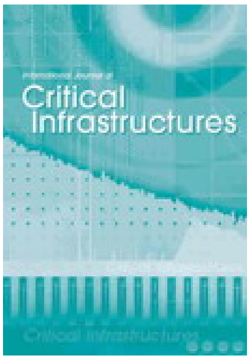
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Critical Infrastructures

ISSN online: 1741-8038 - ISSN print: 1475-3219
https://www.inderscience.com/ijcis

An adaptive recognition of abnormal behaviour in deep excavation support construction site of high-rise buildings

Wei Qi

DOI: 10.1504/IJCIS.2026.10076103

Article History:

Received:	30 September 2025
Last revised:	20 November 2025
Accepted:	23 November 2025
Published online:	10 February 2026

An adaptive recognition of abnormal behaviour in deep excavation support construction site of high-rise buildings

Wei Qi

School of Architecture Management,
Jiangsu Vocational Institute of Architectural Technology,
Xu Zhou, 221116, China
Email: jsjy11015@163.com

Abstract: To address the problems of high target false acceptance rates, low accuracy in abnormal behaviour recognition, and lengthy recognition times in traditional methods, this study proposes an adaptive recognition approach for abnormal behaviour in deep excavation support construction sites of high-rise buildings. Key frames are extracted from surveillance videos using the fractional Fourier transform, and object detection is implemented with spatiotemporal graph convolutional network models. Based on the target detection results, a CNN-LSTM model is used to achieve adaptive recognition of abnormal behaviour by capturing the temporal and spatial features of the target. Experimental results show that the proposed method achieves a minimum target false acceptance rate of 2.43%, a maximum recognition accuracy of 99.12%, and a minimum processing time of 0.19 s.

Keywords: high-rise buildings; deep foundation pit support; construction site; abnormal behaviour; adaptive recognition; key frames; CNN-LSTM.

Reference to this paper should be made as follows: Qi, W. (2026) 'An adaptive recognition of abnormal behaviour in deep excavation support construction site of high-rise buildings', *Int. J. Critical Infrastructures*, Vol. 22, No. 7, pp.1–17.

Biographical notes: Wei Qi received his Master's degree in Engineering Management from Nanjing Forestry University. He is currently a Lecturer at Jiangsu Vocational Institute of Architectural Technology. His research interests include engineering management and engineering mechanics.

1 Introduction

Against the backdrop of rapid urbanisation, the population size of cities is showing an explosive growth trend, and land resources are becoming increasingly scarce. In order to meet people's increasing demand for residential, office, and commercial spaces, high-rise buildings are emerging like new shoots breaking through the city skyline. High-rise buildings, with their unique advantages, can efficiently develop and utilise limited land resources, greatly enhancing the spatial utilisation efficiency of cities. Furthermore, it is widely regarded as a fundamental benchmark for evaluating urban modernisation. However, the construction of high-rise buildings faces many complex and severe technical challenges, and deep excavation engineering is undoubtedly one of the most critical links (Jiang, 2025). As a stable foundation support system for high-rise buildings, the safety performance and stability of deep excavation engineering directly determine the structural safety factor and service life of the entire building. In excavation projects, the support system for deep foundation pits serves a critical function in maintaining overall stability, effectively preventing disasters such as soil collapse and groundwater backflow, and creating a safe and reliable working environment for the smooth construction of underground structures (Zhu et al., 2024; Cheng et al., 2024). It can be seen that ensuring the quality and safety of deep foundation pit support engineering is of great significance for the overall construction of high-rise buildings. However, in actual construction scenarios, due to the comprehensive influence of multiple factors such as the uneven technical level of construction personnel and weak safety awareness, it is common to engage in behaviours that violate construction standards, such as illegal operations and cutting corners. These abnormal behaviours not only affect the construction quality of the support structure (Ji et al., 2024; Zhang, 2024), but also lead to damage during the construction or use of the support structure, causing safety accidents. It is therefore imperative to investigate adaptive behaviour recognition systems to enhance safety protocols on construction sites.

Therefore, Guo et al. (2024) proposed an adaptive recognition method for abnormal behaviour in construction sites based on convolutional neural network algorithm. This article deeply analyses the internal formation mechanism of abnormal behaviour of construction personnel. Based on ultra wideband wireless communication (UWB) high-precision positioning technology, camera self-calibration and calibration technology, and a CNN-based intelligent algorithm for behaviour recognition, an intelligent comprehensive management platform integrating real-time positioning, environmental perception, behaviour analysis, risk warning, and information exchange functions is constructed. The platform adaptively recognises abnormal behaviour on the construction site based on convolutional neural network algorithm and issues warnings to ensure comprehensive monitoring of the construction site. The accuracy of subsequent abnormal behaviour recognition, however, was found to suffer due to the method’s inadequate performance in precise target identification at construction sites. Zhang et al. (2023) proposed a method for identifying abnormal behaviours on construction sites based on YOLOv5. An improved algorithm framework based on YOLOv5 was proposed, and a collaborative object detection model between construction personnel and machinery was constructed to achieve accurate recognition of multi class objects. Furthermore, based on the object detection model, intelligent recognition strategies were designed for typical abnormal behaviours such as intrusion into static hazardous areas, illegal operation of dynamic construction machinery, and lack of personnel safety protective equipment. However, this method demonstrates limited accuracy in identifying abnormal behaviours, falling short of the expected performance target. Zhao et al. (2023) proposed a method for identifying abnormal behaviours on construction sites based on DBN-SVM. Firstly, through a multi-source data collection strategy, the system collected an image dataset covering different lighting conditions, personnel postures, and occlusion situations. The dataset subsequently undergoes a preprocessing pipeline comprising denoising, normalisation, and dimensional adjustment to align with the feature distribution specifications required for deep belief network (DBN) input. On this basis, a hybrid model based on DBN feature extraction and support vector machine (SVM) classification (DBN-SVM) was constructed. DBN is responsible for automatically learning low dimensional feature representations from high-dimensional images, while SVM makes binary classification decisions based on the extracted features, ultimately achieving accurate detection of abnormal behaviour and target localisation. However, building a DBN-SVM hybrid model is relatively complex, resulting in increased recognition time and low efficiency.

