INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Computational Science and Engineering

ISSN online: 1742-7193 - ISSN print: 1742-7185
https://www.inderscience.com/ijcse

Behaviour recognition system of underground drilling operators based on MA-STGCN

Meng Cai, XiChao Wang, Baojiang Li, Haiyan Wang, Xiangqing Dong, Chen Guochu

DOI: 10.1504/IJCSE.2024.10064092

Article History:

Received:	16 August 2023
Last revised:	03 January 2024
Accepted:	02 February 2024
Published online:	21 December 2024

Behaviour recognition system of underground drilling operators based on MA-STGCN

Meng Cai, XiChao Wang*, Baojiang Li,
Haiyan Wang, Xiangqing Dong and Chen Guochu

Shanghai DianJi University,
300 Shuihua Road, Pudong District,
Shanghai 201306, China
Email: 216001010221@st.sdju.edu.cn
Email: wangxc@sdju.edu.cn
Email: libj@sdju.edu.cn
Email: wanghaiyan@sdju.edu.cn
Email: 3083156185@qq.com
Email: chengc@sdju.edu.cn
*Corresponding author

Abstract: In the intricate operations of mining rods, precise behaviour recognition is paramount for operational safety. Addressing target detection and posture feature extraction challenges, this study proposes a method that integrates attention mechanisms with a spatial-temporal graph convolutional network. An efficient channel attention mechanism is introduced during target detection, allocating weights to each channel to adapt to diverse features accurately. Multihead attention modules are incorporated in posture feature extraction, effectively capturing critical behavioural information. Behaviour classification is achieved through the SoftMax function. Experimental results demonstrate the method's accuracy of 95.3% and a recall rate of 91.6% on the custom mining dataset. On the NTU-RGB+D public dataset, the method significantly improves accuracy and recognition speed. This research provides an innovative approach to behaviour recognition in complex environments, ensuring precise identification of various behaviours in real-world scenarios, safeguarding worker safety, and holding crucial implications for applying behaviour recognition technology in industrial fields.

Keywords: drilling operation; attention mechanism; spatial temporal graph convolutional networks; behaviour recognition; pose estimation.

Reference to this paper should be made as follows: Cai, M., Wang, X., Li, B., Wang, H., Dong, X. and Guochu, C. (2025) 'Behaviour recognition system of underground drilling operators based on MA-STGCN', *Int. J. Computational Science and Engineering*, Vol. 28, No. 1, pp.71–86.

Biographical notes: Meng Cai is currently a graduate student at Shanghai Electric Institute. He has filed one invention patent and has approximately two academic papers published. His primary research interests lie in behaviour recognition and machine vision.

XiChao Wang serves as a Lecturer. He graduated from Nanjing University of Aeronautics and Astronautics with a Doctoral degree in Engineering. Since August 2014, he has been teaching at Shanghai Electric Institute, where he has published over ten academic papers, several of which have been indexed in SCI/EI. His main research areas include machine vision and artificial intelligence.

Baojiang Li holds the position of Senior Engineer. He obtained his PhD in Engineering from Nanjing University of Aeronautics and Astronautics. Since May 2019, he has been teaching at Shanghai Electric Institute and serves as the Head of the Intelligent Decision and Control Technology Research Team. He has filed approximately 50 invention patents and published around 20 academic papers. His research focuses on robot control, machine vision, and multi-sensor information fusion.

Haiyan Wang is a PhD graduate from Jiangsu University, teaches at Shanghai Electric Institute, focusing primarily on robot control and artificial intelligence.

Xiangqing Dong is a PhD candidate in Traffic Information Engineering and Control at the School of Civil Aviation, Nanjing University of Aeronautics and Astronautics, China. His main research interests are in vision-based human behaviour recognition.

Chen Guochu is a Professor who currently serves as the Dean of the Electrical Engineering School at Shanghai Electric Institute and the Director of the Electrical Engineering School Committee of Degree Evaluation. He has published over 60 academic papers, with more than 50 indexed in SCI/EI, and holds six authorized invention patents. He is involved in various professional associations, including the Chinese Society of Electrical Engineering, where he serves as a Director. His main research areas include control theory and engineering, as well as the application of intelligent methods.

1 Introduction

Coal, a major global energy source, has historically been associated with significant safety challenges. Remarkably, there has been a considerable decline in coal mining accidents and related fatalities worldwide since the 1970s. As per the International Labour Organisation's 2019 report, the number of fatalities from coal mine accidents was approximately 800, marking a reduction of about 90% in comparison to the mid-1970s.

In recent years, coal mine safety accidents have remained alarmingly frequent, with the number of fatalities surpassing all other accidents combined (Zhang et al., 2016). Analysis of the causes of coal mine accidents reveals that unsafe human behaviours contribute to more than 85% of these incidents (Yu and Li, 2020). Xu et al. (2022) are dedicated to monitoring moving targets underground in coal mines using computer vision. They integrate underground monitoring in coal mines with computer vision processing, employing histogram equalisation processing, dark colour prior dehazing algorithm, and wavelet transform image enhancement methods, respectively (Xu et al., 2022).

However, numerous challenges persist in conducting behaviour recognition in complex mining environments. Firstly, significant variations in lighting conditions pose a substantial problem, ranging from bright light sources to complete darkness. These variations profoundly impact image quality, leading to inaccurate target detection. Secondly, the substantial amount of dust prevalent in mining environments hinders the progress of human body feature extraction. Lastly, operational factors such as equipment, facilities, and ores can wholly or partially obscure workers during operations, complicating the process of behaviour recognition. Moreover, according to the International Labour Organisation, the Asian region is the foremost area for coal mine accidents globally, with China and India recording the highest number of such incidents. However, it is crucial to acknowledge that these nations have undertaken numerous measures to enhance coal miner safety over the past decades. These measures include stricter regulation, improved equipment and technology, and enhanced training (International Labour Organization, 2020).

In analysing a vast array of coal mine safety accident cases, it becomes evident that an increasing number of significant safety incidents are caused by the unlawful behaviour of underground operators, steadily becoming the primary factor in coal mine safety accidents. With the advancement of video surveillance technology and the internet, most coal mining enterprises now employ video surveillance systems to support safety management in the production and scheduling of underground operations. Monitoring violations by underground operators typically involves safety supervisors who use real-time surveillance equipment and analyse historical video data. Two main issues are identified in this approach. The first issue is the limited capacity of safety supervisors, who struggle to effectively supervise each operational scene due to the sheer volume of monitoring images, leading to visual fatigue and a potential oversight of safety hazards posed by operator violations, thus resulting in safety accidents in coal mines; the second issue is the delay inherent in identifying and addressing violations through historical video data, which, while useful for violation management, fails to eliminate safety hazards promptly (Zhang et al., 2019).

Therefore, it is necessary to study efficient, intelligent monitoring method to re-place the inefficient manual monitoring means and be able to implement effective identification and classification of operators' violations to improve the efficiency of safety supervision. In recent years, with the maturity of deep learning methods and the emergence of many fast and efficient target detection, behaviour recognition, and action classification algorithms, deep learning target detection algorithms can meet the needs of practical applications. Behaviour recognition entails analysing, processing, recognising, and understanding human or animal behaviours to acquire insights into their behavioural traits, intentions, and states. This process can leverage various types of sensor data, including images, videos, sounds, and accelerometers. These data sources extract attributes like movement, posture, and motion trajectory of a person or object. These features are essential for various applications, such as behaviour classification, detection, and prediction (Zhang et al., 2018; Wu et al., 2019; Yao et al., 2018).

Belhocine and associates employed the genetic algorithm and particle swarm optimisation in the industrial sector to ascertain the optimal design parameters. Their goal was to improve the fatigue durability of brake discs, as detailed in Belhocine et al. (2021). Kumar and his team have investigated nonlinear dynamical systems comprehensively, employing diverse neural network architectures such as higher-order recurrent neural networks and memory-based Elman recurrent neural networks. Their research pre-dominantly centres on these systems' adaptive control, identification, modelling, and predictive management. This work holds considerable practical and theoretical significance in nonlinear dynamical systems. Furthermore, it provides robust methodologies for human

behaviour recognition by applying neural networks (Belhocine et al., 2021; Kumar, 2023, 2022a, 2022b; Kumar et al., 2017a, 2017b, 2017c, 2018, 2019; Kumar and Srivastava, 2020; Chaturvedi et al., 2023).

The behavioural recognition research presented in this paper holds significant industrial potential, particularly in enhancing real-time safety, productivity, and management efficiency in underground mines. The system's advanced algorithm excels in real-time monitoring, accurately detecting potentially hazardous actions, thus pre-emptively addressing safety risks. Its deep learning-based approach for optimising target detection and gesture feature extraction streamlines workflows, improving productivity. For instance, the system's ability to automatically alert upon detecting risky behaviour demonstrates its proactive safety measures. Furthermore, it aids management in supervising underground activities, reducing manual oversight challenges and costs. Lastly, the system's capability to record staff behaviour data is invaluable for accident investigation and analysis, demonstrating the algorithm's superiority in practical applications.

This study created a dataset capturing human behaviour during mining operations to address the challenge of behaviour recognition in the mine environment. Existing behavioural datasets in this domain must sufficiently represent the conditions beneath the mine. The self-built dataset in this paper bridges this gap, offering a robust experimental foundation for deep learning applications in the field of behaviour recognition in mining environments. Simultaneously, the study introduces algorithm innovations for crucial steps, including target detection, gesture extraction, and behaviour recognition.

