is then simplified as follows:

$$F_c = \bar{G} \left(N + \bar{\alpha} \sqrt{\left(\sum_{i=1}^N \sin \varphi_i \right)^2 + \left(\sum_{i=1}^N \cos \varphi_i \right)^2} \right), \quad (9)$$

where $\bar{\alpha}$ is the mean DLF corresponding to the mean velocity of the crowd in the frame. Therefore, the EDLF of the crowd in a frame is calculated as follows:

$$EDLF = \frac{\bar{\alpha} \sqrt{\left(\sum_{i=1}^N \sin \varphi_i \right)^2 + \left(\sum_{i=1}^N \cos \varphi_i \right)^2}}{N}. \quad (10)$$

The phase-angle distribution of the crowd follows a zero-mean normal distribution [52]. The standard deviation of the phase angle, σ , differs within three crowd density categories: unrestricted traffic (UT), restricted traffic (RT), and exceptionally restricted traffic (ERT). The standard deviations for different crowd densities can be calculated as follows:

$$\sigma = af^2 + bf + c, \quad (11)$$

where f denotes the mean crowd frequency. a , b , and c depend on the crowd density, which can be obtained from Table 3. Hence, the randomness of the EDLF in a frame is determined by the phase angle. This study randomly selected 100 groups of phase angles based on the crowd number and corresponding distribution to calculate the EDLF, and the 95th percentile value was taken as the final EDLF.

Table 3 Parameters for σ in different crowd density

Category	Crowd density (p/m ²)	a	b	c
UT	$\rho < 0.5$	1.4927	-5.4413	5.3624
RT	$0.5 \leq \rho < 1.5$	0.8013	-2.9881	3.2007
ERT	$\rho \geq 1.5$	0	0	0.3642

2.3.4 Load survey procedure

With all the above preparations, the investigation process of the extraordinary load based on surveillance cameras is as follows: 1) the crowd heads are tracked, and a unique tracking ID is assigned to each person; 2) it is determined

whether the head trajectory crosses the intrusion line, by a comparison with the previous frame; 3) the counting changes are recorded at each boundary intrusion line, including the velocities and IDs of pedestrians in the access areas; 4) the crowd quantity is obtained by combining the counting changes in each boundary intrusion line in the survey area; 5) pedestrian velocities are indexed by tracking IDs to obtain the mean velocity, mean movement frequency, and mean DLF using Eqs. (7) and (8); 6) the phase angle distribution is selected based on the crowd density and mean movement frequency; 7) the EDLF is calculated 100 times using the selected phase angle distribution and Eq. (10), and the 95th percentile value is adopted to amplify the crowd size; 8) the steps 1)–4) are repeated for rooms and steps 1)–7) for access areas per frame until the end of extraordinary events.

3 Data set and experiment

This section describes the crowd head detection and HReID data sets used to implement the vision-based extraordinary load survey method. In addition, a crowd quantity investigation experiment and a crowd walking experiment are designed to verify the feasibility of the proposed survey method.

3.1 Head data sets

The CHD and HReID models are implemented based on CNN. Therefore, crowd head image data sets, including crowd head detection and HReID data sets, are collected to train and evaluate the above CNN-based models.

The crowd head detection data set included 3727 images with 59499 head annotations, which are sampled from surveillance videos of teaching buildings. The surveillance images are carefully selected for variance and reduction in the similarity of crowd heads. Each visible head in each image is labeled, including the blocked parts. The SCUT-HEAD data set [53] is also adopted further to enrich the variability of the crowd head data set. Thus, a large-scale head detection data set including 8132 images and 170750 head annotations is used to train and test the CHD model. The final data set is divided into the training and testing sets. The training set contains 6723 images with 139887 head annotations, and the testing set contained 1409 images with 30863 head annotations. Two representative images and annotations are shown in Fig. 5.

The HReID data set is captured using four surveillance cameras in the teaching buildings. Overlapping exists in these cameras. Each ID is captured by at least one camera. This data set contains 7404 head images with 712 identities. The head images are cropped using hand-



Fig. 5 Example images and annotations of the crowd head detection data set: (a) sample-1; (b) sample-2.

drawn bounding boxes, and all the images are normalized to 128×128 pixels. The HReID data set is randomly divided into training and testing sets. The training and testing sets contain 377 identities with 3682 head images, and 335 identities with 3010 head images. In the testing set, one image for each ID in each camera is selected as a query image and the remaining images are taken as gallery images. The sample images of the four identities in the HReID data set are shown in Fig. 6.

3.2 Survey experiments

3.2.1 Crowd quantity investigation

To verify the effectiveness of the vision-based extraordinary load survey method when crowd gathering occurs, four different scenes in teaching buildings are selected for crowd quantity monitoring. Table 4 shows the details of the four scenarios, including the survey area composition, the size of the area, and the number of surveillance cameras used. The frames per second (FPS) of each surveillance camera is 25. Each scenario is observed for 120 s at a high pedestrian density. Before the investigation, a calibration procedure with a leveling rod and frames of surveillance videos is performed to obtain the virtual intrusion lines at each entrance and exit of the survey area. The number of existing crowds in the survey area is obtained by monitoring the changes in pedestrian volume at each intrusion line. For comparison purposes, the change in pedestrian volume at each intrusion line is manually recorded frame by frame to obtain the true value of the crowd quantity in the survey area.

3.2.2 Crowd walking experiment

A crowd walking experiment based on a wooden floor is conducted to evaluate the rationality of the EDLF calculated from crowd velocities for crowd quantity amplification. The static load of the crowd's weight, the static load amplified by the EDLF, and the actual load caused by crowd movement are compared using the displacement response in the midspan of the wooden floor. The size of the wooden floor with the long edges simply supported and short edges free is $6 \text{ m} \times 3 \text{ m}$. A laser displacement sensor with a sampling frequency of 200 Hz is used to measure the displacement response of the wooden floor. Two cameras with 30 FPS are placed at both ends of the wooden floor to obtain the number and velocity of the crowd on the floor. Figure 7(a) shows basic information about the wooden floor and the positions of the displacement sensor and cameras.

