electronics

MDPI

Article

Worker Abnormal Behavior Recognition Based on Spatio-Temporal Graph Convolution and Attention Model

Zhiwei Li, Anyu Zhang, Fangfang Han *, Junchao Zhu and Yawen Wang

School of Electrical Engineering and Automation, Tianjin University of Technology, Tianjin 300384, China
* Correspondence: hanfangfang@tjut.edu.cn

Abstract: In response to the problem where many existing research models only consider acquiring the temporal information between sequences of continuous skeletons and in response to the lack of the ability to model spatial information, this study proposes a model for recognizing worker falls and lays out abnormal behaviors based on human skeletal key points and a spatio-temporal graph convolutional network (ST-GCN). Skeleton extraction of the human body in video sequences was performed using Alphapose. To resolve the problem of graph convolutional networks not being effective enough for skeletal key points feature aggregation, we propose an NAM-STGCN model that incorporates a normalized attention mechanism. By using the activation function PReLU to optimize the model structure, the improved ST-GCN model can more effectively extract skeletal key points action features in the spatio-temporal dimension for the purposes of abnormal behavior recognition. The experimental results show that our optimized model achieves a 96.72% accuracy for recognition on the self-built dataset, which is 4.92% better than the original model; the model loss value converges below 0.2. Tests were performed on the KTH and Le2i datasets, which are both better than typical classification recognition networks. The model can precisely identify abnormal human behaviors, facilitating the detection of abnormalities and rescue in a timely manner and offering novel ideas for smart site construction.

Keywords: smart site; behavior recognition; spatio-temporal graph convolutional network; attention mechanism; Alphapose

Citation: Li, Z.; Zhang, A.; Han, F.; Zhu, J.; Wang, Y. Worker Abnormal Behavior Recognition Based on Spatio-Temporal Graph Convolution and Attention Model. *Electronics* **2023**, *12*, 2915. https://doi.org/10.3390/electronics12132915

Academic Editors: Agostino Cortesi, Eric Matson, Khalid Saeed and Young Im Cho

Received: 18 May 2023
Revised: 22 June 2023
Accepted: 29 June 2023
Published: 3 July 2023

1. Introduction

A smart site is created through a combination of Internet big data, artificial intelligence, and other synergies of various technologies, along with deep integration in physical construction to achieve intelligent and efficient site safety information management in construction sites [1]. Compared with traditional building sites, it is more efficient, less expensive, and allows for the real-time monitoring of worker health and safety. Smart site construction occupies an important position in the modern development process, making construction environment management more secure and reliable [2]. The construction industry, which is one of China's pillar industries, has ushered in a rapid development at the moment; however, the construction site environment is complex, the scope of the project is challenging, and there are risks to workers' safety and health as a result of accidents such as land collapse, falls from great heights, electrocution, and summer construction workers passing out from heatstroke. Therefore, the timely detection of whether a patient has fallen or is lying down and the provision of medical assistance are critical factors to the patient's health, especially for older workers, as a delay could otherwise be fatal. The detection of abnormal human behavior in complex backgrounds at construction sites is a difficult topic because there are still many challenges to doing so, including occlusions, camera movements, the large scale of contemporary engineering projects with large amounts of information, and the similarity between actions.

Traditional behavior recognition algorithms rely on the manual design of feature vectors that can represent behavior [3]. For example, Ahmad et al. [4] extracted specific information from multiple features to identify targeted regions; the authors extracted variables such as the velocity features, color features, and texture features of abnormal behavior and then combined them with particle filtering algorithms for pedestrian tracking. Wang et al. [5] combined three features, namely a gradient histogram, optical flow gradient histogram, and motion boundary histogram, wherein a method based on multi-feature fusion was beneficial for improving the accuracy of behavior recognition. These traditional methods only perform well on specific datasets with weak model generalization, making it difficult to apply to complex, real-world contexts.

Deep learning has quickly grown in recent years and is now frequently utilized in the research of human behavior recognition, bettering traditional algorithms in terms of their performance. For example, Xie et al. [6] proposed inputting RGB images and optical flow information images of video frames to the same convolutional neural network (CNN) separately; the RGB images provide contour information and the optical flow frames provide timing information so as to obtain two prediction results, and the two resultant features are finally fused together for classification. However, the sequence motion features extracted by the dual-stream network are not complete enough, and the temporal convolution can only extract the sequence motion features of adjacent frames and split the time and space, which will reduce the accuracy rate. Shen et al. [7] proposed a random pooling method based on dropout improvement and applied it to a 3D convolutional neural network (3DCNN) for human behavior recognition to solve the overfitting problem and enhance the generalization ability. Ma et al. [8] proposed a dual-stream CNN combined with Bi-LSTM and created a behavior recognition model by adding a convolutional attention module for dual-stream feature extraction, which is capable of adaptively assigning weights and improving the accuracy of unsafe behavior recognition for construction workers. However, these algorithms are susceptible to complex backgrounds in the video, and high-resolution pixel images are used to directly enter the convolutional model calculation, which is computationally extensive and affects the speed of behavior recognition.

Compared with behavior recognition that is based on RGB images and videos, which were studied relatively early, skeletal behavior recognition is more intuitive; has a relatively small amount of data computation; is more resistant to complex environments, body proportions, occlusion lighting changes, and camera shooting angles; and has more research potential, as domestic and foreign research scholars have only drawn attention to behavior recognition based on the human skeleton as of 2017. The essential points of the human body can be retrieved from the video sequence by using a pose estimation technique, or the depth camera can directly acquire the skeleton sequence, where the skeleton sequence information is then fed into graph convolutional networks (GCNs), convolutional neural networks (CNNs), or recurrent neural networks (RNNs) for classification.

The graph convolutional network is an advanced deep learning technique that is based on a graph structure. The human skeleton structure sequence is a naturally occurring topological graph structure, with the skeleton nodes serving as the graph's vertices and the skeleton points acting as its edges. Standard neural networks like CNNs and RNNs have limitations in processing graph inputs because they stack the features of nodes in a specific order [9]. To overcome this constraint, GCNs are separately propagated at each node while ignoring the input order of the nodes, which is more advantageous than CNNs and RNNs in terms of processing graph data and learning feature information more efficiently.

Yan et al. [10] first combined a GCN with skeletal key points recognition and proposed the spatio-temporal graph convolutional network (ST-GCN) model, which had better robustness and novelty and triggered a series of subsequent improvement studies. For example, Li et al. [11] designed a trainable actional links inference module (AIM) to extend the previously set linkage relations on the original graph convolution to model long-range global dependencies, which compensates for the previous deficiency where the more distant linkage relations of ST-GCN were ignored and extracts more global information.

Zhou et al. [12] proposed a PoseC3D-based action recognition model, which first extracts information about the key points of the human skeleton and then generates a 3D heat map, which is stacked and inputted into a 3D-CNN classification network to output the recognition results, improving the robustness of the model to noise-laden skeleton sequences in complex motion backgrounds. All of these techniques have enhanced network performance, but the network structures are intricate and still have drawbacks.

To address the aforementioned issues, this paper adopts the high-precision human posture estimation algorithm Alphapose to extract the key point data of the human skeleton in the video and selects the ST-GCN as the base model for behavior recognition and for monitoring whether there is abnormal behavior in falling and lying down.

In the Section 1, we discuss the significance of the study, the current state of research, and the limitations of behavior recognition. In Section 2, we present the theory related to the conventional methods and pose estimation algorithms for the foreground extraction and action recognition models. In Section 3, we improve the detection accuracy of the Alphapose algorithm and propose a new NAM-STGCN action recognition model that includes an attention mechanism module and whose structure is optimized by using the activation function PReLU. The model's training and validation experiments are conducted in Section 4 of this work, which significantly increases the model's accuracy for identifying abnormal behavior and lowers its loss value. Finally, we compare the model with other classical classification models to draw conclusions and enhance intelligent information security management at construction sites.

2. Related Works

2.1. Traditional Foreground Extraction Methods

Foreground extraction algorithms are widely used in camera surveillance to extract foreground targets (moving targets) from background images in video stream data, which is a prerequisite for target tracking and anomaly monitoring [13]. Our worker aberrant behavior detection system is used to identify human targets and gather motion feature data, which is a crucial step in later behavior recognition. In addition to workers in action, a construction site will contain moving machinery and equipment, elevated planks, and other occlusions, and the complex background environment can interfere with extraction; thus, it is critical to choose a robust foreground extraction method. The two approaches that are most frequently utilized for motion foreground extraction and detection are the optical flow method and the frame difference method.

The optical flow method's primary objective is to determine the optical flow field by assigning each pixel location in the image a corresponding velocity vector, calculating that vector's velocity in the gradient direction. If the velocity of the point exceeds a threshold value, then the region in which it is located belongs to the foreground; otherwise, it will be considered as a background region.

The frame difference method for motion target extraction is computationally simple and does not require background modeling; it simply performs a difference operation on the pixels between frames. If the gray value of the pixel point exceeds an artificially set threshold, the region where the point is located is considered as the foreground image.

2.2. Alphapose Algorithm

There are two types of multi-person human pose estimation techniques: one is top-down estimation, where all human positions are first detected from the sequence images, and then all single targets are separately estimated for pose estimation to extract skeletal key points; the other method is bottom-up estimation, which detects the skeletal key points of every character in the video frames before matching the key points to connect them into a graph, removing the incorrect connections for each pose through graph optimization. While the latter method recognizes people faster, it does not fully utilize the global spatial information of human posture, resulting in a less accurate recognition than when using the top-down method.

Alphapose is an open-source pose estimation algorithm developed by Cegou Lu's team at Shanghai Jiao Tong University that allows for the accurate estimation of multi-person poses [14]. When using the Alphapose algorithm to detect the key points of the human skeleton, the main program called Tiny-YOLOv3 is used to detect the human body area and extract the target frame. The extracted target box location information is then fed into a symmetric space transformation network (SSTN), which is capable of extracting individual targets from inexact bounding boxes. By only retaining the skeleton pose with the highest confidence level, the novel parametric pose non-maximum suppression (PPNMS) technique solves the issue of redundant poses produced by existing skeleton models as a result of positioning errors and achieves accurate skeleton key point recognition. The literature [15] was tested on the MSCOCO standard dataset and was compared with several other bone detection models, where the Alphapose model was shown to have certain advantages in terms of accuracy and recognition speed, according to the results. The top-down representative algorithm Alphapose technique was used in this study because the subsequent behavior identification model necessitates the input of high-quality skeletal sequence data.

