sustainability

Article

Unsafe Mining Behavior Identification Method Based on an Improved ST-GCN

Xiangang Cao [1,2,*], Chiyu Zhang [1,2], Peng Wang [1,2], Hengyang Wei [1,2], Shikai Huang [1,2] and Hu Li [1,2]

[1] School of Mechanical Engineering, Xi'an University of Science and Technology, Xi'an 710054, China
[2] Shaanxi Provincial Key Laboratory of Intelligent Testing of Mine Mechanical and Electrical Equipment, Xi'an 710054, China
* Correspondence: caoxg@xust.edu.cn

Abstract: Aiming to solve the problems of large environmental interference and complex types of personnel behavior that are difficult to identify in the current identification of unsafe behavior in mining areas, an improved spatial temporal graph convolutional network (ST-GCN) for miners' unsafe behavior identification network in a transportation roadway (NP-AGCN) was proposed. First, the skeleton spatial-temporal map constructed using multi-frame human key points was used for behavior recognition to reduce the interference caused by the complex environment of the coal mine. Second, aiming to solve the problem that the original graph structure cannot learn the association relationship between the non-naturally connected nodes, which leads to the low recognition rate of climbing belts, fighting and other behaviors, the graph structure was reconstructed and the original partitioning strategy was changed to improve the recognition ability of the model for multi-joint interaction behaviors. Finally, in order to alleviate the problem that the graph convolution network has difficulty learning global information due to the small receptive field, multiple self-attention mechanisms were introduced into the graph convolution to improve the recognition ability of the model for unsafe behaviors. In order to verify the detection ability of the model regarding identifying unsafe behaviors of personnel in a coal mine belt area, our model was tested on the public datasets NTU-RGB + D and the self-built datasets of unsafe behaviors in a coal mine belt area. The recognition accuracies of the proposed model in the above datasets were 94.7% and 94.1%, respectively, which were 6.4% and 7.4% higher than the original model, which verified that the proposed model had excellent recognition accuracies.

Keywords: miners; unsafe behavior; skeleton spatial-temporal map; spatial-temporal graph convolution; self-attention

check for updates

Citation: Cao, X.; Zhang, C.; Wang, P.; Wei, H.; Huang, S.; Li, H. Unsafe Mining Behavior Identification Method Based on an Improved ST-GCN. *Sustainability* **2023**, *15*, 1041. https://doi.org/10.3390/su15021041

Academic Editors: Ke Xing and Bin Huang

Received: 28 November 2022
Revised: 18 December 2022
Accepted: 27 December 2022
Published: 6 January 2023

1. Introduction

In recent years, China has gradually attached importance to coal mine safety in production, but accidents still occur frequently. The death toll caused by coal mine accidents exceeds that of all other accidents combined [1]. By analyzing the causes of coal mine accidents, it was found that man-made unsafe behaviors accounted for more than 85% of accidents [2]. At present, the management of miners in coal mines mainly involves manually supervising the real-time behaviors of coal miners through surveillance video [3]. This method makes it difficult to have a timely response to emergencies, and a large number of cameras are unattended, resulting in a waste of resources. How to identify the unsafe behavior of miners under the influence of fuzzy monitoring imaging, uneven illumination and complex human behavior is an urgent problem to be solved.

Early image-based human behavior recognition is mainly achieved using feature extraction carried out on a single image. Researchers hope to use object detection and other methods for behavior recognition [4,5], but this ignores the correlation between successive actions, and thus, it is difficult to accurately describe the complicated movement and the recognition accuracy is generally not high. Subsequently, more and more people proposed video-based behavior recognition methods that are mainly based on a two-stream convolutional neural network [6–9] and a long short-term memory network [10–12]. The method based on a two-stream convolutional neural network uses the spatial information and multi-frame optical flow information extracted from the video that is fed into the convolutional neural network as the input of spatial and temporal flow to realize behavior recognition. The method based on a long short-term memory network uses the feature of a recurrent neural network that can save the previous time series information to learn the feature information of the current moment and the feature information of the previous moment to realize the behavior recognition of video. When applied to a coal mine, the above methods cause difficulty in behavior recognition due to the influence of dust and illumination in the coal mine environment and the problem of human shielding. However, the use of human joint information for behavior recognition will greatly reduce the interference caused by the environment due to the characteristics of high robustness and insensitivity to light conditions in the general way of acquiring human joint information [13,14]. Therefore, some people apply the method of acquiring human joint information in the harsh environment of a coal mine as the pre-processing of subsequent tasks [15–17].

With the introduction of advanced human pose estimation algorithms, such as OpenPose [18], and the popularity of depth sensors, such as Kinect [19], it is increasingly easier for us to obtain human joint node data, which also makes some research on behavior recognition change from the traditional method of using RGB image information recognition to the method of joint point data recognition. Compared with RGB images, the advantage of the joint data is that it can better filter out the interference of background obscuration, illumination, imaging conditions and other noises to make the data cleaner. In addition, the joint data only need to store the joint coordinates and confidence of the human body, which greatly reduces the difficulty of data collection and storage. At present, the research on the behavior recognition of joint data is mainly based on manual features and deep learning. In the deep learning method, three different directions are derived through different methods of processing joint node data, which are a convolutional neural network (CNN), a long short-term memory network (LSTM) and a graph convolutional network (GCN). (1) A CNN represents joint node data as pseudo-images and then performs convolution, pooling and other operations on the images to realize behavior recognition. For example, Li et al. [20] proposed a method based on a 3D skeleton data mapped into RGB images through multi-scale neural networks, and Le et al. [21] used a CNN architecture that went from thin to coarse to simultaneously extract the temporal and spatial features of joint nodes to learn the correlation of skeleton information in different periods. (2) An LSTM network can learn long temporal information while retaining the characteristics of a recurrent neural network, which enables it to realize the recognition of long temporal behavior by learning the vector sequence constructed from key point data. For example, Shahroudy et al. [22] proposed a recurrent neural network that could learn the long-term temporal correlation of the features of each node while proposing the NTU RGB + D dataset. In order to solve the problem of structural information loss when transforming the node data to fit the input format of CNN or RNN, Zheng et al. [23] proposed an attention cycle relationship network that modeled the temporal and spatial dynamics and added an adaptive attention module for behavior recognition. (3) The method based on a GCN can transform the data of the joint points by considering the human body joint points and limbs as the vertices and edges of the topological graph, respectively; this method can better retain the feature information of the joint points compared with other methods. The ST-GCN proposed by Yan et al. [24] is the first network that uses graph convolution for behavior recognition. It uses the spatial and temporal information in the graph structure constructed by human joint nodes to realize the recognition of continuous actions. Many subsequent methods based on graph convolution take it as the baseline or improve it. For example, Shi et al. [25] proposed a two-stream adaptive graph convolution network (2S-AGCN), which fused bone point information and bone length and direction information ignored by the ST-GCN through a two-stream framework to improve the model performance. Zhang et al. [26] added more advanced semantic information, including joint type and frame index, to the input information to enhance the feature representation ability. Alsawadi et al. [27] improved the partitioning strategy of an ST-GCN so that the network no longer only aimed at neighboring nodes but learned through multi-layer nodes, which improved the detection performance. Yang et al. [28] introduced time and channel attention into the ST-GCN to enable the network to better learn important nodes. Wu et al. [29] proposed the addition of dense connections and spatial residual layers into the spatial-temporal graph convolution network to improve the efficiency of the model at processing spatial-temporal information. Liu et al. [30] improved the feature extraction ability of the model for the time dimension by adding a residual network to the time information processing of the node. The abovementioned work greatly improved some of the problems that exist in an ST-GCN, but few people have considered the problems that exist in the network identification of miners' behavior. Shi et al. [31] used an ST-GCN for behavior recognition in a coal mine, but without improving the network, there were some problems of unsatisfactory behavior recognition.