To obtain the stiffness between the midspan displacement and the uniform load, uniform static load testing using steel plates is first performed. The loading procedure is performed five times with an increment of 3.4 kN each time. Figure 7(b) shows the loading process. Next, crowd walking tests are conducted with 27 pedestrians. The average mass and average height of the 27 test subjects are 64 kg and 170 cm, respectively. Before the crowd walking tests, a calibration procedure is performed based on the mean height to obtain the intrusion lines for crowd counting and velocity estimation. Considering the size of the wooden floor, the distance between the first and second horizontal intrusion lines is set to 2 m to calculate the velocity, as shown in

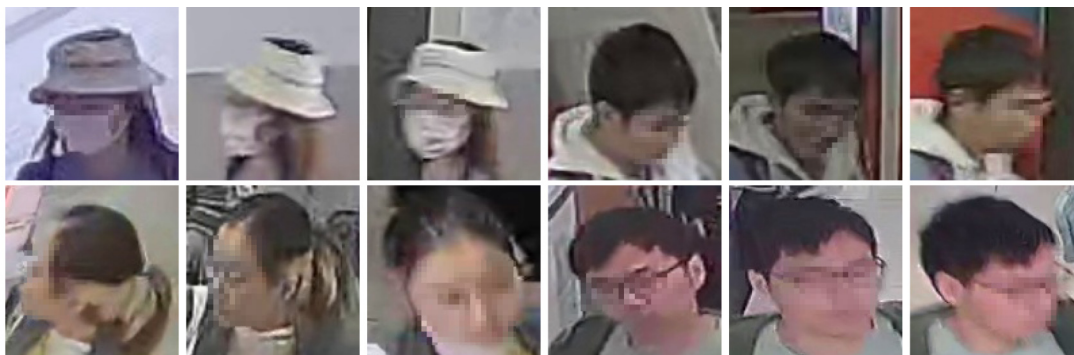


Fig. 6 Example images of the HReID data set.

Table 4 Details of investigation scenes

Scene	Survey area	Area size (m ²)	Camera number	Time (s)
Scene1	classroom	79	1	120
Scene2	corridor	25	2	120
Scene3	classroom, corridor	247	4	120
Scene4	classroom, corridor, lobby, stair	633	6	120

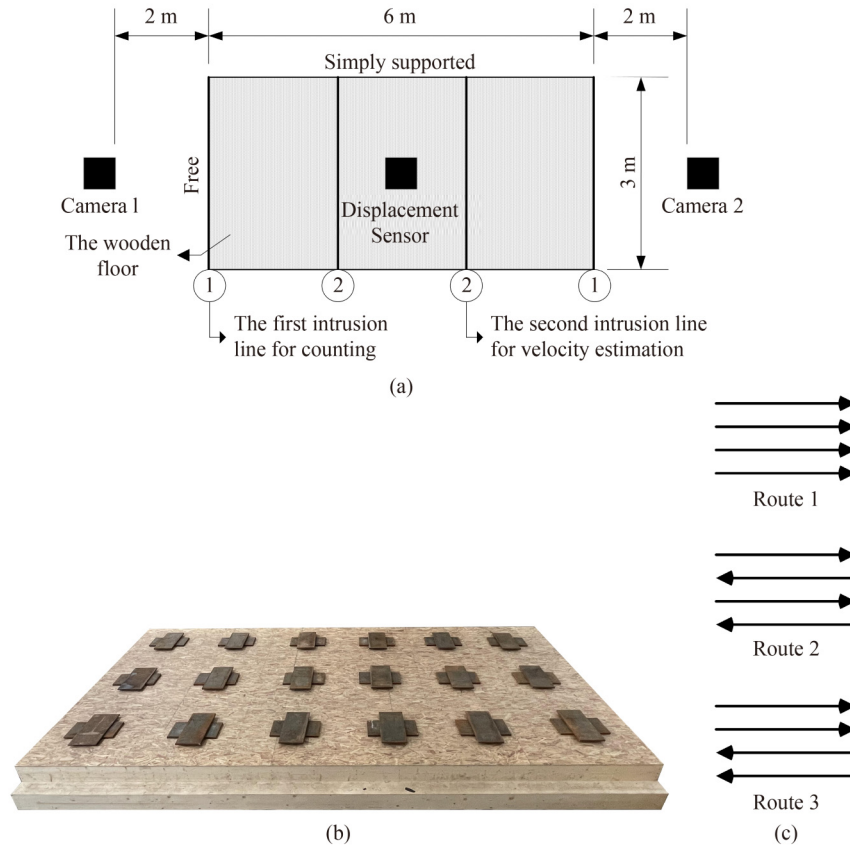


Fig. 7 (a) Basic information regarding the wooden floor; (b) loading process of the uniform static load testing; (c) walking route of the crowd.

Fig. 7(a). The test subjects are asked to walk freely to simulate unrestricted and restricted traffic. In each traffic status, the walking route of the crowd is divided into three types: one-way traffic and two types of two-way traffic, as shown in **Fig. 7(c)**. In addition, the speed of the crowd is divided into slow, medium, and fast in UT, whereas only slow-speed conditions are adopted in RT for safety reasons.

4 Results and discussion

4.1 Testing results of convolutional neural networks based models

The CHD and HReID models were trained and tested on the crowd head detection and HReID data sets,

respectively. The final models for the subsequent implementation of the vision-based extraordinary load survey method were selected based on evaluation metrics. Experiments on both CNN-based models were conducted using a GeForce GTX 1070Ti 8 GB GPU.