To solve the problems existing in the above methods, an adaptive recognition method of abnormal behaviour in deep excavation support construction site of high-rise building is proposed.

2 Adaptive recognition of abnormal behaviour in the construction site of deep foundation pit support for high-rise buildings

2.1 Key frame extraction of monitoring video for deep foundation pit support construction site of high-rise buildings

The construction of deep foundation pit support is the fundamental guarantee for the safety of high-rise buildings, and the acquisition of monitoring videos needs to meet the requirements of real-time, accuracy, and traceability. Due to the complex and ever-changing environment of the construction site, there is a large amount of dynamic interference, such as the movement of mechanical equipment, frequent personnel activities, changes in lighting conditions, and dust effects, resulting in a large number of redundant and repetitive frames in the video sequence. At the same time, it is necessary to ensure that the extracted key frames can accurately capture the moments of key safety status changes such as deformation of support structures and personnel violations; Its technical feature lies in the use of fractional Fourier transform and focusing on the phase spectrum information of the image, utilising the strong representation ability of the phase spectrum for the overall structure and edge contour of the image. By calculating the mean square error (MSE) of the phase spectrum between adjacent frames and forming an MSE curve, it adaptively detects local peak points to screen out key frames that truly reflect significant changes in scene content. This method effectively overcomes the shortcomings of traditional extraction methods that overly rely on surface features such as colour and texture, and improves the ability of key frames to represent key events in complex construction environments. The procedure for obtaining such monitoring videos during the construction of these support systems is detailed below:

Firstly, anti shake and waterproof cameras are installed at key locations around the foundation pit, such as support pile tops, crown beam nodes, and slope top deformation monitoring points. They are arranged in a circular pattern with a 5-metre spacing to ensure full coverage, and the coordinates of the monitoring points are calibrated through RTK positioning. During the construction process, the top of the tower crane is equipped with a 200° wide-angle ball camera for overhead shooting, and a dust-proof camera (IP67 rating) with fill lights is deployed at the joint of the ground wall. All equipment is connected to a gigabit fibre optic ring network through on-site industrial switches, and 1,080P video is transmitted in real-time to the project monitoring centre at a frame rate of 30 fps. For nighttime construction, a combination of thermal imaging cameras and laser night vision devices is used to ensure clear capture of subtle deformations in the support structure even in low light environments. The video data is synchronously stored in the local NVR and cloud platform, and marked with timestamps, coordinate information, and construction conditions (such as excavation depth and support stage), providing structured data support for subsequent intelligent analysis.

The keyframe extraction methodology for monitoring videos of high-rise building foundation pit construction sites, based on fractional Fourier transform, proceeds as follows:

1 Video frame sequence generation: Extract complete surveillance footage from the monitoring system and segment it according to temporal parameters to generate a continuous video frame sequence (Lu et al., 2025a). This process utilises actual monitoring records from deep foundation pit support construction sites.

2 Perform image preprocessing on the surveillance video frame sequence generated in step (1), starting with greyscale processing. Greyscale processing can convert colour images into greyscale images containing only brightness information, and the processed images no longer have colour data. The reason for this step is that in terms of storage, the amount of greyscale image data is much smaller than that of colour images. Colour images require multi-channel recording of colours, while greyscale images only need to store brightness in a single channel, which is more conducive to large-scale surveillance video storage. In terms of computational efficiency, subsequent algorithms such as image analysis and object detection can reduce colour information processing steps, accelerate computation speed, and improve method efficiency when processing greyscale images due to their simple data structure (Huang and Zhang, 2024).

3 According to the adaptive method based on the golden ratio point, select the appropriate transformation order p, and finally determine the value of transformation order p as 0.6.

4 Perform the two-dimensional fractional Fourier transform on the video frame sequence using the order p determined in step (3), yielding the transformed result $F^{p_1,p_2}(u,v)$. The transform is defined as follows:

$$F^{p_1,p_2}(u,v)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)K^{p_1,p_2}(x,y,u,v)dxdy \tag{1}$$

In this expression, $f(x, y)$ corresponds to the phase spectrum of the video frame, with (x, y) and (u, v) indicating spatial and fractional frequency domain coordinates, respectively. The term K refers to the fractional Fourier transform kernel, while the transform orders are set as $p_1 = p_2 = 0.6$ according to Shekhar and Agrawal (2024).

5 The phase spectrum captures essential edge contours and global structural information of images, making it suitable for subsequent analytical steps (Zheng et al., 2025). Based on two-dimensional Fourier transform principles, the fractional Fourier phase spectrum $\varnothing^{p_1,p_2}(u,v)$ is extracted from $F^{p_1,p_2}(u,v)$ using the following expression:

$$\varnothing^{p_1,p_2}(u,v)=\arctan\left(\frac{I^{p_1,p_2}}{R^{p_1,p_2}}\right) \tag{2}$$

Here, $I^{p_1,p_2}(u,v)$ corresponds to the real component of $F^{p_1,p_2}(u,v)$; $R^{p_1,p_2}(u,v)$ represents its imaginary part.

6 With the phase spectrum information of the surveillance video frame sequence obtained, the MSE between adjacent frames is computed. The MSE serves as a metric for image quality assessment, determining the degree of difference between a reference image and its distorted version. By using the MSE as an evaluation metric, it is possible to more accurately determine the changes in targets in surveillance video frame sequences. Assuming that the length and width of a video frame are M and N, respectively, and the phase spectrum of the current video frame is denoted as $f(x, y)$ and the phase spectrum of the previous video frame is denoted as $g(x, y)$, the MSE of adjacent frames can be obtained. The calculation formula is as follows:

$$MSE = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} \left(f(x, y) - g(x, y) \right)^2 \tag{3}$$

7 Using the MSE values of adjacent video frames obtained in step (6), form a MSE curve, termed the MSE curve.

8 Detect peak values on the peak curve formed in step (7), and use the peak points detected during this period as the basis for analysing important frames, in order to determine the optimal keyframe points.