However, an ST-GCN is limited by its structure and has a small receptive field, where it can only learn the behavior characteristic information through the learning of adjacent nodes and it is difficult to learn the mutual relationship between different limbs. When recognizing the behavior of coal miners, it is difficult to detect the unsafe behaviors of hands and feet, such as when using climbing equipment. To solve the above problems, a new partition self-attention spatial temporal graph convolutional network (NP-AGCN) was proposed. The main contributions of this study are summarized as follows:

- We proposed a completely new partition strategy that connects some non-naturally connected nodes, divides new partitions and assigns weights to improve the learning ability of the model for association relationships between non-naturally connected joints.
- We introduced a multi-head self-attention mechanism into the graph convolution module of an ST-GCN to increase the receptive field. Therefore, the feature learning ability of the model was improved to improve the recognition ability of the model for all actions.
- A spatial-temporal graph convolution network with self-attention was constructed by using a new partitioning strategy and introducing a multi-head self-attention module, which can effectively solve the problems encountered in coal mine environmental behavior recognition.

2. Methods to Study

2.1. Definition of Unsafe Behavior of Miners in a Transport Roadway

The unsafe behavior of miners refers to the behavior miners exhibit that fails to strictly comply with safety rules and regulations in the production process, which may have a negative impact on organizational and personal safety [32]. For the staff of a transport roadway, these behaviors can be generally divided into two categories: "operator error" and "venturing into dangerous places" [33]. Operator error refers to the behavior of the operator that does not conform to the safety rules and regulations and results in injuring themself and damaging equipment, mainly via the behaviors of damaging equipment, throwing sundry equipment, fighting and running. Venturing into dangerous places is defined as people approaching or contacting large equipment in a coal mine without compliance, including mechanical, hydraulic and electrical equipment, which is mainly divided into climbing equipment and riding belts. The definition of unsafe behavior should also include passive abnormal behavior after a person is injured, such as falling and getting in contact with equipment. The unsafe behaviors of miners in the belt area are shown in Table 1

Table 1. Unsafe behavior of miners.

Category	Behavior
Operator error	Throwing sundry equipment
	Fighting
	Running
Venturing into dangerous places	Climbing equipment
	Hitting equipment
	Riding belts
Passive abnormal behavior	Falling
	Body entering a device

2.2. Constructing a Spatial-Temporal Map of a Human Skeleton

Since the ST-GCN realizes behavior recognition based on the graph structure constructed from the human joint node data, it is necessary to transform the discrete joint node coordinates into topological graphs, that is, to form a spatial skeleton graph by connecting the joint nodes in pairs. The spatial skeleton visualization is shown in Figure 1.

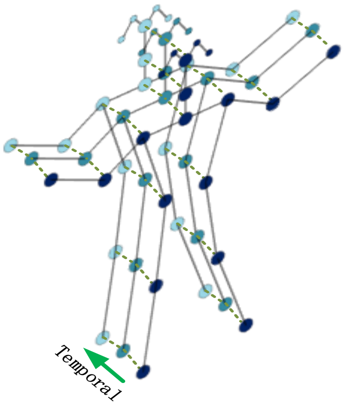


Figure 1. Spatiotemporal diagram of the skeleton sequence. Blue dots indicate body joints. Interarticular connections are defined by natural connections in the human body, with the dotted lines connecting identical joints between successive frames. The green arrow indicates the time dimension. The joint coordinates are used as the input to an ST-GCN.

Then, the adjacency matrix required by the graph convolution was constructed. An undirected graph $G = (V, E)$ was constructed on a skeleton sequence with N nodes and T frames, with characteristics of V for the node (the node) and E (edge) for edge features such that $V = \{V_{ti}|t = 1, \ldots, T, i = 1, \ldots, N\}$ is the set of features of a node, where V_{ti} represent the different node characteristics, T is a different frame node (that is, the time domain) and i is the different body key points in the same frame (nodes). The dimension of V_{ti} is (*x*, *y*, *confidence*), where *x* and *y* are the coordinates of the key point and confidence is the confidence of the key point. The V node can be composed of a single frame graph. The edge set E consists of two parts, namely, E_S and E_F, to construct the inter-frame graph. E_S represents the relationship between different joints in the same frame, denoted as $E_S = \{v_{ti}v_{tj}|(i,j) \in H\}$, where H is the maximum number of human joint connections; E_F represents the relationship between the same joints in different frames, denoted as $E_S = \left\{v_{ti}v_{(t+1)j}\right\}$, where i and j indicate the different frame indexes. The two sets eventually jointly construct the spatial-temporal map of the human skeleton so that the spatial-temporal map convolution can obtain spatial and temporal information simultaneously.

