# Identifying unsafe behavior of construction workers: a dynamic approach combining skeleton information and spatiotemporal features

Han Wu[a], Yu Han[b,*], Meng Zhang[c], Bihonegn Dianarose Abebe[d], Molla Betelhem Legesse[e], Ruoyu Jin[f]

[a] Graduate Student, Faculty of Civil Engineering and Mechanics, Jiangsu University, 301 Xuefu Road, Zhenjiang, 212013, Jiangsu, China. Email: wu1024502851@163.com

[b,*] Professor, Faculty of Civil Engineering and Mechanics, Jiangsu University, 301 Xuefu Road, Zhenjiang, 212013, Jiangsu, China. Email: hanyu85@yeah.net

[c] Graduate Student,, Faculty of Civil Engineering and Mechanics, Jiangsu University, 301 Xuefu Road, Zhenjiang, 212013, Jiangsu, China. Email:2221923026@stmail.ujs.edu.cn

[d] Graduate Student,, Faculty of Civil Engineering and Mechanics, Jiangsu University, 301 Xuefu Road, Zhenjiang, 212013, Jiangsu, China. Email: Dianaroseabebe27@gmail.com

[e] Graduate Student,, Faculty of Civil Engineering and Mechanics, Jiangsu University, 301 Xuefu Road, Zhenjiang, 212013, Jiangsu, China. Email: betelhemlegesse2@gmail.com

[f] Associate Professor, School of Built Environment and Architecture, London South Bank University. 103 Borough Rd, London SE1 0AA, UK. Email: jinr@lsbu.ac.uk

## Abstract

Vision-based methods for action recognition are valuable for the supervision of construction workers' unsafe behaviors. However, existing methods are limited due to the lack of ability to extract worker action information from video streams. Using spatiotemporal relationships between workers' skeletal points to identify hazardous action remains a huge challenge for safety management of construction sites.. In this study, an improved dynamic skeleton model, named Attention Module Spatial-Temporal Graph Convolutional Neural Network (AM-STGCN) is built from the modality data of 2D skeleton points, and a combination of designed human partitioning strategies and non-local attention mechanisms are adopted to extract global information during worker movement to automatically identify unsafe behaviors on construction sites. The method includes three basic modules, namely video data acquisition, workers' skeleton information extraction, as well as recognition and classification of hazardous actions. The test accuracy reached 93.66% in the laboratory, and 90.50% and 87.08% in typical working scenarios (i.e., high-altitude working scenarios with close-up and far views) respectively. The promising test results indicated that the developed AM-STGCN model could be more widely applied in wider construction scenarios, such as foundation excavation.

## Keywords

Hazard scenario; Unsafe behavior; Construction safety; Skeleton modality data; Action recognition; Dynamic model;

## 1. Introduction

The construction industry has been a global concern due to its high risk measured accident rates. The death toll due to accidents in construction accounts for more than 20% of occupational deaths every year. According to the statistics of the Health and Safety Executive of the UK, a total of 123 workers died due to accidents in the UK in 2021, and 30 of them were related to the construction industry, accounting for 24.4% of the total death toll of accidents (HSE 2022). In China, the death toll in the construction industry reached 794 in 2020 alone (MOHURD 2022). Site accidents are often the result of a combination of factors, the Heinrich Accident Causation Theory stated that unsafe behavior of workers is the core cause of accidents (Heinrich 1941), among them the lack of safety protection equipment and workers' hazardous actions are the main causes. To facilitate safety inspection on construction sites, numerous studies on unsafe behavior of workers have been conducted.

In recent years, computer vision technology has become one of the mainstream themes in construction safety research (Fang et al. 2020). Machine vision technology based on color and contour feature extraction (e.g., HOG) and deep learning-based target detection technology such as Faster-R CNN, YOLO, and SSD were among the earlier major approaches on the detection of safety gear of workers. These methods were used to evaluate the interaction between workers and safety helmets, such as space relationship, geometric information and color feature (Park et al. 2015) or to evaluate workers' safety helmets and safety belts under different operating conditions, e.g., scene, weather, light, etc. (Trabelsi et al. 2019; Fang et al. 2018). These studies showed that the research on the inspection of workers' safety protective gear has achieved promising outcomes, but the study of workers' hazardous actions had been still insufficient.

The essence of hazardous action is the change of skeleton joints, which is composed of the state of multiple consecutive frames. Focusing on video clips allows a better and more accurate reflection of the features of the action. The aforementioned detection methods for safety protection gear are only suitable for the detection of static target states, and the use of these methods for the recognition of workers' hazardous actions will significantly reduce the accuracy and increase the false alarm rate. At present, sensor technology is commonly used to detect angular ratios and spatially varying signals between skeleton points to assess and classify the behavior of construction workers. This requires the installation of sensors on each worker and piece of equipment and is a heavy burden for construction site applications. Some researchers (Kim et al., 2016; Fang et al., 2019) have applied vision-based methods to study the relationship between construction workers and targets (e.g., machinery, equipment, materials, etc.) at a given moment, such as assessing the risk conditions of workers who have just stepped into a danger zone or are in the blind spot of construction machinery. However, these studies had not addressed workers' movements during non-conforming operations. Combining long and short-term memory (LSTM) networks and other neural network approaches for time series prediction of unsafe behaviors (Kong et al. 2021; Tang et al. 2020) has made good progress in behavioral regulation. However, action analysis through temporal variation alone ignores spatial variation in human posture, and this a challenge to tackle for recognition of complex construction actions.

The analysis method based on skeleton modality data for action recognition has achieved positive results in extracting and application of motion information. It could extract information about workers' skeleton points, analyze the spatial relationships between workers' skeleton points, and evaluate workers' behaviors. Currently, many studies focus on combining the skeleton with sensors or vision devices, dividing the acquired video into multiple static skeleton images to obtain the change information about workers' motion angle, acceleration and skeleton length. Isolating the dynamic process of motion into static images to process the information effectively extracts the spatial relationship of the skeleton at a single frame, but this approach ignores the temporal relationship of the skeleton information in different frames. The lack of utilization of spatiotemporal information in the dynamic process of workers' construction movements ultimately affects the accuracy of motion recognition. In general, there are two limitations in the current recognition methods for construction workers' actions, specifically: (1) The spatiotemporal information between the skeletons in the construction activities is difficult to use, and a large amount of effective information would be lost; and (2) the lack of effective detection of dynamic action processes affects the recognition accuracy of hazardous actions.

