Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **9**, Number **4**, November **2024**

Enhanced Spatio-temporal Graph Convolutional Power Operator Violation Recognition Method Based on Dual-flow Structure

Yang Xi*, Zi-Hao Zhang, Hao Wang

School of Computer Science
Northeast Electric Power University, Jilin 132012, P. R. China
474465389@qq.com, 531999659@qq.com, 471116753@qq.com

Si-Yu Meng, Jia Fu

Yongji Power Supply Company
State Grid Jilin Electric Power Co., Ltd, Jilin 132012, P. R. China
771742082@qq.com, 809818269@qq.com

Zhen-Yu Wu

Department of Orthopedics of Affiliated Hospital of Beihua University
Beihua University, Jilin 132012, P. R. China
myemail780216@sina.com

Jie Cao

School of Computer Science
Northeast Electric Power University, Jilin 132012, P. R. China
379688648@qq.com

*Corresponding author: Yang Xi
Received December 12, 2023, revised March 25, 2024, accepted June 29, 2024.

Abstract. *The safety accidents caused by the illegal behaviors of the operators climbing the equipment, climbing over the fence, smoking, stepping on the equipment, and making phone calls during the electric power operation occur frequently, resulting in serious casualties and economic losses. At present, there is no identification method for the above violations of electric power operators, and the conventional spatio-temporal map convolution method ignores the association of non-directly connected joints, with poor temporal feature extraction and low identification accuracy. We propose an enhanced spatio-temporal graph convolutional power operator violation behavior recognition method based on the dual-stream structure, firstly, we use the helmet, safety belt as the features, and use YOLOv5 to detect the power operators. We enhance the spatial and temporal features respectively, firstly using an adaptive graph convolution to enhance the correlation of non-directly connected joints; then a multi-scale temporal convolution module based on channel attention mechanism is proposed to extract the temporal features of the violation behavior more adequately. Finally, considering that the direction and length of bones also contain rich behavioral information, a dual-stream structure violation recognition model is constructed. Through experiments on the private dataset of electric power operation, the method can accurately identify the violation behaviors of electric power operators climbing equipment, overcoming fences, smoking, stepping on equipment, and making phone calls, and the recognition accuracy reaches 94.7%, which is a substantial improvement com-pared with the mainstream model, and it can effectively reduce the safety risk of the electric power operation site and decrease the probability of safety accidents.*

Keywords: dual-stream structure, YOLOv5, enhanced spatio-temporal graph convolutional, electric power operation, violation recognition modeldual-stream structure, YOLOv5, enhanced spatio-temporal graph convolutional, electric power operation, violation recognition model

1. **Introduction.** Electricity safety accidents and staff violations have a direct relationship. Electricity operators safety awareness is weak, fatigue lax, the operation process of climbing equipment, over the fence, smoking, stepping on the equipment, telephone violations occur from time to time, these behaviors have a direct impact on the operation of the probability of accidents, once an accident occurs, it will result in significant property damage and casualties.

At present, most of the supervision of electric power operators are from the perspective of personal protective equipment to determine whether the illegal operation, such as whether to wear helmets, insulated gloves, safety belts and so on. There is no effective method to recognize and control the illegal behaviors of electric power operators. Therefore, we address the practical needs and propose a violation behavior recognition method for electric power operators using the skeletal spatio-temporal characteristics of behavior. The method can accurately identify the unruly behaviors of electric power operators in terms of climbing equipment, scaling fences, smoking, stepping on equipment, and making phone calls, which regulates the operation behavior, effectively reduces the probability of accidents, and ensures the safe and normal operation of the power system. Our main contributions are summarized below:

1. Characterized by the helmets and safety belts of power operators, use YOLOv5 to detect personnel and locate power operators.

2. An adaptive graph convolution is used to enhance the correlation of non-directly connected joints such as hands and feet.

3. A multi-scale temporal convolution module based on the channel attention mechanism is designed to more adequately extract temporal features of behavior.

4. Considering that the direction and length of the skeleton also contain rich behavioral information, a dual-stream structure violation recognition model is constructed to improve the recognition accuracy of the violation.

2. **Related Work.** For behavior recognition, Chen and Guan [1] screened images with representative actions from students' classroom videos as samples, and used the YOLO model to detect transient actions and then determine behaviors. YOLO is a target detection method for behavior recognition, which can only determine the behavior by identifying a transient action, but not the whole coherent action, and the robustness is poor.