2.3. Spatial-Temporal Graph Convolutional Networks

As a behavior recognition network, an ST-GCN can autonomically learn the spatial and temporal characteristics of data through its graph convolution block and temporal convolution block to realize the recognition of dynamic behaviors. An ST-GCN transforms the temporal and spatial dimensions through two parts: spatial graph convolution (SGCN) and temporal convolution (TCN). Finally, it uses average pooling and a fully connected layer to classify features. According to the constructed human skeleton diagram, the operation of spatial graph convolution can be defined as follows:

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

where $B(v_{ti})$ is the set of sampling centers v_{ti}; for the elements with a path length less than or equal to constant D, the value of D is 1. Z_{ti} is equal to the cardinality of the corresponding subset. f_{in} is the input feature data, namely, the human body joint data. W is the weight function that provides the weight vector. $l_{ti}(v_{tj})$ is the mapping function between the root node of the partitioning policy and the labels of its neighboring nodes. Since the number of weight vectors in W is unchanged, it is used to assign weights to the feature vectors, which are represented as

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

where r_i represents the average coordinates of all joints in a coordinate system, namely, the center of gravity.

After constructing the spatial graph convolution, the spatial graph structure was connected in the time dimension. The time graph structure was constructed by connecting the same points between different frames, and the set of data of different frames of the same point was obtained:

$$B(v_{ti}) = \{v_{qj} | d(v_{tj}, v_{ti}) \leq K, |q - t| \leq \Gamma/2\} \tag{3}$$

where Γ is the time range of adjacent graphs, that is, the incoming frame number information. q and t are the frame index numbers and $d(v_{tj}, v_{ti})$ is the minimum distance between nodes and adjacent frames. Then, the label grouping mapping function was changed and denoted as

$$l_{st}(v_{qj}) = l_{ti}(v_{tj}) + (q - t + \Gamma/2) \times K \tag{4}$$

where $l_{ti}(v_{tj})$ is the label mapping result of the node. Finally, the formula for the convolution of the space-time graph was obtained.

3. Identification Network of Unsafe Behavior of Miners

In order to improve the defects of the graph convolutional network in the identification of miners' unsafe behaviors, we proposed an improved spatial temporal graph convolutional network for miners' unsafe behaviors in a transportation roadway, which was named a new partition self-attention spatial temporal graph convolutional network. In this section, we introduce the overall network and some components of our proposed new partition self-attention spatial temporal graph convolutional network (NP-AGCN) in detail. This includes the construction of graph structure and partition strategy, the overall construction of the network and the improved self-attention module; these parts constitute our proposed NP-AGCN.

3.1. Model Structure

Our model took multi-frame human points as the data input. First, the skeleton graph sequence was constructed with human joints as the graph vertices and bones as the graph edges. The initial feature of the vertices was the 3D coordinates corresponding to the human joints to obtain the graph structure data of dimension $V \times T \times C$, where V is the number of joint points, T is the number of frames of input video data, and C is the information of a single node. Then, using a partitioning strategy, different moving nodes were classified so that the network could learn behavior information from the graph structure. We set the center of gravity of the whole skeleton as the root node, and the nodes with different distances from the root node were put into different partitioning subsets. Different from the traditional ST-GCN partitioning, we connected some non-naturally connected joints and divided them into different subsets by distance. Finally, the graph structure and the obtained learnable adjacency matrix were input into the multi-head self-attention spatial-temporal graph convolutional network.

Each basic unit of the self-attention spatial temporal graph convolution network consisted of two parts, namely, a spatial graph convolution module and a temporal convolution module. The spatial graph convolution module included a spatial graph convolution layer for extracting spatial features, a batch normalization layer and a ReLU activation function layer, and a multi-head self-attention module was added to part of the spatial graph convolution layer. The temporal convolution module is located after the spatial graph convolution module, which contained a temporal convolution layer for extracting temporal features, a batch normalization layer and a ReLU activation function layer. The temporal graph convolution layer was similar to the traditional two-dimensional convolution on images, that is a convolution with the size of 1 × 1 was used to carry out the convolution operation on the feature map to extract features in the temporal dimension. Finally, residual connections were used in each basic unit to make the training more stable. The overall network structure was composed of the nine basic units mentioned above, as shown in Figure 2. The number of channels of input data was three. Before being input into the network, the data was normalized through the BN layer to accelerate the convergence and then input to the subsequent network. The number of output channels of the first three basic units was 64; the number of output channels of the three basic units in the middle was 128. The number of output channels of the last three cells was 256, and the residual mechanism was used in each basic cell. In the fourth and seventh layers of the model, the time convolution step was set to 2, and in order to reduce the loss of graph structure information, multiple self-attention modules are used only in the last six layers. After nine basic units, the output feature map was sent to the pooling layer for global average pooling to get a fixed-size feature vector. At the end of the network was a softmax classifier, which could realize the classification of actions and was used to predict the final result.

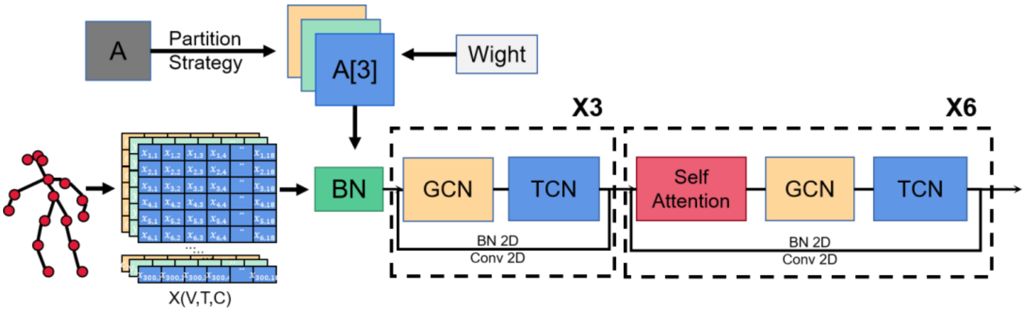


Figure 2. NP-AGCN structure. X(V,T,C) is the input node information, where V is the number of nodes, T is the number of video frames, and C represents the characteristics of different nodes. A is an adjacency matrix, which was divided into a three-dimensional matrix A[3] after a new partitioning strategy and given a learnable weight. After the data were normalized, a GCN was used to extract spatial information and a TCN was used to extract time information. The self-attention module was added to the last six layers of the network to learn the global information about the human skeleton.

3.2. Structure Construction of the In-Frame Interaction Graph

The connection of the ST-GCN was realized by referring to the natural connection of human nodes in reality. In this way, although the spatial variation relationship between various joint nodes in the process of human movement can be extracted well, regarding the movement of multiple limbs cooperating with climbing equipment in an underground coal mine, it is likely to be confused with other movements if only the changes in the limbs themselves are used for the recognition. At the same time, the increase in joint connections can also improve the discrimination ability of the model for similar actions, which can significantly reduce the probability of misjudgment in the identification process of coal mine unsafe behaviors. The connection is shown in Figure 3.