The specific innovations outlined in this paper are as follows:

1 Enhancing the YOLOX (Redmon and Farhadi, 2021) target detection algorithm with an efficient channel attention module. This module dynamically adjusts channel weights to adapt to varying characteristics of different areas and targets, accurately extracting human features in various lighting conditions.

2 In the AlphaPose pose recognition model, a lightweight 'SimpleBaseline' structure replaces the traditional Openpose framework. Using a direct convolutional structure for key point localisation, it focuses on extracting key point positions directly from images, capturing human posture information.

3 A multi-head attention module is introduced into the spatial-temporal graph convolutional network (STGCN) to parallel process behavioural data across different dimensions. By focusing on different aspects of behavioural features, this mechanism enhances the model's understanding of complex behaviour patterns, thus improving the dimensionality and depth of data analysis.

These innovations lend a novel aspect to algorithm design and performance enhancement. The integrated approach significantly improves the accuracy and comprehensiveness of behaviour recognition in subterranean mining environments.

2 Related works

This section explores human pose estimation and behaviour recognition, specifically emphasising techniques based on STGCN.

2.1 Human posture estimation technology

Faranak Shamsafar and colleagues developed an approach that combines holistic and partial-based predictions for human pose estimation. They utilised convolutional neural networks, modelling them sequentially as regression and classification tasks within three frameworks: multi-task, tandem, and parallel (Shamsafar and Ebrahimnezhad, 2021). Wandt and Rosenhahn (2019) introduced the rep net framework, employing a generative adversarial network (GAN) for positional pose estimation. This framework uses alternating training of poses through generators and discriminators and camera position estimation via camera network-derived parameters. However, when mapping RGB images directly to 3D key point markers, certain limitations arise, such as multiple 2D projections of different 3D poses corresponding to the same 2D pose, indicating a shortcoming of this monocular image-based pose estimation method (Wandt and Rosenhahn, 2019). Utilising staged processing with the reconstruction method reduces bias in image prediction during data acquisition. However, images representing 2D poses as an intermediate step often include noise. The accuracy of 3D pose estimation networks is highly contingent on 2D pose estimation data quality. Li et al. (2021) proposed the TAG-Net model within an augmented training dataset to address this. This deep architecture, comprising an accurate 2D joint detector and a novel 2D-3D cascade network, employs a cascade residual network for obtaining 3D key point coordinates and a pose optimisation network to mitigate noise impact, facilitating the prediction of 3D human skeletons without ground truth (GT) annotations by utilising a priori knowledge for evolution operator definition. The synthetic skeleton is projected onto a 2D heatmap and formed into 2D-3D pairs for data augmentation in the 2D-3D network. Berlin and John (2022) formulated an efficient deep learning-based Siamese framework for human fall detection, configuring two frameworks: one incorporating a standard 2D convolutional filter and the other a depth-separable convolutional filter. Despite improvements in effectiveness, detection accuracy remains a challenge. Xu et al. (2020) integrated human motion following 2D-3D correspondence and kinematic laws into their depth model, designing it subject to a 2D key point optimisation scheme constrained by perspective projection. This approach refines 2D pose using perspective projection and corrects kinematic structure for noisy 2D inputs to exclude unreliable nodes while completing 3D trajectory reconstruction using more

reliable components. However, the depth of information is only sometimes apparent, potentially reducing 3D estimation accuracy. Xu and Takano (2021) proposed graph stacking hourglass networks that utilise pooling and anti-pooling for data down-sampling and up-sampling, respectively. By continuously repeating the encoding process in the graph hourglass network and fusing multi-level intermediate features, they achieve high-accuracy 2D-3D detection, although the process is somewhat time-consuming.

The paragraph summarises various approaches to human pose estimation using deep learning techniques. It highlights that each method, while innovative, has its own set of limitations. For instance, Shamsafar et al.'s use of convolutional neural networks needs more detailed discussion. At the same time, Wandt and Rosenhahn (2019) rep net with GAN is limited in mapping monomer images to 3D vital points. Similarly, the TAG-Net model by Li et al. (2021) and the fall detection model by Berlin and John (2022) show potential but have areas needing improvement. Xu et al. (2020) integration of 2D-3D correspondence in the depth model and Xu and Takano (2021) graph stacking hourglass net-work demonstrate advancements in pose estimation but face clarity and processing speed challenges, respectively. These methods aim to enhance pose estimation accuracy and efficiency but underscore the need for further refinement.

2.2 Behaviour recognition technology

Behaviour recognition is a dynamic area of research with diverse applications spanning surveillance, healthcare, human-computer interaction, and robotics. Various methodologies have been developed for behaviour recognition, encompassing deep learning-based methods, approaches using hand-crafted features, and hybrid models that integrate both. In recent years, deep learning-based methods have demonstrated impressive results and gained significant traction in behaviour recognition tasks. Prominent among these are deep learning approaches like convolutional neural networks, recurrent neural networks, and their derivatives, including long short-term memory and gated recurrent unit networks (Karpathy et al., 2014; Donahue et al., 2015; Chung et al., 2014).

One of the critical challenges in behaviour recognition is the need for large-scale annotated datasets. Nonetheless, there are some notable datasets commonly used in this research field, such as the UT interaction dataset (Ryoo and Matthies, 2013), the HMDB51 dataset (Kuehne et al., 2011), and the UCF101 dataset (Soomro et al., 2014). These datasets comprise videos of human activities, each annotated with the corresponding behaviour. Deep learning methods have gained prominence in behaviour recognition due to their high accuracy and automatic feature extraction capabilities. Convolutional neural networks are typically employed to derive features from video frames, followed by the use of long short-term memory networks to model the temporal dynamics of these activities. Various CNN-LSTM architectures have been proposed, including two-stream CNNs, 3D CNNs, and attention-based networks (Simonyan and Zisserman, 2014; Tran et al., 2015; Wang et al., 2016). Despite advancements in behaviour recognition, significant challenges persist. A primary issue is the lack of labelled data, which hampers the effectiveness of supervised learning methods. Another challenge is accurately modelling complex behaviours, particularly those with long-term dependencies, such as human activities in natural settings (Wang et al., 2019). The traditional methods for behaviour recognition also need to be improved, including inadequate feature representation. Manual feature extraction methods require extensive design efforts and often fail to capture essential action characteristics, leading to compromised recognition accuracy.

Additionally, the complexity of models presents a challenge. Traditional machine learning models or shallow CNNs struggle with complex action sequences, while deeper models demand more data and computational resources. Furthermore, the lack of generalisability and robustness in traditional methods means they are often confined to specific scenarios and action categories, with poor performance in new environments and unfamiliar actions.

Behaviour recognition methods face key challenges: lack of extensively annotated datasets hinders deep learning effectiveness; traditional approaches struggle with capturing complex behaviours due to limited feature extraction capabilities; deep models require substantial computational power and more versatility across various environments and actions. Our approach integrates attention mechanisms with graph convolutional networks to overcome these drawbacks. This method in our paper shows improvements in feature representation, model complexity, and generalisability.

2.3 Behaviour recognition based on STGCN

Behaviour recognition leveraging STGCN has garnered significant interest in recent years. Several studies have investigated methods to encode spatial-temporal data into graphs and utilise graph convolutional networks for extracting and classifying features. This method conceptualises a video sequence as a spatial-temporal graph, applying GCN for aggregating and classifying node features (Zhang and Zhu, 2019). Liu et al. (2019) developed a behaviour recognition technique that combines inflated cavity convolution with STGCN. Their method employs expanded cavity convolution to extract spatial and temporal features from video sequences, followed by GCNs for feature aggregation and classification. Furthermore, Zhao et al. (2020) proposed a method for behaviour recognition based on STGCN and bipartite graphs. This approach bifurcates the video sequence into spatial-temporal feature maps and frame feature maps, utilising GCNs to aggregate and classify features in both components.

While state-of-the-art performance in action recognition has been achieved using STGCN, which processes human skeletal data, there are still challenges specific to certain environments. STGCN relies on a set of body joints and their 3D coordinates over time, constructing a

spatial-temporal map to model joint relationships. Features are then extracted from this map using a graph convolutional network for action label prediction. For instance, Shi et al. recently introduced an STGAN for action recognition. This method employs graph attention networks to capture long-term dependencies between body joints (Shi et al., 2019). Recent studies have sought to enhance ST-GCN's performance further by integrating it with other techniques like attention mechanisms, data augmentation, and multimodal fusion. Chen et al. (2020) for example, introduced a spatial-temporal attention residual module to bolster the discriminative power of STGCN features. Zhang et al. (2020) proposed a spatial-temporal data augmentation approach to create more varied training samples, thereby increasing model robustness.

Additionally, Huang et al. (2020) developed a multimodal STGCN method that merges skeleton and RGB data for improved results. However, in the underground mining environment, behaviour recognition accuracy can be compromised due to sensor data being impacted by noise. This noise often originates from underground excavation activities and mechanical vibrations. The primary issue in underground mines is also the limited data collection. Challenges stemming from equipment constraints, personnel limitations, and factors difficult to observe lead to inadequate training and testing datasets, subsequently impacting the accuracy of behaviour recognition. Ye et al. (2023) study on human behaviour analysis uses a spatial-temporal dual-stream heterogeneous convolutional neural network, showcasing innovative approaches in behaviour analysis. Wang et al. (2022) and colleagues focus on skeleton-based gait recognition using multi-stream part-fused graph convolutional networks. Qin (2022) applies deep neural network-based recognition in English teaching, demonstrating the technology's versatility. Hu et al. (2022) explore traffic forecasting with spatial-temporal graph convolutional neural networks. Ye et al. (2022) develop a human interactive behaviour recognition method through multi-feature fusion. Finally, Gawande et al. (2022) and the team enhanced pedestrian detection using scale and illumination invariant Mask R-CNN, contributing to technological advancement.