Aiming to address these limitations, the main purpose of this study is to propose a deep learning method that combines skeleton information and spatiotemporal features for the recognition of construction workers' hazardous actions. Researchers established an improved dynamic skeleton model, which takes into account the temporal and spatial relationship of adjacent skeleton joints. The model can be used for the analysis of the dynamic process of construction workers' hazardous construction actions, and for achieving automatic recognition and detection of hazardous actions. Unlike CNN, which segments all videos into frame-by-frame pictures and inputs them into the network, this method directly inputs the skeleton and joint data of workers, hence greatly reducing the number of parameters. It can be used in video surveillance systems to effectively prevent on-site safety accidents.

## 2. Literature review

Traditionally, the recognition methods for construction workers' hazardous actions can be divided into two categories in terms of implementation, namely sensor-based and vision-based methods.

### 2.1 Sensor-based action recognition

Sensor-based action recognition methods typically focus on workers' gestures and fall postures to observe changes in limbs during worker's action. Cheng et al. (2013) obtained position parameters and chest posture data from sensors mounted on workers over a sequence of time to identify workers' actions during activity. Fang and Dzeng (2014) mounted workers' vests and helmets with motion sensors and brainwave sensors to detect workers' fall risk by correlating changes in these two externally transmitted signals over time. Jebelli et al. (2016) used an inertial measurement unit to record the characteristics of changes in sensor parameters over this time period during a worker's fall to comprehensively assess the fall risk of rebar workers. Akhavian and Behzadan (2016) captured workers' body movements by using embedded accelerometers and gyroscopic sensors and simulated various types of construction activities in the laboratory through time-series changes in parameters. As the use of sensors for construction action recognition requires manual processing of large amounts of data, inflexible methods, complex operations and an unprotected user experience, researchers have been gradually combining machine learning and deep learning methods with sensors for construction action recognition.

For example, Gong et al. (2022) adopted a machine learning approach to analyze data from wearable sensors over multiple time periods to identify and classify construction behavior based on parametric temporal features. Bangaru et al. (2021) combined wearable EMG and MU sensors with ANN artificial neural networks to perform data mining in the form of time series on sensor parameters installed on multiple parts of the workers' body to achieve automatic recognition of their' construction actions, and the test results showed good robustness. Ogunseju et al. (2021) used an Inception v1 network to acquire time-series data signals from wearable sensors on workers' lower arms to identify and classify worker actions such as carpentry. However, even though machine learning and deep learning methods improved the speed and accuracy of detection, the results were still with a large amount of data and graphics. Manual analysis of the data was required to define a range of parameter values for the action features. This recognition process ignored the spatial characteristics and feature associations of worker actions. In addition, it demanded a high level of knowledge from managers and was not practical for on-site safety supervision.

### 2.2 Vision-based action recognition

Vision-based action recognition method is a popular research trend in construction safety in recent years. It mainly analyzes workers' construction actions by collecting construction images and videos. Using a single frame picture to recognize action cannot effectively obtain coherent time information in the process of hazardous actions, hence often leading to misjudgment (Guo and Lai 2014). Using RGB video as the research object could obtain the spatial and temporal information of workers' limbs, and that could significantly improve the recognition accuracy (Zhang 2019; Zhao 2019).

#### 2.2.1 Human action recognition

For vision-based methods, action recognition often requires the extraction of action features. Manual feature extraction methods are the main way to extract action features. Feature descriptors such as HoG, HOF and MBH are introduced into the iDT algorithm to obtain the trajectory of feature points and to describe human behavior. But generating local descriptors to describe human behaviors by manually extracting spatiotemporal interest points will take longer computation time and lose more valuable information in videos (Laptev 2005).

Deep learning methods have emerged due to its excellent extraction features and inspection efficiency in image and video processing. There are three main types of methods: two-stream CNN methods, 3D-CNN (3D convolutional neural network) methods and skeleton-based methods. Simonyan and Zisserman (2014) divided the neural network into two parts, one for capturing the spatial features of images and the other for analyzing the temporal information contained in videos. Since then, many scholars have improved this method (Lan et al. 2017; Zhou et al. 2018). For example, Wu et al. (2015) proposed a method based on LSTM and CNN, they applied CNN to perform feature extraction on video clips, then used LSTM to classify long-term span temporal features, and finally extracted multiple manually defined different action features. However, the result was the textual output form of the corresponding action. In addition, the time domain information in the dual-stream network was all derived from the inter-frame optical flow, which was not good for grasping information for a long time, and could be easily affected by many factors, e.g., background, light, and shadow, etc. Compared to 2DCNN, 3DCNN has one more dimension for capturing temporal information, so that long-term information in the action process can be effectively utilized. Tran D et al. (2015) proposed a C3D architecture based on 3D CNN, which could capture spatiotemporal information for human action recognition and improve the recognition accuracy greatly. However, due to a large number of parameters, it was a heavy burden for the actual application effect.

Since the skeleton modality data is not affected by the above-mentioned factors and the connection effect between skeleton points can visually represent the action information, it is more suitable for action recognition. However, RNN and CNN networks treat the skeleton point data input by RGB video as a long-term sequence or 2D matrix to extract features, hence making it difficult to understand the connection information between human skeleton joints (Li et al. 2018; Si et al. 2019), resulting in poor recognition of actions. As one type of graph neural network, GCN analyzes data by using generalized topological graph structure, and is good at processing the relationship between such non-Euclidean data and modeling nodes, which is suitable for the extraction of human skeleton information (Yan et al. 2018).

### 2.2.2 Hazardous action recognition method in construction

For the study of construction workers' hazard actions, previous studies usually focused on the detection, location and tracking of workers. Memarzadeh et al. (2013) detected construction workers and equipment by analyzing construction activities in videos using directional gradients and color histograms. Kim et al. (2016) combined computer vision with fuzzy inference method and augmented reality technology to monitor workers' contact with dangerous areas and evaluate workers' behavioral safety conditions when they worked nearby heavy equipment.

In recent years, some scholars have focused their research on the process of workers' actions. Yang et al. (2016) extracted the location of workers under continuous time series by dense trajectory method to identify workers' behavior with positive results under MBH descriptors. Ding et al. (2018) integrated CNN and LSTM networks to achieve automatic recognition of workers' unsafe behaviors by extracting visual features in a video stream. It actually recognized actions by obtaining different time series information of multiple key points in the video through LSTM and being unified by CNN after extracting data features. However, this approach ignored the spatial feature changes in images of different frames attention, and therefore the result generated was a textual description matching the action category.