Compared with an instantaneous action, coherent action frames are more responsive to real behavior. Wang and Schmid [2] proposed IDT (Improved Dense Trajectories) by improving the traditional dense trajectory method, removing the interference caused by the camera movement and using the appearance features of the moving objects for filtering. However, with the development of deep learning, this traditional behavior recognition method was gradually eliminated. Behavior recognition methods based on deep learning are mainly divided into three branches according to the network structure: Two-Stream method, 3D convolution method, and spatio-temporal map convolution method. Simonyan and Zisserman [3] proposed Two-Stream Network, which adopts a two-branch network architecture and captures the spatial and temporal information of the video respectively. The spatial domain uses the RGB image as input to extract appearance features, and the temporal domain uses the optical flow information as input to extract timing features, the two branches of the network judge the categories of the actions separately, and finally the results of the two networks are fused to obtain the behavioral categories. Feichtenhofer et al. [4] proposed a network architecture that can be more finely grained to fuse spatial and temporal information, considering the fact that there is a certain connection between the feature maps of the Two-Stream Network. fusion of spatio-temporal information, which further improves the recognition accuracy. Berlin and John [5] used joint entropy to calculate the optical flow features, which makes the optical flow feature extraction more accurate and efficient. The dual-stream method has high recognition accuracy, but the computation of optical flow features on the temporal sequence is large, resulting in a slow overall network speed. Tran et al. [6] used 3D convolution for the first time in behavioral recognition, which can better extract spatio-temporal features by adding another temporal dimension to the original spatial dimension. Although 3D convolution can capture both temporal and spatial information, it consumes too much arithmetic and graphics memory, Qiu et al. [7] proposed a pseudo 3D convolutional layer instead of the regular 3D convolutional layer, and constructed a deep network by using the ResNet residual linkage method, which achieves the effect of parameter reduction. Feichtenhofer et al. [8] argued that the categorical attributes of a behavior are slowly changing, while the process of a behavior is usually fast, proposed the SlowFast network with two branches capturing semantic information at low frame rates and action information at high frame rates to further improve the recognition accuracy of 3D convolution. As far as the current research shows, both the dual-flow and 3D convolution methods mostly use RGB images and optical flow as data modalities, and the 3D convolution method is not as accurate as the dual-flow method for recognition.

Biological observations have shown that the positions of a small number of joints can effectively represent human behavior even without appearance information. Skeletal features have purer features and less noise compared to raw RGB images and optical flow information. Since the human skeletal map conforms to the topological map structure, Yan et al. [9] proposed a spatio-temporal map convolution method for behavior recognition, which uses the first-order skeletal point coordinates to firstly extract the spatial

features of the skeleton by map convolution, and then obtain the temporal features by temporal convolution, and finally fuses the features to arrive at the classification results. This method has high recognition accuracy and fast algorithm speed, which is the mainstream recognition idea in the field of behavior recognition at present.

3. **Method.** The recognition method firstly uses YOLOv5 to detect power operators, then fully extracts the spatio-temporal features of violation behaviors through adaptive graph convolution and multi-scale temporal convolution modules, and finally constructs a two-stream structure recognition model of violation behaviors based on skeletal points and skeletal ones. The overall research flow is shown in Figure 1:

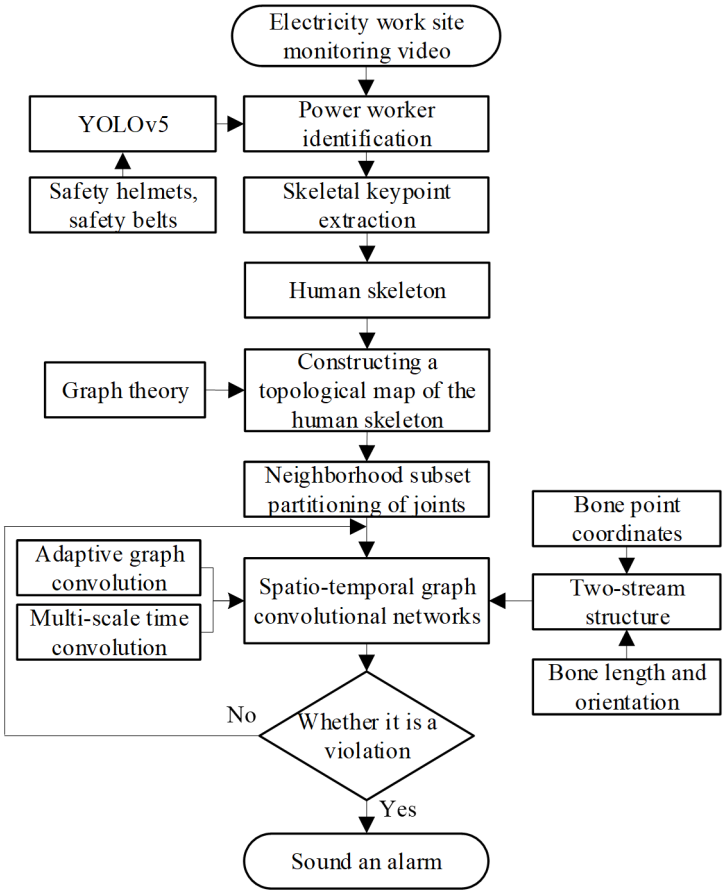


Figure 1. Research Flowchart.