The CHD models were trained on the training set with 139887 head annotations, and the implementation details are as follows. The CHD model was initialized with pre-trained weights of the MS-COCO data set [54], and the Adam optimizer [55] was adopted to update the parameters. The mini-batch size was set to four owing to the limited GPU memory, and the total number of training epochs was fixed at 60. To obtain a stable gradient momentum in each iteration, gradual warmup [56] and cosine annealing [57] schedules were used to adjust the learning rate smoothly. The maximum and minimum learning rates were set to 10^{-3} and 10^{-6} ,

respectively. The warmup phase contained 2 epochs, and the cosine annealing phase consisted of 58 epochs. Routine data augmentation such as flipping, affine, and photometric distortions was adopted to increase the input image variability. In each batch iteration, the training image sizes were randomly set to 384, 448, 512, 576, or 640 pixels. The aforementioned training procedures were implemented for all five models (Table 1).

Each CHD model was tested on an independent testing set with 30863 head annotations. The size of the test images varied from 384 pixels to 1152 pixels in increments of 128 pixels. The average precision with an IoU threshold of 0.5 (AP50) [58] and the inference time metrics were utilized to evaluate the performance of these CHD models. The test results for the five CHD models are shown in Fig. 8. With an increase in the parameters and depth of the models, the head detection performance could be steadily improved, while the inference time will then gradually increase. However, the increase in the model parameters had a negligible effect on the improvement of AP50 after the depth and width factors reached 1.0. Regarding an individual model, the head detection accuracy improved, and the inference time increased with the increase in the input image size. The detection performance of these CHD models leveled off after the image size reached 896 pixels. The CHD-s model exhibited an AP50 of 94.6% with an inference time of 6.4 ms, and the CHD-m model exhibited an AP50 of 95.4% with an inference time of 15.6 ms. The CHD-s and CHD-m models demonstrated a better balance of inference time and accuracy and are thus preferred for subsequent crowd head tracking.

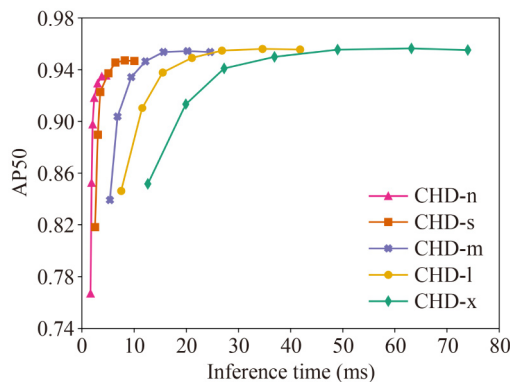


Fig. 8 Testing results of the CHD models with different parameters.

The HReID model with softmax loss was trained to discriminate the identities in the training set of the HReID data set. The HReID model was initialized with pretrained weights of the ImageNet data set [63]. The mini-batch size and total epochs were set to 32 and 60, respectively. The optimizer for parameter updating and the learning rate schedule applied in the HReID model training process were the same as those used for the CHD

models. A data augmentation procedure, including random crop and photometric distortions, was used to enhance the generalization ability of the model. After training, the softmax classifier was stripped of the network, and distance queries were conducted using cosine similarity in the final layer of the HReID model. A comparison with other lightweight networks, as shown in Table 5, was performed to demonstrate the effectiveness of the HReID model. For a fair comparison, the above training procedures for the other lightweight models were the same as those for the HReID model.

Table 5 Testing results and params of different lightweight models

Models	Params (M)	MFLOPs	Rank-1 (%)	MAP (%)	Inference time (ms)
OSNet [59]	0.3	44.0	96.1	85.4	14.8
ShuffleNetV2 [60]	0.3	11.5	92.5	75.9	6.1
SqueezeNet [61]	1	229.6	92.5	74.6	2.2
MobileNetV3 [62]	0.7	37.2	92.8	78.8	7
ResNet18 [27]	5.3	499.1	94.6	82.6	2.6
HReID	1.9	204.4	96.7	88.2	1.4

Each lightweight model was tested on a test set with 335 identities and 3010 head images. Cumulative matching characteristics [64], Rank-1, and mean average precision (MAP) [65] metrics were adopted to evaluate the performance of the lightweight models. Table 5 lists the million floating point operations (MFLOPs), params, inference time, and testing results of different lightweight models on the HReID data set. The HReID model achieved the best performance among all the lightweight models and provided an MAP of 88.2% and Rank-1 of 96.7%. In terms of the MAP metric, HReID exhibited clear advantages. Although the MFLOPs and params of the HReID model are not minimal in Table 5, it provided the least inference time. The complicated multi-branch and depth-wise convolution could substantially reduce the params and MFLOPs of other lightweight models, but they also increased the memory cost and slow down the inference. The HReID model without any branches proved to be simple, offering a favorable balance of inference speed and accuracy for subsequent crowd head tracking.

4.2 Experiment results

4.2.1 Crowd quantity monitoring results

Based on the implementation of the CHD and HReID models, a crowd-head tracking task can be achieved using the CHT framework. The matched times were set to 3 to determine the head track state, and the max-age for the Kalman filter prediction of each confirmed head track was set to 50. The latest 75 feature embeddings were reserved in the head feature gallery for each track and utilized for data association. Because the surveillance

cameras are fixed, the motion information and feature embeddings were both used in the combined metric, and the hyperparameter λ was set to 0.5. Figure 9 shows several crowd head tracking results of the CHT in Scene2. The same color in the bounding box and tracking trajectory represents the same person. CHT was found to be capable of detecting and tracking crowd heads stably in crowded situations. Even after longer occlusions, the CHT did not cause any ID switch and effectively preserved the ID, as shown in Figs. 9(a)–9(c). The continuous head trajectory of each ID was used to determine whether the ID had entered the survey area. Figure 9(b) illustrates a person crossing the red vertical intrusion line and entering the corridor survey area, while Fig. 9(d) shows a person crossing the red horizontal intrusion line and walking out of the corridor survey area. As shown in Figs. 9(a) and 9(b), the counts changed accordingly as the

person crossed the intrusion lines.