To capture the spatial features of workers' movement more clearly, Escorcia et al. (2012) adopted an RGB-D camera to collect motion skeleton data of workers under construction in a building and then used a discriminative classifier to detect workers' actions. Similarly, Han and Lee (2013) extracted 3D skeleton data of workers from videos and effectively used spatial features in the images to identify hazardous actions. However, the spatial features of consecutive frames in the video are often redundant, and the use of 3D convolution introduces repetitive spatial features. Such a form of action recognition actually discriminates from a particular 3D skeleton pose by ignoring the changing relationship between the skeletons during the action, and has low accuracy in recognizing actions with similar postures. Yu et al. (2017) applied a static recognition method based on the image skeleton and tested the accuracy of worker's climbing action by changing the values of the joint parameters, avoiding the redundancy of the parameters. Manually recording the parameter changes in joint angle values was cumbersome and was limited to static skeleton information. The detection accuracy for the three hazardous actions was only 81.44%, which would be further reduced when used for the identification of multiple unsafe action types at construction sites. Guo et al. (2018) also simplified the dynamic skeleton movement process to a static process in order to achieve real-time detection of unsafe actions. They described the static pose by a few parameters for action recognition, without considering the relationship between the skeleton information in time and space, and that resulted in a high false alarm rate for the action measurement.

Based on skeleton modality data, this research is devoted to dealing with the spatiotemporal connection between the overall skeleton data of workers, in order to achieve the dynamic detection of workers' hazard actions and to facilitate the application in actual construction scenarios. A spatiotemporal graph convolutional neural network (ST-GCN) is proposed by adopting a deep learning method for automatically extracting worker skeleton information and identifying the dynamic process of construction workers' hazardous actions.

## 3. Methodology

The methodology to implement a deep learning model for dynamic recognition of workers' hazardous actions consists of four steps, namely: (1) extraction of worker's skeleton points; (2) selection of action recognition algorithms; (3) data collection and model building; and (4) recognition of dynamic process of actions. The research is intended to provide a general methodological basis for subsequent studies of workers' hazardous construction actions. Fig. 1 shows the method flow designed in this study, with each step including specific implementation details.

### 3.1 Openpose-based skeleton extraction method

Due to the complex environmental factors of construction sites, workers' bodies are often obscured by construction materials, equipment or structures. One of the current mainstream methods for extracting the skeleton requires a top-down approach, which first detects the human beings by target detection algorithm, and then detects the key points of single human skeleton. However, this kind of method is difficult to perform worker's skeleton identification and extraction when the construction worker's body is more than 30% occluded. Unlike most top-down methods for extracting workers' skeleton joints, the Openpose network uses a bottom-up structure to extract workers' skeleton joints, which first detects each skeleton joint of workers and then connects all the identified skeleton points to generate a complete image of the worker's skeleton. This way of extracting workers' skeleton can effectively reduce the reliance on personnel detectors, improve the timeliness of skeleton point extraction, and enable the recognition of key points of workers' skeleton in the case of multi-person construction. Hence this method is suitable for applications in construction sites with a large number of personnel movements. Therefore, the Openpose network was designed for skeleton extraction of construction workers in this study.

VGG-16 was used as a pre-base network in this study, which was able to perform feature extraction and generate feature maps for the prediction and connection of skeleton points via two channels. The specific implementation process of extraction of workers' skeleton point using Openpose pose estimation network is as follows: first, the CPM operation method is adopted to predict the skeleton joint points of all workers in the video collected at the construction sites, and then to detect the heatmap of the skeleton points of the workers (Wei et al. 2016). Each joint generates a corresponding Gaussian peak, and the location of the peak is the worker's skeleton joint. After completing the worker skeleton point prediction, the isolated skeleton points of the workers are connected by regressing the PAFs.

### 3.2 ST-GCN-based construction action recognition method

Construction action is composed of multiple pose graph structures such as the action of a worker climbing on scaffolding, and it is difficult to effectively identify specific action categories only by considering the spatial location between skeleton points (Zhou et al. 2020). The spatiotemporal graph convolutional neural (ST-GCN) network changed the form of spectral-based convolution of GCN networks. The network adds a temporal convolution module, it combines the location information of skeleton joints and temporal information and introduces the graph convolution in the spatiotemporal domain for capturing the variation patterns among nodes (Yan et al. 2018). In this way, it is easy to identify the types of workers' actions with large changes in spatial location of the skeleton. Therefore, the ST-GCN network was adopted as the action recognition used for this study. Fig. 2 describes the main structure of the ST-GCN introduced.

Different from CNN network that performs the sampling method and assign weights to the convolution principle, researchers in this study replaced nodes with image pixel points in ST-GCN networks. Then the sampling function $p(v_{ti},v_{tj})$ was used to represent the distance between the first-order neighboring nodes involved in the convolution process, where $v_{ti}$ is a point in a sequence of joint points, $v_{tj}$ denotes an adjacent node, and the weight function $w(v_{ti},v_{tj})$ is applied to represent the weight vector of the nodes and their neighbors. After the weighted average of the standard normalized $Z_{ti}(v_{tj})$, the updated graph convolution equations were expressed as Eq. (1) and Eq. (2).

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

$$Z_{ti}\left(v_{tj}\right) = \left|\{v_{tk} | l_{ti}(v_{tk}) = l_{ti}\left(v_{tj}\right)\}\right| \tag{2}$$

The sampling function and weight function mentioned above were designed for the spatial graph structure only, without considering the temporal factor. Therefore, researchers defined the spatiotemporal graph by recomputing the label grouping mapping function. The equation of the spatiotemporal graph structure is shown in Eq. (3).

$$l_{ST}\left(v_{qj}\right) = l_{ti}\left(v_{tj}\right) + (q - t + \lfloor \Gamma/2 \rfloor \times K) \tag{3}$$

Where $l_{ti}(v_{tj})$ is the label map for the single frame case at $v_{ti}$, $\Gamma$ is the temporal kernel size, and $l_{ST}$ represents the labeling map.