The Alphapose algorithm is used to detect human skeletal key points in a video, and 18 skeletal key points are obtained in each frame, $\{(ki, ci), i \in (1, 18)\}$, where ki and ci denote the position coordinates and confidence level of the ith key point, respectively. The key points of the human body are the important skeletal parts on the human body topology, such as the 18 key points in Figure 1: nose, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right span, right knee, right ankle, left span, left knee, left ankle, right eye, left eye, right ear, and left ear (numbered 0, 1, . . . , 17 in order). The human skeleton is a representation in which the 18 key points detected by the Alphapose algorithm are connected in the manner shown in Figure 1, with each key point corresponding to a location coordinate.

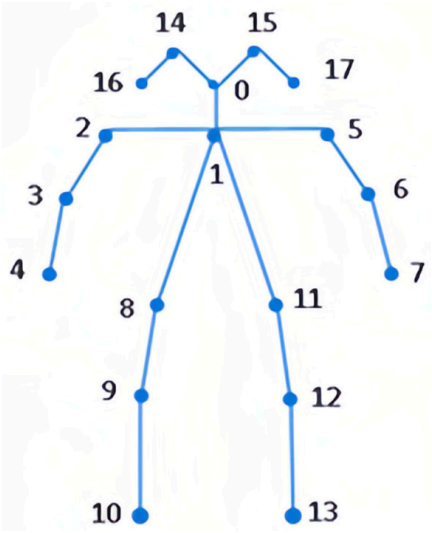


Figure 1. Structure of the human skeleton with 18 key points.

2.3. Comparison with Traditional Extraction Methods

The foreground extraction methods optical flow method and frame difference method were compared with the deep learning human pose estimation algorithm using the same video for detection, and Figure 2 displays the outcome of the experiment. The advantages and disadvantages of the method are summarized in Table 1.

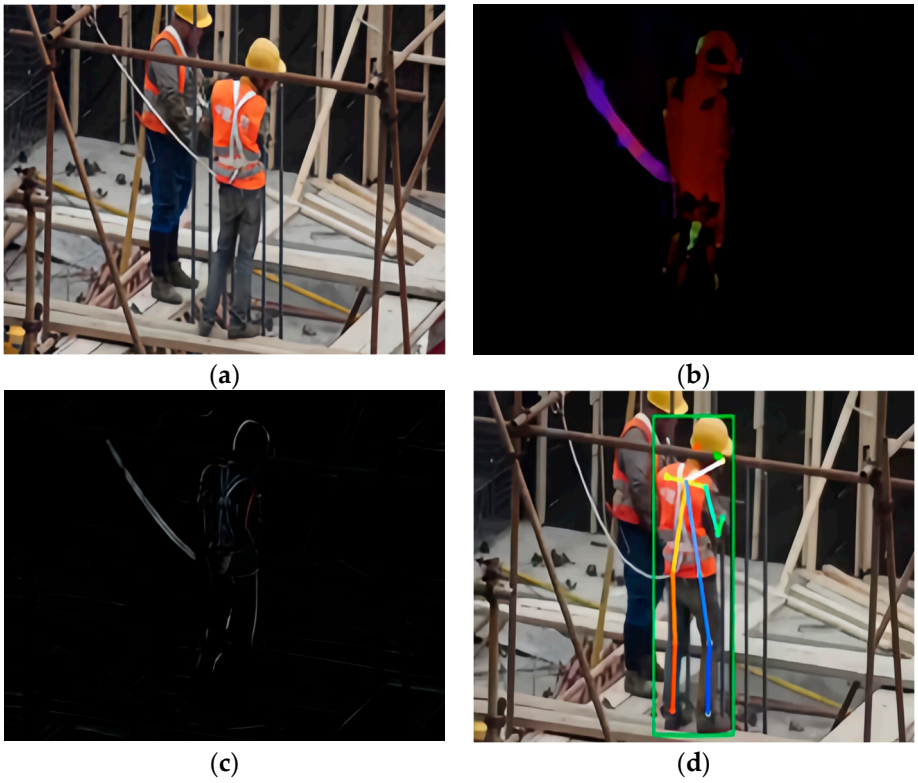


Figure 2. Comparison of the extraction effect between the traditional method and the pose estimation method. (**a**): Video keyframe original; (**b**): optical flow method extraction; (**c**): frame difference method extraction; (**d**): Alphapose algorithm extraction.

Table 1. Comparison of the advantages and disadvantages using traditional methods.

Prospect Extraction Method	Definition	Advantages	Disadvantages
Optical flow method	The velocity field is estimated based on the gradient direction of the pixel points in the image sequence, and the moving targets and scenes are detected and segmented by analyzing the changes in the velocity vector.	More flexible; no other a priori information is needed to detect moving objects	Large calculation volume; time consuming; vulnerable to noise, shadow obscuration, and other external factors
Frame difference method	The pixel values between two or three adjacent frames of an image sequence are subjected to a difference operation to extract the motion regions in the image.	Small calculation, simple algorithm; not easily affected by ambient light	Not suitable for the case of background motion; prone to edge hollowness and incomplete extraction of motion targets
Alphapose algorithm	A popular top-down detection algorithm that first detects all people in the image sequence and then extracts the skeletal key points for all single targets separately.	High detection accuracy; low interference from external environment background; supports multi-person attitude estimation; good real-time performance	The algorithm is relatively complex; constrained by the target detection task

The contrast experiment uses dense optical flow, and the foreground of the person in the video is extracted by using the optical flow method as shown in Figure 2b. The optical flow information of the worker on the left was not extracted in the multi-person target detection, but the optical flow information of a non-personal foreground target, a safety

rope tied behind the worker that moved with the worker, was extracted; the extracted feature information's edge accuracy was also low. Although the complicated background can substantially impede the calculation of the optical flow field, the visible optical flow technique can manage the case of background motion. Furthermore, the optical flow approach performs poorly in real-time, making it hard to meet the demand for monitoring.

Figure 2c illustrates how to extract the person's foreground using the frame difference method of detection. In the multi-person target detection, the same workers on the left were missed, the extracted human target outline was incomplete, and the human target was truncated in chunks, and some non-human target information is also extracted.In unsafe behavior detection tasks such as worker falls and fainting, faster movement speeds may lead to a hollow condition of the extracted human target, and the human silhouette cannot be extracted correctly.

The foreground extraction method used in this study was the Alphapose estimation algorithm, and the accuracy of motion information of human key points was higher compared with the person outline information; the algorithm's effect is shown in Figure 2d. In multi-person target detection, there are no missed or incorrect detections, and the use of deep learning methods is more robust for human target extraction in complex backgrounds such as construction sites. This shows that the light green detection box accurately frames the two workers in the figure, and the key points of the person are extracted intact. Compared with traditional processing methods, human foreground extraction based on deep learning will be more effective and will help subsequent abnormal behavior recognition tasks.

2.4. ST-GCN Algorithm Principle

The spatio-temporal graph convolutional network (ST-GCN) is used to extract the temporal and spatial features of the skeleton, which are inputted into softmax to predict action classes. In ST-GCN, a spatio-temporal graph with N skeletal key points and T-frames can be constructed, $G = (V, E)$, where V is the set of skeletal points and E is the set of skeletal edges. This skeleton sequence includes connections between the same key points that naturally occur within the human body as well as links between neighboring frames. The set of all skeletal key points in this spatiotemporal graph is:

$$V = \{v_{ti} | t = 1, \ldots, T, i = 1, \ldots, N\} \quad (1)$$

The undirected spatio-temporal diagram constructed on the skeleton sequence is shown in Figure 3 [16], where the blue lines connect the spatial dimensions and the light green lines connect the temporal dimensions on different frames, graphically representing the dynamic changes of the extracted human skeleton in time and space. The location coordinates of the ith joint on the frame and the confidence level make up the feature vector on node $F(v_{ti})$, which serves as the input to the ST-GCN network model.

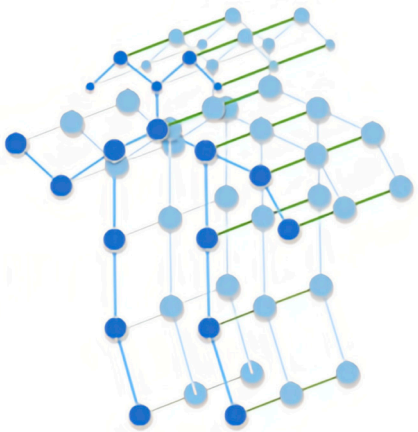


Figure 3. Spatio-temporal diagram of ST-GCN.

In the spatio-temporal map of the skeleton, the mapping operation is performed by mapping the adjacent regions of the skeletal point V_{ti}. $l_{ti} : B(v_{ti}) \rightarrow \{0, \cdots, K-1\}$ assigns different weight parameters to the adjacent regions of the skeletal point V_{ti} to construct the weight function w. By using the partitioning strategy, the adjacent region $B(v_{ti})$ of the skeletal point V_{ti} can be decomposed into K subregions $\{0, \cdots, K-1\}$, which simplifies the mapping changes. The weight function $w(v_{ti}, v_{tj})$ is

$$w(v_{ti}, v_{tj}) = w(l_{ti}(v_{tj})) \tag{2}$$

where $l_{ti}(v_{tj})$ denotes the label that v_{tj} belongs to in the molecular set labeling with v_{ti} as the central node; $w(l_{ti}(v_{tj}))$ denotes the weight value corresponding to the label $l_{ti}(v_{tj})$.