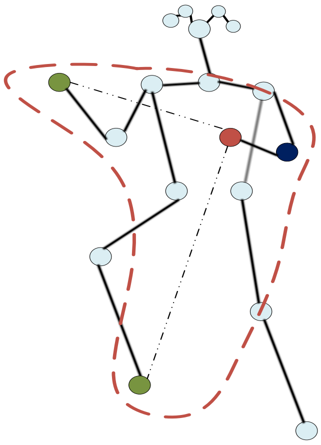


Figure 3. The proposed new partitioning strategy. Different colors are divided into different partitions.

In this study, it was considered that different non-adjacent nodes with potential correlation were connected in the process of graph structure construction so that the ST-GCN could obtain the association information between different joints. However, after testing, the detection performance of the model could not be improved simply by connecting nodes. In the original partitioning strategy, after the aggregation of the node information, the ST-GCN divided nodes into root nodes, centripetal groups and centrifugal groups to simulate the concentric and eccentric movements of actual body parts. However, there was no such relationship between nodes that were not physically connected. As a result, the network aggregated the wrong information. Therefore, the original partitioning strategy was improved, where the nodes that were not physically connected were divided into regions via a distance judgment and re-assigned weights. The partitioning of regions is shown in Figure 3. Finally, the neighbor set was divided into three subsets: (1) the root node itself, (2) adjacent nodes that were closer to the skeleton barycenter than the root node or non-naturally connected nodes that were longer than the root node and (3) adjacent nodes that were farther from the barycenter than the root node or non-naturally connected nodes that were shorter than the neighboring nodes. The center of gravity of the skeleton was obtained by averaging the coordinates of all joints. Finally, the mapping function of the labels of the root node and neighboring nodes was as follows:

$$l_{ti}(v_{tj}) = \begin{cases} 0 \text{ if } r_j = r_i \\ 1 \text{ if } r_j < r_i \ or \ \ r_j > r_k \\ 2 \ \text{ if } r_j > r_i \ or \ \ r_j < \ r_k \end{cases} \tag{5}$$

where r_i represents the average coordinates of all joints in a coordinate system, namely, the center of gravity. r_k is another node connected to the root node.

3.3. Multi-Head Self-Attention Module

In order to solve the problem of "short-sightedness" during feature learning of the spatiotemporal graph convolutional network and reduce the interference of additional connected nodes on some behaviors, multiple self-attention mechanisms were introduced into the graph convolutional network to enlarge the receptive field and improve the network's learning ability for global information in the process of spatial feature extraction. The module structure is shown in Figure 4.

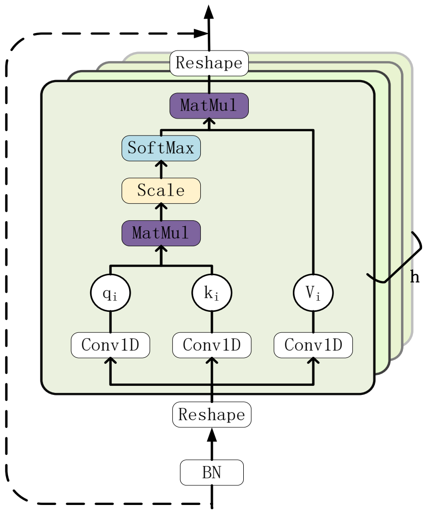


Figure 4. Structure diagram of multi-head self-attention module.

First, the input was normalized by the BN layer. Then, the input data structure was changed to merge the joint data of all frames to obtain the global joint information. Through the learnable parameter matrices $W_q \in \mathbb{R}^{C_{in} \times d_q}$, $W_k \in \mathbb{R}^{C_{in} \times d_k}$ and $W_v \in \mathbb{R}^{C_{in} \times d_v}$, the query vector $q_n^t \in \mathbb{R}^{d_q}$, key vector $k_n^t \in \mathbb{R}^{d_k}$ and value vector $k_n^t \in \mathbb{R}^{d_k}$ of each node j in time t were obtained, where d_q, d_k and d_v are the dimensions of the query vector, key vector and value vector, respectively, and C is the number of input features. For each pair of joint connection points (j_i^t, j_j^t), the score α_{ij}^t was obtained via the dot product of q_i^t and k_j^t transposed, which was divided by $\sqrt{d_k}$ to prevent gradient explosion. Then, each v_j^t was weighted by the obtained score and the weighted sum of all nodes was calculated to obtain the final output z_i^t. The formula can be expressed as follows:

$$z_i^t = \sum_j softmax_j(\frac{\alpha_{ij}^t}{\sqrt{d_k}})v_j^t \tag{6}$$

In order to prevent the self-attention mechanism from excessively focusing its attention on its position in the operation process and producing overfitting, the multi-head attention mechanism was added. Instead of using a single attention mechanism, we transformed queries, keys and values by obtaining h different linear projections from learnable parameter matrices $\mathrm{W_q} \in \mathbb{R}^{C_{in} \times N_h \times d_q^h}$, $\mathrm{W_k} \in \mathbb{R}^{C_{in} \times N_h \times d_k^h}$ and $\mathrm{W_v} \in \mathbb{R}^{C_{in} \times N_h \times d_v^h}$. Then, the transformed queries, keys and values of the h-group were pooled in parallel. The formula can be expressed as follows:

$$head_i = Softmax(\frac{(q_i w_{qi})(k_i w_{ki})}{\sqrt{d_k^{N_h}}})(v_i w_{vi}) \tag{7}$$

Finally, the output of the h attention pool was concatenated and transformed using another parameter matrix W^o that was learned to produce the final output. The formula can be expressed as

$$z_i^t = Concat\left(head_1, \ldots, head_{N_h}\right) W^o \tag{8}$$

At the same time, in order to avoid information loss in the process of self-attention, a residual structure was added to the module to retain the original input information. After the final output of the self-attention module was obtained, the data were separated into different frames of joint data and passed to the temporal convolution module to extract temporal features.