## 3.3 Algorithm adjustment and optimization

### 3.3.1 Human partitioning strategy adjustment

Considering the complexity of workers' operation actions, the action process often involves not only the location changes of the limbs, but also the changes of the torso which are equally important. Researchers set the domain span of the nodes to 2 based on the Spatial partitioning strategy and re-divided the skeleton point neighborhood into 3 subsets, namely root nodes, centripetal groups and centrifugal groups. The new partitioning strategy could expand the extraction range of workers' skeleton features, which could extract features of key points and improve the accuracy rate for construction workers' construction actions. The new partitioning strategy is shown in Fig. 3 where the root node is shown in purple, the skeleton point near the center of gravity of the skeleton and adjacent to the root node (green) is the centripetal group, and the centrifugal group is the neighboring nodes far from the centre of gravity of the skeleton (yellow).

Researchers then assigned weights to the skeleton points of each region according to the new partitioning strategy. The new weight assignment is shown in Eq. (4), where $r_j$ represents the distance from the skeleton point j to the centre of gravity of the worker's body, and $r_i$ is the average distance from the center of gravity to the skeleton point.

$$l_{ti}(v_{tj}) = \begin{cases} 0 & if\ \ r_j = r_i \\ 1 & if\ \ r_j < r_i \\ 2 & if\ \ r_j > r_i \end{cases} \tag{4}$$

In Eq .(4), the centre of gravity is the average coordinate of all joints in a body (black cross in Fig. 3), and 0 indicates that no weight is assigned to joints where $r_j$=$r_i$ (i.e., no change in the joint during movement), 1 infers that less weight is assigned to joints where the $r_j$<$r_i$ (i.e., closer to the centre of gravity), 2 denotes more weight is assigned to joints where the $r_j$>$r_i$ (i.e., farther from the centre of gravity).

### 3.3.2 Non-local attention mechanism

The impact of different body parts on the accuracy of a worker's action recognition during construction varies. Workers rely primarily on the limb parts of the body during construction work, while parts such as the head and neck are not as involved and provide little effective information for movement recognition.

Furthermore, the relationships between skeletal joints during construction actions are not restricted to adjacent joints. For example, for many action processes such as probing and climbing, the connection between the joints of the arms, legs and the trunk cannot be ignored. However, in the original ST-GCN network, the perceptual domain of the convolution operation was the neighboring nodes of the root node, which was only used to capture local features of the action process, such as joint changes at the calf and joint changes at the arm. Such a feature extraction approach cannot simultaneously analyze the joint changes at the calf and arm during an action to identify the type of action (Simonyan Zisserman 2015). Despite the adaptation of the human partitioning strategy in the previous section, there was a skill gap in extracting the motor features of architectural actions.

To solve this problem, the researchers introduced a non-local neural network module to optimize the action recognition network. The non-local network is usually embedded in vision models as a simple and efficient general-purpose module that can improve the classification accuracy of images and videos (Wang et al. 2018; Kong et al. 2019). Based on this module, researchers modified the original ST-GCN network and designed a new dynamic skeleton model based on a non-local attention mechanism for hazard action recognition. The new model was named attention module spatiotemporal graph convolutional neural (AM-STGCN) network. The workflow of this new model was that it first focused on features of all joints, rather than only local features of certain joints. It changed the way in which local information about workers' actions was extracted to one in which global information is extracted. After analyzing the global information relationships between the nodes, more effective features were obtained for key regions according to the human body partitioning strategy. The network structure of AM-STGCN is shown in Fig. 4, where the model consists of nine layers of spatiotemporal graph convolution operators. The first three layers had 64 output channels, the middle three layers had 128 output channels, and the last three layers had a total of 256 output channels. Each layer included spatial convolution operations (Conv *S*) and temporal convolution operations (Conv *T*), and residual connections were added on each layer. Three attention modules were added to the temporal convolution (Conv *T*) in the third layer of ST-GCN network in order to achieve optimal performance in the recognition of workers' construction actions.

### 3.4 The running process of the model

After designing and improving the above method, a complete operation flow chart was constructed. The data input of AM-STGCN model was skeleton sequence information, so in order to realize the classification of workers' behavior, it would be necessary to combine Openpose skeleton point extraction algorithm and AM-STGCN spatiotemporal graph convolutional neural network to jointly construct the framework of construction workers' hazardous action recognition model.

The model utilized video streams as input, and first used Openpose for worker skeleton data extraction to establish the spatiotemporal dimensional information in the human skeleton data and to construct a human skeleton sequence map. Subsequently, the extracted worker skeleton information was passed into AM-STGCN for learning and training of behavioral states. Finally, a SoftMax classifier was implemented for behavioral result output. The specific operation flow is shown in Fig. 5.

## 4. System building

The main purpose of this study was to apply a new dynamic skeleton method to detect the action characteristics of workers at construction sites. The focus of the study shifted from the static characteristics of workers to the dynamic characteristics. This study was not limited to the detection of single-frame targets, but also investigated the spatiotemporal relationships through video sequences for worker construction safety. Therefore, all the model constructions in this study used video sequences as the data source.

## 4.1 Datasets collection

### 4.1.1 Action selection for recognition

In order to achieve dynamic recognition of workers' construction action processes in real construction scenes, a new construction dataset based on the weights of the existing general scene of dataset was established. Therefore, researchers selected the high-altitude scaffolding scenario, where fall-at-height accidents are common, as a practical construction application case to study. This scenario can be adopted as the basis for the study of the full-scene hazardous actions. Statistics on the causes of work-at-height injuries and deaths point to unauthorized climbing, probing, leaning and crouching movements made by workers on scaffolding as the main causes of accidents (HSE 2022; MOHUD 2022). Researchers selected some of the hazardous actions for identification. However, it should be noted that the expansion of data types could be implemented in the future. Videos were obtained for seven types of actions, including normal walking, running operation, lean-over operation, scaffold climbing, hazard crossing operation, sitting on scaffolding, and material handling. Fig. 6 shows some partial video intercepted segments of various types of construction actions.

### 4.1.2 Video acquisition and data processing

Handheld cameras and drones were used to capture the types of workers' construction actions. Many studies based on image data have shown that the difference in shooting conditions would have an impact on the recognition effect (Jegham et al. 2020; Wen et al. 2022). Therefore, video clips were collected that reflected the entirety of the construction operation and the camera or drone angle met the requirements of multiple perspectives. Images and videos were collected under different weather and lighting conditions, and effects of other factors (e.g., far view, close-up view, single-target, multi-objective, etc.) were considered to avoid affecting the training effect. 2

The acquired videos needed to be processed in a uniform format. Researchers expanded the number of samples and enhanced the data by horizontal mirror flip, by using toolkit to crop the videos into action sequences of 5s each. Each action sequence contained at least one action category. A total of 8,000 action sequences were obtained for skeleton extraction, containing 12,215 worker targets. The number of each category of action sequences was basically kept balanced. The dataset was divided into a training set, a validation set and a test set in the ratio of 6:3:1, where the test set was the original video without skeletal point annotation. Table 1 shows the number of datasets for each type of action in the training set.