Applying STGCN for behaviour recognition in underground mining encounters several distinct challenges:

1 Noise generated from mining activities adversely affects sensor data quality.

2 The constraints imposed by the mining environment and equipment limitations result in limited data collection, which hinders effective model training and reduces recognition accuracy.

3 There is a pressing need for comprehensive pre-processing and the creation of more diverse datasets to improve recognition performance significantly.

To enhance the STGCN for effective behaviour recognition in underground mining environments, the following improvements have been made: The pre-processing of sensor data has been refined to mitigate data noise and elevate data quality, thereby boosting the accuracy of behaviour recognition. Additionally, the volume of data collection has been increased, encompassing a broader range of behaviour types and a larger number of data instances. This expansion aims to improve the comprehensive-ness of the training and testing datasets, subsequently enhancing the accuracy of behaviour recognition.

3 Method

Figure 1 presents the behaviour recognition system designed for drilling pipe operations in mining environments. The process begins with the input of video data. The enhanced you only look once X (YOLOX) algorithm, augmented with the efficient channel attention mechanism (ECA-Net), detects miners within the challenging mining environment, effectively pinpointing target subjects. Following this, the AlphaPose algorithm is applied to discern human body poses, thereby capturing essential pose information of the miners accurately. This crucial data is then processed by the STGCN, which analyses spatiotemporal features to deduce miners' behaviour patterns. The final step involves using SoftMax functions to assign specific behaviour labels, completing the journey from video input to behaviour recognition. In this study, the YOLOX target detection network enhanced with ECA-NET is termed YOLOX-ECA. Additionally, the STGCN, when integrated with a multi-head attention mechanism, is referred to as MA-STGCN.

Figure 1 Algorithm framework for mining personnel behaviour recognition

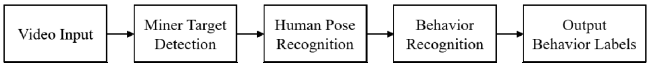
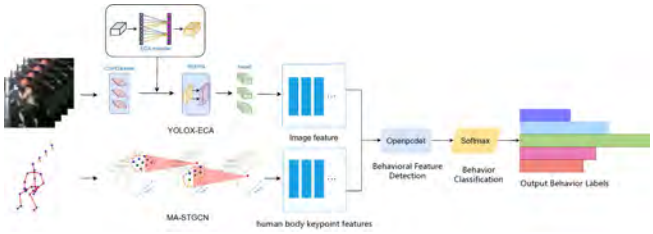


Figure 2 depicts the setup of the behaviour recognition model that incorporates a multi-head attention mechanism. This model consists of two primary algorithms: YOLOX-ECA for object detection and MA-STGCN for behaviour recognition. The YOLOX-ECA algorithm is tasked with generating image features, while the MA-STGCN algorithm concentrates on producing human body key points. Following this, the model utilises a SoftMax function, a normalised exponential function, for classifying and recognising behaviours. The ensuing section provides an in-depth examination of the operational principles underpinning each algorithm within this framework.

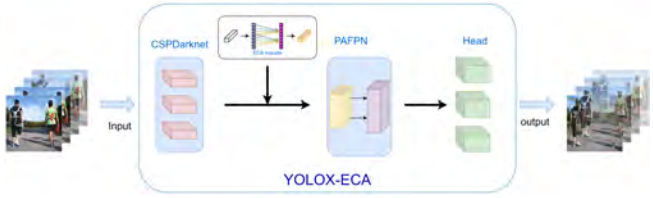
Figure 2 Algorithm framework for mining personnel behaviour recognition (see online version for colours)



3.1 YOLOX-ECA

In this paper, the YOLOX network, augmented with the ECA-Net, is employed for enhanced performance in detecting mining personnel. ECA-Net, an advanced version of the SE-Net (Li et al., 2018), implements the channel attention mechanism. In conventional self-attention mechanisms, dimension reduction often introduces undesirable side effects to the channel attention process, and capturing dependencies across all channels can be inefficient and unnecessary. The ECA attention mechanism module innovatively employs a 1x1 convolution layer right after the global average pooling layer, bypassing the need for a fully connected layer. This design eliminates dimension reduction, enabling effective cross-channel interaction while maintaining efficiency with minimal parameter involvement. ECA-Net utilises one-dimensional convolution for cross-channel information exchange, where a specific function adaptively varies the size of the convolution kernel. This allows layers with more channels to engage in more extensive cross-channel interactions. As illustrated in Figure 3, the enhanced YOLOX network architecture incorporates an attention module within the block layer, aiming to reduce the parameter count and boost computational efficiency.

Figure 3 YOLOX-ECA network structure (see online version for colours)



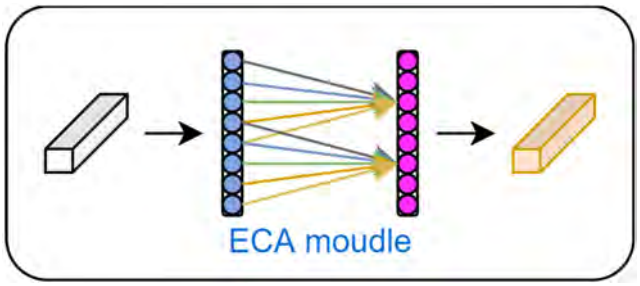
Notes: The backbone feature extraction network used by YOLOX is CSP Darknet. The input is an image frame or a video frame, and the output is a picture with a detection box.

The attention module selected for this study is the ECA module, as depicted in Figure 4. The ECA module considers each channel and its neighbouring channels after global average pooling in the self-attention process. It rapidly computes the channel weights using one-dimensional convolution. Here, k signifies the count of neighbouring channels considered when a channel's weight is computed. The number of k is crucial in influencing the efficiency and effectiveness of the ECA's computational process. The adaptive function employed by ECA-NET is as follows:

$$k = \left| \frac{\log_2(C)}{\gamma} \frac{+b}{\gamma} \right|, \gamma = 2, b = 1 \text{ Wang et al. } (2020\text{b}) \quad (1)$$

ECA-Net begins its process by conducting global average pooling on the input feature map, transforming it from a matrix of dimensions [*h*, *w*, *c*] to a vector of [1, 1, *c*]. Following this, it adaptively calculates the size of the one-dimensional convolution kernel, denoted as kernel_size. This kernel size is then utilised in a one-dimensional convolution operation to derive weights for each channel of the feature map. The concluding step involves normalising these weights and applying them channel-wise to the original input feature map, producing a weighted feature map.

Figure 4 ECA module (see online version for colours)



Algorithm ECA-NET

```
1   class ECA-NET:
2       def __init__(self, channel, gamma=2, b=1):
3           t = “int(abs((log2(channel) + b) / gamma))”
4           k_size = “t if t is odd else t + 1”
5           self.avg_pool = “AdaptiveAvgPool2d(1)”
6           self.conv = “Conv1d(1, 1, kernel_size=k_size,
            padding=(k_size // 2), bi-as=False)”
7             self.sigmoid = “Sigmoid()”
8           def forward(self, x):
9             y = “avg_pool(x)”
10  y =
    “conv(y.squeeze().transpose()).transpose().unsqueeze()”
11  y = “sigmoid(y)”
12  return “x * y.expand_as(x)”
```

The ECA-NET layer adeptly ascertains inter-channel interactions through the adaptive determination of convolution kernel dimensions. During the forward pass, it initially conducts global average pooling on the input feature map, condensing spatial information into channel-specific descriptors. Subsequently, it learns inter-channel dependencies via one-dimensional convolution, and channel-specific weights are computed using the sigmoid activation function. Ultimately, these calculated weights are superimposed on the original input, resulting in a weighted feature map. This map is crucial for augmenting the model's efficacy in object detection tasks.