9 After detecting the peak points in step (8), extract all local peak points from the MSE curve. Include all extracted local peak points into the candidate keyframe sequence and define it as the candidate keyframe set. The number of keyframes in this candidate keyframe set is represented as S.

10 Select the final required keyframe from the already defined candidate keyframes. Firstly, calculate the numerical differences between all local peak points and their previous and next frames, and calculate their average. Then, the following relationship holds:

$$\begin{cases} avga = \sum_{i}^{n} a_i / (S - 1) \\ avgb = \sum_{i}^{n} b_i / (S - 1) \end{cases} \tag{4}$$

In the above formula, a_i and b_i represent the numerical difference between the peak point and the previous and next frames, respectively. Here, *avga* and *avgb* denote the average differences between the peak point and the preceding and subsequent frame values, respectively.

Extract video frames with a value difference greater than Navgb from the previous video frame and a value difference greater than Navgb from the next video frame as keyframes. According to the above operation, extract all keyframes on the MSE curve in sequence and determine them as the final keyframes.

2.2 *Target detection at the construction site of deep foundation pit support for high-rise buildings*

After extracting key frames from monitoring videos at the construction site of deep foundation pit support in high-rise buildings, target detection is carried out on the construction site of deep foundation pit support in high-rise buildings. In this process, the spatiotemporal graph convolutional network is innovatively applied to target recognition in the construction scene. By constructing a skeleton point spatiotemporal graph model that combines spatial and temporal edges, and proposing a three zone strategy based on the centroid, the nodes are divided into root nodes, centripetal groups, and centrifugal groups to more accurately describe the dynamic posture of construction personnel and equipment operation trajectories; On this basis, a new graph convolution weight function and partition mapping rule were designed to achieve collaborative extraction and adaptive fusion of multi-objective spatiotemporal features in complex construction scenes, significantly improving the recognition accuracy and robustness of support structures, personnel, and mechanical targets in complex situations such as occlusion and overlap.

The spatiotemporal graph convolutional network model combines graph convolutional networks and temporal convolutional networks (TCN) to extend to spatiotemporal graph models (Lu et al., 2025b), and designs a universal representation $G = (V, E)$ for skeleton point sequences used in behaviour recognition. This model represents the human skeleton as a graph, and the joints of the human body correspond to the nodes in the graph. There are two types of edges in the figure, one is the spatial edge $E_S = \{v_{ti}v_{tj} \mid (i, j) \in H\}$ connected by human joints at the same time, and the other is the temporal edge $E_F = \{v_{ti}v_{(t+1)j}\}$ connected by the same joint at adjacent times. Afterwards, a formulaic definition is given for ordinary convolutions. Given a feature map f_{in}, the single channel output at spatial position x is represented by the following formula:

$$f_{out}(x) = \sum_{h=1}^{K} \sum_{\omega=1}^{K} f_{in}\left(p(x, h, \omega) \cdot w(h, \omega)\right) \tag{5}$$

In the formula above, p denotes the sampling function of x, and ω represents the inner product of the c-dimensional channel input feature vector, respectively. The definition of graph convolution is derived from the extension of ordinary convolution, and the calculation formula for sampling function p is as follows:

$$v_{tj} = p\left(v_{ti}, v_{tj}\right) \tag{6}$$

The weight function of graph convolution is calculated by dividing the neighbour set $B(v_{ti})$ of a certain joint point v_{ti} into a fixed set of K subsets. The graph convolution's weight function is formulated as follows:

$$w\left(v_{ti}, v_{tj}\right) = w'\left(l_{ti}\left(v_{tj}\right)\right) \tag{7}$$

Introduce p and weight function w to change formula (4) to graph convolution form. The improved formula is as follows:

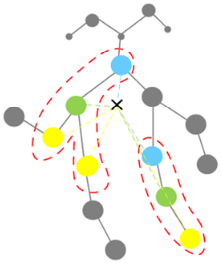
$$f_{out}\left(v_{ti}\right) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti(v_{tj})}} f_{in}\left(p\left(v_{ti}, v_{tj}\right)\right) \cdot w\left(v_{ti}, v_{tj}\right) \tag{8}$$

Combining formulas (5), (6), and (7) yields the following result:

$$f_{out}\left(v_{ti}\right) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti(v_{tj})}} f_{in}\left(v_{ti}\right) \cdot w\left(l_{ti}\left(v_{tj}\right)\right) \tag{9}$$

A new partitioning strategy is incorporated in the spatiotemporal graph convolutional network model, illustrated in Figure 1.

Figure 1 Partition strategy of spatiotemporal graph convolutional network (see online version for colours)



Divide the nodes in the skeleton diagram into three parts: root nodes, centripetal groups, and centrifugal groups. Firstly, calculate the average coordinates of all nodes, represented as centroids, as indicated by the black cross in the figure. The root node itself is divided into the first subset, as shown by the green nodes in the figure. The centripetal group is divided into a second subset, which contains adjacent nodes closer to the centroid than the root node, as shown by the blue nodes in the figure (Jiang et al., 2025). The centrifugal group is classified as the third subset, in which adjacent nodes are further away from the centroid compared to the root node, as can be seen intuitively from the nodes marked in yellow in the graph. The calculation formula for partition strategy is as follows:

$$l_{ti}\left(u_{tj}\right) = \begin{cases} 0, & \text{if } r_j = r_i \\ 1, & \text{if } r_j < r_i \\ 1, & \text{if } r_j > r_i \end{cases} \tag{10}$$

where r_i denotes the average distance between the skeleton centroid and node i in the surveillance video's key frames. The target detection calculation for the high-rise building deep foundation pit support construction site is expressed as follows:

$$f_{out} = \sum\nolimits_j \text{ì}_j^{-\frac{1}{2}} \left(A_j \otimes M\right) \text{ì}_j^{-\frac{1}{2}} f_{in} W_j \tag{11}$$