4. Experiment

4.1. Introduction to Experimental Datasets

NTU-RGBD: This is one of the most widely used datasets in action recognition tasks. The dataset was collected using three Kinect V2.0 sensors, which took pictures of the target from different angles at the same height. The NTU-RGBD dataset contained a total of 60 different action categories and a total of 56,880 action sample sequences. Each sample sequence contained no more than two action implementation objects, which were composed of the 3D coordinates of 25 human joints from all frames. The actions mainly included the daily behaviors of a single person or pair. There were two evaluation benchmarks for this dataset: (1) a CS (cross-subject) benchmark, where the datasets in this benchmark were divided into a training set and test set, whose action execution subjects were different; (2) a CV (cross-view) benchmark, which uses the data collected on the second and third devices as the training set and the data collected on the first device as the test set.

Datasets of unsafe behavior of miners in transportation roadway: This is a self-built dataset that was built based on some common unsafe behaviors of miners in a coal mine belt working area. It contains 10 action categories, specifically the eight unsafe behaviors defined above and the normal operation of equipment and carrying items as positive samples, giving a total of 2897 video samples. These data samples were obtained from field video collection and Internet collection. The length of each video sample was about 10 s. If the length was short, it was completed via stitching. If the length was longer, the timeout part was trimmed. When making the datasets, the OpenPose human pose estimation algorithm was used to extract the position and confidence information of 18 joint points of people in a video, and the joint point information of the two people with the highest confidence in the video was selected for saving. The datasets were divided into two parts: a training set (2000 samples) and a testing set (897 samples). Part of the sample is shown in Figure 5.

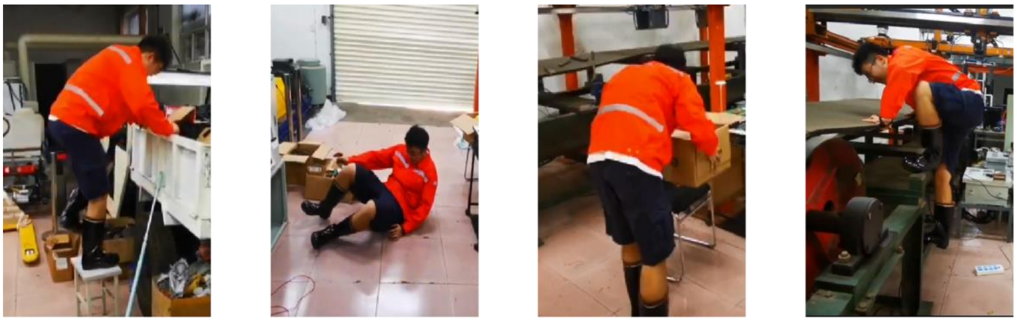


Figure 5. Dataset partial action samples.

4.2. Experimental Parameter Setting

All experiments were carried out on the PyTorch deep learning framework and the hardware platform used two CPUs (Intel Xeon 4214R) and a GPU (NVIDIA A100 40 GB). Stochastic gradient descent (SGD) with Nesterov momentum (0.9) was used as the optimization strategy. The cross-entropy was chosen as the loss function of the backpropagation gradient, and the weight decay was set to 0.0001, which is defined as shown in the equation

$$L_s = -\sum_{i=1}^{m} \ln \frac{e^{y_i}}{\sum_{j=1}^{N} e^{y_j}} \tag{9}$$

where y is the ith output of the full connection and represents the probability of the ith class, m is the input bich_size and N is the number of categories. The performance of the behavior recognition network was evaluated using the recognition accuracy of all targets, which can be calculated using the following formula:

$$Acc. = \frac{\sum_{k=0}^{N} TP_k}{\sum_{k=0}^{N} (TP_k + FN_k)} \tag{10}$$

where TP is the number of correctly identified samples, FN is the number of samples identified as other categories and k is the category index. During training, bich_size was set to 8, the number of heads of the multi-head attention was set to 8, and in all these experiments, the d_q, d_k and d_v dimensions of each layer were 0.25 × Cout. A variable learning rate was used for learning. For the training of the NTU-RGBD dataset, the initial learning rate of the model was set to 0.1 and the number of iterations was set to 80. When the number of iterations reached 30 and 70, the learning rate was adjusted to decay, where the decay rate was 0.1. For the training of the dataset of miners' unsafe behavior, the initial learning rate of the model was set to 0.1 and the number of iterations was set to 120. When the number of iterations reached 20, 40 and 70, the learning rate was attenuated and adjusted, where the decay rate was 0.1.

4.3. Ablation Experiments

In order to verify the promotion effect of different improvements on the original model, comparative experiments were conducted on the NTU-RGBD datasets using different partitioning strategies and adding self-attention modules in different network layers. The specific experimental results are shown in Table 2, where X-Sub and X-View respectively represent the results obtained from different targets or different camera angles for test samples. In order to verify whether the new partitioning policy can improve the performance of the model, the original partitioning policy was replaced by the new partitioning policy in the original ST-GCN and our NP-AGCN, namely, ST-GCN$_{\mathrm{NP}}$ and NP-AGCN$_{\mathrm{NP}}$ in the table. A self-attention module was added after each of layers 1, 3 and 7 to verify the influence of multiple self-attention modules added to different layers in the model.

Table 2. Comparative experimental results of different model structures.

Methods	X-Sub (%)	X-View (%)
ST-GCN	81.5	88.3
ST-GCN$_{\mathrm{NP}}$	82.7	89.1
NP-AGCN$_{\mathrm{NP}}$	87.1	94.7
NP-AGCN$_1$	85.5	92.1
NP-AGCN$_3$	86.7	93.5
NP-AGCN$_7$	86.1	92.9

The experimental results showed that the new partitioning strategy could improve the recognition accuracy of the model when added to the original model and the improved model, and it was shown that the multi-joint connection could improve the recognition ability of the model for some behaviors. The experimental results of adding a self-attention module to different layers showed that no matter which layer the module was added to, it could obtain better recognition accuracy than the original model. At the same time, it was found that adding the self-attention module after the third layer of the network could obtain better results than other layers because adding the self-attention mechanism too early may lose part of the graph structure information and adding the self-attention mechanism only in the last three layers led to the failure of the maximum performance of the module. Therefore, the self-attention module was added to the third layer in the subsequent experiments.

4.4. Ablation Experiments

In this part, in order to show the advantages of our model compared with other models, it was shown that the improvement of the model was necessary for miners' behavior recognition. Therefore, the comparison test of different models was conducted on the NTU-RGBD datasets and the self-built dataset of miners' unsafe behavior, mainly from the complexity and accuracy of the model to compare with ST-GCN and its improved model. First, we compared the proposed NP-AGCN with the original ST-GCN and the improved networks 1S-AGCN and 2S-AGCN using the NTU-RGBD 60 dataset. The reason why it was compared with the ST-GCN was that as a baseline network, it could display the improved effect of the model the most intuitively. The reason for comparison with the 1S-AGCN was that it was more robust than the ST-GCN and it could also reflect the advantages of the self-attention mechanism term over the adaptive graph convolutional network. As an excellent two-stream STGCN model, the 2S-AGCN was compared with it in order to reflect the performance of the model more objectively.