Figure 5 The numbers corresponding to each bone point in the human body from 0 to 17 are nose, chest, left shoulder, left elbow, left hand, right shoulder, right elbow, right hand, left hip, left knee, left foot, right hip, right knee, right foot, left brow bone, right brow bone, left ear, right ear

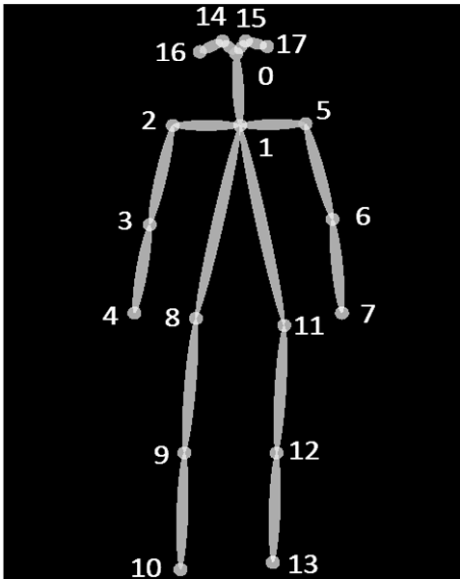
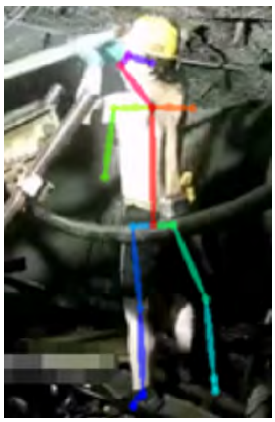


Figure 6 Illustration of extracted underground worker skeletal point data (a) is the behaviour of the top rod and (b) is the behaviour of the operating rod (see online version for colours)



(a)

(b)

3.2 Simple pose

Simple Baseline (Shahroudy et al., 2016) is a proficient pose estimation algorithm that utilises deep convolutional neural networks for keypoint detection and pose estimation. Its merits are manifold: firstly, it possesses a single-stage network structure that is both computationally economical and memory-efficient; secondly, it leverages multi-resolution feature representation, extracting feature maps of various resolutions through multi-scale convolutional operations and cascade fusion, thereby enhancing its capacity for detailed modelling in pose estimation tasks. Additionally, it optimises the key point matching strategy, allowing for multi-scale matching and increasing the accuracy of keypoint detection. When tested, the algorithm proves efficient, requiring a single forward pass to estimate the poses of all individuals in an image. Owing to these advantages, Simple Baseline excels in pose estimation tasks and offers strong scalability for both single – and multi-person pose estimation and instance segmentation tasks. To enhance the pose estimation capabilities of AlphaPose, the OpenPose algorithm has been replaced with Simple Baseline, subsequently termed SimplePose. SimplePose is utilised to detect human joints, where the coordinates (x, y) of the 18 skeletal joints in each frame are obtained. Following this, the angle of each joint in the human body for that particular frame is calculated. Figure 5 shows the 18 key points of the human skeleton.

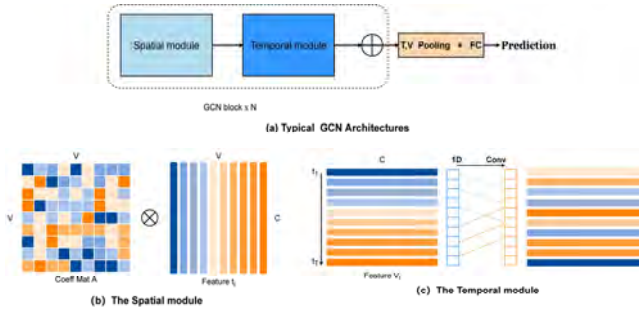
The key points of the human body are extracted by the SimplePose algorithm, which extracts the locations and categories of key points of the human body from the video by two parallel convolutional neural networks, as shown in Figure 6:

3.3 MA-STGCN

This paper's baseline model is STGCN, a deep learning model for action recognition that uses convolutional neural networks to process sequences in time and uses graph convolution to handle spatial relationships. The main idea of STGCN is to convert a video or image sequence into a graph and use graph convolution to capture spatial and temporal relationships. Specifically, STGCN splits the video or image sequence into frames and then converts each into a graph. In the graph, each node represents a pixel or a set of pixels, and edges represent the spatial relationships between the nodes. Then, STGCN uses graph convolution to process these graphs, capturing spatial and temporal relationships. Finally, STGCN uses a fully connected layer to predict the action category. The advantage of STGCN is that it can effectively handle spatial and temporal relationships in video or image sequences and automatically learn features without using manual features. In addition, STGCN is also adaptable to different types of graph structures so that it can be used in various application scenarios.

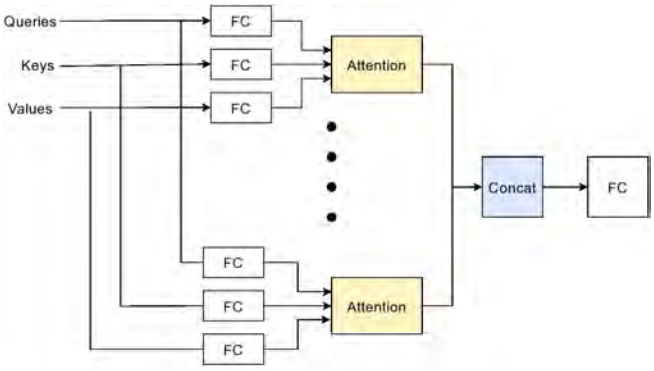
Figure 7 is a typical framework of STGCN for action recognition based on skeleton.

Figure 7 (a) a GCN is composed of n stacked GCN blocks, each of which consists of a spatial module and a temporal module (b) the spatial module performs feature fusion between joints using coefficient matrices (c) the temporal module learns temporal features through one-dimensional temporal convolution (see online version for colours)



This study's enhanced multi-head attention mechanism is a fusion of multiple self-attention structures. Each entry into the attention layer enables learning varied spatial node feature representations. By focusing attention in diverse ways, this process augments the model's fitting capability. In implementing MA-STGCN, with a consistent set of queries, keys, and values, the model is designed to discern different behaviours through the same attention mechanism and subsequently amalgamate these varied behaviour insights. For instance, it captures dependencies within sequences. Rather than relying on a singular attention pool, the approach involves learning h distinct linear projections independently for transforming queries, keys, and values. These h-transformed elements are then processed in parallel for attention pooling. The outputs from these h attention pools are concatenated and passed through another learnable linear projection, culminating in the final output. This architecture is termed multi-head attention, with each h output representing a 'head.' Figure 8 illustrates this multi-head attention, employing a fully connected layer for the learnable linear transformation.

Figure 8 Multiple attention mechanism (see online version for colours)

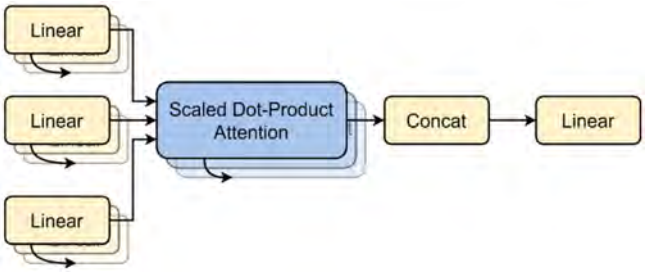


If this multi headed attention model is formally described in mathematical language, it is that given query q $\in R^{dq}$, key $k \in R^{dk}$ and value $v \in R^{dv}$, each attention head $h_i (i = 1 \ldots h)$, the calculation method of each attention head is:

$$\boldsymbol{h}_i = \mathrm{f}\left(\mathbf{W}_i^{(q)}\mathbf{q}, \mathbf{W}_i^{(k)}\mathbf{k}, \mathbf{W}_i^{(v)}\mathbf{v}\right) \in \mathbb{R}^{\mathrm{pv}} \text{ Vaswani et al. } (2017) \quad (2)$$

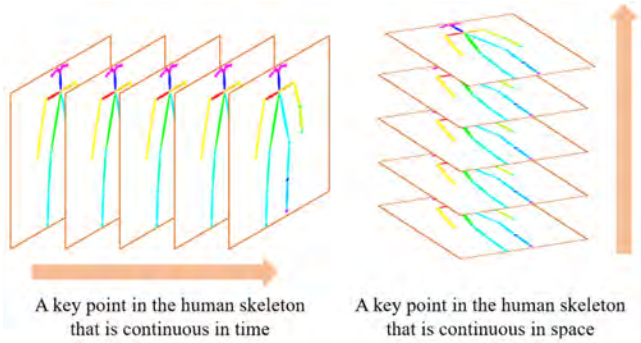
The parameters that can be learned include $\mathbf{W}_{\mathrm{i}}^{(\mathrm{q})} \in \mathbb{R}^{\mathrm{Pq_q} \times \mathrm{d_q}}$, $\mathbf{W}_{\mathrm{i}}^{(\mathrm{k})} \in \mathbb{R}^{\mathrm{pk^k} \times \mathrm{d_k}}$, $\mathbf{W}_{\mathrm{i}}^{(\mathrm{v})} \in \mathbb{R}^{\mathrm{pv} \times \mathrm{d_v}}$. And the function representing attention pooling can be additive attention and scaled 'point sum product' attention. The output of multi head attention needs to go through another linear transformation, which corresponds to the result of h head splicing, so its learnable parameter is $\mathbf{W}_{\mathrm{o}} \in \mathbb{R}^{\mathrm{p_o} \times \mathrm{hpv}}$:$\mathbf{W}_{\mathrm{o}} \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_{\mathrm{h}} \end{bmatrix} \in \mathbb{R}^{\mathrm{p_0}}$. The mathematical model is shown in Figure 9.

Figure 9 Mathematical model of multiple attention mechanism (see online version for colours)



This design allows each head to concentrate on distinct aspects of the input, enabling the representation of functions more complex than mere weighted averages. The concept of masked multi-head attention plays a pivotal role here. Specifically, subsequent elements should not influence the process when decoding an element within a sequence. This exclusion is achieved through the application of a mask. For instance, while computing the x_l output, the procedure simulates the current sequence length as I, ignoring subsequent sequence elements.