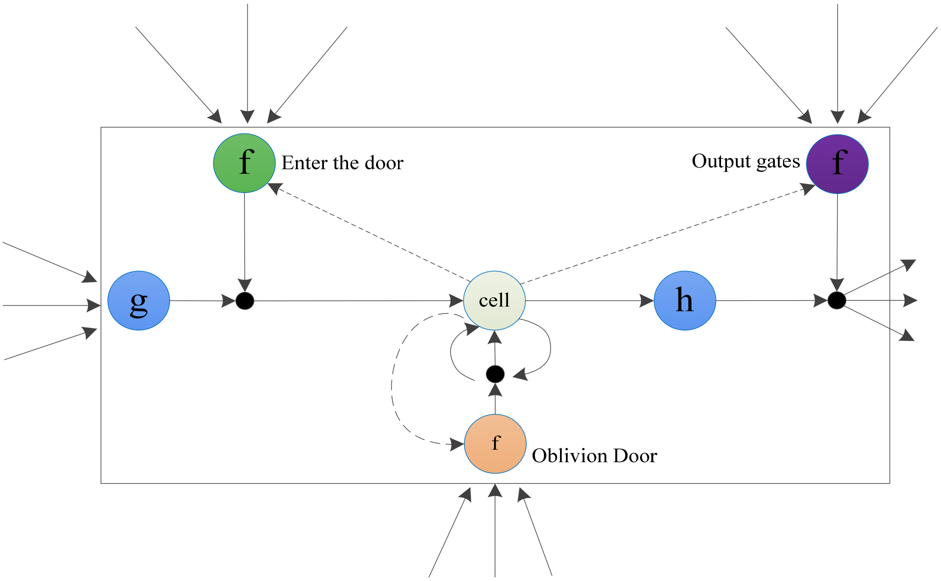
where W_j denotes the weight of target feature points in the surveillance video's key frames (Zhang et al., 2024), M is the on-site target feature matrix, and A_j is the feature weight coefficient matrix.

2.3 Adaptive recognition method for abnormal behaviour based on CNN-LSTM

In the CNN-LSTM-based adaptive recognition method for abnormal behaviour, this section innovatively constructs an end-to-end recognition architecture that deeply integrates spatiotemporal features. By seamlessly coupling the deep spatial features extracted by the Inception-v3 network with the long-term time-dependent model constructed by the LSTM network, the serialisation modelling and adaptive recognition of abnormal behaviour are achieved for the first time in deep excavation construction scenarios; This method innovatively uses the temporal mean vector of LSTM hidden state sequence as the global behaviour representation, and designs a Softmax classification mechanism for construction abnormal behaviour, effectively solving key problems such as spatiotemporal feature separation and insufficient capture of long-term dependencies in traditional methods. It significantly improves the identification accuracy and scene adaptation ability for typical construction abnormal behaviours such as illegal operations and area intrusion.

As a variant of recurrent neural networks (RNN), LSTM demonstrates distinctive advantages in capturing long-range dependencies. Its design effectively addresses gradient instability and supports sophisticated sequence modelling. The fundamental LSTM architecture is shown in Figure 2.

Figure 2 Basic structure of LSTM (see online version for colours)



Assuming a time step of t, the calculation formula for updating the unit state is as follows:

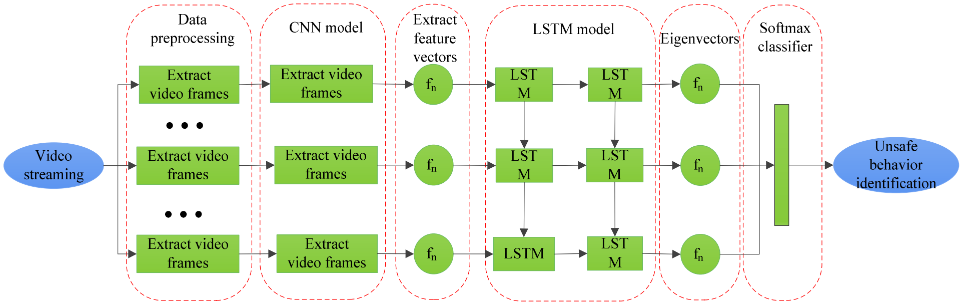
$$\begin{cases} i_t = \delta\left(W_{xi}x_t + V_{hi}h_{t-1} + b_i\right) \\ f_t = \delta\left(W_{xf}x_t + V_{hf}h_{t-1} + b_f\right) \\ o_t = \delta\left(W_{xo}x_t + V_{ho}h_{t-1} + b_o\right) \\ g_t = \tanh\left(W_{xc}x_t + V_{hc}h_{t-1} + b_c\right) \\ c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \\ h_t = o_t \otimes \tanh\left(c_t\right) \end{cases} \tag{12}$$

In this expression: i_t, f_t, o_t and c_t denote the temporal outputs of the input gate, forget gate, output gate, and memory unit; h_t, b_t, b_f, b_o and b_c represent different bias vectors; W_{xi}, W_{xf}, W_{xo}, W_{xc}, V_{hi}, V_{hf}, V_{ho} and V_{hc} are respective coefficient matrices; and δ is the sigmoid function, defined by:

$$\delta(x) = \left(1 + e^{-x}\right)^{-1} \tag{13}$$

This model has strong adaptive capabilities, which can automatically and accurately identify abnormal behaviour patterns based on the feature values of input data. In the complex and uncertain environment of deep excavation support construction sites for high-rise buildings, timely and accurate recognition of abnormal behaviour is crucial. The application of LSTM models to adaptive abnormal behaviour recognition in this context demonstrates considerable practical utility, offering robust technical support for construction site safety management. Figure 3 illustrates the proposed CNN-LSTM adaptive recognition framework for identifying anomalous behaviours at deep foundation pit support sites of high-rise buildings.