Combining the characteristics of the graph data and the properties of the convolutional network, the sampling function and weight function are defined using Equation (2) to obtain the spatial graph convolution formula, as follows:

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{tj}) \cdot w(l_{ti}(v_{tj})) \tag{3}$$

where f is the feature graph and V_{ti} is the node of the graph. $B(v_{ti})$ is the set of neighboring nodes of V_{ti} and is divided into K subsets, and $Z_{ti}(V_{tj})$ is the normalization term with a normalization constraint on the subset bases.

In constructing the skeleton spatio-temporal graph, the connection of the same key points on a sequence of continuous skeleton diagrams can be used to describe the time dimension and be selected to apply spatial graph convolution on a sequence of skeleton frames with a time range of Γ; Equation (3) can be expanded to the temporal dimension. The spatio-temporal graph convolution equation is

$$B(v_{ti}) = \{ \ v_{qj} | d(v_{tj}, v_{ti}) \leq D, |q-t| \ \leq \lfloor \Gamma/2 \rfloor \ \} \tag{4}$$

where Γ is the time range in the adjacent graph; that is, the original spatial distance constraint is less than D (take D = 1 in equation), and add the time constraint of before and after a time of less than or equal to $\lfloor \Gamma/2 \rfloor$ frames. $d(v_{tj}, v_{ti})$ is the minimum path from v_{ti} to v_{tj}, and $|q-t|$ is the inter-frame difference on the time axis.

The spatio-temporal graph convolutional network sampling function is the same as the spatial graph convolutional network, and the graph sampling function is

$$p(v_{ti}, v_{tj}) = v_{tj} \tag{5}$$

For the weight function, because the time axis is ordered, the result of updating the mapping of the neighborhood of v_{ti} skeletal points l_{ST} is

$$l_{ST}(v_{qj}) = l_{ti}(v_{tj}) + (q - t + \lfloor \Gamma/2 \rfloor) \times K \tag{6}$$

Where $l_{ti}(v_{tj})$ denotes the label mapping at a single frame v_{tj}. K is the number of subsets to be divided, and $(q - t + \lfloor \Gamma/2 \rfloor) \times K$ denotes the adjustment of the label grouping mapping after joining the time domain. In this way, the spatio-temporal map convolution based on the skeleton spatio-temporal map is constructed.

We used a spatial configuration partitioning approach to divide the neighborhood set of key points so that the GCN can extract spatial features on the skeleton, which assigns three different weights to the neighborhood $B(v_{ti})$ of the root node V_{ti}, and this spatial partitioning strategy works the best. There are three subsets created from the set of neighbors: the root node itself; the centripetal group, which consists of nodes that are next to and closer to the skeleton's center of gravity compared with the root node; and the

centrifugal group, which consists of the remaining nodes. Take K = 3; the space allocation partitioning policy is

$$l_{ti}(v_{tj}) = \begin{cases} 0 & \text{if rj} = \text{ri} \\ 1 & \text{if rj} < \text{ri} \\ 2 & \text{if rj} > \text{ri} \end{cases} \quad (7)$$

where $l_{ti}(v_{tj})$ is the weight mapping function of all the subsets of the division $t = 1, 2, \cdots, T$, $i = 0, 1, \cdots, 17$ and v_{tj} is the key point of the j ($j = 0, 1, 2$) subset of the t frame; r_j is the distance from v_{ti} to the point at the center of gravity, and r_i is the average distance from v_{ti} to the skeleton's center of gravity.

3. Proposed Method

3.1. Alphapose Algorithm-YOLOv5

Because the accuracy of the top-down pose estimation method's detection target location directly influences the accuracy of the algorithm, we consider a human detection model with higher detection accuracy and better results. The flowchart of the algorithm is shown in Figure 4. YOLOv3-SPP is used by default in the official Alphapose repository, but its detection effect needs to be improved and its anti-obscuring ability is poor. The human posture estimation algorithm in the literature [17] uses Tiny-YOLOv3 to generate a detection frame for the recognition of the unsafe ladder climbing behavior of construction workers, which reduces the model computational complexity and improves the detection speed; however, there is still much room for improvement in the model accuracy, which further affects the accuracy of subsequent behavior recognition.

With its capacity to handle more complicated background settings (light changes and occlusions), the emergence of YOLOv5 in recent years has achieved new heights in the field of target recognition. Therefore, several experiments were conducted using videos of construction workers at construction sites, and the Tiny-YOLOv3 model was replaced in this study with the YOLOv5x model of the YOLOv5 series to detect human body regions to generate target frames, which improved the model's detection accuracy.

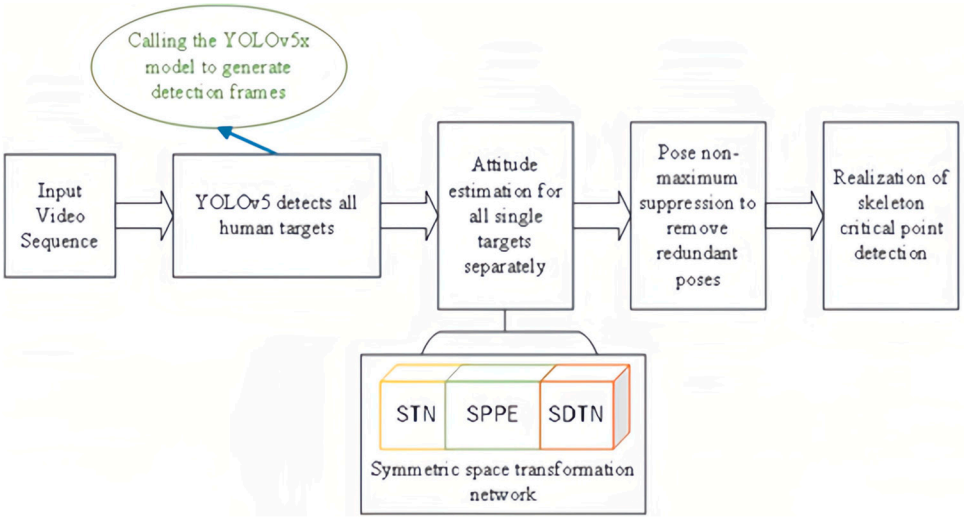


Figure 4. Flow chart of the Alphapose algorithm used in this paper.

The detection effect comparison chart is shown in Figure 5. Figure 5 shows the same video detection process, where the Tiny-YOLOv3 algorithm model in Figure 5a misses detection of the left worker; Figure 5c is impacted by the background interference of the site elevation, where the detection frame is unable to fully frame the workers, leading to part of the extracted skeleton key points outside the frame and other issues. After replacing the model with the YOLOv5x algorithm, the detection effect is improved.

When the YOLOv5x model is used for human detection, as shown in Figure 5b, there is no target miss detection, and in Figure 5d, the key points extracted from the multi-person pose are relatively complete and all within the detection frame. The outcomes demonstrate that our chosen Alphapose algorithm performs better at detection tasks and is more resilient to complicated background interference.

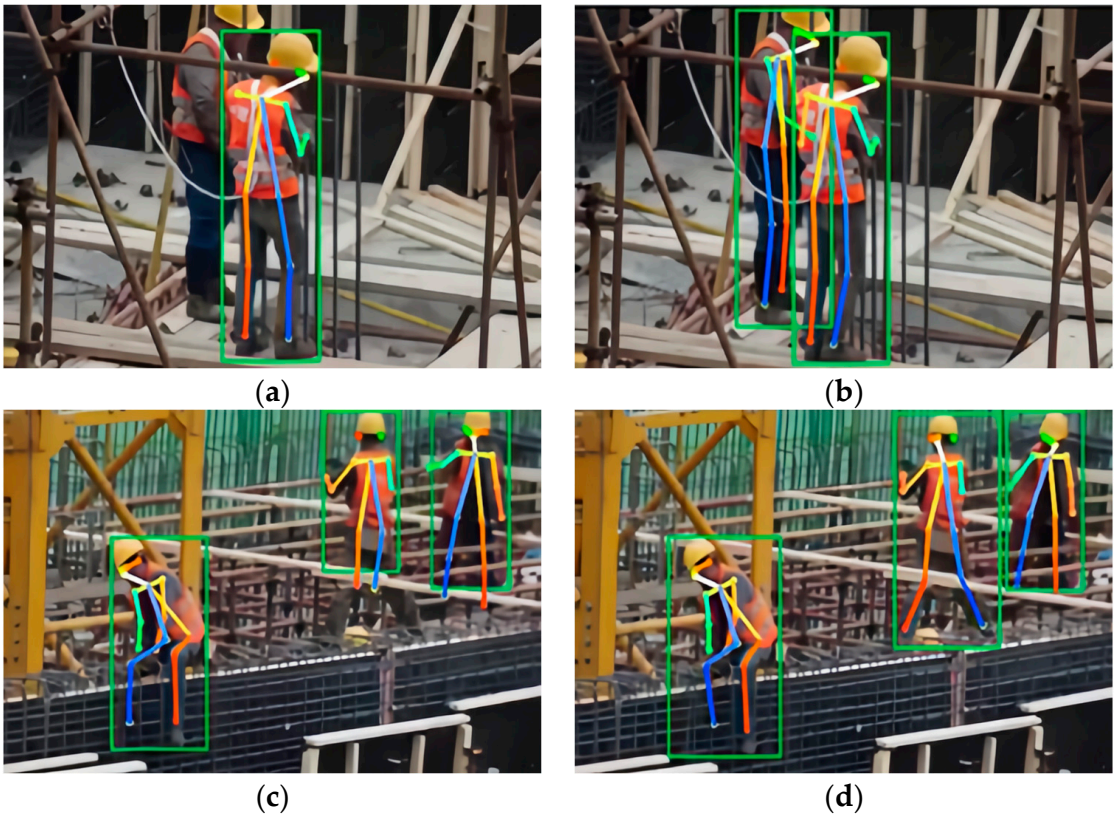


Figure 5. Comparison of the effect before and after model improvement. (**a**): Scenario I, the original Tiny-YOLOv3 model; (**b**): Scenario I, the YOLOv5x model; (**c**): Scenario II, the original Tiny-YOLOv3 model; (**d**): Scenario II, the YOLOv5x model.