Figure 10 The continuous action of the key points of the human skeleton respectively in time and space (see online version for colours)



Behaviour recognition requires MA-STGCN to extract key point features from human skeleton information. The network takes human skeleton key point coordinates extracted by AlphaPose as the model input, and constructs a spatiotemporal graph with joints as graph nodes and natural

connections and time relationships between the same joints as graph edges, integrating information in the time and space domains. MA-STGCN is divided into spatial graph convolution and temporal graph convolution. Space graph convolution is a space graph convolution within a frame constructed based on the natural connectivity of human joint. Spatial graph convolution is the construction of intra-frame spatial graph convolution based on the natural connectivity of human joints, which can be denoted as $G_S = (V_S, E_S)$. Where $V_S = \{v_{ti} \mid i = 1, \ldots, N_S\}$ denotes all the joints in a skeleton, $E_S = v_{ti}v_{tj} \mid (i, j) \in H$ denotes the connections between joints, and each node is characterised by a feature vector $F(V_{ti})$ describing the spatial feature, which is obtained by spatial graph convolution. Temporal graph convolution, on the other hand, connects the same joints in consecutive multi-frame images on the spatial graph to form a spatiotemporal graph of the skeleton sequence, denoted as $G_T = (V_T, E_T)$. $\{v_{ti} \mid t = 1, \ldots, N_t\}$ denotes the sequence of joints in the same part, and $E_T = \{v_{ti}v_{(t+1)i}\}$ denotes the connections between them, as shown in Figure 10.

The dimensions of these key point data are generally (N, C, T, V, M): N represents the number of videos; usually, a batch has 64 video segments; C represents the feature of the joint; usually, a joint contains three features such as x, y, acc, x, y is the position coordinates of the node joint, acc is the confidence, T represents the number of keyframes, generally a video has 150 frames, V represents the number of joints, usually one person annotates 18 joints, M represents the number of people in a frame, and generally the average confidence of the top 2 people is selected.

The human key point features extracted by STGCN are:

$$\boldsymbol{F}_s = \{\boldsymbol{f}_{si}\} = \left\{\sum_{v_j \in Bi} \boldsymbol{w} p(v_j)\right\} \text{ Shi et al. } (2020) \tag{3}$$

The features of the human key points extracted by STGCN are represented by f_{si}, where i is the index of the key point; v_j is the neighbour node of the key point v_i (the centre node), $j = 1, 2, ..., n$, and $j \neq i$; B_i is the set of neighbour nodes of v_i; $\boldsymbol{w}$ is the weight to be learned; $p(v_j)$ is the sampling function of the neighbour node v_j, indicating the range of nodes involved in the convolution. In this paper, it is set to 1, indicating that only the centre node and its connected neighbour nodes are used. After obtaining the features of the human joint nodes, the action classification result can be obtained through the FC fully connected layer and the SoftMax function.

This paper's MA-STGCN model commences using the STGCN layer to extract spatial and temporal elements from the input data. Subsequently, it enhances its capability to process time-series data by incorporating a multi-head attention mechanism. This sequence of operations ultimately converts the processed features into outputs for behaviour recognition via a linear layer. This integrated method leverages STGCN's strength in processing spatiotemporal data and augments the model's ability to discern temporal details, thanks to the multi-head attention mechanism. As a result, this approach significantly improves the accuracy and efficiency of behaviour recognition. Here is the pseudo-code for the MA-STGCN model implemented in Python.

Algorithm MA-STGCN

```
class MultiHeadAttention:
    def __init__(self, query_dim, key_dim, num_units, num_heads):
        self.W_query = “Linear(query_dim, num_units)”
        self.W_key = “Linear(key_dim, num_units)”
        self.W_value = “Linear(key_dim, num_units)”
        self.num_heads, self.key_dim = num_heads, key_dim
    def forward(self, query, key, mask=None):
        querys, keys, values = “ W_query (query)”, “W_key(key)”, “W_value(key)”
        querys, keys, values = “Split and stack for each head”
        scores = “Matmul(querys, keys) / sqrt(key_dim)”
        if mask: scores = “Apply mask”
        return “Matmul(Softmax(scores), values)”, “scores”
class STGCNWithMultiHeadAttention:
    def __init__(self, num_nodes, in_channels, temporal_feature_dim, num_heads):
        self.stgcn_layers = “STGCNLayer(...)”
        self.mha = “MultiHeadAttention (temporal_feature_dim, temporal_feature_dim, num_heads)”
        self.final_layer = “Linear(temporal_feature_dim, num_classes)”
    def forward(self, x, mask=None):
        x = “mha(stgcn_layers(x), x, mask)”
        return “final_layer(x)”
```

4 Experiments

This study's innovation and significance will be thoroughly elucidated, encompassing aspects such as dataset preparation, configuration of the coding environment for experimental purposes, pseudo-code for object detection and behaviour recognition algorithms, performance evaluation metrics for these algorithms, ablation studies, and analysis of results. This paper's experiment includes the hardware devices, code development software, and deep learning framework.

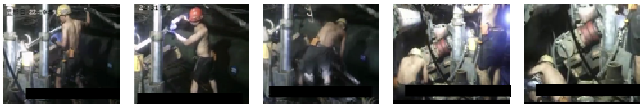
This paper's experiment includes the hardware devices, code development soft-ware, and deep learning framework. The experimental setup utilises a 64-bit Ubuntu 22.04 operating system and an Intel Core i7-12700KF processor, complemented by two NVIDIA 2080Ti GPUs. It operates on the CUDA 11.4 computing platform and employs the PyTorch 1.13.0 framework for training. Regarding the NTU RGB-D and NTU RGB-D 120 datasets, the configured

batch size is 64, and the dimension of each sample is standardised to 64 frames.

4.1 Datasets

- *NTU-RGB+D:* This dataset represents a comprehensive resource for human action recognition, encompassing 56,880 skeleton action sequences. These sequences were enacted by 40 participants, encompassing a diverse range of 60 action categories. Each sequence is characterised by a distinct action involving up to two objects. The dataset's unique feature is its capture methodology, utilising three Microsoft Kinect v2 cameras positioned at varied angles, enabling simultaneous multi-view data acquisition.
- *NTU-RGB+D 120 (Xia et al., 2019):* The NTU-RGB+D 120 dataset is the most expansive dataset with 3D joint annotations dedicated to human action recognition. It extends the original NTU RGB+D dataset by incorporating 57,367 skeleton sequences, spanning 60 additional action categories. 113,945 samples were enacted by 106 individuals across 120 classes, captured using three cameras. The dataset comprises 32 settings, each depicting a unique location and background context.
- *Mining behaviour:* The dataset focusing on drill pipe behaviour in coal mines encompasses 500 videos, segmented into five distinct behavioural categories. This dataset is methodically divided into three sections, aligning with the training requirements of the STGCN model: The training subset comprises 350 videos aimed at model learning and optimisation; the validation subset consists of 75 videos intended for model calibration and overfitting prevention; and the test subset, also containing 75 videos, is specifically purposed for model performance evaluation. Each video spans 3–5 seconds, providing adequate dynamic behaviour data to enhance the model's accuracy and generalisability. A segment showcasing the behaviour of drill pipe personnel in a mining setting is depicted in Figure 11, and the specific division of the dataset is shown in Table 1.

Figure 11 Shows an example video dataset of the five behaviours of the drill crew, such as ascending the rod, moving the rod, unloading the rod, falling, and sitting (see online version for colours)



When dividing datasets for object detection and behaviour recognition in coal mine scenarios, ensuring a balanced distribution of behaviour types in both the training and test sets is imperative. This balanced approach should encompass various behaviour patterns in the dataset, such as ascending the rod, moving the rod, unloading the rod, falling, and Sitting, guaranteeing representative samples in both sets. Such a distribution is critical for effective algorithm training on diverse behaviours and for its accurate identification in practical applications. Furthermore, the unique conditions of coal mine environments necessitate special consideration of data representation under varying lighting and environmental factors. Coal mine settings vary greatly, from dimly lit underground mines to brightly illuminated open-pit mines, and these lighting conditions significantly influence the performance of target detection and behaviour recognition algorithms. Including images and videos from various lighting scenarios in the dataset can enhance the algorithm's adaptability to environmental shifts, thereby improving its stability and accuracy in practical use. Additionally, factors like dust and humidity in coal mines can affect the performance of vision systems. Thus, the chosen dataset should encompass scenarios under diverse environmental conditions to ensure the algorithm's effective operation in real-world coal mine environments.

Table 1 Video dataset segmentation of mining personnel's drill pipe operation behaviour

Behaviour label	*Train*	*Val*	*Test*
Ascending the rod	110	15	15
Moving the rod	120	15	15
Unloading the rod	80	15	15
Falling	100	15	15
Sitting	40	15	15

In the field of worker object detection in coal mines, 15,000 images derived from cropped video data are meticulously segregated into a training set, a validation set, and a test set. This distribution is essential for the model's training and subsequent evaluation. The training set includes approximately 10,500 images geared towards model education and refinement; the validation set comprises 2,250 images, intended for parameter optimisation and overfitting prevention during the training process. Furthermore, the test set comprises around 2,250 images allocated to evaluate the model's efficacy comprehensively.