#### 4.1.3 Skeleton extracting and data labeling

Different from the dataset format for static target annotation, the skeleton point sample data needs to extract the human skeleton information of each frame image from the video. Referring to the format of the kinematics skeleton dataset, a total of 18 key points of information on worker skeleton were extracted, and then, JSON files of different action categories were generated through built-in data transformation algorithm module for transmission to the AM-STGCN model. The file format of normal walking is shown in Fig. 7, where "frame_index" is the frame index representing the skeleton data of a specific frame, and "skeleton" is the skeleton joint point information of the frame. Finally, the json file was converted into npy and pkl format to form a skeleton sample dataset.

### 4.2 Recognition of workers' skeleton

In the data preparation stage, one second of video was divided into 30 frames. So for a complete action of video sequence, more than one hundred frames were generated. In another word, for one action, more than 100 skeleton data as shown in Fig. 7 would be generated. In these data, the extracted skeleton information of each frame was different from the previous frame. By extracting the 2D coordinate information of the 18 nodes of the worker's skeleton frames, the extraction degree of the workers' skeleton information could be maximized to ensure the accuracy of the spatiotemporal sequence when passed into ST-GCN network for analysis. Fig. 8 shows the effect of extracting information about experimenter's skeleton at a certain frame during the execution of the three movements (e.g., Frame 89 for the sitting on scaffolding; Frame 143 for the scaffold climbing; and Frame 67 for the lean-over operation).

### 4.3 System testing

After extracting the human skeleton, the Pytorch platform was used for training the dynamic recognition model. The platform used Windows 10 64 as the operating system, with a built-in NVIDIA GeForce RTX 1050ti graphics card and an Intel i7 processor. At the same time, CUDA and other operators were installed to accelerate the model on the graphics processor (GPU). Before the model training, The training parameters were configured according to the hardware device performance of the training platform. The specific parameter settings are shown in Table 2, the learning rate was reduced with step decay during training to enhance the training effect (Feng and Li 2018).

In order to save training time and improve the training effect, migration learning was used in the training process, and the weights of ST-GCN fully trained in the Kinetics dataset were loaded as the initial training weights (Weiss et al. 2016). The training is set for 300 epochs, and the loss value was calculated for each completed epoch. The variation of the loss values is shown in Fig. 9 (a). It can be seen that the training loss value decreased rapidly in the first 30 rounds, when the training reached 40 rounds, the learning rate decreased to 0.01 to continue the training. As can be seen from Fig 9 (b), after 30 rounds of training, the Top1 accuracy exceeded 90% and the accuracy started to converge. After 300 rounds of training, the learning rate decreased to 0.00001 and the loss value also decreased from 1.19 to about 0.14 converging. The Top1 accuracy could be stably maintained at about 91%, indicating that the model for hazardous action recognition was completely trained.

### 4.4 Recognition of workers' action

Scaffolding climbing in the laboratory was selected as an example. Fig.10 illustrates the test process of an unsafe climbing action in the laboratory, and the video results were generated from the modified ST-GCN model. Figure 10(a) shows the worker skeleton information extraction using Openpose for a 5s video sequence containing a worker target. It learned the process of a worker making a complete climbing movement and subsequently extracted the skeleton information changes of the worker's legs and waist during leg lifts and drops. Fig. 10(b) shows a visual representation of workers' skeleton extraction, which is a continuous frame of human skeleton map with a duration of 5s, indicating that the video stream with skeleton information is fed into AM-STGCN for model learning and action classification. Fig. 10(c) and Fig. 10(d) show the key point extraction and action recognition results after using the human body partitioning strategy and the non-local neural network attention module. After the whole action sequence was made, the classifier was used to evaluate the action type and to output the final result.

### 4.5 Model validation and analysis

After the model was trained, divided sample test set was adopted to test the model. The essence of worker action recognition was to classify types according to the set action object, and the classification task commonly used Accuracy (A), Precision (P) and Recall (R) as evaluation indicators. The calculation process of precision rate and recall rate is shown in Eq. (5) and Eq. (6).

$$Precision = \frac{T_p}{T_P + F_p} \tag{5}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{6}$$

In the equations, $T_p$ represents the number of workers whose actions are correctly identified, $F_p$ represents the number of workers whose incorrect actions are mistakenly considered correct, and $F_N$ denotes the number of workers whose correct construction actions are evaluated to be wrong. The specific laboratory test results indicators are shown in Table 3.

As can be seen from Table 3, the model did not miss any recognition of actions, and for the sample test set, the overall accuracy of worker action recognition reached 93.39%. Among the different actions tested, the method achieved high recognition recall for scaffold climbing and sitting on scaffolding, by reaching 95.48% and 96.18% respectively. The recognition recall for the other five actions was lower, but the recall was around 93%. This could be caused by the fact that scaffold climbing and sitting on scaffolding had a distinct limb performance during the movement and more characteristic changes in skeletal information. In contrast, for actions such as normal walking, the temporal and spatial information of the skeleton did not change significantly during the action of a few seconds, and there was partial identity of skeletal information during the action. Overall, the models could achieve promising action recognition results in a laboratory setting.

### 4.6 Case study

To verify the practicality of the method, application and testing work were carried out in combination with real construction scenarios. Three construction projects in Zhenjiang China were selected as real-world test sites to obtain different test videos. The high-altitude scaffolding actions of workers in the close-up view and far view were acquired for testing, and the action sequences included single-person and multi-person targets. A total of 3,000 action sequences were acquired for method testing in both close-up and far views respectively.

For each action, the model extracted the skeleton information every 1 frame and combined the skeleton information of 10 frames to complete the result output once. After outputting multiple skeleton information in this manner, the action classification was finally completed by recognizing the results of the whole action sequence. Fig. 11 shows the recognition results of workers' work video collected on site. The recognition effectiveness of the method was then evaluated using the accuracy A, precision P and recall R, and the average recognition time T for each frame of the video.