The four scenes in Table 4 were examined under high crowd flow conditions based on the stable tracking of crowd heads. For the scenes in which multiple cameras were used to conduct surveys, the surveillance data from each camera were combined by time synchronization to obtain the crowd quantity in the survey area. Figure 10 shows the time histories of the crowd quantity changes over 120 s in the four survey areas using the CHD-m model and an input image size of 768 pixels. The number of people gradually gathering in the survey area can be accurately recorded when only one surveillance camera is present, as in Scene1. When the survey area is small, as in Scene2, the crowd density is high and the quantity of crowd changes substantially. In such scenarios, crowded conditions lead to errors in monitoring, such as two people occluding each other but simultaneously passing through an intrusion line. As the area surveyed is

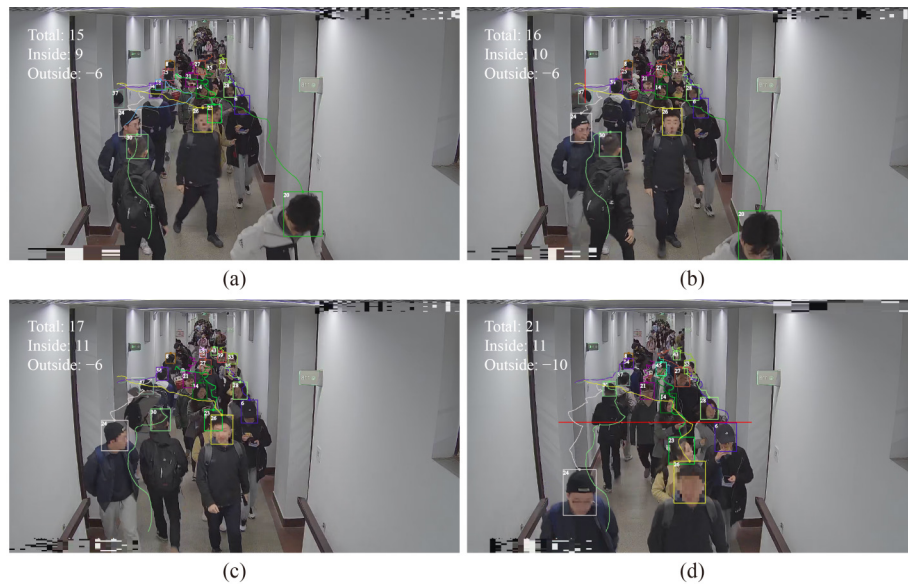


Fig. 9 Sample tracking results visualization of CHT in a crowded situation: (a) result-1; (b) result-2; (c) result-3; (d) result-4.

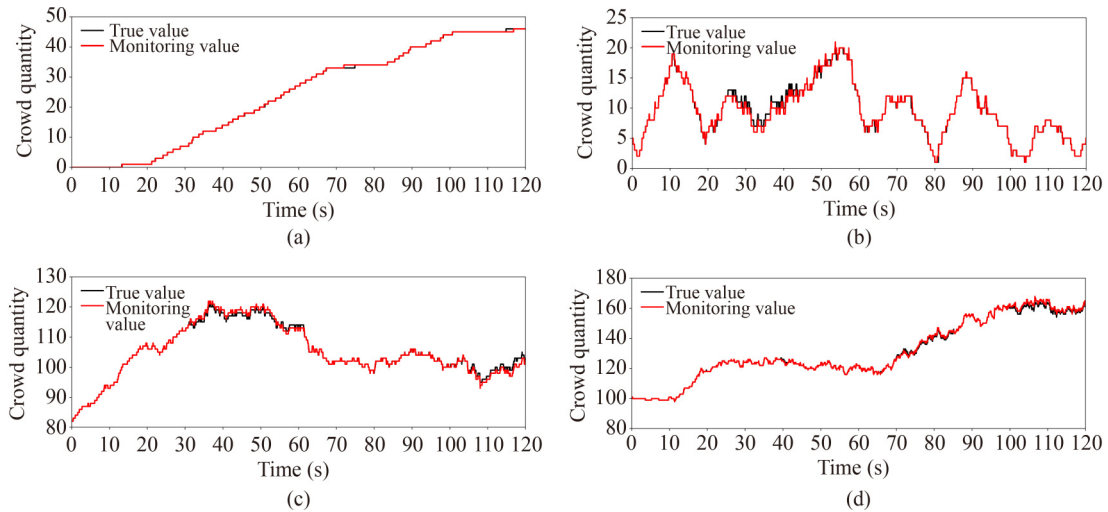


Fig. 10 Time histories of the crowd quantity changes in four survey areas: (a) scene1; (b) scene2; (c) scene3; (d) scene4.

increased, the number of surveillance cameras used increases, and the proposed method can monitor quantity changes effectively at the beginning of crowd quantity growth. However, with the gradual increase in crowd size, the homogeneous appearance of heads in the distance view makes monitoring prone to errors, as in Scene3 and Scene4. Further enhancement of the head detector and tracker can help alleviate this problem. Overall, the monitoring values fluctuate, but generally match well with the true values for all scenes. The results demonstrate that the proposed survey method can effectively capture changes in crowd quantity in the survey area when a crowd gathering situation occurs.