Figure 3 Adaptive recognition model for abnormal behaviour of deep excavation support construction site in high-rise buildings based on CNN-LSTM (see online version for colours)



The CNN-LSTM based adaptive behaviour recognition model for deep foundation pit construction sites employs a dual-stage architecture. Initially, convolutional neural networks process preprocessed action videos to extract spatial feature representations. Subsequently, LSTM networks capture temporal feature sequences, which are integrated into the CNN's final layer through a Softmax classifier for adaptive anomaly detection. The Softmax function used in this classification layer operates as a probability mapping function, formulated as follows:

$$P\left(y^{(i)}=n \mid x^{(i)};W\right)=\begin{bmatrix} y^{(1)}=1 \mid x^{(1)};W \\ y^{(2)}=1 \mid x^{(2)};W \\ \vdots \\ y^{(i)}=1 \mid x^{(i)};W \end{bmatrix}=\frac{1}{\sum_{j=1}^{n} w^{WTx(i)}}\begin{bmatrix} e^{W_1^T}x^{(i)} \\ e^{W_2^T}x^{(i)} \\ \vdots \\ e^{W_n^T}x^{(i)} \end{bmatrix} \tag{14}$$

where P corresponds to the i^{th} training instance from n video keyframe samples of high-rise building foundation pit support construction, indicates its assigned class among j categories with weight W and $W_j^t x^{(i)}$ denotes the input to the Softmax layer. The implementation workflow of the model proceeds as follows:

1 Random temporal segments are sampled from surveillance footage of high-rise building foundation pit support construction to represent key time nodes. For the task of extracting target features from these key frames, video sequences are initially processed through a CNN architecture utilising the Inception-v3 framework. This framework parallelises multi-scale convolution kernels and efficiently extracts

features. After the video frame enters the model, the convolutional layer uses a learnable convolution kernel to perform convolution operations on each local region of the input video frame, that is, to calculate the dot product of the convolution kernel and the covered small region. The convolution kernel slides through the video frame at a specific step size to generate multiple feature maps. After multiple convolutional and pooling layers, the fully connected layer finally outputs a 2,048 dimensional feature vector. This vector contains key features such as shape, texture, and edges of the target space dimension, which can accurately represent the spatial characteristics of keyframe targets and provide data support for subsequent tasks.

2 The spatial characteristics of keyframe targets are extracted by iterating the initial processing step, generating feature vectors for each video segment. These vectors are subsequently fed into an LSTM network for temporal feature learning. The temporal feature generation process initiates from the second LSTM module in the sequence.

3 Input all time features into the Softmax classifier, take the average over the time period, and obtain a probability representation.

During forward propagation in CNN architectures, each processing stage sequentially performs two core operations on preceding layer outputs: convolutional filtering through sliding kernels for local feature extraction, followed by activation function transformation to enhance representational capacity. The mathematical implementation is formalised as follows:

$$h_{ij}^{k} = f\left((W^{k}x)_{ij} + b_{k}\right) \tag{15}$$

where f represents the activation function, b_k represents the offset of the feature map, and W^k represents the value connected to the k^{th} feature map.

The abnormal behaviour recognition model for deep foundation pit support construction sites employs a dual-layer LSTM architecture to process temporal dynamics from Inception-v3's final pooling layer. Let $\{f_1, f_2, \ldots, f_n\}$ denote the n features extracted by Inception-v3 from n video frames. For each input sequence $\{f_1, f_2, \ldots, f_n\}$, the memory units in both LSTM layers produce a representation sequence $\{m_1, m_2, \ldots, m_n\}$. The temporal feature vector F is then obtained by temporally averaging this sequence, expressed as:

$$F = \frac{m_1 + m_2 + \cdots\cdots + m_n}{n} \tag{16}$$

The resulting feature vector F is subsequently processed by the Softmax layer for the adaptive detection of anomalous activities on the deep foundation pit support construction site of high-rise buildings.

In summary, the proposed method has multiple technical characteristics for identifying abnormal behaviours in the construction site of deep foundation pit support for high-rise buildings. This method first collects construction site videos through multi-source devices such as anti shake waterproof cameras, thermal imaging cameras, and laser night vision devices, and uses keyframe extraction technology based on fractional Fourier transform to effectively capture the phase spectrum reflecting the overall state of the structure and edge profile information in the video sequence. By calculating the MSE curve of the phase spectrum between adjacent frames and extracting local peak points, representative keyframes are adaptively selected to significantly reduce data redundancy and enhance change sensitivity. In the target detection stage, a spatiotemporal graph convolutional network model is used to model the human skeletal joint points as graph structures. The behaviour of construction personnel is dynamically described by combining spatial and temporal edges, and a partitioning strategy based on the centroid is introduced to accurately extract spatiotemporal features of support structures, machinery, and personnel targets. Subsequently, the method constructs a CNN-LSTM hybrid recognition model, extracts deep spatial features of targets in keyframes through the Inception-v3 framework, and uses the triple gating mechanism of input gate, forget gate, and output gate of the LSTM network to model the long-term temporal dependence of the behaviour sequence, effectively capturing the temporal patterns of illegal operations. Finally, the aggregated temporal feature vector is input into the Softmax classifier to achieve adaptive recognition of abnormal behaviour. The entire method is optimised for the complexity and dynamics of deep excavation construction scenes in video capture, keyframe extraction, object detection, and behaviour classification. It not only has high-precision and high-efficiency recognition performance, but also demonstrates strong environmental adaptability and engineering practicality.