The obtained videos of construction workers from high-altitude scaffolding were used for method testing and statistical analysis. A total of 1,500 videos of workers' operations with 2,137 worker targets were selected for the close-up view test, and a confusion matrix was introduced to evaluate the model effects (Yang et al., 2016). The test results are shown in Table 4 and Figure 12. The method can recognize all the workers' targets and evaluate the actions in the close-up view. There was no omission in the identifying of actions, and the average recognition time for a single frame video was 127.61ms. The classification results for different actions varied slightly, with the highest recall rates for scaffold climbing actions and sitting on scaffolding, which was generally consistent with laboratory. In terms of classification results, the highest number of actions were misclassified as normal walking and material handling, while the lowest number of actions were misclassified as scaffold climbing and sitting on scaffolding. Compared to the test results in the laboratory, the overall recognition accuracy of the method in the close-up view has decreased, but it can still reach 90.50%, indicating that the method can better recognize hazardous actions of high-altitude scaffolding work in the close-up view.

A total of 1,500 videos were selected for the far-view test, containing 2,291 worker targets. The effect of worker action recognition in far view is shown in Table 5 and Fig. 13. In the test of the far view, due to the smaller size of the worker targets within the video, the feature acquisition ability of the worker skeleton information was reduced, resulting in a decrease in the recognition effect of the method compared to that in close-up view. The average recognition time of a single frame action was 132.54 ms. The model still had the highest recall for scaffold climbing and sitting on scaffolding, with 90.40% and 89.78% respectively. The model continued to have the lowest number of actions misclassified as these two action types when identifying other action types. For all other action types, the recall rate decreased to varying degrees, which was roughly the same as the test results in the close-up view. The average accuracy in the far view was able to maintain at 87.08%. There was no omission in the action recognition, indicating that the method was also robust for high-altitude scaffolding in the far view.

## 5. Discussion

### 5.1 Accuracy variation analysis

The difference in accuracy may be due to the fact that the spatiotemporal information of the skeleton in scaffold climbing and sitting on scaffolding was more obvious. But in actions such as normal walking and running operation, the change in skeletal information during the action was approximately the same, and that caused recognition errors. In order to analyze the influence of the spatiotemporal information extracted by the dynamic skeleton model on the classification of different action types, two types of actions with high recognition accuracy and two types of actions with low recognition accuracy were selected for comparison. The effect of the spatiotemporal features of the skeleton on the accuracy of action recognition was analyzed by comparing the weight parameters assigned to the body parts in different frames. The comparison results are shown in Table 6 and Table 7.

From Tables 6 7, it can be seen that for scaffold climbing and sitting on scaffolding, where the skeleton features vary significantly, the weight parameters assigned to each body part in different frames varied significantly. The skeleton features were easily observed during the continuous 5s movements, so the recognition accuracy was higher. For the normal walking and running operation movements, the differences of weights assigned to the body parts in the different frames of the two action types were less significant, resulting in similar skeletal features in several action types and hence leading to recognition errors. Overall, the improved dynamic skeleton model had high recognition accuracy for complex construction actions, and the recognition accuracy for actions with low complexity was also higher than previous methods.

### 5.2 Performance evaluation

In this study, a new dynamic skeleton model (AM-STGCN) was designed to identify hazardous actions of construction workers. The model analyzed the spatial characteristics of workers' skeletons between different frames by two convolutional modules, namely spatial convolution and temporal convolution. Researchers extracted the global information of key joints in human partitioning strategies and non-local neural network modules to identify complex worker actions. To verify the performance of the improved dynamic skeleton model algorithm, researchers selected three models for comparisons to the improved method, including the baseline model ST-GCN network(ⅰ); ST-GCN network adjusted by human body partitioning strategy only (ⅱ); and ST-GCN network modified by non-local attention mechanism only(ⅲ). The experiments used the same data and training parameters. The comparison results are shown in Table 8. For each worker action type, the improved algorithm outperformed the baseline model and the other partially improved methods measured by recognition accuracy.

### 5.3 Potential limitation

In terms of the overall effect, the method could achieve the recognition of the dynamic process of construction workers' hazardous actions under different shooting viewpoints. Combining the results of action recognition under two different viewpoints, researchers found that in the scaffold climbing recognition, the method did not miss the action target, but there were cases of misjudgment. The accuracy decreased with the increasing number of recognized targets and the reduction of the target size. Overall, the recognition effect of scaffold climbing and sitting on scaffolding was promising. It was also found that there were mainly the following reasons for misjudgment: (1) the influence of occlusions; (2) the interaction between actions.

The high-altitude scaffolding scenario selected for this study was complex. The junction parts of horizontal, vertical and diagonal bars of scaffolding would form a complex interwoven structure of bars. When construction workers were at multiple scaffolding junctions and their body parts are covered by large areas, it would be difficult for the method to integrate whole-body skeleton information for action recognition (Sahoo et al. 2022). In the video captured under the far view, the method sometimes misidentified the body parts of construction workers as actions, as shown in Fig. 14.

Actions were composed of a series of consecutive behavioral gestures. There were situations where different action processes had partially similar gestures, and other actions might also be interspersed when engaging in specific actions, resulting in method misclassification (Vasconez et al. 2021). The detection case shown in Fig. 15 was generated from a video sequence containing multiple actions, researchers added a fast process of hazard crossing to the normal walking process. The model first judged the feature as a normal walking but then misjudged it as a lean-over operation in the second half of the video sequence. Similarly, the rare cases where a fast normal walking process was added to the lean-over operation process also caused misjudgment.

Although the recognition result was based on the integration of the recognition results of a large number of stage frames, a misjudgment in a single video frame had little impact on the overall recognition result. It has been reported that for the detection of construction workers' hazardous actions, a very small number of misjudgments may lead to serious injuries and fatalities (Pinto et al. 2011). Therefore, the occurrence of miscalculations needed to be avoided as much as possible. In terms of the causes of miscalculation of the method, factors such as obstacles and occlusions in the recognition of skeleton features by the method could affect the extraction of skeleton features (Li and L 2022), while in the analysis and classification of actions, the correlation and interpolation between actions could also disrupt the recognition and judgment of the method for specific actions (Yang 2018).