The discrepancy between the monitored and true values was further evaluated using the peak relative error (*PRE*), Pearson correlation coefficient (*PCC*), and root mean square error (*RMSE*) metrics, which represent the survey performance. Figure 11 shows the average evaluation results for the four scenes obtained using the vision-based survey method with different CHD models and image sizes. In small image size cases, the proposed method presented low *PCC* and high *RMSE* values, because the low resolution makes the CHD models less effective in providing stable head detections for head tracking. As the image size increases, the performance of the proposed method can be improved, and the proposed method using the CHD-m model provides the highest *PCC* of 99.7%

and the lowest *RMSE* of 0.579 at an image size of 768 pixels, as shown in Figs. 11(a) and 11(b). The *PRE* metric was stable within 5%, as shown in Fig. 11(c), demonstrating the robustness of the proposed survey method for capturing the maximum crowd quantity under different image sizes. Furthermore, the time required by the proposed method to process each surveillance video in the four scenes was investigated, as shown in Fig. 11(d). The mean time cost of the proposed survey method increased with the image size and CHD model size. For the same image size, the proposed method adopting the CHD-m model generally provided better survey results than the proposed method using the CHD-s model while requiring a higher time cost. In actual extraordinary load surveys, the proposed method using the CHD-m model and an image size of 768 pixels can be selected to realize a real-time and accurate survey, whereas the proposed method adopting the CHD-s model is a good choice for devices with limited computing powers.

4.2.2 Crowd walking experiment results

The proposed survey method using the CHD-m model and an image size of 768 pixels was selected to process the test videos. The crowd quantity on the wooden floor in each frame was obtained using the same method as in the previous experiment. Figure 12 shows the tracking

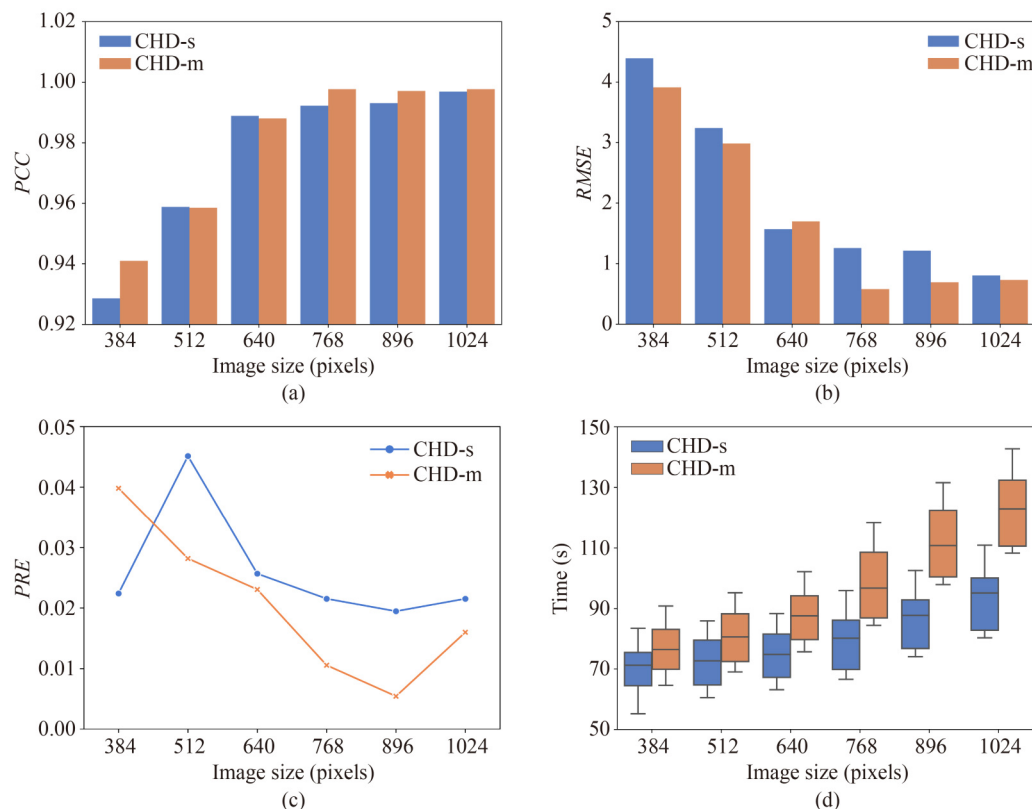


Fig. 11 Evaluation metrics of the vision-based survey method for different CHD models and image sizes: (a) *PCC*; (b) *RMSE*; (c) *PRE*; (d) survey time.



Fig. 12 Tracking numbers and velocities of pedestrians on the wooden floor: (a) result-1; (b) result-2.

results of pedestrians on the wooden floor under UT conditions, where the green lines represent intrusion lines. The velocity of each pedestrian was estimated based on the actual distance between two intrusion lines and the time across two intrusion lines. Each pedestrian on the wooden floor had a velocity and unique ID, as shown in Fig. 12(a). The absent velocity in this frame could be obtained from the subsequent frame, as shown in Fig. 12(b). For each camera data, the tracking ID of each pedestrian on the wooden floor in every frame was recorded, and the velocity of each pedestrian was saved to a collection with tracking IDs as indices. Therefore, after processing the test videos, the velocity of each pedestrian on the wooden floor in every frame was obtained using a unique tracking ID. In this experiment, the velocity of each pedestrian from two camera data was adopted to calculate the mean velocity of the crowd on the wooden floor for each frame.

The EDLF in every frame can be obtained using Eq. (10), based on the quantity and mean velocity of the crowd on the wooden floor. Figure 13(a) shows the time history of the EDLF for pedestrians walking in Route 2 under a RT status. The midspan displacements of the wooden floor in every frame were calculated using the static load (the mean weight multiplied by the crowd quantity) and amplified load (the mean weight multiplied by the amplified crowd quantity using the EDLF) and then compared with the measured displacements. The stiffness between the midspan displacement and uniform load obtained from uniform static load testing was 338.3 kN/m^3 . Figure 13(b) shows the time history of the midspan displacements from the laser displacement sensor, static load, and amplified load under the RT status and Route 2. Evidently, the displacement calculated from the static load was lower than that caused by the actual crowd movement, demonstrating that the actual load is higher than the static load. In contrast, the displacements calculated from the amplified load covered the actual displacements in most cases.