3.2. The Proposed NAM-STGCN Model

This study is based on a spatio-temporal graph convolutional neural network (ST-GCN) used for action recognition, which incorporates the idea of a temporal convolutional network (TCN) into a graph convolutional neural network (GCN). The essence of the work of graph neural networks is feature extraction. The working schematic of the ST-GCN is shown in Figure 6. The input videos are first processed by the Alphapose pose estimation algorithm to obtain the position coordinates data of human skeletal key points for each frame; then a spatio-temporal graph with key point coordinates as graph nodes and connections of human body structure and connections of consecutive multiple time frames as edges can be constructed as the input of ST-GCN according to the given key point connection order. The input data is subjected to multi-layer spatio-temporal graph convolution operation and extracts feature information, aggregates the node features to represent the features of the whole graph, and generates a higher-level feature map on the graph, as indicated by the red circle in the figure. Finally, action classification is performed using softmax to achieve action recognition.

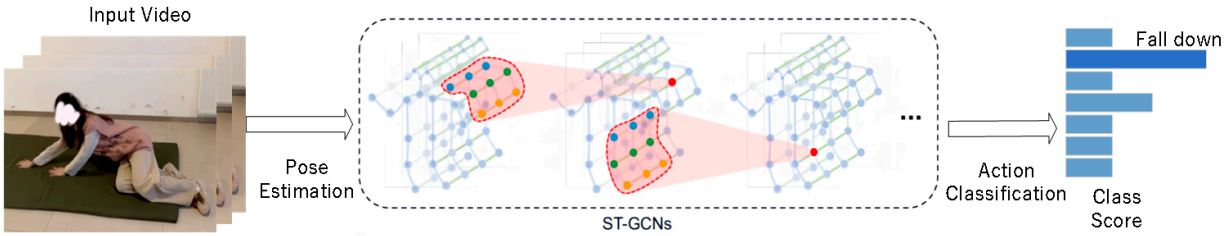


Figure 6. Workflow graph of action recognition model.

Because the ST-GCN only learns the local information of a certain neighborhood and does not learn or capture the relevant information between all joints (global information), it is also difficult for the network to connect new key points that are far away or on the non-existent edges, which affects the model recognition accuracy. Therefore, this paper proposes the fusion of a normalization-based attention module (NAM), an attention mechanism that preserves channel and spatial aspects to enhance the importance of cross-latitude interactions and reduces the weight of less significant features; when implemented as the improved NAM-STGCN model, the model can combine global and local information to build a rich hierarchical structure.

The attention mechanism module of NAM applies a sparse weight penalty to make the weight calculation more efficient while guaranteeing the same performance and while improving the classification accuracy of the network; the scale factor of batch normalization is also used to indicate the importance of the weights so that the computational burden of adding fully connected and convolutional layers such as SE (squeeze-and-excitation), BAM (bottleneck attention module) and CBAM (convolutional block attention module) can be avoided.

The principle of the post-fusion model is shown in Figure 7. It consists of the Batch-Normal layer, the spatio-temporal feature extraction unit layer (containing 9 ST-GCN units, each of which alternately uses a graph convolution (A-GCN) fused with a graph attention mechanism and time-domain convolution (TCN), where the size of the spatio-temporal convolution kernel is 5×9), a NAM attention mechanism module, an average pooling layer (av-pool), and a fully connected layer (FC).

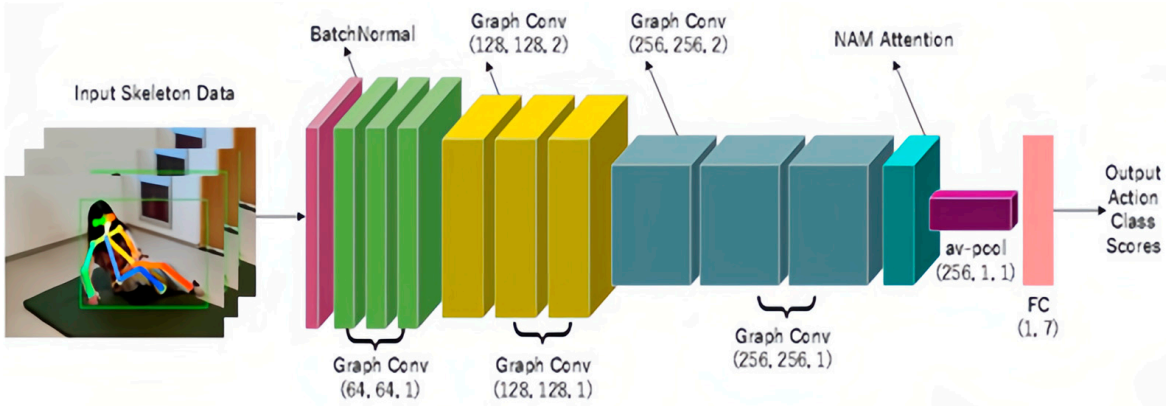


Figure 7. Structure of the improved NAM-STGCN network.

The distribution of the skeletal sequence data is normalized using the BatchNormal layer, which improves the model's capacity to generalize. To enable feature fusion across regions, each ST-GCN unit employs the fusion of feature residuals. To avoid overfitting the model, a dropout mechanism of 0.5 is also utilized. Finally, the features are downsampled using a pooling layer operation with a step of two after the fourth and seventh ST-GCN units. The NAM channel attention mechanism module is fused behind the final layer of the spatio-temporal graph convolution with 256 input channels to make it more effective at extracting action features in the spatio-temporal dimension; the NAM attention mechanism

module is also introduced in the GCN for forward propagation so that the node information under the extracted spatial features is strengthened by the attention mechanism to improve the model accuracy. An av-pool layer is used to reduce the dimensionality of high-dimensional action features and aggregate node features to represent the entire graph's features. The FC layer is the fully connected layer. The softmax function is used to output the class and confidence level of the model's recognition behavior.

The fused NAM network is a lightweight and efficient attention mechanism that uses the modular integration of CBAM to redesign the channel attention module by using a batch normalization (BN) scale factor [18]. As in Equation (8), the scale factor is the variance in the BN, reflecting the magnitude of the variation in each channel and the corresponding importance.

$$B_{out} = BN(B_{in}) = \gamma \frac{B_{in} - \mu_B}{\sqrt{\sigma_B^2 + \in}} + \beta \tag{8}$$

where μB and σB are the mean and typical deviation of mini batch B, respectively; γ and β are the scale and displacement parameters that can be trained for learning.

Thus, the channel attention mechanism module is shown in Figure 8 and Equation (9), with M_C denoting the final output features obtained and γ denoting the factor that scales with each channel, such that the weights of each channel can be gained; the weights are obtained from $W_\lambda = \frac{\gamma_i}{\sum_{j=0} \gamma_j}$.

$$M_c = sigmoid(W_\gamma(BN(F_1))) \tag{9}$$

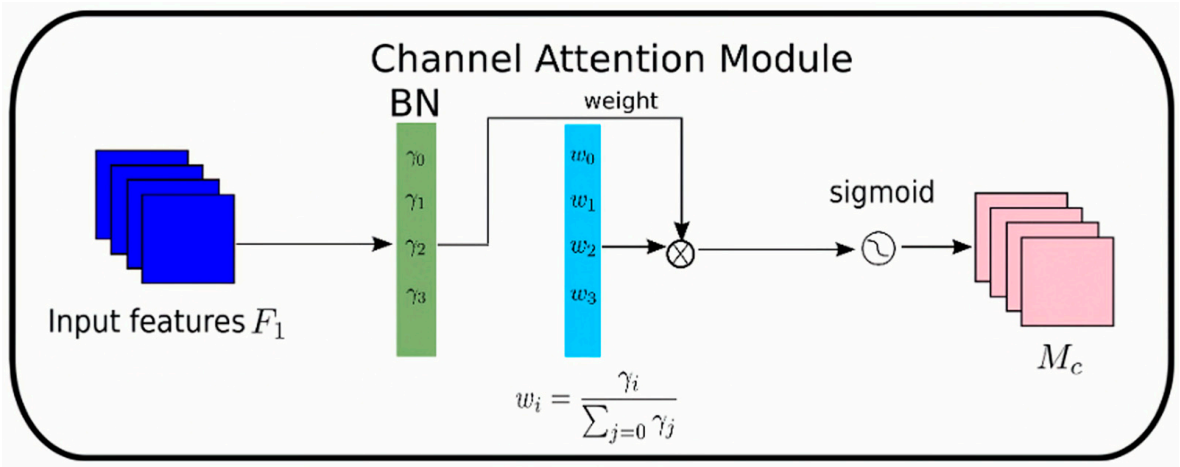


Figure 8. Channel attention module structure.

In this paper, a tensor of [B, C, T, V, M] is used to represent the initial input data for human skeletal behavior recognition, where B is the training batch, C is the feature dimension of the key points, T is the sequence preservation length of the skeleton keyframes, V is the amount of recognition key points, and M is the amount of people with high average confidence in the human body retained in the keyframes. The inputs in this study were (16, 3, 155, 18, 2). The average human confidence level is

$$c = \frac{\sum_{i=0}^{17} ci}{18} \tag{10}$$

where c_i is the confidence level of the ith key point.

4. Experiments and Results Analysis

4.1. Dataset Creation

Due to the lack of behavioral datasets with construction sites as the application setting and due to their poor generality, we created a behavioral recognition dataset and identified and classified the abnormal behaviors that construction workers are likely to engage in.

Given that workers often carry out construction work, directing machine operations, equipment inspection, and other activities in the summer, they are prone to heat stroke or fainting from illness. This study defines five types of normal movements, such as sitting down, waving hands, sitting, standing up, and walking, and two types of abnormal movements, such as falling and lying down, forming a total of seven categories. Four of the fall videos are from the publicly available Le2i fall detection dataset; the rest were taken by eight volunteers from the lab in two locations, the hallway and the living room, to simulate normal and abnormal actions that may occur during worker construction, and they were captured on video using cell phones.