3 Experimental design

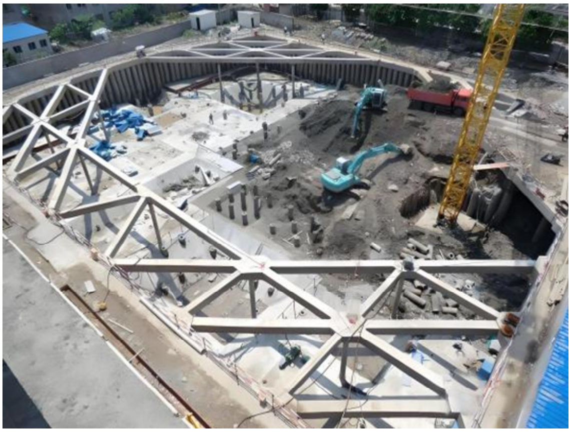
3.1 Experimental scheme

To validate the practical efficacy of the proposed framework in detecting anomalous activities at high-rise building foundation pit support sites, a series of experimental evaluations were implemented.

3.1.1 Experimental setup and data collection

The physical construction environment used for abnormal behaviour monitoring is depicted in Figure 4, which illustrates the operational context for all subsequent data acquisition and analysis.

Figure 4 Construction site (see online version for colours)



The acquisition of surveillance videos follows the following professional process.

Firstly, in the selection and installation of monitoring equipment, a high-definition camera with a resolution of not less than 1,080P should be selected to ensure clear capture of key details such as cracks in the support structure and personnel violations. At the same time, equipped with cameras with infrared night vision or low light function to ensure normal monitoring during nighttime construction or low light environments; In addition, cameras with a protection level not lower than IP66 should be selected to adapt to harsh environments such as dust and rainwater on deep excavation construction sites; Prioritise the use of cameras with intelligent functions such as behaviour analysis and object detection to effectively reduce the pressure of manual monitoring. In terms of equipment installation, fixed brackets should be installed on supporting structures, surrounding buildings, or dedicated monitoring poles to ensure camera stability. Provide stable power supply for the camera and connect to the on-site network through wired or wireless means. After installation, it is necessary to adjust the camera focal length, angle, and brightness to ensure clear and distortion free images, and calibrate time synchronisation. Secondly, in the process of video data collection and transmission, cameras should be setup to record continuously 24 hours a day to ensure coverage of all construction periods. Adopting a dual stream storage method of main stream and sub stream to balance storage space and image quality. In terms of data transmission, video data should be transmitted to the on-site monitoring centre through fibre optic or Ethernet cables to ensure transmission stability and bandwidth. In areas where wiring is not possible, 4G/5G wireless transmission technology is used to ensure real-time uploading of video data. Deploy edge computing equipment on the camera side to preliminarily divide the video data to reduce the amount of data transmission. In terms of storage solutions, NVRs or disk arrays should be deployed in the monitoring centre to store video data for nearly 30 days. Synchronise and upload video data to the cloud for long-term storage and remote backup to prevent data loss.

To ensure the accuracy of simulation results, the initial simulation parameters are strictly set according to the monitoring conditions and equipment performance indicators of the actual high-rise building deep foundation pit support construction site. The order parameters of the fractional Fourier transform in the keyframe extraction stage are adaptively optimised based on the golden ratio method and fixed at 0.6 to balance time-frequency resolution. The spatiotemporal graph convolutional network part sets the number of bone graph nodes to 18 and adopts a three partition strategy, In the CNN-LSTM model, the Inception-v3 input size is set to 299 × 299 and the LSTM hidden layer dimension is set to 1,024. During the training phase, the batch size is fixed at 32 and the initial learning rate is set to 0.001, using an exponential decay strategy. All experiments were conducted on the same hardware platform and deep learning framework, and the systematic configuration of the above parameters effectively ensured the data consistency and algorithm comparability between the simulation process and real construction scenarios.

3.1.2 Experimental indicators

The Guo et al. (2024) method, Zhang et al. (2023) method, and the proposed method were selected for comparative evaluation. Their performance was assessed using three key metrics: target false acceptance rate (FAR), abnormal behaviour identification accuracy, and computational time required for behaviour recognition at high-rise building foundation pit support construction sites.

The target false detection rate at the construction site of deep foundation pit support for high-rise buildings specifically refers to the proportion of non-target objects incorrectly identified as target objects by intelligent monitoring systems at the construction site of deep foundation pit support for high-rise buildings. The lower the false detection rate, the higher the target recognition accuracy.

The accuracy of identifying abnormal behaviours on the construction site of deep foundation pit support for high-rise buildings refers to the proportion of intelligent monitoring systems that correctly identify abnormal behaviours (such as illegal operations and safety hazards) on the construction site of deep foundation pit support for high-rise buildings. The higher this value, the more it can ensure that abnormal behaviours are identified in a timely and accurate manner.

The time required for identifying abnormal behaviour on the construction site of deep foundation pit support in high-rise buildings refers to the total time from capturing on-site data to completing abnormal behaviour detection, which directly affects the timeliness of safety response. The shorter the identification time, the higher the identification efficiency.

3.2 Experimental result

Table 1 presents the comparative results of the target FAR at the construction site, as obtained by the three evaluated methods.