The mean EDLF and the peak displacement at midspan of the 30 s steady-state were further compared under different working conditions. For convenience, the working conditions were represented as an abbreviated combination of the walking route and speed. The three walking routes were abbreviated as R1, R2, and R3, and

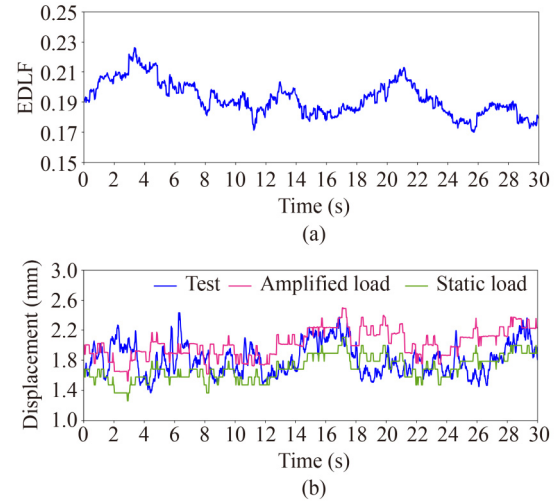


Fig. 13 Time history of the EDLF and displacement response under RT status and route 2: (a) variations in EDLF; (b) displacement of midspan.

the three walking speeds were abbreviated as S, M, and F. R1S denotes the working condition of the crowd walking on Route 1 at a slow speed. Figure 14(a) shows the mean value of the EDLF for each working condition. It was observed that the static load was amplified by approximately 27% on average for an unconstrained status, whereas the static load was amplified by approximately 20% on average for the constrained status. Based on the obtained EDLF, the peak values of the midspan displacements calculated using the static load and amplified load were compared with the peak value of the measured displacement for every working condition, as shown in Fig. 14(b). The actual displacements were larger than the static displacements in all working conditions, demonstrating the necessity of considering the dynamic effect of crowd movement in extraordinary load surveys. Under UT status, the peak displacements of the amplified load were closer to the measured values than the peak displacements of the static load, but most cases were on the small side. Because it is unlikely to produce a uniform load when the crowd quantity is small, the concentrated loading in the midspan is evident during crowd walking. Under a RT status, which is also a common situation during extraordinary events, the peak displacements of the amplified load were larger than the measured values.

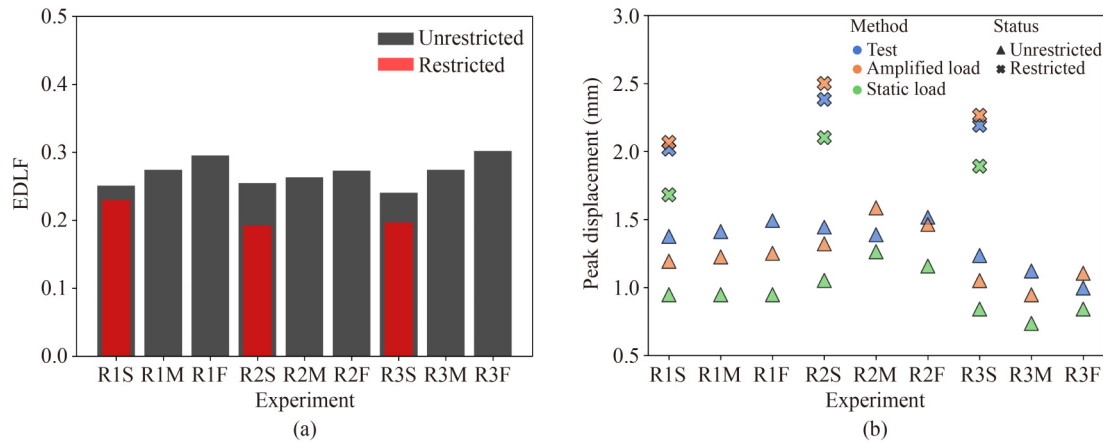


Fig. 14 (a) The mean value of EDLF and (b) the peak displacement of midspan in different working conditions.

The results prove the rationality of the EDLF for crowd quantity amplification, and the amplified load can incorporate the dynamic effects due to crowd movement.

4.3 Discussion

In addition to investigating crowd gathering events, the vision-based survey method is also applicable for collecting load data due to furniture stacking events. For that task, the extraordinary load due to furniture stacking events can be described as a series of different load types (i.e., different furniture). Two factors of each load type need to be determined in the load survey: the weight and number. The former can be obtained through web data, while the latter can be collected by the proposed survey method here. When furniture stacking events occur, the vision-based survey method can be employed to track each type of furniture and determine the quantity of furniture in the survey area. Accordingly, an image data set with various furniture types is necessary to extend the detection and reidentification model to detect and discriminate furniture effectively.

The proposed scheme can be utilized in two manners for long-term and continuous load survey tasks: group and individual surveys. In the former task, the proposed vision-based survey approach needs to be deployed on a powerful cloud server to automatically process surveillance videos uploaded from a group of buildings. In the latter case, the algorithms can run on a local computer to process surveillance videos of an individual building. Given current building surveillance and web bandwidth conditions, there are no technical obstructions to the implementation.

Since the proposed survey method uses surveillance cameras, applicable survey scenarios include office buildings, shopping malls, hospitals, schools, and public transportation hubs. In contrast, the proposed approach is less suitable for spaces where surveillance cameras have limited coverage or areas where surveillance cameras pose a risk of violating community privacy, such as in

residential buildings. In the case of dense crowds, the proposed survey method shows erratic tracking due to object occlusions and abrupt appearance changes. To obtain more accurate load data, it may be necessary to focus on both data set and algorithm improvements, such as by establishing a large image data set to enhance the performance of the detector and reidentification model and developing a tracking algorithm that incorporates scene understanding.