In addition to selecting near and far views as the study scenes, researchers also selected single and multi-person targets as the sample data set. In the test process, researchers found that the single-person action recognition effect was slightly better than the multi-person action recognition effect, but the difference was marginal. This might be due to the fact that most of the selected multi-person construction targets were two workers, and the difference in the number of targets was not obvious. In fact, there were usually many worker targets in a construction work area. To further investigate the effect of the number of workers' actions on the action feature extraction ability and action recognition effectiveness, it is necessary to select construction videos containing more workers' actions to study the variation of method performance in the future.

## 6. Conclusions

Researchers proposed a framework for recognizing workers' hazardous actions by fusing skeleton extraction and spatiotemporal features. Openpose (i.e., skeleton point extraction network) and ST-GCN (i.e., spatiotemporal graph convolutional neural network) were designed to jointly build a dynamic skeleton model, which could analyze the spatiotemporal relationship of workers' skeletons and automatically recognize construction workers' hazardous actions. In order to achieve an enhanced performance of the model method for construction site application, researchers made algorithm adjustments and built a dataset of real-life construction scenes based on the existing public dataset. The high-altitude scaffolding scene was used as a test case and tested under several challenging situations such as viewpoint change (close-up view and far view) and target change (single target and multiple targets). The results showed that the method recognition accuracy reached 90.50% and 87.08%, respectively, and the single frame recognition time could be controlled between 127~133ms with good robustness.

Compared to previous studies on construction workers' hazardous actions, the contributions of this research are as follows: (1) by introducing a dynamic skeleton model to analyze the spatiotemporal relationship between skeleton points during workers' construction actions, researchers effectively utilized the information of workers' movement characteristics, which complemented the defects of previous research methods that ignored dynamic action information; (2) combining the Openpose skeleton extraction algorithm and the improved ST-GCN, an operational framework for hazardous action recognition was constructed to automate and visualize the process of hazardous action recognition of construction workers. This provided a technical basis for managers to check workers' hazardous actions through real-time monitoring in the future; and (3) through the adjustment of partitioning strategy and the addition of attention mechanism, the method enabled the extraction of global features of workers' skeleton information, which effectively improved the accuracy of action recognition.

Given the high accuracy of the method during testing, researchers believe that the proposed method has good application prospects. The method provides managers a new perspective on construction site safety management, rather than just focusing on the status of workers wearing safety gear or the state of worker-object (e.g., material, equipment, area) interaction. This detection method for capturing hazardous construction actions can be used in parallel with target detection methods such as safety protective equipment to capture both dynamic and static hazards present on construction site, expanding the scope of site safety detection and assisting in screening for more types of unsafe behavior events to improve worker safety. In addition, the unsafe information captured could be used for job training and warning education to better regulate construction workers' behavior. In the follow-up work, the feature extraction algorithm needs to be further improved to reduce the influence of occlusions and different actions, to better understand the spatiotemporal relationships between skeleton points, and to improve the recognition accuracy under complex influencing factors.t is also necessary to improve the parameters of the dynamic skeleton model to enhance the recognition speed of the model, and to expand the number of action types and worker targets to enhance the applicability of the model scenes for future real-time monitoring of general scenes.

## Acknowledgement

This research is supported by the National Natural Science Foundation of China (Grant No. 72071097), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Grant No.20YJAZH034), and Foundation of Jiangsu University (Grant No. SZCY-014).

## Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

## References

Akhavian, R., and Behzadan, A.H. 2016. “Smartphone-based construction workers' activity recognition and classification.” *Automation in Construction*, 71, 198-209.

Bai, S., Kolter, J.Z., and Koltun, V. 2018. “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.” *Computer Science*, 1803.01271.

Bangaru, S.S., Wang, C., Busam, S.A., and Aghazadeh, F. 2021. “ANN-based automated scaffold builder activity recognition through wearable EMG and IMU sensors.” *Automation in Construction*, 126, 103653.

Cheng, T., Teizer, J., Migliaccio, G.C., and Gatti, U.C. 2013. “Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data.” *Automation in Construction*, 29, 24-39.

Detrembleur C, Van den Hecke A, and Dierick F. (2000)."Motion of the body centre of gravity as a summary indicator of the mechanics of human pathological gait." Gait & posture, 12(3), 243-250.

Ding, L., Fang, W., Luo, H., Love, P.E.D., Zhong, B., and Ouyang, X. 2018. “A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory.” *Automation in Construction*, 86, 118-124.

Escorcia, V., Dávila, M.A., Golparvar-Fard, M., and Niebles, J.C. 2012. “Automated vision-based recognition of construction worker actions for building interior construction operations using RGB-D cameras.” *Construction Research Congress 2012: Construction Challenges in a Flat World*, 879-888.

Fang, W., Ding, L., Love, P.E.D., Luo, H., Li, H., and Peña-Mora, F. 2020. “Computer vision applications in construction safety assurance.” *Automation in Construction*,110, 103013.

Fang, W., Ding, L., Luo, H., and Love, P.E.D. 2018. “Falls from heights: A computer vision-based approach for safety harness detection.” *Automation in Construction*, 91,53-61.

Fang, W., Zhong, B., Zhao, N., Love, P.E.D., Luo, H., and Xue, J. 2019. “A deep learning-based approach for mitigating falls from height with computer vision: Convolutional neural network.” *Advanced Engineering Informatics*, 39, 170-177.

Fang, Y.-C., and Dzeng, R.-J. 2014. “A Smartphone-based Detection of Fall Portents for Construction Workers.” *Procedia Engineering*, 85, 147-156.

Feng, Y., and Li, Y. 2018. “An overview of deep learning optimization methods and learning rate attenuation methods.” *Hans Journal of Data Mining*, 8(4),186-200.

Gong, Y., Yang, K., Seo, J., and Lee, J.G. 2022. “Wearable acceleration-based action recognition for long-term and continuous activity analysis in construction site.” *Journal of Building Engineering*, 52, 104448.

Guo, G., and Lai, A. 2014. “A survey on still image based human action recognition.” *Pattern Recognition*, 47(10), 3343-3361.

Guo, H., Yu, Y., Ding, Q., and Skitmore, M. 2018. “Image-and-skeleton-based parameterized approach to real-time identification of construction workers’ unsafe behaviors.” Journal of construction engineering and management, 144(6), 04018042.

Han, S., and Lee, S. 2013. “A vision-based motion capture and recognition framework for behavior-based safety management.” *Automation in Construction*, 35, 131-141.

Heinrich, H.W. 1941. “Industrial Accident Prevention. A Scientific Approach (Second Edition).” McGraw-Hill, New York. https://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.22.1.119-b

HSE. 2022. “Work-related fatal injuries in Great Britain.” Accessed August 3, 2022. https://www.hse.gov.uk/statistics/fatals.html.