We used both horizontal flip and luminance enhancement to enhance the video data; there was a total of 303 videos with a video resolution of 640 $\times$ 480 and a frame rate of 25 fps and about 45,400 frames of skeletal sequence maps for training and testing, which contained 195 normal action videos and 108 abnormal action videos. The duration of each video shot is 5–7 s. The data were shot utilizing various angles, distances, and periods of acquisition while accounting for the actual project camera angle, a light background, and other aspects. Some of the self-built dataset keyframe images are shown in Figure 9.

The self-constructed behavioral dataset consists of a total of 303 videos, with 80% of the dataset used being randomly divided into the training set and 20% being divided into the test set. Before training and evaluating the model, the human skeleton information in the action video was extracted frame-by-frame using the Alphapose algorithm and was saved at a frame rate of 25 fps, with each video being shot for 5–7 s and about 45,400 frames of skeletal sequence maps being created. The extracted skeletal key point information was saved as a key-value pair (JavaScript object notation, JSON), which was used for the training and testing of the anomalous behavior recognition network. Given that the real shooting videos were limited and that most of them were used for training, there were 242 JSON files in the training set and 61 JSON files in the test set.

4.2. Experimental Platform and Model Performance Experiments

In order to verify the feasibility and practical effect of the method used in this paper, the improved model was trained and tested. The experiments were conducted based on a 64-bit Windows 10 operating system with the following hardware configurations: the CPU was an Intel(R) Core (TM) i5-8300H @ 2.3GHZ processor, and the GPU was an NVIDIA GeForce GTX 1050Ti with 8GB of RAM. The equipment is all HP Pavilion Gaming Laptop from Baoding, China; software environment: the graphics processing gas pedal was CUDA11.6 and cudnn8.2.1, the programming language was Python3.9, and the deep learning framework was Pytorch1.12.1. The model was trained using the SGD optimizer, the momentum was 0.9, the batch size was set to 16, the number of training iterations (epochs) was set to 150, the base learning rate was 10^{-2}, and the weight decay factor was 10^{-4}.

4.2.1. Discussing the Performance Impact of Different Activation Functions

We optimized the activation functions of the TCN and residual modules in the ST-GCN network. The results on the test set are shown in Table 2, and the model's identification accuracy was greatly increased by using the PReLU function. The original ST-GCN model uses the ReLU activation function to solve the gradient disappearance problem caused by sigmoid, but it converts to zero when the input is negative, which causes some neurons to "die" easily during training. To solve this problem, we replaced ReLU with the ParametricReLU (PReLU) function, which has more learning ability and effective features.

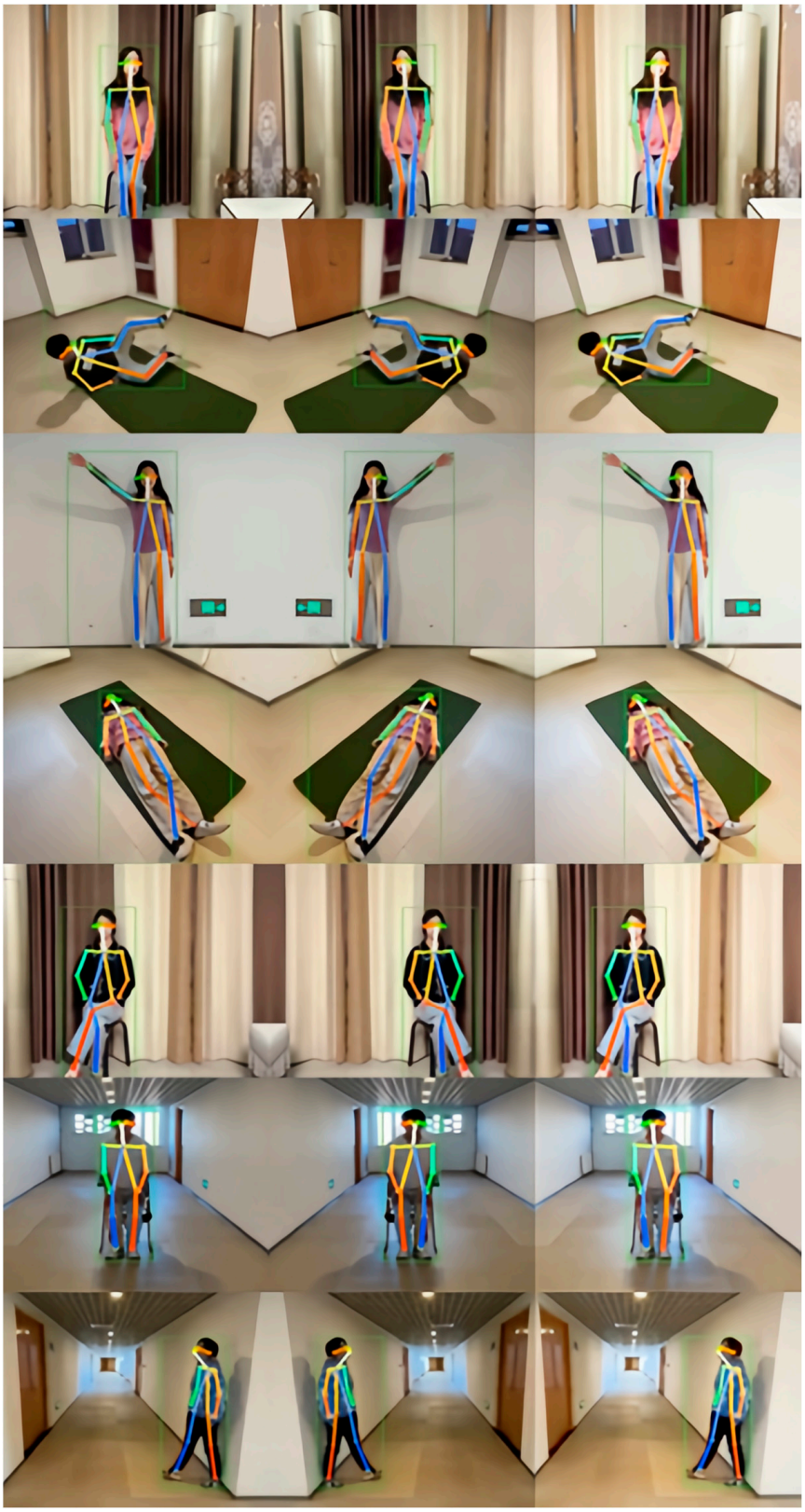


Figure 9. Sample self-built behavioral dataset. The first column of the image is the original video keyframe, the second column is the keyframe after horizontal flipping, and the third column is the keyframe after brightness increment processing.

Table 2. Effect of different activation functions on model performance.

Activation Function	Accuracy/% (Top 1)	Accuracy/% (Top 5)
ReLU	91.80	98.36
LeakyReLU	93.44	100.00
PReLU	95.08	100.00

Although the LeakyReLU function was used to adjust the gradient in the training experiments by providing a very small constant slope value in the region where the input was negative, which improved the situation where some neurons were never activated, the top 1 accuracy improved by 1.64% relative to the ReLU function of the original model, and the top 5 accuracy reached 100%. However, its gradient depends on manual adjustment and the settings are fixed, which is not flexible; this could result in a poor performance in some cases.

To circumvent these restrictions, we used the PReLU activation function, which converts the gradient values of the negative half-axis into dynamically learnable parameters. This provides us with more freedom to dynamically modify the parameter values during the training process in order to achieve the best training results. This study builds a deeper network layer, and the use of PReLU function is beneficial for ensuring that the gradient runs through the entire model structure while speeding up the convergence of the algorithm and improving the network performance. It was experimentally verified that the accuracy of the top 1 using the PReLU function improved by 3.28% relative to the ReLU function of the original model and by 1.64% relative to the LeakyReLU function; the accuracy of the top 5 also reached 100% with the best results.

4.2.2. Validating the Model Recognition on the Dataset

The performance comparison experiments of the model verified the effectiveness of the optimized model in this paper for recognition on the self-built behavioral dataset. The overall test results of the method in this paper were a 96.72% accuracy for the top 1 and a 98.36% accuracy for the top 5; the accuracy of the top 1 was selected as the final accuracy judging criterion in this paper.

Figure 10 shows a graph of the change in accuracy of the top 1 test set, which was plotted according to the accuracy of the model saved once every 10 iterations. The accuracy curve of the original classical ST-GCN model with the ReLU activation function starts to converge after 50 iterations, and the accuracy rate fluctuates around 90%. However, the curve showed local mutations in the early training period, indicating unstable results, and only reached a maximum accuracy of 91.80% after 140 iterations. The NAM-STGCN model incorporating the normalized attention mechanism module has a higher accuracy than the original model during the first 10 iterations, and the accuracy rate increases slowly in the early stage but gradually converges after 40 iterations; finally, the accuracy rate after convergence is smoothly maintained at 93.44%, and the model performance is improved.

Our final model is a further optimization of the proposed NAM-STGCN model with the activation function being replaced by the PReLU function. It was tested on a self-built dataset with no local mutation in the early stages and started to converge after 40 iterations with a faster convergence rate. Its test accuracy is 96.72%, which is 4.92% better than the original ST-GCN model and 3.28% better than the NAM- STGCN model, and it has the properties of smooth convergence and the best model performance.