The final experimental results are shown in Table 3, where Params in the table indicates the total number of network parameters, expressed in millions (M), and FPS indicates the number of inference frames of the model. Our improvement increased the detection accuracy of the original network ST-GCN by 5.6% and 6.4% relative to the 1S-AGCN and 2S-AGCN, respectively. Compared with other single-stream networks, the detection accuracy was improved by 1.1% and 1.0%, respectively. Compared with the 2S-AGCN, the overall accuracy was slightly lower than that of this network. However, our model was much more lightweight because it had 11% fewer parameters and a shorter runtime. It was shown that the introduction of multiple self-attention modules could improve the graph convolution and reduce the number of parameters while improving the network performance.

Table 3. Comparison results of different models using the NTU-RGBD dataset.

Methods	Params (M)	FPS	X-Sub (%)	X-View (%)
ST-GCN	3.141	105	81.5	88.3
1S-AGCN	3.47	96	86.0	93.7
NP-AGCN (ours)	3.06	107	87.1	94.7
2S-AGCN	3.54	89	88.5	95.1

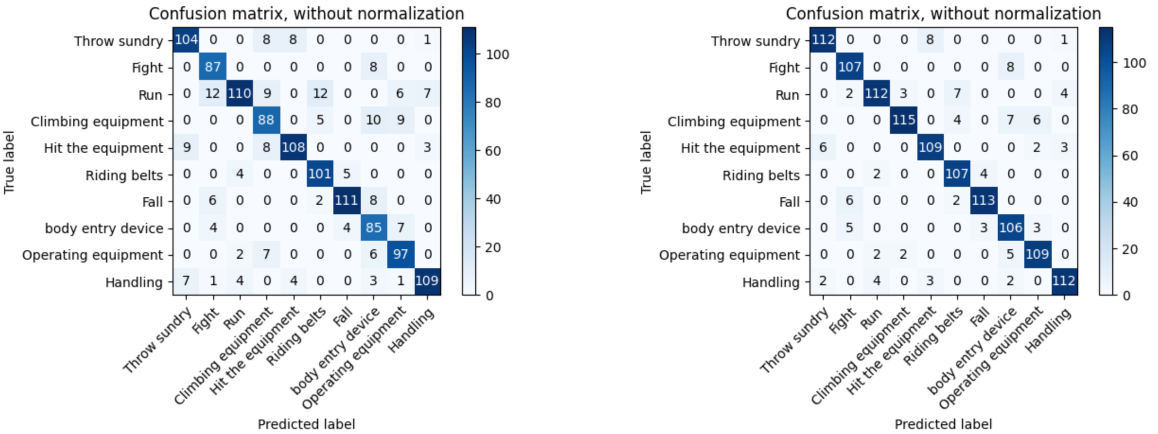
In order to verify the performance of the improved network for the identification of miners' unsafe behavior, different networks were compared on the dataset of miners' unsafe behavior. The final experimental results are shown in Table 4. Compared with the original network, the accuracy of our model on this dataset was improved by 7.4%. Moreover, our comparison model achieved the best result among all the comparison models. It can be seen that the introduction of a multi-head self-attention mechanism could improve the recognition performance of the model for some unsafe behaviors of miners.

Table 4. Comparison results of different models on the dataset of miners' unsafe behavior.

Methods	Acc. (%)
ST-GCN	86.7
1S-AGCN	89.29
2S-AGCN	94.06
NP-AGCN (ours)	94.1

4.5. Validation of Model

The validity of the proposed model was verified on a validation set of 1200 data samples of miners' unsafe behavior, which was independent of the dataset and the test set to ensure the objectivity of the data. Figure 6a,b represent the identification results of the ST-GCN and our proposed NP-AGCN, respectively, which are represented by a confusion matrix. Through the confusion matrix, we can visually see that the blue in Figure 6b is darker than in Figure 6a, which represents the improvement of the recognition accuracy of the improved model compared with the original model and it had a better recognition effect for each category.



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

Figure 6. Confusion matrix of two network recognition results: (**a**) behavior recognition confusion matrix of the ST-GCN and (**b**) behavior recognition confusion matrix of the NP-AGCN.

In order to clearly show the improvement of each category, the histogram in Figure 7 shows the accuracy of the ST-GCN and NP-AGCN for different behaviors in the verification set of miners' unsafe behaviors. The horizontal axis represents all categories in the dataset, while the vertical axis represents the accuracy. It can be seen that the ST-GCN is not ideal for multi-limb motion detection of some movements, such as fighting, climbing equipment and bodies entering a device. After the improvement of the model, the recognition accuracy of these behaviors was greatly improved. The detection rate of equipment involvement is generally lower than other behaviors, mainly because there are too few samples in this part of the dataset construction process. The proposed two improvements to the model could greatly enhance the weight allocation of global information and important nodes in the recognition process of the model. Meanwhile, for some similar behaviors, the accuracy could be effectively improved by learning the features between different nodes. The accuracy of different behaviors increased after the model was improved. In the case of throwing sundry equipment and bodies entering a device, the accuracy rate increased by 6.7% and 18.3%, respectively.

The comparison of actual detection effects is shown in Figure 8. The ST-GCN result is on the left and the result from the NP-AGCN proposed by us is on the right. Each group of images is divided into two parts: the upper part is the original input video frame and the lower part is the confidence graph of the skeleton output by the network. It can be seen from the confidence graph that compared with the ST-GCN, thanks to the introduction of the new partitioning strategy, the NP-AGCN could mobilize more information about the node during the identification process, and thus, it could have a good recognition effect on the climbing actions in Figure 8a,b. As can be seen from Figure 8b, the original confidence degree was no longer focused on a certain joint that was mainly involved in the action but dynamically distributed the attention among multiple nodes that participated in the whole action. This was basically the same as the original intention of introducing a multi-head self-attention module, and also showed that a multi-head self-attention module was well used in the behavior recognition network. It can be seen from Figure 8c,d that both models displayed good recognition performance on important nodes for actions such as falling and carrying because the features of these two categories were more obvious; however, the ST-GCN still had an error in identifying Figure 8d.