4.2 YOLOX-ECA performance assessment

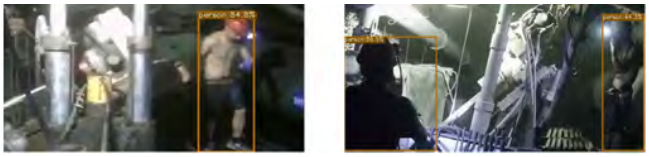
This study assesses the model's performance using the average precision (AP) values from the COCO dataset (Lin et al., 2014). AP values evaluate target detection accuracy across varying intersection ratios and confidence thresholds. For specified intersection ratios and confidence thresholds, detection outcomes are categorised as true positives (TP) and false positives (FP). TP represents the count of accurately detected targets, while FP denotes targets detected inaccurately. These metrics jointly establish the precision rate of target detection under the given conditions, denoted as '*precision iou*' and is computed as demonstrated in equation (3):

$$precison_{conf}^{iou} = \frac{TP}{TP + FP} \tag{4}$$

AP_{iou}^{101} The metric measures the effect of 101 different confidence thresholds on the detection accuracy at the same cross-ratio threshold and is calculated as shown in equation (4):

$$AP_{iou}^{101} = \frac{\sum_{conf\in\{0,0,01,\ldots,1\}} precision_{conf}^{iou}}{101} \tag{5}$$

Figure 12 Real-time image of an underground mining crew detected using YOLOX-ECA, which shows that the target can be accurately detected in a complex mining environment (see online version for colours)



The AP metric is derived from the indicator, which assesses how varying intersection ratio thresholds impact detection accuracy, as detailed in equation (5). Depending on the size of the detected targets, the AP metric is further segmented into APs, APM, and APL. APs calculate the AP for small targets (pixel area less than 332), APM for medium targets (pixel area between 332 and 962), and APL for large targets (pixel area greater than 962). In evaluating the YOLOX-ECA detection, AP metrics at cross-merge ratio thresholds of 0.5 and 0.75, along with other AP values, were utilised to gauge model performance.

Table 2 Comparison of target detection performance in mining scenarios

Evaluation indicators	*Backbone*	*Size*	*AP*	AP_{50}^{101}	AP_{75}^{101}	AP^{S}	AP^{M}	AP^{L}
YOLOv3	Darknet-53	416	39.0	65.6	41.2	16.5	34.2	43.5
YOLOV4	CSP Darknet-53	416	47.5	66.2	50.2	20.3	48.3	58.2
YOLOV5	Modified CSP v5	640	50.4	68.2	51.3	18.3	49.1	61.3
YOLOX	Modified CSP v5	640	50.3	68.5	52.4	20.1	36.2	62.1
YOLOX-ECA(ours)	Modified CSP v5	640	55.1	71.2	56.1	21.5	38.2	65.5

To evaluate the detection performance of the improved YOLOX-ECA model, a comparative analysis was conducted against classical target monitoring models, including YOLOv3 (Redmon and Farhadi, 2018), YOLOV4 (Bochkovskiy et al., 2020), YOLOV5 (Wang et al., 2020a), and the original YOLOX. The results, as depicted in Table 2, indicate that while the original YOLOX model achieved an AP value of 50.3%, surpassing other baseline models, the YOLOX-ECA algorithm further elevated this metric to 55.1% in identifying underground workers, marking a 4.8% improvement over the original YOLOX. This enhancement and improvements in AP, AP50, APS, APM, and APL values highlight the superior performance of YOLOX-ECA compared to other classical target detection algorithms.

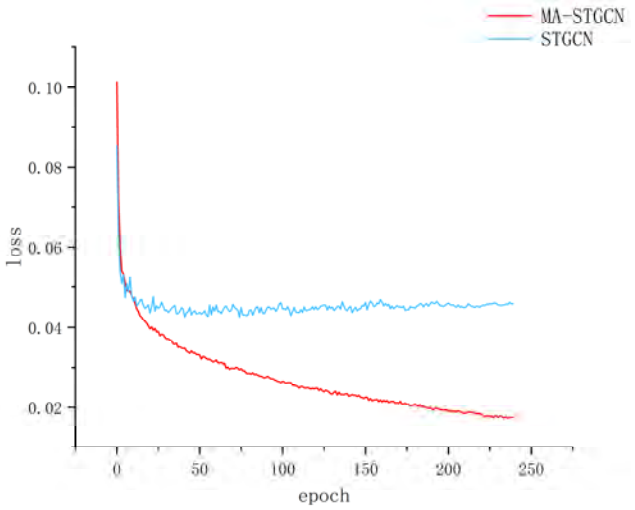
The remarkable efficacy of the YOLOX-ECA model is attributed to integrating the ECA network. This network refines the channel attention mechanism, bypassing the dimensionality reduction seen in SE-Net, and effectively captures inter-channel interactions. These improvements allow the YOLOX-ECA to discern critical features in image data processing more accurately, particularly in complex miner behaviour recognition contexts. Specifically, the ECA network significantly enhances the model's ability to distinguish personnel from the background in detecting drill pipe personnel, thereby improving average accuracy and other key performance indicators.

4.3 MA-STGCN performance assessment

4.3.1 NTU RGB+D dataset experiments

Experiments to evaluate MA-STGCN were carried out on the NTU-RGB D dataset, focusing on behavioural recognition using the MA-STGCN approach. The model's initial learning rate was established at 0.1, with the number of iterations set to 250. Adjustments for learning rate decay were made at the 80th and 100th iteration marks. Figure 13 illustrates the variation curve of the loss rate. To maintain consistency in loss rates during the initial 20 training rounds, both MA-STGCN and ST-GCN employed a cross-entropy loss function.

Figure 13 MA-STGCN and ST-GCN loss variation curves at training time (see online version for colours)



To assess the efficacy of the proposed MA-STGCN model, comparative analyses were conducted against baseline models such as STGCN and other prevalent behaviour recognition models, including deep LSTM (Zhu et al., 2016), TCN (Hou et al., 2018), HCN (Li et al., 2018), AS-GCN (Li et al., 2019a), and ST-GR (Li et al., 2019b). Results, presented in Table 3, utilised cross subject (CS) and cross-view (CV) evaluation protocols. The CS protocol involved training on actions from ten subjects and testing on actions from another ten subjects, while the CV protocol

used footage from the first camera pair for training and the rest for testing. The MA-STGCN model, with its multi-headed attention mechanism, enhanced the original model's accuracy to 88.1%, accelerated convergence, reduced network fitting time, and improved training speed. As Table 3 shows, MA-STGCN achieved accuracies of 88.1% and 94.0% on the Mining Behaviour dataset under CS and CV protocols, respectively.

Table 3 The verification results of NTU-RGBD dataset

Models	*Cross subject*	*Cross-view*
Deep LSTM	62.8	67.1
TCN	76.3	80.5
ST-GCN	82.1	83.3
HCN	86.3	88.2
AS-GCN	86.9	90.1
ST-GR	87.1	93.2
MA-STGCN(ours)	88.1	94.0

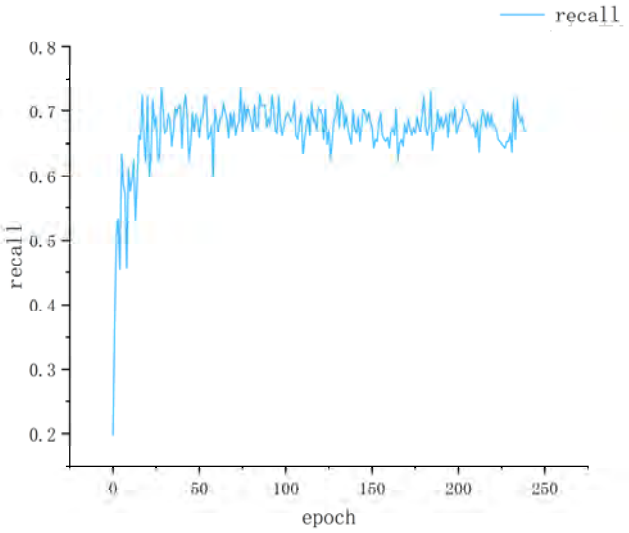
The MA-STGCN model's exceptional performance in detecting drill pipe behaviour is attributable to its innovative algorithmic framework. By incorporating the Simple Baseline algorithm, the model significantly improves posture recognition precision, which is crucial for analysing complex miner movements. Integrating a multi-head attention mechanism further augments the STGCN's ability to process spatio-temporal data, tracking dynamic miner movements over time and space. This dual strategy enhances the model's sensitivity to subtle movements. It boosts its interpretative accuracy in the variable and complex conditions of mining environments, as evidenced by its high accuracy in both CS and CV evaluation metric.

4.3.2 Mining behaviour dataset experiments

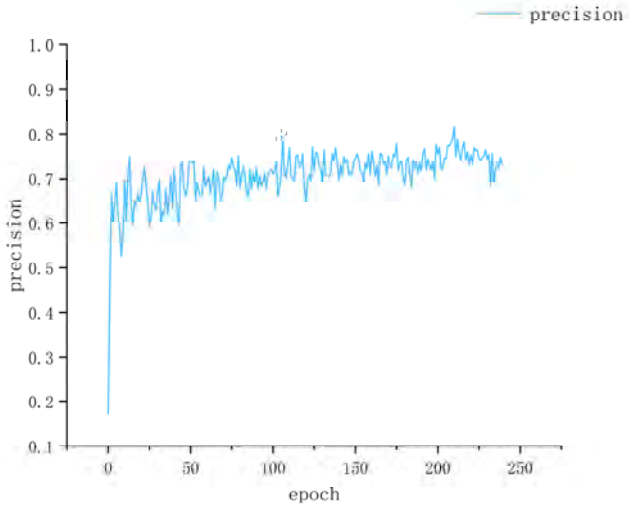
The MA-STGCN model's experimental validation on the Mining Behaviour dataset focused on accuracy and recall as the primary metrics. Figure 14 illustrates the model's precision and recall throughout the training process. Table 4 details the dataset's optimal performance across five behaviours, notably its effective recognition of sitting behaviour. The results emphasise MA-STGCN's proficiency, especially its superior accuracy and recall in identifying falling behaviour. MA-STGCN's performance was benchmarked against leading methods in behaviour recognition. Noteworthy is that the model was trained on the dataset without employing data augmentation techniques.