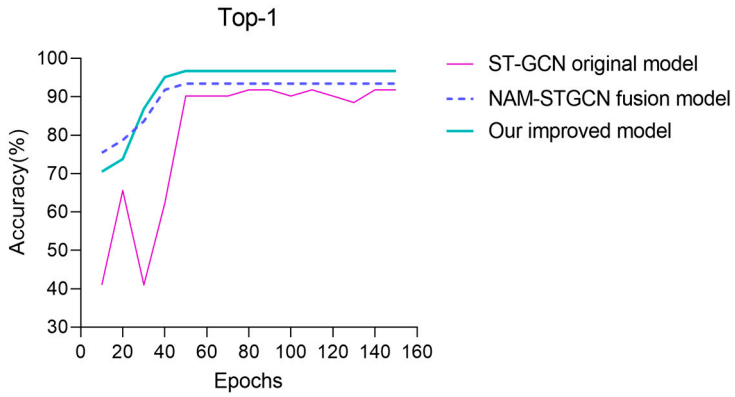
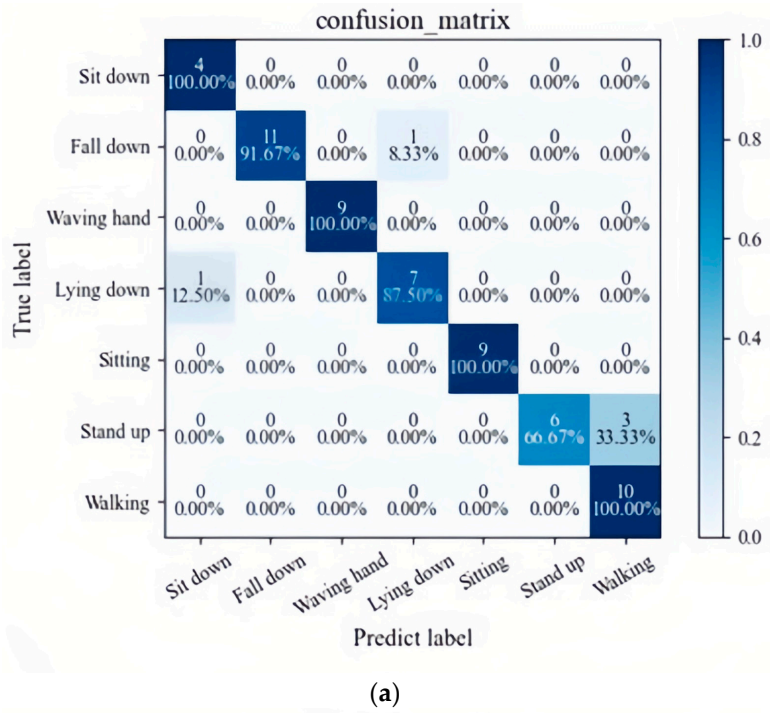


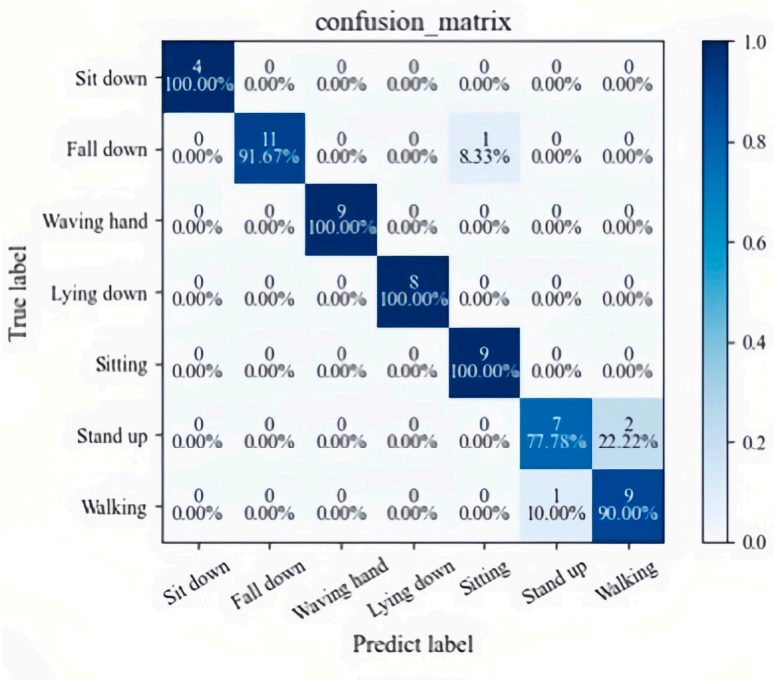
Figure 10. Comparison of model accuracy variation.

Figure 11 illustrates the validation of the modified model on the test set and shows its confusion matrices for identifying the seven classes of actions with the best accuracy. Our model clearly has the highest recognition accuracy, and only the transient behavior of standing up has a lower recognition accuracy. Considering the temporal information extraction of the ST-GCN model, the duration of each action of the self-built dataset is around 5–7 s, but for transient behavior such as standing up, where the subject can only try to get up slowly to ensure the singularity of the data label action, a video corresponds to only one label; thus, standing up contains standing action it is easy to confuse with walking action. The recognition rate of the remaining six types of behaviors all reached 100%, respectively. These results indicate that the improved model has a high recognition rate as well as good robustness.

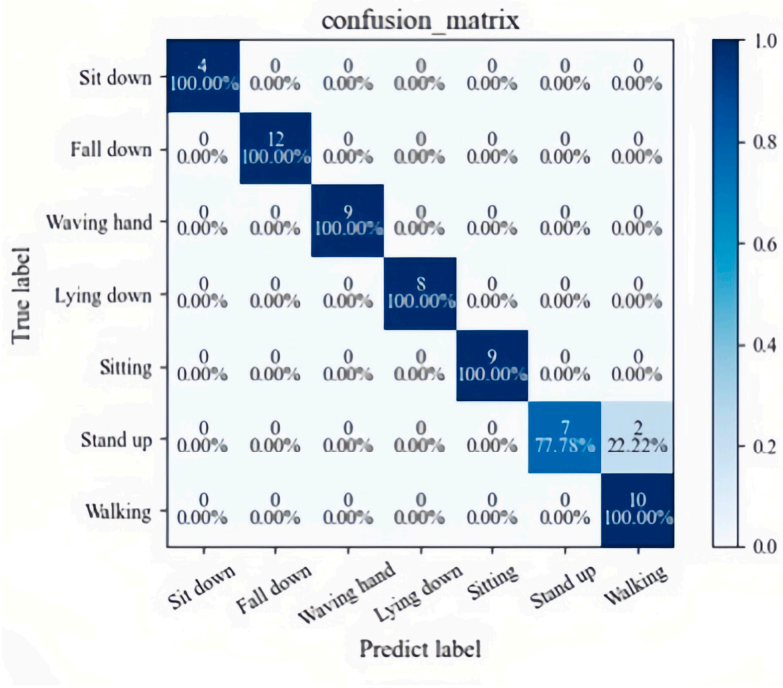


(**a**)

Figure 11. *Cont.*



(**b**)



(**c**)

Figure 11. Confusion matrix diagram for model testing. (**a**): Original ST-GCN model; (**b**): the NAM-STGCN model; (**c**): our model.

Figure 12 shows the comparison of the loss value curves obtained from the improved before and after models that were validated on the self-built behavioral dataset. As can be seen from the figure, the original ST-GCN model starts to converge around 50 iterations, and the loss value floats around 0.6. However, the NAM-STGCN model with a fused normalized attention and our improved final model had a lower drop in the loss value. Compared with the original model, our modified model starts to converge after 45 iterations, converging faster and more smoothly after convergence, and the loss values all converge

to below 0.2 with little change. This indicates that our modified model has a reduced loss value and better learning ability.

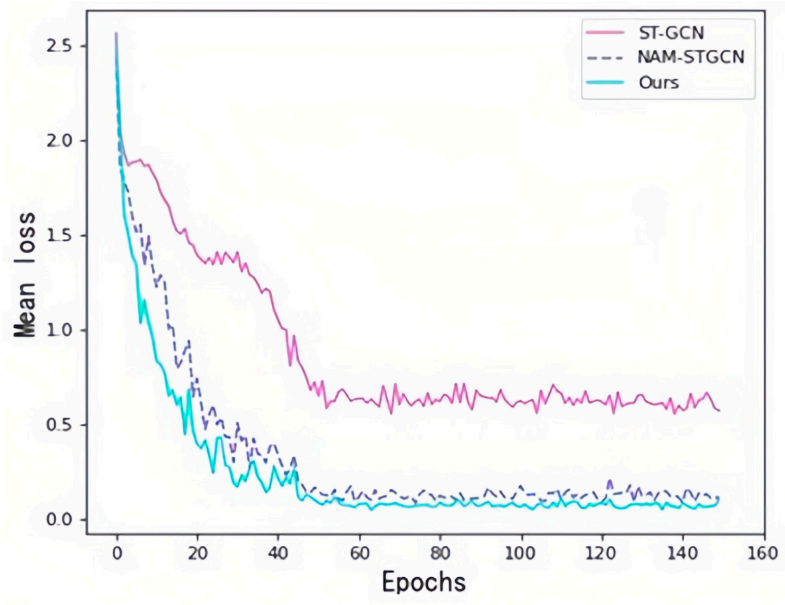


Figure 12. Comparison of the model loss value changes.

4.3. Algorithm Comparison Analysis

The study model was evaluated using the KTH public dataset and the Le2i fall detection dataset, and the model was compared with other model algorithms in the literature to better validate the effectiveness of the improved model proposed here. The hardware device and software environment used in the comparison experiment are the same as above; the batch size was 8, the initial learning rate was 10^{-3}, and the weight decay coefficient was 10^{-4}.

4.3.1. Testing on the KTH Public Data Set

The KTH dataset is a milestone in the field of computer vision and selects 6 types of actions performed by 25 people in 4 different scenes and shooting angles (outdoor S1, outdoor near and far scale change S2, outdoor different clothes S3, indoor S4). With the dataset including boxing, hand clapping, hand waving, jogging, running, walking, we selected boxing as abnormal behavior and the remaining five types of action as normal behavior. There are a total of 600 videos, and each frame in the video is preprocessed as a 640 $\times$ 480 pixel image; we placed 480 videos in the training set and 120 videos in the test set.

The results of the proposed model tested on the KTH dataset in this study are plotted on the confusion matrix shown in Figure 13, with a test accuracy (top1) of 94.96%. The results show that jogging and running, two behaviors with similar movements, are more difficult to distinguish and are easily confused, resulting in a low accuracy rate of about 85%. Even so, the model still managed to reach 100% accuracy for regular actions such as hand clapping and hand waving and a 96% accuracy for walking, which is similar to running. These results indicate that our proposed improved ST-GCN model still has a high recognition rate for behaviors with significant variances.