Data analysis from Table 1 reveals significant performance differences in target FARs. The Guo et al. (2024) method exhibits FAR values ranging from 18.45% to 22.36%, while Zhang et al. (2023) method shows a range of 12.01% to 18.96%. In contrast, the proposed method demonstrates substantially superior performance with FAR between 2.43% and 3.69% – representing reductions of 16.02%–18.67% and 9.58%–15.27% compared to Guo et al. (2024) method and Zhang et al. (2023) method, respectively. This comparative analysis confirms the proposed method's enhanced capability in accurately identifying on-site targets, thereby establishing a reliable foundation for subsequent abnormal behaviour detection in deep foundation pit support construction environments. The key reason why the proposed method can achieve a lower target false detection rate in the construction site of deep foundation pit support for high-rise buildings is that it comprehensively uses fractional Fourier transform to extract key frames from monitoring videos. This method calculates the MSE between adjacent video frame phase spectra and accurately locates local peak points, thereby selecting the key frames that best represent the changes in scene structure, effectively filtering out image redundancy and instantaneous interference caused by factors such as lighting fluctuations, personnel movement, and mechanical obstruction; On this basis, a spatiotemporal graph convolutional network model is adopted for object detection. The model constructs the human skeletal joint points as a spatiotemporal graph structure, models the dynamic behaviour of the target through spatial and temporal edges, and uses a partitioning strategy based on the centroid to distinguish root nodes, centripetal groups, and centrifugal groups, enhancing the feature recognition ability of core targets such as support structures, construction personnel, and mechanical equipment in complex scenes, and avoiding the misjudgement caused by traditional methods relying on apparent features or static models; Through the close integration of the above process, the distinguishability and robustness of target features were significantly improved, enabling the method to more accurately identify real targets in complex construction site environments, minimising the probability of misidentifying background disturbances or non-critical objects as construction targets.

Table 1 Test results of FAR

Number of experiments	*False acceptance rate/%*		
	Guo et al. (2024) method	*Zhang et al. (2023) method*	*Proposed method*
10	20.33	15.63	2.48
20	20.14	12.47	2.63
30	22.36	18.96	2.47
40	19.62	12.38	3.15
50	18.45	16.75	3.69
60	20.39	12.36	2.43
70	21.87	13.78	2.55
80	22.18	14.63	2.96
90	24.16	12.01	3.14

The accuracy performance of the three methods in identifying anomalous behaviours is presented in Table 2.

Data from Table 2 indicates notable performance variations in abnormal behaviour identification accuracy. The Guo et al. (2024) method achieves accuracy rates between 78.56% and 84.75%, while the Zhang et al. (2023) method ranges from 80.23% to 86.33%. In comparison, the proposed method demonstrates markedly superior performance with accuracy values of 96.38% to 99.12% - representing improvements of 16.15-17.82% over Zhang et al. (2023) method and 12.79-14.37% over Guo et al. (2024) method. This comparative evaluation confirms the proposed method's enhanced capability in accurately detecting anomalous behaviours at high-rise building foundation pit support sites, thereby fulfilling the research objectives. The reason why the proposed method can achieve higher accuracy in identifying abnormal behaviours is because it uses a CNN-LSTM hybrid model to deeply integrate spatial and temporal features; Firstly, a convolutional neural network based on Inception-v3 is used to efficiently extract deep spatial features of construction personnel, mechanical equipment, and support structures from keyframes extracted by fractional Fourier transform, including key visual information such as posture shape and edge texture; Subsequently, these spatial feature sequences are input into the LSTM network, and its unique input gate, forget gate, and output gate mechanisms are used to accurately model the long-term temporal dependencies of behavioural actions, thereby effectively capturing the complete evolution patterns of abnormal behaviours such as violations and regional intrusions in the temporal dimension; The collaborative mechanism of spatial feature extraction and temporal dynamic modelling enables the model to not only accurately identify static abnormal signs in a single frame, but also judge dynamic patterns that conform to abnormal behaviour rules from continuous action sequences. Finally, the aggregated spatiotemporal features are accurately classified by the Softmax classifier, significantly improving the overall recognition accuracy and discrimination reliability of various abnormal behaviours in complex construction scenes.

Table 2 Results of accuracy test for abnormal behaviour recognition

Number of experiments	*Recognition accuracy/%*		
	Guo et al. (2024) method	*Zhang et al. (2023) method*	*Proposed method*
10	80.66	80.23	98.52
20	78.56	81.47	97.48
30	81.23	85.26	98.23
40	84.56	86.33	97.46
50	82.31	85.28	99.12
60	84.75	84.17	98.63
70	78.63	85.69	98.52
80	79.61	82.17	96.38
90	81.34	83.64	96.71

Figure 5 Time consumption test results for abnormal behaviour recognition (see online version for colours)

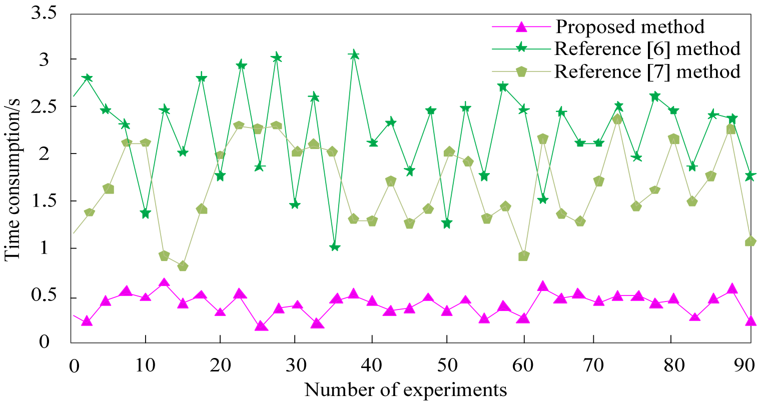


Figure 5 presents the time consumption comparison for abnormal behaviour identification across the three evaluated methods.