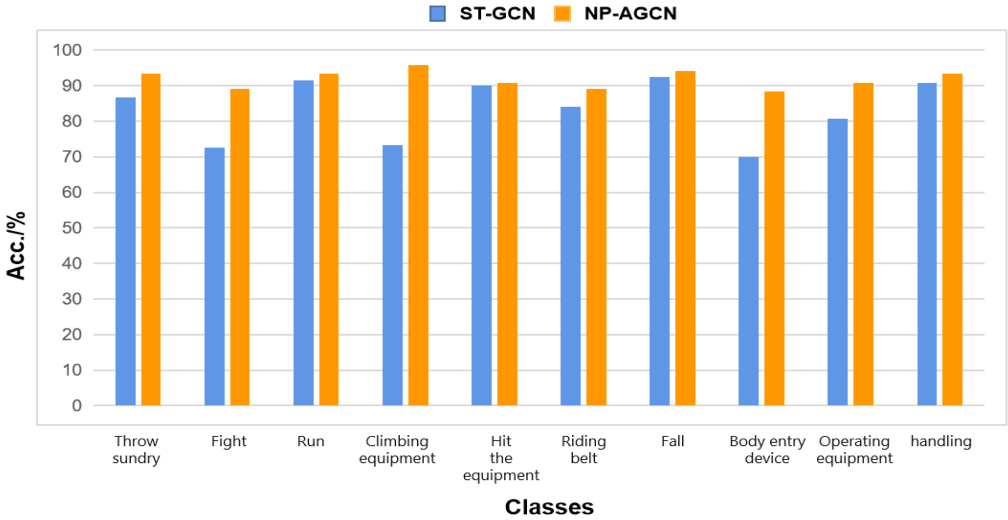
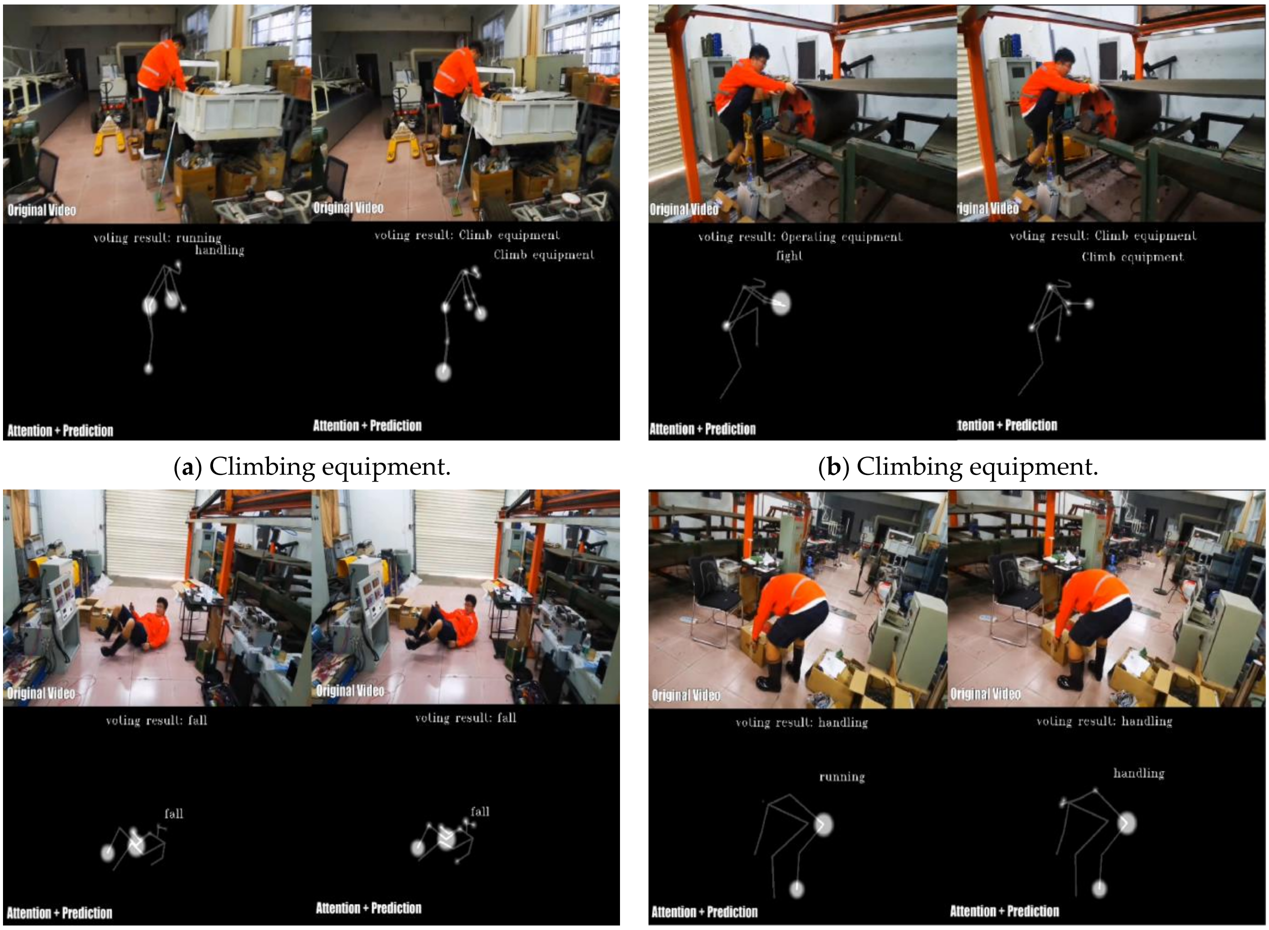


Figure 7. Comparison of the recognition accuracy.



(**a**) Climbing equipment.

(**b**) Climbing equipment.

(**c**) Fall.

(**d**) Handing.

Figure 8. Partial action recognition results.

5. Conclusions

In this study, we proposed a network model for unsafe behavior identification for miners in a coal mine transportation roadway. The model was based on an ST-GCN with two improvements. On the one hand, a new partitioning strategy was proposed to connect some joints that were not physically connected and repartition all joints. In the accuracy experiment of a single category, it was found that the addition of the new zoning strategy could greatly improve the networks' ability to identify the miners' movements of multiple limbs, such as when climbing equipment. On the other hand, the self-attention mechanism was introduced into the graph convolution structure, and its learning ability for global information was utilized to enable the model to combine global nodes when learning local nodes. After visualizing the confidence graph of recognition results, it can be seen that the self-attention mechanism was dynamically distributed among multiple nodes that participated in the whole action, thus improving the detection accuracy. Finally, in order to verify the improvement of the model, our model was compared with the original network and some improved networks using the NTU-RGBD and miners' unsafe behavior datasets. The accuracies of the model on two different benchmarks, namely, 1S-AGCN and 2S-AGCN, using NTU-RGBD were 87.1% and 94.7%, respectively, which were 5.6% and 6.4% higher than that of the original model, respectively. An accuracy of 94.1% was obtained using the dataset of miners' unsafe behavior, which was the best result among all the comparison models. These results demonstrated the performance improvement of the original model and the applicability of our improved model for the identification of the unsafe behavior of miners.