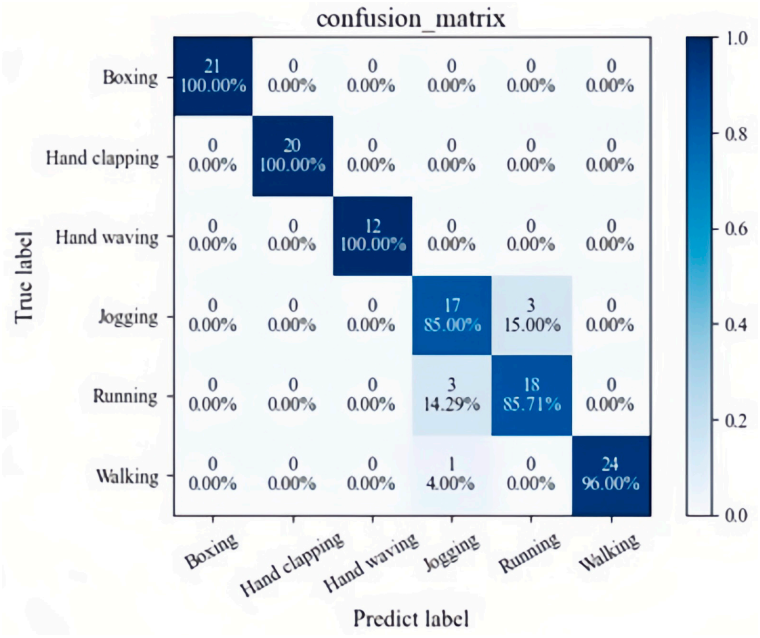


Figure 13. Confusion matrix plot for testing on the KTH dataset.

The research approach used in this study is contrasted with the typical behavior recognition method used in recent years in order to further validate the efficacy of our proposed model, and the results are displayed in Table 3. It is clear that the model approach in this paper improves the recognition accuracy by 11.56%, 2.42% and 0.86% compared to HigherHRnet-VGG16, VGG16-CNN and PARNet models, respectively. The literature [19] used a CNN model based on the VGG16 backbone network, and in order to solve the problem of inefficient CNN caching of feature information per layer, the CAE module was introduced to enhance the information interactivity between different layers so that the extracted feature information is richer, reducing the feature redundancy. The average accuracy on the KTH dataset is 92.54%, and the CNN network structure is only 5 layers, while the deeper layers of our model can better extract the temporal information of continuous video sequences. The literature [20] used a pose-appearance relational network (PARNet) to identify 14 skeletal key points of the human body and a temporal attention mechanism-based LSTM model (TA-LSTM) for action recognition to capture long-term contextual information in action videos to improve the robustness of the network. And the Spatial Appearance (SA) module was used to improve the aggregation between adjacent frames, with an accuracy of 94.10% tested on the dataset. This author divided the skeleton key points into five body parts (head, left and right arms, left and right legs) and then aggregated them to get the human posture feature information. Compared to the spatial type division strategy in this paper, it's easy to ignore the action feature information carried by the local skeleton key points. The literature [21] used Higher HRnet to extract skeletal key points, and the redundant background information was filtered out through the fusion of spatio-temporal information of key points to reduce the dimensionality and retain the action trajectory information, which was finally identified by Resnet101 network classification. Although our method is slightly less accurate compared to the HigherHRnet-ResNet101 method, it has a greater advantage in terms of the number of model parameters and the complexity of the network structure.

Table 3. Identification results of different methods on the KTH dataset.

Model Methodology	Parameters/M	Accuracy/%
HigherHRnet-VGG16	30.00	83.40
VGG16-CNN [19]	—	92.54
PARNet [20]	—	94.10
HigherHRnet-ResNet101 [21]	29.30	95.10
Ours	16.55	94.96

4.3.2. Testing on the Le2i Fall Detection Public Data Set

Database Description

In the proposed work, we considered the Le2i fall detection dataset [22]. This dataset contains 191 videos in 4 different contexts, which comprised a total of 382 videos after we expanded the dataset with horizontal flipping. Among the data, there are 252 videos of falling behavior and 130 videos of normal behavior. The latter is a video of activities of daily living, containing movements such as walking, sitting, standing up, and squatting. Based on a fixed camera, the photos were taken by the actors in four different locations (coffee room as in Figure 14, home as in Figure 15, lecture room as in Figure 16, and office as in Figure 17). The actors wore a variety of clothing and attempted to simulate various kinds of normal daily activities and falls to increase the diversity of the dataset. Furthermore, this dataset presents occlusions, shadows, and variations in illumination. Eighty percent of the videos from the dataset were randomly selected as training videos, and the remaining twenty percent were used as test videos. The categories of the dataset and the number of videos for training and testing are shown in Table 4.

(**a**) (**b**) (**c**)

Figure 14. Example of Le2i fall detection dataset in a coffee room scenario: (**a**) standing: (**b**) falling: (**c**) having fallen.

(**a**) (**b**) (**c**)

Figure 15. Example of Le2i fall detection dataset (with masking) in a dark scene at home: (**a**) walking; (**b**) falling; (**c**) having fallen.

(**a**) (**b**) (**c**)

Figure 16. Example of Le2i fall detection dataset (with masking) in a lecture room scenario: (**a**) squatting; (**b**) walking; (**c**) standing.

(**a**) (**b**) (**c**)

Figure 17. Example of Le2i fall detection dataset (with masking) in an office scenario: (**a**) walking; (**b**) sitting and bending down; (**c**) standing up.

Table 4. Quantitative description of the dataset used in the experiment.

Behavior Categories	Number of Training Set Videos	Number of Test Set Videos	Total
Falling behavior	207	45	252
Normal behavior	98	32	130
Total	305	77	382

Dataset Evaluation Experiment

The test evaluation was based on the top 1 accuracy, sensitivity, and specificity, which are the most commonly used performance indicators. As shown in Equations (11) and (12), the performance measures were derived using the concepts of true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*). In this study, the videos containing falling behavior were identified as positive samples, and the videos with other normal behavioral activities were identified as negative samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

In this paper, four different scenes in the Le2i fall detection dataset are tested separately for training, and the best trained model in each scenario has an accuracy of 100%. To better test the generalization ability of the evaluated models, the trained models in each scenario are tested separately for the validation set of the overall Le2i dataset (containing different scenes). The results of the test accuracy, sensitivity, and specificity experiments are shown in Figure 18. The test accuracies of the model trained in coffee room, home, lec-ture room, and office scenes were 97.40%, 96.10%, 97.40%, and 94.81%, respectively, and the average accuracy of the test results for the four scenes was calculated to be 96.43%, the

average sensitivity was 95.65%, and the average specificity was 97.75%, proving that our model still has high detection accuracy under different scenes, different viewpoints, and different fall poses.

The sensitivity of the test in the lecture room is 100%, and the sensitivity in the home and office scenes is lower at 93.30%. The videos in these two scenes have many overhead shots and fall on the ground with their backs to the camera, and the key points in some frames should not be extracted completely, which in turn affects the spatio-temporal graph convolutional network model for action recognition. The test specificity of the mod-el trained in the coffee room and home scenes was 100%. The specificity rates of the mod-els trained in lecture room and office scenes are 97% and 94%, respectively, which means that normal behaviors are incorrectly predicted as fall behaviors. The two scenes in the dataset contain many behaviors with high similarity to falls, such as squatting and tying shoelaces, sitting, and bending to pick up things, which may lead to misjudgment of the model, but the sensitivity of the lecture room is as high as 100%, indicating that abnormal behaviors such as falls are correctly predicted.

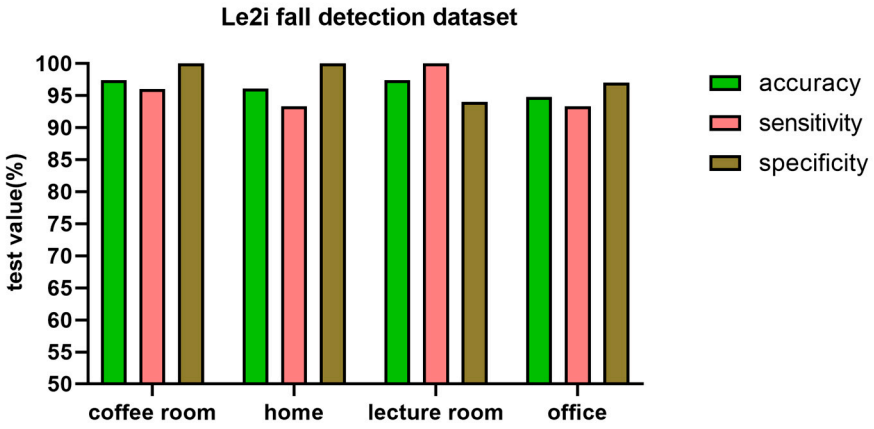


Figure 18. Test model accuracy, sensitivity, and specificity in different scenes on the dataset.

We conducted test experiments on the Le2i fall detection dataset after overall training, and the accuracy was 98.70%, sensitivity was 97.78%, specificity was 100%, and the average error was as low as 0.02. On the other hand, a comparison with similar works is given in Table 5. The results of this paper's model are compared with those of other authors who tested their proposed method on the dataset.