As shown in Figure 15, this study conducts comprehensive field tests to ascertain the efficacy of the MA-STGCN model within a real-world mining environment. Employing meticulously crafted experimental methodologies, the research focuses on documenting and analysing five characteristic miner behaviours during coal mine operations: ascending the rod, moving the rod, unloading the rod, sitting and falling. The findings are showcased via an extensive collection of images, encompassing visual comparisons of posture recognition, authentic on-site photographs, and conclusive behaviour recognition results.

Figure 14 Evaluation of metric changes when training on the mining behaviour dataset (a) precision (b) recall (see online version for colours)



(a)



(b)

Table 4 The verification results of mining behaviour dataset

Behaviour	*Precision*	*Recall*
Ascending the rod	92.6%	91.5%
Moving the rod	91.3%	90.2%
Unloading the rod	93.1%	92.1%
Falling	96.5%	98.0%
Sitting	97.8%	97.2%

Figure 15 The behaviour recognition effect of ascending the rod, moving the rod, unloading the rod, siting and falling in the actual coal mine scene (see online version for colours)

4.4 Ablation study

To assess the efficacy of the modules in this model, ablation experiments were con-ducted using a custom dataset. These experiments focused on mainstream self-attention, multi-head self-attention, and time-based attention mechanisms. As indicated in Table 5, the model incorporating the multi-head attention module demonstrated superior behavioural recognition accuracy compared to the other models. This outcome validates the effectiveness of the multi-head attention mechanism. Employing the multi-head attention mechanism aids in eliminating redundant information that could impact results, allowing for a more efficient fusion of image features and human key point features, notably enhancing recognition accuracy.

There is no one-size-fits-all answer regarding the choice between Adam and RMSprop optimisers. The selection depends on the specific task and data characteristics. Generally, Adam optimiser tends to be more effective for deep neural network training, while RMSprop is preferable for larger or more complex models that require a more robust convergence mechanism. Both optimisers are offered to accommodate varying requirements.

Table 5 Comparison of classification accuracy in NTURGB+D and NTU-120 uplink

	Optimiser	*NTU60-XSub*	*NTU60-XView*	*NTU120-XSub*	*NTU120-XSet*
ST-GCN	Adam	81.5	88.3	70.7	73.2
ST-GCN + SA	Adam	82.6	88.4	71.0	73.
ST-GCN + SA	RMSprop	82.3	87.6	71.4	73.1
ST-GCN + TA	Adam	81.3	88.7	71.2	73.8
ST-GCN + TA	RMSprop	83.2	89.0	72.1	74.6
MA-STGCN	Adam	84.7	89.3	74.7	75.9
MA-STGCN	RMSprop	85.1	89.9	75.3	75.6

The results detailed in Table 5 demonstrate the performance of the MA-STGCN model. According to the table, the MA-STGCN model attains an 85.1% accuracy rate on the NTU60-XSub dataset, surpassing the basic ST-GCN model's 81.5% accuracy under the same evaluation criteria. This indicates a 3.6% advantage for MA-STGCN over ST-GCN. In the more challenging NTU120 dataset, the MA-STGCN model achieved a 75.3% accuracy in the XSub assessment, outperforming the STGCN mod-el, which scored 70.7%, marking a 4.6% improvement. These results validate the efficacy of the multi-head attention mechanism in complex behaviour recognition tasks, particularly in niche applications like miner behaviour analysis.

The superiority of the MA-STGCN model is largely ascribed to its multi-head attention mechanism and the optimiser selection process. While self-attention mechanisms concentrate on the relationships within individual input sequences, multi-head attention mechanisms simultaneously engage multiple representation subspaces, thus offering enhanced feature representation and comprehensive information integration. The time-based attention mechanism is pivotal in pinpointing crucial moments in time series data. In scenarios like coal mine behaviour recognition, the MA-STGCN model adeptly captures the intricacies and spatio-temporal nuances of miners' movements, courtesy of the multi-head attention mechanism's capability to parallel process diverse information patterns, making it exceptionally suitable for complex scene analysis. The choice of the Adam and RMSprop optimisers largely depends on the specific application and data characteristics. Typically, Adam is favoured for training deep neural networks, whereas RMSprop is preferred for larger models or those with convergence challenges. In this context, either optimiser could be chosen based on distinct requirements.

5 Conclusions

This paper focuses on constructing a behaviour recognition system for underground drilling operators. Previous studies largely relied on video-based target detection, often conducted in simulated scenarios, lacking consideration for the actual underground mining environment. This study constructs a real underground behaviour dataset. It innovates the algorithm for such a mining scenario, addressing issues like poor visibility of personnel features due to lighting, dust, and obstructions, which hinder effective behaviour recognition. The main contributions of this research include creating a genuine underground behaviour dataset and filling a gap in behaviour recognition research in the mining field. Secondly, integrating an efficient channel attention mechanism into the YOLOX target detection algorithm effectively deals with issues like lighting variations, heavy dust, and severe obstructions in mine environments, enhancing the accuracy of target detection. Thirdly, replacing the traditional OpenPose structure with a lightweight Simple Baseline convolutional

network to directly extract human key points from images, improving processing efficiency and accuracy. Finally, introducing a multi-head attention mechanism into the STGCN enables the model to process and learn multidimensional behavioural traits of drilling personnel in parallel, enhancing the comprehensiveness of behaviour understanding and analysis in complex environments.

In conclusion, this paper's research significantly contributes to behaviour recognition and the industrial sector. First, it enhances the safety of mining operations by monitoring worker behaviour to prevent potential accidents. Second, intelligent behaviour recognition can improve production efficiency and management levels. In addition, this technology can provide critical data support for accident cause analysis. Although this method has challenges, such as large model parameters and high hard-ware requirements, future research will continue to focus on lightweight behaviour recognition methods, further advancing practical applications and technological progress in this field.

References

Belhocine, A., Shinde, D. and Patil, R. (2021) 'Thermo-mechanical coupled analysis based design of ventilated brake disc using genetic algorithm and particle swarm optimization [J/OL]', *JMST Advances*, pp.41–54, http://dx.doi.org/10.1007/s42791-021-00040-0.

Berlin, S.J. and John, M. (2022) 'Vision based human fall detection with Siamese convolutional neural networks', *J. Ambient. Intell. Human Comput.*, Vol. 13, pp.5751–5762, https://doi.org/10.1007/s12652-021-03250-5.

Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M. (2020) *'YOLOv4: Optimal Speed and Accuracy of Object Detection*, arXiv preprint arXiv:2004.10934.

Chaturvedi, S., Kumar, N. and Kumar, R. (2023) 'A PSO optimized novel PID neural network model for temperature control of jacketed CSTR: design', *Simulation, and a Comparative Study*, Vol. 28, No. 2024, pp.4759–4773

Chen, Y., Li, C., Li, Y., Pan, J. and Loy, C.C. (2020) 'Spatial temporal attention residual network for action recognition', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11252–11261.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014) *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*, arXiv preprint arXiv:1412.3555.

Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K. and Darrell, T. (2015) 'Long-term recurrent convolutional networks for visual recognition and description', in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp.2625–2634.

Gawande, U., Hajari, K. and Golhar, Y. (2022) 'Robust pedestrian detection using scale and illumination invariant Mask R-CNN', *International Journal of Computational Science and Engineering*, Vol. 25, No. 6, pp.607–618.

Hou, J., Wang, G., Chen, X. et al. (2018) 'Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition', *Proceedings of the European Conference on Computer Vision (ECCV) Workshops.*

Hu, N., Zhang, D., Xie, K. et al. (2022) 'Graph learning-based spatial-temporal graph convolutional neural networks for traffic forecasting', *Connection Science*, Vol. 34, No. 1, pp.429–448.

Huang, C., Li, Y., Li, W. and Li, T. (2020) 'Multimodal graph convolutional networks for human action recognition', in *Proceedings of the 28th ACM International Conference on Multimedia*, pp.3089–3097.

International Labour Organization (2020) *Safety and Health in Mines: A Global Perspective*, ILO, Geneva.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014) 'Large-scale video classification with convolutional neural networks', in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp.1725–1732.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T. (2011) 'HMDB: a large video database for human motion recognition', in *Proceedings of the International Conference on Computer Vision*, pp.2556–2563.

Kumar, R. (2022a) 'Memory recurrent Elman neural network-based identification of time-delayed non-linear dynamical system', *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 53, No. 2, pp.753–762.

Kumar, R.A. (2022b) 'Lyapunov-stability-based context-layered recurrent pi-sigma neural network for the identification of nonlinear systems', *Applied Soft Computing*, Vol. 122, No. 108836, pp.1568–4946.

Kumar, R. (2023) 'Double internal loop higher-order recurrent neural network-based adaptive control of the nonlinear dynamical system', *Soft Computing*, Vol. 27, No. 22, pp.717313–17331.