Figure 5 reveals substantial differences in processing time for abnormal behaviour identification. The Guo et al. (2024) method requires 0.88s~3.08s per detection, while Zhang et al. (2023) method needs 0.75 s~2.27 s. In contrast, the proposed approach demonstrates significantly faster performance with processing times of only 0.19 s~0.67 s – representing reductions of 0.56 s~2.39 s compared to the benchmark methods. This comparative analysis confirms the computational efficiency of the proposed method, achieving the research objective of real-time abnormal behaviour monitoring at high-rise building foundation pit support construction sites. The reason why the proposed method can significantly reduce the time consumption of abnormal behaviour recognition is mainly due to the efficient optimisation of its overall process and the reasonable allocation of computing resources; This method first extracts keyframes from surveillance videos through fractional Fourier transform, and quickly locates representative frames using the phase spectrum MSE curve, greatly reducing the amount of video data that needs to be processed later and compressing the computational load from the source; In the object detection stage, the spatiotemporal graph convolutional network avoids redundant calculations of all pixels in the image by using pre-defined skeleton point spatiotemporal graph structures and partitioning strategies, achieving fast and accurate localisation of construction targets; The CNN-LSTM model subsequently adopted fully utilised the advantages of end-to-end architecture, where the Inception-v3 network extracts spatial features in parallel through multi-scale convolutional kernels, while the LSTM network efficiently processes temporal dependencies through gating mechanisms. The two work together to avoid the time overhead caused by complex feature engineering and multi model concatenation in traditional methods; At the same time, the entire processing flow has achieved pipeline optimisation from keyframe extraction to behaviour classification, with efficient and compact data transmission between each loop, ultimately enabling the method to quickly identify and respond to abnormal behaviours while maintaining high accuracy.

4 Conclusions

Deep foundation pit engineering is a key link in the construction of underground structures such as high-rise buildings, subway stations, and underground complexes. Its construction environment is complex and affected by multiple factors such as geological conditions, hydrological conditions, construction techniques, and external loads, making it prone to safety accidents. Therefore, an adaptive recognition method of abnormal behaviour in deep excavation support construction site of high-rise building is proposed. The experimental results show that the proposed method has a minimum target FAR of 2.43%, a maximum abnormal behaviour recognition accuracy of 99.12%, and a minimum time consumption of 0.19s for the construction site of deep foundation pit support in high-rise buildings. The proposed framework demonstrates high-precision recognition capabilities coupled with significant computational efficiency. This research contributes to enhanced safety and reliability in deep foundation pit support engineering, effectively reducing accident risks while improving construction productivity and cost-effectiveness. Looking forward, the convergence of artificial intelligence, Internet of Things, and digital twin technologies will propel excavation monitoring toward more intelligent and proactive paradigms. This technological evolution will ultimately support the achievement of 'zero-accident' objectives in intelligent construction practices.

Declarations

All authors declare that they have no conflicts of interest.

References

Cheng, X., Yin, S., Li, X. et al. (2024) 'Optimisation for sandy pebble deep foundation pit support based on multi-objective fuzzy grey relation projection method', *Australian Journal of Civil Engineering*, Vol. 22, No. 2, pp.215–222.

Guo, F., Kong, H. and Qiao, G.G. (2024) 'Research on intelligent recognition of worker's unsafe behavior in urban rail transit based on convolutional neural network algorithm', *Urban Mass Transit*, Vol. 27, No. 3, pp.230–233+239.

Huang, G. and Zhang, F. (2024) 'The fast computation of multi-angle discrete fractional Fourier transform', *Signal Processing*, Vol. 218, No. 1, pp.1–11.

Ji, X., Zhao, S. and Li, J. (2024) 'An algorithm for abnormal behavior recognition based on sharing human target tracking features', *International Journal of Intelligent Robotics and Applications*, Vol. 8, No. 3, pp.583–595.

Jiang, F., Han, X., Wen, S. et al. (2025) 'Spatiotemporal interactive learning dynamic adaptive graph convolutional network for traffic forecasting', *Knowledge-Based Systems*, Vol. 311, No. 1, pp.1–10.

Jiang, N. (2025) 'Application of deep foundation pit construction technology in civil engineering construction', *Journal of Architectural Research and Development*, Vol. 9, No. 1, pp.46–51.

Lu, L., Sun, H.F., Li, D. et al. (2025a) 'Fractional Fourier transform ridge line extraction and structural instantaneous frequency identification based on improved hill climbing method', *Structures*, Vol. 74, No. 1, pp.1–11.

Lu, T., Gu, H., Gu, C. et al. (2025b) 'A multi-point dam deformation prediction model based on spatiotemporal graph convolutional network', *Engineering Applications of Artificial Intelligence*, Vol. 149, No. 1, pp.18–26.

Shekhar, A. and Agrawal, N.K. (2024) 'Inequality and estimate of generalized pseudo-differential operators involving fractional Fourier transform', *AIP Conference Proceedings*, Vol. 3087, No. 1, pp.1–15.

Zhang, D. (2024) 'Laboratory abnormal behavior recognition method based on skeletal features', *International Journal of Advanced Computer Science & Applications*, Vol. 15, No. 8, pp.15–26.

Zhang, S.R., Liang, B.J., Ma, Z.G. et al. (2023) 'Unsafe behavior recognition method of construction workers in water conservancy project', *Journal of Hydroelectric Engineering*, Vol. 42, No. 8, pp.98–109.

Zhang, Y., Xue, L., Zhang, S. et al. (2024) 'A novel spatiotemporal graph convolutional network framework for functional connectivity biomarkers identification of Alzheimer's disease', *Alzheimer's Research & Therapy*, Vol. 16, No. 1, pp.126–139.

Zhao, H., Hu, S.G., Xu, J.X. et al. (2023) 'Abnormal detection of safety helmet in electric smart construction site based on DBN-SVM', *Automation & Instrumentation*, Vol. 12, No. 5, pp.92–95.

Zheng, C., Tang, J., Li, M. et al. (2025) 'Gas well productivity prediction based on fractional Fourier transform', *Journal of Petroleum Exploration and Production Technology*, Vol. 15, No. 5, pp.1–10.

Zhu, Y., Qin, H., Zhang, X. et al. (2024) 'Innovative three-row pile support system of ultra-deep foundation pit and cooperative construction technology with basement for high-rise tower structures', *Buildings (2075-5309)*, Vol. 14, No. 4, pp.1–10.