Jebelli, H., Ahn, C.R., and Stentz, T.L. 2016. “Fall risk analysis of construction workers using inertial measurement units: Validating the usefulness of the postural stability metrics in construction.” *Safety Science*, 84, 161-170.

Jegham, I., Ben Khalifa, A., Alouani, I., and Mahjoub, M.A. 2020. “Vision-based human action recognition: An overview and real-world challenges.” *Forensic Science International: Digital Investigation*, 32, 200901.

Kim, H., Kim, K., and Kim, H. 2016. “Vision-based object-centric safety assessment using fuzzy inference:

Monitoring struck-by accidents with moving objects.” *Journal of Computing in Civil Engineering*, 30(4), 04015075.

Kong, T., Fang, W., Love, P.E.D., Luo, H., Xu, S., and Li, H. 2021. “Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction.” *Advanced Engineering Informatics*, 50, 101400.

Kong, Y., Li, L., and Zhang, K. 2019. “Attention module-based spatial–temporal graph convolutional networks for skeleton-based action recognition.” *Journal of Electronic Imaging*, 28(4), 043032.

Lan, Z., Zhu, Y., Hauptmann, A.G., and Newsam, S. 2017. “Deep local video feature for action recognition.” *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1-7.

Laptev, I. 2005. “On space-time interest points.” *International journal of computer vision*, 64(2), 107-123.

Li, J., Li, B., Xu, J., Xiong, R., and Gao, W. 2018. “Fully connected network-based intra prediction for image coding,” *IEEE Transactions on Image Processing*, 27(7), 3236-3247.

Li, Z., and Li, D. 2022. “Action recognition of construction workers under occlusion.” *Journal of Building Engineering*, 45, 103352.

Memarzadeh, M., Golparvar-Fard, M., and Niebles, J.C. 2013. “Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors.” *Automation in Construction*, 32, 24-37.

MOHURD. 2022. “Circular on the production safety accidents of housing and municipal engineering in 2020.” Accessed 19 June 2022. https://www.mohurd.Gov.cn/gongkaifdzdgknr/tzgg/202006/20200624_246031.html

Park, M.-W., Elsafty, N., and Zhu, Z. 2015. “Hardhat-Wearing Detection for Enhancing On-Site Safety of Construction Workers.” *Journal of Construction Engineering and Management*, 141(9), 04015024.

Pinto, A., Nunes, I.L., and Ribeiro, R.A. 2011. “Occupational risk assessment in construction industry-Overview and reflection.” *Safety Science*, 49(5), 616-624.

Sahoo, S.P., Modalavalasa, S., and Ari, S. 2022. “DISNet: A sequential learning framework to handle occlusion in human action recognition with video acquisition sensors.” *Digital Signal Processing*, 131, 103763.

Si, C., Chen, W., Wang, W., and Tan, T. 2019. “An attention enhanced graph convolutional lstm network for skeleton-based action recognition.” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1227-1236.

Simonyan, K., and Zisserman, A. 2015. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” *Computer Science*, 1409.1556.

Tang, S., Golparvar-Fard, M., Naphade, M., and Gopalakrishna, M. 2020. “Video-based motion trajectory forecasting method for proactive construction safety monitoring systems.” *Journal of Computing in Civil Engineering*, 34(6), 04020041.

Trabelsi, R., Jabri, I., Melgani, F., Smach, F., Conci, N., and Bouallegue, A. 2019. “Indoor object recognition in RGBD images with complex-valued neural networks for visually-impaired people.”

*Neurocomputing*, 330, 94-103.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. 2015. “Learning spatiotemporal features with 3d convolutional networks.” *Proceedings of the IEEE international conference on computer vision*, 4489-4497.

Vasconez, J.P., Admoni, H., and Auat-Cheein, F. 2021. “A methodology for semantic action recognition based on pose and human-object interaction in avocado harvesting processes.” *Computers and Electronics in Agriculture*, 184, 106057.

Wang, X., Girshick, R., Gupta, A., and He, K. 2018. “Non-local neural networks.” *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*,7794-7803.

Wei, S.E., Ramakrishna, V., Kanade, T., and Sheikh, Y. 2016. “Convolutional pose machines.” *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 4724-4732.

Weiss, K., Khoshgoftaar, T.M., and Wang, D. 2016. “A survey of transfer learning.” *Journal of Big data*, 3(1), 1-40.

Wen, J., Shen, Y., and Yang, J. 2022. “Multi-view gait recognition based on generative adversarial network.” *Neural Processing Letters*, 54(3), 1855-1877.

Wu, Z., Wang, X., Jiang, Y.G., Ye, H., and Xue, X. 2015. “Modeling spatial-temporal clues in a hybrid deep learning framework for video classification.” *Proceedings of the 23rd ACM international conference on Multimedia*, 461-470.

Yan, S., Xiong, Y., and Lin, D. 2018. “Spatial temporal graph convolutional networks for skeleton-based action recognition.” *Thirty-second AAAI conference on artificial intelligence*, 180.07455.

Yang, J., Shi, Z., and Wu, Z. 2016. “Vision-based action recognition of construction workers using dense trajectories.” *Advanced Engineering Informatics*, 30(3), 327-336.

Yang, J. 2018. “Enhancing action recognition of construction workers using data-driven scene parsing.” *Journal of Civil Engineering and Management*, 24(7), 568-580.

Yu, Y., Guo, H., Ding, Q., Li, H., and Skitmore, M. 2017. “An experimental study of real-time identification of construction workers' unsafe behaviors.” *Automation in Construction*, 82, 193-206.

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., and Zheng, N. 2019. “View adaptive neural networks for high performance skeleton-based human action recognition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8),1963-1978.

Zhao, R., Xu, W., Su, H., and Ji, Q. 2019. “Bayesian hierarchical dynamic model for human action recognition.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7733-7742.

Zhou, B., Andonian, A., Oliva, A., and Torralba, A. 2018. “Temporal relational reasoning in videos.” *Proceedings of the European conference on computer vision*, 803-818.

Zhou, K., Wu, T., Wang, C., Wang, J., and Li, C. 2020. “Skeleton Based Abnormal Behavior Recognition Using Spatio-Temporal Convolution and Attention-Based LSTM.” *Procedia Computer Science*, 174, 424-432.