5 Conclusions

This study proposes a new vision-based survey method using surveillance cameras to collect load data during extraordinary events, including crowd quantity and velocities. A crowd head tracking framework consisting of a CHD model and a HReID model was developed to track the crowd and obtain the crowd head trajectory. Then, the trajectory analysis based on virtual intrusion lines was performed to obtain crowd quantity and velocities in the survey area. For survey areas with frequent crowd movements, the EDLF was calculated based on crowd velocities to amplify the crowd quantity to account for actual loading situations. The feasibility and effectiveness of the proposed survey method were investigated through crowd quantity monitoring and crowd walking experiments. The main findings of this study are.

1) The proposed vision-based survey framework, integrating the CHD-m model with AP50 of 95.4% and the HReID model with *MAP* of 88.2% and Rank-1 of 96.7%, provides good survey accuracy and low computational cost, thus being suitable for real-time survey tasks.

2) The dynamic effects caused by crowd movements in access areas should be considered in extraordinary load surveys and could be reasonably quantified using the EDLF calculated by the proposed survey method.

3) Compared with traditional methods based on questionnaires and questioning, the new survey method

can automatically and continuously collect extraordinary load information to form a reliable data basis for design load revaluation.

In the future, an extensive and continuous load survey using the proposed method can be conducted to provide a significant amount of data to update information on extraordinary and design loads for buildings.

Acknowledgements The authors acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 52178151). The helpful suggestions provided by Dr. Jinping Wang are greatly appreciated.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Competing interests The authors declare that they have no competing interests.

References

1. Peir J. A stochastic live load model for buildings. Dissertation for the Doctoral Degree. Cambridge: Massachusetts Institute of Technology, 1972
2. Peir J C, Cornell C A. Spatial and temporal variability of live loads. *Journal of the Structural Division*, 1973, 99(5): 903–922
3. McGuire R K, Cornell C A. Live load effects in office buildings. *Journal of the Structural Division*, 1974, 100(7): 1351–1366
4. Mitchell G R, Woodgate R W. Floor Loadings in Office Buildings: The Results of A Survey. Watford: Building Research Station, 1971
5. Culver C G. Live-load survey results for office buildings. *Journal of the Structural Division*, 1976, 102(12): 2269–2284
6. Jin X Y, Zhao J D. Development of the design code for building structures in China. *Journal of the International Association for Bridge and Structural Engineering*, 2012, 22(2): 195–201
7. Andam K A. Floor live loads for office buildings. *Building and Environment*, 1986, 21(3–4): 211–219
8. Choi E C C. Live load in office buildings point-in-time load intensity of rooms. *Proceedings of the Institution of Civil Engineers. Structures and Buildings*, 1992, 94(3): 299–306
9. Kumar S. Live loads in office buildings: Point-in-time load intensity. *Building and Environment*, 2002, 37(1): 79–89
10. Ellingwood B, Culver C. Analysis of live loads in office buildings. *Journal of the Structural Division*, 1977, 103(8): 1551–1560
11. Chalk P L, Corotis R B. Probability model for design live loads. *Journal of the Structural Division*, 1980, 106(10): 2017–2033
12. Harris M E, Bova C J, Corotis R B. Area-dependent processes for structural live loads. *Journal of the Structural Division*, 1981, 107(5): 857–872
13. Paloheimo E, Ollila M. Research in the Live Loads in Persons. Finland: Ministry of Domestic Affairs, 1973
14. Choi E C C. Live load for office buildings: Effect of occupancy and code comparison. *Journal of Structural Engineering*, 1990, 116(11): 3162–3174
15. Choi E C C. Extraordinary live load in office buildings. *Journal of Structural Engineering*, 1991, 117(11): 3216–3227
16. Andam K A. Variability in stochastic sustained and extraordinary load processes. *Building and Environment*, 1990, 25(4): 389–396
17. Kumar S. Live loads in office buildings: Lifetime maximum load. *Building and Environment*, 2002, 37(1): 91–99
18. Corotis R B, Jaria V A. Stochastic nature of building live loads. *Journal of the Structural Division*, 1979, 105(3): 493–510
19. Sentler L. A Live Load Survey in Domestic Houses. Lund: Division of Building Technology, Lund Institute of Technology, 1974
20. Issen L A. A literature review of fire and live load surveys in residences. Final Report National Bureau of Standards, 1978
21. Ho L V, Trinh T T, de Roeck G, Bui-Tien T, Nguyen-Ngoc L, Abdel Wahab M. An efficient stochastic-based coupled model for damage identification in plate structures. *Engineering Failure Analysis*, 2022, 131: 105866
22. Tran V T, Nguyen T K, Nguyen-Xuan H, Abdel Wahab M. Vibration and buckling optimization of functionally graded porous microplates using BCMO-ANN algorithm. *Thin-walled Structures*, 2023, 182: 110267
23. Dang B L, Nguyen-Xuan H, Wahab M A. An effective approach for VARANS-VOF modelling interactions of wave and perforated breakwater using gradient boosting decision tree algorithm. *Ocean Engineering*, 2023, 268: 113398
24. Cha Y J, Choi W, Büyüköztürk O. Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 2017, 32(5): 361–378
25. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In: *Proceedings of 2017 IEEE International Conference on Image Processing*. Beijing: IEEE; 2017, 3645–3649
26. Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137–1149
27. He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV: IEEE, 2016, 770–778
28. Xu Y, Bao Y Q, Chen J H, Zuo W M, Li H. Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images. *Structural Health Monitoring*, 2019, 18(3): 653–674
29. Ghosh Mondal T, Jahanshahi M R, Wu R T, Wu Z Y. Deep