Author Contributions: Conceptualization, X.C.; Methodology, C.Z.; Data curation, C.Z.; Writing—original draft, C.Z.; Writing—review & editing, P.W. and H.W.; Funding acquisition, S.H. and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number: 51975468.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: NTU RGB+D action recognition dataset: https://github.com/shahroudy/NTURGB-D.

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

References

1. Zhang, Y.; Shao, W.; Zhang, M. Analysis 320 coal mine accidents using structural equation modeling with unsafe conditions of the rules and regulations as exogenous variables. *Accid. Anal. Prev.* **2016**, *92*, 189–201. [CrossRef] [PubMed]
2. Yu, M.; Li, J. Psychosocial safety climate and unsafe behavior among miners in China: The mediating role of work stress and job burnout. *Psychol. Health Med.* **2019**, *25*, 793–801. [CrossRef]
3. Di, H.; Sbeih, A.; Shibly, F.H.A. Predicting safety hazards and safety behavior of underground coal mines. *Soft Comput.* **2021**, 1–13. [CrossRef]
4. Wang, H.; Klaser, A.; Schmid, C. Action recognition by dense trajectories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 3169–3176.
5. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, NSW, Australia, 1–8 December 2013; pp. 3551–3558.
6. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576.
7. Wang, L.; Ge, L.; Li, R.; Fang, Y. Three-stream CNNs for action recognition. *Pattern Recognit. Lett.* **2017**, *92*, 33–40. [CrossRef]
8. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recog nition(CVPR), Las Vegas, NV, USA, 27 June 2016–30 June 2016; pp. 1933–1941.
9. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In *Computer Vision: ECCV 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 20–36.

10. Ng, J.Y.H.; Hausknecht, M.; Vijayanarasimhan, S. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
11. Du, W.; Wang, Y.; Qiao, Y. RPAN: An end-to-end recurrent pose-attention network for action recognition in videos. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3725–3734.
12. Long, X.; Gan, C.; Melo, G.; Liu, X.; Li, Y.; Li, F.; Wen, S. Multimodal keyless attention fusion for video classification. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
13. Ding, C.; Liu, K.L. Spatiotemporal weighted posture motion features for human skeleton action recognition research. *Chin. J. Comput.* **2020**, *43*, 29–40.
14. Tölgyessy, M.; Dekan, M.; Chovanec, L'.; Hubinský, P. Evaluation of the Azure Kinect and Its Comparison to Kinect V1 and Kinect V2. *Sensors* **2021**, *21*, 413. [CrossRef] [PubMed]
15. Shang, W.; Cao, X.; Ma, H. Kinect-Based Vision System of Mine Rescue Robot for Low Illuminous Environment. *J. Sens.* **2016**, *2016*, 8252015. [CrossRef]
16. Tran, T.-H.; Le, T.-L.; Hoang, V.-N.; Vu, H. Continuous detection of human fall using multimodal features from Kinect sensors in scalable environment. *Comput. Methods Programs Biomed.* **2017**, *146*, 151–165. [CrossRef] [PubMed]
17. Kim, H.; Choi, Y. Development of a 3D User Interface based on Kinect Sensor and Bend-Sensing Data Glove for Controlling Software in the Mining Industry. *J. Korean Soc. Miner. Energy Resour. Eng.* **2019**, *56*, 44–52. [CrossRef]
18. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
19. Shotton, J.; Sharp, T.; Kipman, A.; FitzGibbon, A.; Finocchio, M.; Blake, A.; Cook, M.; Moore, R. Real-Time Human Pose Recognition in Parts from Single Depth Images. *Commun. ACM* **2013**, *56*, 116–124. [CrossRef]
20. Li, B.; Dai, Y.; Cheng, X.; Chen, H.; Lin, Y.; He, M. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017.
21. Minh, T.L.; Inoue, N.; Shinoda, K. A Fine-to-Coarse Convolutional Neural Network for 3D Human Action Recognition. *arXiv* **2018**, arXiv:1805.11790.
22. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the IEEE Computer Society, Pittsburgh, PA, USA, 27–30 June 2016; pp. 1010–1019.
23. Zheng, W.; Li, L.; Zhang, Z.; Huang, Y.; Wang, L. Relational Network for Skeleton-Based Action Recognition. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019.
24. 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, 2–7 February 2018.
25. Shi, L.; Zhang, Y.F.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12026–12035.
26. Zhang, P.; Lan, C.; Zeng, W. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
27. Alsawadi, M.S.; Rio, M. Skeleton Split Strategies for Spatial Temporal Graph Convolution Networks. *Comput. Mater. Contin.* **2022**, *71*, 4643–4658.
28. Yang, H.; Gu, Y.; Zhu, J.; Hu, K.; Zhang, X. PGCN-TCA: Pseudo Graph Convolutional Network with Temporal and Channel-Wise Attention for Skeleton-Based Action Recognition. *IEEE Access* **2020**, *8*, 10040–10047. [CrossRef]
29. Wu, C.; Wu, X.-J.; Kittler, J. Spatial Residual Layer and Dense Connection Block Enhanced Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019.
30. Liu, S.; Bai, X.; Fang, M.; Li, L. Mixed graph convolution and residual transformation network for skeleton-based action recognition. *Appl. Intell.* **2021**, *52*, 1544–1555. [CrossRef]
31. 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]
32. Yang, L.; Wang, X.; Zhu, J.; Qin, Z. Influencing Factors, Formation Mechanism, and Pre-control Methods of Coal Miners′ Unsafe Behavior: A Systematic Literature Review. *Public Health* **2020**, *10*, 792015. [CrossRef]
33. Yang, L.; Birhane, G.E.; Zhu, J.; Geng, J. Mining Employees Safety and the Application of Information Technology in Coal Mining: Review. Front. *Public Health* **2021**, *9*, 709987. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.