Kumar, R. and Srivastava, S. (2020) 'Externally recurrent neural network based identification of dynamic systems using Lyapunov stability analysis', *ISA Transactions*, Vol. 98, No. 2020, pp.292–308.

Kumar, R., Srivastava, S. and Gupta, J.R.P. (2017a) 'Modeling and adaptive control of nonlinear dynamical systems using radial basis function network', *Soft Computing*, Vol. 21, Nos. 1–4, pp.4447–4463.

Kumar, R., Srivastava, S. and Gupta, J.R.P. (2017b) 'Lyapunov stability-based control and identification of nonlinear dynamical systems using adaptive dynamic programming', *Soft Computing*, Vol. 21, No. 2017, pp.4465–4480.

Kumar, R., Srivastava, S. and Gupta, J.R.P. (2017c) 'Diagonal recurrent neural network based adaptive control of nonlinear dynamical systems using lyapunov stability criterion', *ISA Transactions*, Vol. 67, No. 2017, pp.407–427.

Kumar, R., Srivastava, S. and Gupta, J.R.P. et al. (2018) 'Self-recurrent wavelet neural network–based identification and adaptive predictive control of nonlinear dynamical systems', *International Journal of Adaptive Control and Signal Processing*, Vol. 32, No. 9, pp.1326–1358.

Kumar, R., Srivastava, S., Gupta, J.R.P. et al. (2019) 'Temporally local recurrent radial basis function network for modeling and adaptive control of nonlinear systems', *ISA transactions*, 2019, 87: 88-115.

Li, C., Zhong, Q., Xie, D. et al. (2018a) 'Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation', arXiv preprint arXiv:1804.06055, 2018.

Li, Y., Wang, S. and Ouyang, W. (2018b) 'CrowdPose: efficient crowded scenes pose estimation and a new benchmark', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.6752–6761.

Li, M., Chen, S., Chen, X. et al. (2019a) 'Actional-structural graph convolutional networks for skeleton-based action recognition', Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.3595–3603.

Li, B., Li, X., Zhang, Z. et al. (2019b) 'Spatio-temporal graph routing for skeleton-based action recognition', Proceedings of the *AAAI Conference on Artificial Intelligence*, Vol. 33, No. 1, pp.8561–8568.

Li, S.C., Ke, L., Pratama, K. et al. (2021) 'Cascaded deep monocular 3D human pose estimation with evolutionary training data', *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Seattle: IEEE, DOI: 2020:6172-6182.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... and Zitnick, C.L. (2014) 'Microsoft COCO: common objects in context', in *European Conference on Computer Vision*, pp.740–755.

Liu, X., Han, Z., Wang, Y. and Zhang, M. (2019) 'Attention-aware spatial-temporal graph convolutional networks for skeleton-based action recognition', *IEEE Transactions on Multi-Media*, Vol. 21, No. 10, pp.2580–2590.

Qin, M. (2022) 'Application of efficient recognition algorithm based on deep neural network in English teaching scene', *Connection Science*, Vol. 34, No. 1, pp.1913–1928.

Redmon, J. and Farhadi, A. (2018) *YOLOv3: An Incremental Improvement*, arXiv preprint arXiv:1804.02767.

Redmon, J. and Farhadi, A. (2021) *YOLOX: Exceeding YOLO Series in 2021*, arXiv pre-print arXiv:2107.08430.

Ryoo, M.S. and Matthies, L. (2013) 'First-person activity recognition: What are they doing to me?', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2730–2737.

Shahroudy, A., Liu, J. and Wang, G. (2016) 'NTU RGB+D: A large scale dataset for 3D human activity analysis', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.1010–1019.

Shamsafar, F. and Ebrahimnezhad, H. (2021) 'Uniting holistic and part-based attitudes for accurate and robust deep human pose estimation', *J. Ambient. Intell. Human. Comput.*, Vol. 12, pp.2339–2353, https://doi.org/10.1007/s12652-020-02347-7

Shi, L., Ji, Y., Wang, Y. and Ye, J. (2019) 'Skeleton-based action recognition with spatial temporal graph attention networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.11812–11820.

Shi, L., Zhang, Y., Cheng, J. et al. (2020) 'Skeleton-based action recognition with multi-stream adaptive graph convolutional networks', *IEEE Transactions on Image Processing*, Vol. 29, pp.9532–9545.

Simonyan, K. and Zisserman, A. (2014) 'Two-stream convolutional networks for action recognition in videos', *Advances in Neural Information Processing Systems*, Vol. 1, No. 9, pp.568–576.

Soomro, K., Zamir, A.R. and Shah, M. (2014) 'UCF101: a dataset of 101 human action classes from videos in the wild', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3738–3745.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015) 'Learning spatiotemporal features with 3d convolutional networks', *Proceedings of the IEEE International Conference on Computer Vision*, pp.4489–4497.

Vaswani, A., Shazeer, N., Parmar, N. et al. (2017) 'Attention is all you need', *Advances in Neural in-Formation Processing Systems*, p.30.

Wandt, B. and Rosenhahn B. (2019) *RepNet:Wealdy SupanisediTrsning of an Atversaial Reprjction Network for 3D Human Pose Eistimation (2)/IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*, Long Beach: IEEE, DOI: 2019:7774-7783.

Wang, C., Zhang, X. and Li, Z. (2020a) 'YOLOv5: improved real-time object detection with one-stage object detectors', arXiv preprint arXiv:2011.08036.

Wang, X., Zhang, T., Gong, D., Liu, C. and Zheng, N. (2020b) 'ECA-Net: efficient channel attention for deep convolutional neural networks', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11575–11585.

Wang, L., Chen, J., Chen, Z. et al. (2022) 'Multi-stream part-fused graph convolutional networks for skeleton-based gait recognition', *Connection Science*, Vol. 34, No. 1, pp.652–669.

Wang, L., Xiong, Y. and Lin, D. (2019) 'Deep learning for video analysis: A comprehensive review', *ACM Computing Surveys (CSUR)*, Vol. 52, No. 3, pp.1–42.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L. (2016) 'Temporal segment networks: Towards good practices for deep action recognition', *European Conference on Computer Vision*, pp.20–36.

Wu, X., Wang, Y., Wang, Y. et al. (2019) 'Deep learning for sensor-based activity recognition: a survey', *Pattern Recognition Letters*, Vol. 119, No. 33, pp.3–11.

Xia, L., Chen, C.C. and Aggarwal, J.K. (2019) 'NTU RGB+D 120: A Large-Scale Bench-mark for 3D human activity understanding', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.12026–12036.

Xu, J.W., Yu, Z.B., Ni, B.B. et al. (2020) 'Deep kinematics analysis for monocular 3D human pose estimation', *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Seattle:IEEE, pp.896–905.

Xu, P., Zhou, Z. and Geng, Z. (2022) 'Safety monitoring method of moving target in underground coal mine based on computer vision processing', *Scientific Reports*, Vol. 12, No. 1, p.17899.

Xu, T. and Takano, W. (2021) 'Graph stacked hourglass networks for 3d human pose estimation', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.16105–16114.

Yao, Y., Li, Y., Zhang, J. et al. (2018) *Human Behaviour Recognition Based on Convolutional Neural Networks//International Conference on Neural Information Processing*, Springer, Cham, pp.55–63.

Ye, Q., Li, R., Yang, H. et al. (2022) 'Human interactive behaviour recognition method based on multi-feature fusion', *International Journal of Computational Science and Engineering*, Vol. 25, No. 3, pp.262–271.

Ye, Q., Zhao, Y. and Zhong, H. (2023) 'Human behaviour analysis based on spatio-temporal dual-stream heterogeneous convolutional neural network', *International Journal of Computational Science and Engineering*, Vol. 26, No. 6, pp.673–683.

Yu, M. and Li, J. (2020) 'Psychosocial safety climate and unsafe behaviour among miners in China: the mediating role of work stress and job burnout', *Psychology, Health and Medicine*, Vol. 25, No. 7, pp.793–801.

Zhang, L.Y., Zhao, Z.Y. and Chen, KL. (2018) 'Research on behaviour recognition based on deep learning', *Computer Science*, Vol. 45, No. 12, pp.1–8.

Zhang, S. and Zhu, X. (2019) 'Spatial temporal graph convolutional networks for skeleton-based action recognition', in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp.4383–4390.

Zhang, S., Song, X. and Sun, C. (2019) 'Analysis of coal mine safety management based on video surveillance', *IEEE Access*, Vol. 7, pp.108403–108410, DOI: 10.1109/ACCESS.2019.2934089.

Zhang, X., Zhu, X., Dai, Y., Xiong, H. and Li, H. (2020) 'Spatio-temporal data augmentation for action recognition with dynamic images', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp.1503–1504.

Zhang, Y., Shao, W., Zhang, M. et al. (2016) 'Analysis 320 coal mine accidents using structural equation modeling with unsafe conditions of the rules and regulations as exogenous variables', *Accident Analysis and Prevention*, Vol. 92, No. 2016, pp.189–201.

Zhao, Y., Zhang, Z., Yang, S., Huang, H. and Sun, Y. (2020) 'Action recognition using two-part temporal convolutional graph neural network', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 31, No. 1, pp.223–235.

Zhu, W., Lan, C., Xing, J. et al. (2016) 'Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30, No. 1.