- learning-based multi-class damage detection for autonomous post-disaster reconnaissance. *Structural Control and Health Monitoring*, 2020, 27(4): e2507
30. Nguyen D H, Wahab M A. Damage detection in slab structures based on two-dimensional curvature mode shape method and Faster R-CNN. *Advances in Engineering Software*, 2023, 176: 103371
 31. Nath N D, Behzadan A H, Paal S G. Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction*, 2020, 112: 103085
 32. Li Y, Chen J. Computer vision-based counting model for dense steel pipe on construction sites. *Journal of Construction Engineering and Management*, 2022, 148(1): 4021178
 33. Luo X C, Li H, Wang H, Wu Z Z, Dai F, Cao D P. Vision-based detection and visualization of dynamic workspaces. *Automation in Construction*, 2019, 104: 1–13
 34. Xiao B, Xiao H R, Wang J W, Chen Y. Vision-based method for tracking workers by integrating deep learning instance segmentation in off-site construction. *Automation in Construction*, 2022, 136: 104148
 35. Xiao B, Kang S C. Vision-based method integrating deep learning detection for tracking multiple construction machines. *Journal of Computing in Civil Engineering*, 2021, 35(2): 4020071
 36. He K M, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice: IEEE, 2017, 2961–2969
 37. Wang Z C, Zhang Q L, Yang B, Wu T K, Lei K, Zhang B H, Fang T W. Vision-based framework for automatic progress monitoring of precast walls by using surveillance videos during the construction phase. *Journal of Computing in Civil Engineering*, 2021, 35(1): 4020056
 38. Jocher G, Stoken A, Chaurasia A, Borovec J. YOLOv5: v6.0. 2021. Available at the website of GitHub
 39. Liu S, Qi L, Qin H F, Shi J, Jia J Y. Path aggregation network for instance segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018, 8759–8768
 40. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*. Lille: JMLR, 2015, 448–456
 41. Elfwing S, Uchibe E, Doya K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 2018, 107: 3–11
 42. Wang C Y, Liao H M, Wu Y H, Chen P Y, Hsieh J W, Yeh I H. CSPNet: A new backbone that can enhance learning capability of CNN. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Seattle: IEEE, 2020, 390–391
 43. He K M, Zhang X Y, Ren S Q, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904–1916
 44. Lin T Y, Dollár P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017, 2117–2125
 45. Kalman R E. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 1960, 82(1): 35–45
 46. Kuhn H W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955, 2(1–2): 83–97
 47. Ding X H, Zhang X Y, Ma N N, Han J G, Ding G G, Sun J. Repvgg: Making vgg-style convnets great again. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021, 13733–13742
 48. Wang J P, Chen J. A comparative study on different walking load models. *Structural Engineering and Mechanics*, 2017, 63(6): 847–856
 49. Tubino F, Piccardo G. Serviceability assessment of footbridges in unrestricted pedestrian traffic conditions. *Structure and Infrastructure Engineering*, 2016, 12(12): 1650–1660
 50. Chen J, Wang H Q, Peng Y X. Experimental investigation on Fourier-series model of walking load and its coefficients. *Journal of Vibration and Shock*, 2014, 33(8): 11–15 (in Chinese)
 51. Zhao D S, Chen J. Tests for correlation and modeling of individual 3-D continuous walking load. *Journal of Vibration and Shock*, 2019, 38(11): 166–172 (in Chinese)
 52. Wang J P, Chen J, Yokoyama Y, Xiong J C. Spectral model for crowd walking load. *Journal of Structural Engineering*, 2020, 146(3): 04019220
 53. Peng D Z, Sun Z K, Chen Z R, Cai Z R, Xie L L, Jin L W. Detecting heads using feature refine net and cascaded multi-scale architecture. In: *Proceedings of 2018 24th International Conference on Pattern Recognition*. Beijing: IEEE, 2018, 2528–2533
 54. Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C L. Microsoft COCO: Common objects in context. In: *Proceedings of Computer Vision—ECCV 2014: 13th European Conference*. Zurich: Springer, 2014, 740–755
 55. Kingma D P, Ba J. Adam: A method for stochastic optimization. 2014, arXiv: 1412.6980
 56. Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, Tulloch A, Jia Y Q, He K M. Accurate, large minibatch sgd: Training imagenet in 1 hour. 2017, arXiv: 1706.02677
 57. Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts. 2016, arXiv: 1608.03983
 58. Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, 88(2): 303–338
 59. Zhou K Y, Yang Y X, Cavallaro A, Xiang T. Omni-scale feature learning for person re-identification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019, 3702–3712
 60. Ma N N, Zhang X Y, Zheng H T, Sun J. Shufflenet v2: Practical guidelines for efficient CNN architecture design. In: *Proceedings of the European Conference on Computer Vision*. Munich: Springer, 2018, 116–131
 61. Iandola F N, Han S, Moskewicz M W, Ashraf K, Dally W J, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and < 0.5 MB model size. 2016, arXiv: 1602.07360
 62. Howard A, Sandler M, Chu G, Chen L C, Chen B, Tan M X, Wang W J, Zhu Y K, Pang R M, Vasudevan V, Le Q V, Adam H.

- Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019, 1314–1324
63. Deng J, Dong W, Socher R, Li L J, Li K, Li F F. Imagenet: A large-scale hierarchical image database. In: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009, 248–255
64. Wang X G, Doretto G, Sebastian T, Rittscher J, Tu P. Shape and appearance context modeling. In: Proceedings of 2007 IEEE 11th International Conference on Computer Vision. Rio de Janeiro: IEEE, 2007, 1–8
65. Zheng L, Shen L Y, Tian L, Wang S J, Wang J D, Tian Q. Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2015, 1116–1124