The literature [23] used a lightweight human pose estimation model to extract 14 skeletal key points of the human body using only 30 consecutive frames of skeletal sequence input and a simplified ST-GCN network using only 6 layers of feature extraction modules, including the ATT (attention) module, which improved the detection speed of the model but affected the recognition accuracy. The attention mechanism added by our method is more novel and efficient than the attention module of Lightweight ST-GCN, with a 2.6% increase in test accuracy, a 5.28% increase in sensitivity, and a 4.3% increase in specificity. The literature [24] detects human skeleton information used V2V-PoseNet and then used a dynamic time warping (DTW) algorithm to calculate the variability of action execution speed between adjacent sequences and thus classify whether they have fallen or not. The method is relatively simple, but it is easy to misjudge movements with large speed differences between adjacent frames, such as squatting and bending, which leads to a low accuracy rate. Our spatio-temporal convolutional neural network not only considers the confidence level and the spatio-temporal location coordinates for key skeleton points, but it also collects these action features effectively after nine ST-GCN convolutional layers and the NAM attention mechanism module. Despite the relatively low sensitivity, the test specificity of our method improved by 13.0% and the accuracy by 5.03%. The literature [25] used an exposure correction technique that pre-treats all the under-illuminated videos in the Le2i dataset with dual illumination estimation (DUAL) before detecting them using

YOLOv7 and then uses deep SORT tracking to determine human behavior. Although this can improve the detection accuracy, the practical applicability is poor, especially on a site with such a complex background affected by the weather. Light time is difficult to put into application. Compared to our model that directly handles continuous multi-frame video input, it still has high accuracy and sensitivity for difficult videos that are obscured by tables and chairs, as well as poor lighting, and has some applicability. Our method improves accuracy by 4.2% and specificity by 3.0% compared to YOLOv7-DeepSORT.

Table 5. Testing performance of different state-of-the-art methods on the Le2i dataset.

Model Methodology	Accuracy/%	Sensitivity/%	Specificity/%
Lightweight ST-GCN [23]	96.10	92.50	95.70
V2V-PoseNet [24]	93.67	100.00	87.00
YOLOv7-DeepSORT [25]	94.50	98.60	97.00
Ours	98.70	97.78	100.00

5. Conclusions

An improved ST-GCN model combined with the Alphapose posture estimation algorithm was proposed to intelligently monitor and identify the abnormal behavior of workers' health and safety conditions, which is in line with the advanced concept of smart sites and helps to reduce major safety accidents. The main contributions and experimental findings of this research paper are as follows:

(1) Most of the existing studies on the identification of unsafe behaviors at construction sites are directed at the unsafe behaviors of workers not wearing helmets, safety undershirts, or other protective equipment; meanwhile we have paid more attention to the abnormal health behaviors of workers falling and lying down. Identifying the abnormal behaviors of workers is helpful for getting workers medical assistance in times when they are in danger and guarantees safe construction.

(2) The top-down, high-precision Alphapose pose estimation algorithm model is used to detect key points of the human skeleton on image sequences. Among them, we use the YOLOv5x model for human target detection, which improved the problems of human miss detection and the weak anti-interference ability of the model in complex environments. We also validated the self-built dataset, the KTH dataset, and the Le2i fall detection dataset, which all achieved high accuracy in behavior recognition.

(3) This paper proposes the NAM-STGCN model, which is built on the spatio-temporal graph convolutional neural network and a novel fusion of normalization-based attention modules. This combination increases the model identification accuracy and more efficiently extracts action information in the spatiotemporal dimension without adding to the computational load; it also adjusts the appropriate learning rate for the phenomenon where deep neural networks are prone to overfitting when loaded with many parameters, and it addresses the problem of high model complexity in the process of behavior classification. We also improved the NAM-STGCN model by replacing the ReLU activation function with the PReLU activation function. After optimization, our training network was tested on the self-built dataset with an accuracy of 96.72%, which is a 4.92% improvement relative to the original model and a 3.28% improvement relative to the NAM-STGCN model; the model loss value converges to below 0.2.

(4) To better validate the effectiveness of the model proposed in this paper, it is compared with other advanced methods proposed in the literature on the publicly available KTH dataset and Le2i dataset. The results show that our model has an accuracy of 94.96% tested on the KTH dataset, with a small number of model parameters, and still has a high recognition accuracy for walking movements with a high similarity to jogging; the test was conducted on the Le2i dataset and achieved 98.7% accuracy, 97.78% sensi-

tivity, and 100% specificity despite the inclusion of difficult scenes such as occlusion and dim light. It shows that the improved model displays better performance in abnormal behavior recognition than other models.

Although the Alphapose pose estimation algorithm used in this paper has strong robustness to complex background environments, it is not easy to extract the complete key points of the human body after a fast fall when the human body falls with its back to the camera, and the algorithm performance of the pose estimation algorithm is considered to be further improved in the future. Some application scenarios of behavior recognition require high accuracy, in which case using multimodal data is a better choice to guarantee performance. At present, the research on behavior recognition on three modal data sets—video, depth image sequence, and skeleton sequence—is relatively independent. In the future, with the upgrade of the arithmetic power of hardware devices, it will be more feasible to consider the fusion of data from these modalities and then realize high-precision behavior recognition.

Author Contributions: Conceptualization, Z.L. and J.Z.; methodology, A.Z. and F.H.; software A.Z.; validation, A.Z. and Z.L.; formal analysis, F.H. and A.Z.; investigation, A.Z. and J.Z.; data management, Z.L., A.Z. and Y.W.; writing—original draft preparation, Z.L. and A.Z.; writing—review and editing, A.Z. and F.H.; visualization, A.Z.; supervision, Z.L., F.H. and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weijin, J.; Yongxia, S.; Haoran, Z.; Pingping, C.; Wanqing, Z.; Junpeng, C. Surveillance video behavior recognition mechanism based on ST-GCN under edge-cloud collaborative computing. *J. Nanjing Univ. (Nat. Sci.)* **2022**, *58*, 163–174.
2. Liu, Y.; Zhang, S.; Li, Z.; Zhang, Y. Abnormal behavior recognition based on key points of human skeleton. *IFAC-Pap.* **2020**, *53*, 441–445. [CrossRef]
3. Cai, Q. Human behavior recognition algorithm based on hog feature and SVM classifier. In Proceedings of the 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 18–20 October 2019; pp. 233–236.
4. Ahmadinia, M.; Alinejad-Rokny, H.; Ahangarikiasari, H. Data aggregation in wireless sensor networks based on environmental similarity: A learning automata approach. *J. Netw.* **2014**, *9*, 2567. [CrossRef]
5. Wang, X.; Song, H.; Cui, H. Pedestrian abnormal event detection based on multi-feature fusion in traffic video. *Optik* **2018**, *154*, 22–32. [CrossRef]
6. Xie, H.; Shin, H. Two-stream small-scale pedestrian detection network with feature aggregation for drone-view videos. *Multidimens. Syst. Signal Process.* **2021**, *32*, 897–913. [CrossRef]
7. Hanxu, S.; Yue, L.; Hao, C.; Qiongyang, L.; Xiaonan, Y.; Yongquan, W.; Jun, G. Research on human action recognition based on improved pooling algorithm. In Proceedings of the 2020 Chinese Control And Decision Conference (CCDC), Hefei, China, 22–24 August 2020; pp. 3306–3310.
8. Li, M.; Zhuo, W.; Xinguan, D.; Ronghao, J. Lightweight unsafe behavior recognition model of construction workers based on two-stream CNN and Bi-LSTM. *J. Xi'an Univ. Sci. Technol.* **2022**, *42*, 809–817.
9. Zhang, Z.; Cui, P.; Zhu, W. Deep learning on graphs: A survey. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 249–270. [CrossRef]
10. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–3 February 2018.
11. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3316–3333. [CrossRef]
12. Shengru, Z.; Zhigang, C.; Yiqin, D. A tennis action recognition and evaluation method based on Pose C3D. *Comput. Eng. Sci.* **2023**, *45*, 95.
13. Chandrakar, R.; Raja, R.; Miri, R.; Sinha, U.; Kushwaha, A.K.S.; Raja, H. Enhanced the moving object detection and object tracking for traffic surveillance using RBF-FDLNN and CBF algorithm. *Expert Syst. Appl.* **2022**, *191*, 116306. [CrossRef]
14. Shaojie, W.; Yongxia, Z. Human Body Fall Detection Model Combining Alphapose and LSTM. *J. Chin. Comput. Syst.* **2019**, *40*, 1886–1890.
15. Huayong, K.; Zhiyong, N.; Lilin, S.; Jinlu, Z. Personnel dress detection method with human keypoints and attention mechanism. *J. Chongqing Univ. Technol. (Nat. Sci.)* **2023**, *37*, 206–214.

16. Shi, X.; Huang, J.; Huang, B. An underground abnormal behavior recognition method based on an optimized alphapose-st-gcn. *J. Circuits Syst. Comput.* **2022**, *31*, 2250214. [CrossRef]
17. Liu, Y.J.S. Application of ST-GCN in unsafe action identification of construction workers. *China Saf. Sci. J.* **2022**, *32*, 30.
18. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based attention module. *arXiv* **2021**, arXiv:2111.12419.
19. Cai, X.; Su, W.; Han, G. Human action recognition based on multi-level feature fusion. In Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020; pp. 1393–1397.
20. Cui, M.; Wang, W.; Zhang, K.; Sun, Z.; Wang, L. Pose-Appearance Relational Modeling for Video Action Recognition. *IEEE Trans. Image Process.* **2022**, *32*, 295–308. [CrossRef]
21. Bowen, L.I.; Qing, P.; Nili, T.; Qiongqiong, W. Human action recognition method based on joint point space time information fusion and dimension reduction. *Microelectron. Comput.* **2022**, *39*, 26–30.
22. Charfi, I.; Miteran, J.; Dubois, J.; Atri, M.; Tourki, R. Optimized spatio-temporal descriptors for real-time fall detection: Comparison of support vector machine and Adaboost-based classification. *J. Electron. Imaging* **2013**, *22*, 041106. [CrossRef]
23. Weiting, H.; Bi, Z.; Wenxuan, C. Real-Time Fall Detection Based on Light-weight Human Pose Estimation and Graph Convolution Network. *Comput. Sci. Appl.* **2021**, *11*, 783.
24. Youssfi Alaoui, A.; Tabii, Y.; Oulad Haj Thami, R.; Daoudi, M.; Berretti, S.; Pala, P. Fall detection of elderly people using the manifold of positive semidefinite matrices. *J. Imaging* **2021**, *7*, 109. [CrossRef]
25. Zi, X.; Chaturvedi, K.; Braytee, A.; Li, J.; Prasad, M. Detecting Human Falls in Poor Lighting: Object Detection and Tracking Approach for Indoor Safety. *Electronics* **2023**, *12*, 1259. [CrossRef]