3.1. **YOLOv5-based detection method for electric power operators.** In order to accurately differentiate between power operators, the detection of people at the power operation site is performed based on YOLOv5. Firstly, the person in the image is detected, then the helmet or seat belt is detected in the human anchor frame, if the helmet or seat belt is detected and 90% of the anchor frame exists in the human anchor frame, the person is judged as an electric power operator, and then the person is subjected to subsequent behavioral recognition.

3.2. **Adaptive graph convolution.** In order to better represent the high-dimensional mapping of the graph convolution, we use a spatial structure partitioning strategy that conforms to the human body's movement pattern. Different from the convolution kernel of ordinary CNN, here the neighborhood with a distance of 1 from the node is used as

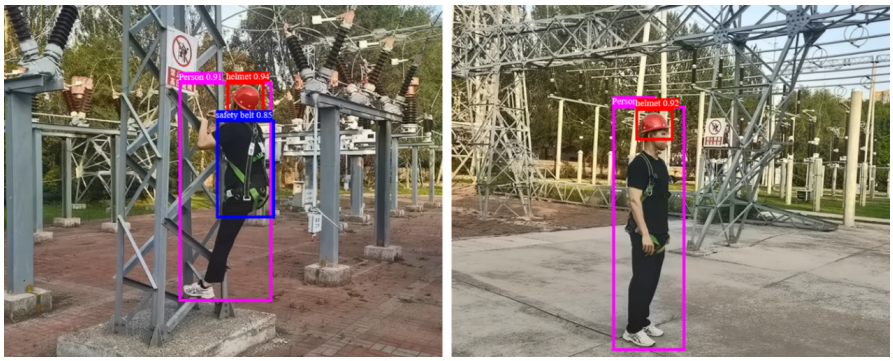


Figure 2. Electricity Operator Testing

the sensory field of the graph convolution, as shown in Figure 3(a). The spatial structure partitioning strategy is shown in (b), using the human center of gravity in the action as a reference, the neighborhood with a distance of 1 from the node is divided into three subsets, the blue centripetal subset near the center of gravity, the orange centrifugal subset far from the center of gravity, and the green node itself. The average of the coordinates of all joints in each frame of the skeleton sequence is taken as the center of gravity of the human skeleton, and the division strategy is as follows:

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

Where the distance of the node itself from the center of gravity of the body is r_i, the distance of the neighboring nodes from the center of gravity of the body is r_j, l_{ti} is the label to which the node belongs.

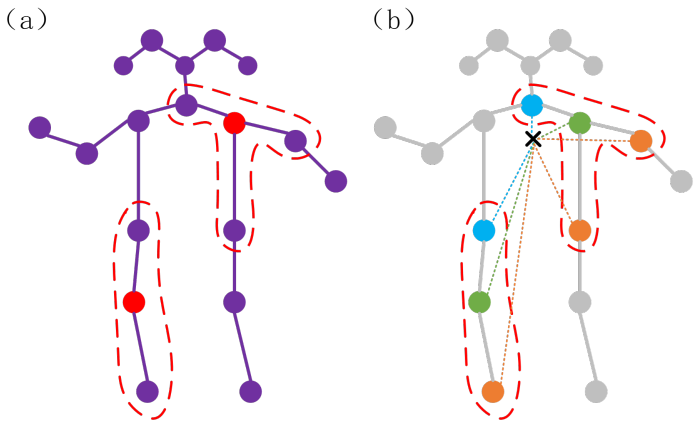


Figure 3. Graph convolution partitioning strategy.

The topology of graph convolution in ST-GCN is constructed based on the physical connections of human skeleton, and its graph convolution formula is:

$$f_{out} = \sum_{k}^{K_v} W_k \left(f_{in} A_k\right) \odot M_k \tag{2}$$

where W_k is the weights, A_k is the adjacency matrix, M_k is the attention mask, and $K_v = 3$ is the number of subset categories. The mask M_k is directly multiplied with the

adjacency matrix by the elements of A_k. If some of the elements inside the adjacency matrix A_k are 0, then the final result is 0 regardless of the sum of the elements corresponding to M_k, so no new connections are created that do not exist in the original physical map. So no new connection will be created that does not exist in the original physical map. However, in human skeletal behavior recognition, motion features do not necessarily exist only in the directly connected bones, but also between the joints of bones that are not directly connected, such as climbing and overstepping, and there is a strong correlation between hands and feet. We use the adaptive graph convolution shown in Equation (2) [10], in which the neighbor matrix in the original graph convolution is improved, where the neighbor matrix is mainly divided into three parts A_k, B_k, C_k:

$$f_{out} = \sum_{k}^{K_v} W_k f_{in} \left(A_k + B_k + C_k\right) \tag{3}$$

The first part A_k is the same as A_k in Equation (2) and represents the physical structure of the human body. The second part, B_k, is also an $N \times N$ adjacency matrix. Unlike A_k, B_k is a trainable weight matrix whose values are not constrained, which means that B_k is a data-driven adjacency matrix completely learned from the training data, and can represent not only whether two nodes are connected or not, but also the strength of the connection. In contrast to M_k, addition is used instead of multiplication, which in turn creates new connections that do not exist in the original human physical connections. The third part, C_k, is a sample-based graph adjacency matrix, which is formed by embedding a normalized Gaussian function to compute the similarity between two joints, as shown in Equation (4):

$$f(v_i, v_j) = \frac{e^{\theta(v_i)^T \phi(v_i)}}{\sum_{j=1}^{N} e^{\theta(v_i)^T \phi(v_i)}} \tag{4}$$

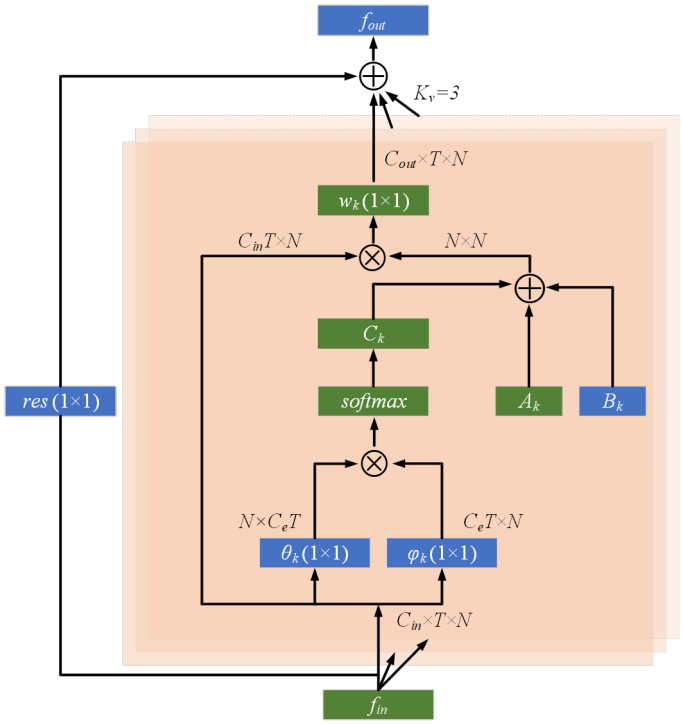


Figure 4. Structure of adaptive graph convolution.

where N is the number of joints, when the input feature map is $f_{in} \in \mathbb{R}^{C_{in} \times T \times N}$, using two 1×1 convolutional embedding functions θ and ϕ, transform f_{in} into the embedding space $\mathbb{R}^{C_e \times T \times N}$, transforming these two matrices into the matrix $M_{\theta k} \in \mathbb{R}^{N \times C_e T}$ and matrix $M_{\phi k} \in \mathbb{R}^{C_e T \times N}$, where C_e is the dimension of the embedding channel. The two matrices are then multiplied. The two matrices are then multiplied to obtain a similarity matrix C_k for $N \times N$, where C_k^{ij} denotes the similarity between joints v_i and v_j, whose value is normalized to between 0 and 1, is used as a soft connection between the two joints. Since the normalized S matching has *softmax* operations, Equation (5) is used to compute C_k.

$$C_k = \text{softmax}\left(f_{in}^T W_{\theta k}^T W_{\phi k} f_{in}\right) \tag{5}$$

where W_θ and W_ϕ are the trainable parameters of the embedding functions θ and ϕ respectively. The adaptive graph convolution model is shown in Figure 4, with graph convolution kernel of size 1 and number 3, and $K_v = 3$ denotes the number of subsets.

3.3. **Multiscale temporal convolution module based on channel attention mechanism.** From the time dimension, some complex behaviors are composed of consecutive sub-behaviors, and the sub-behaviors at different stages have different weights on the determination of the whole behavior, and there are also dependencies between the sub-behaviors at the previous and previous stages. However, the size of the sensory field in the standard temporal convolutional network is fixed, and if it is used for the extraction of temporal dependencies, it can only represent the information in a time scale, and it cannot fully extract the temporal information of behavior. However, if used to extract behavioral temporal dependencies, it can only represent information on a time scale, which is not sufficient for extracting behavioral temporal information. Therefore, we propose a multi-scale temporal convolution module based on the channel attention mechanism in the temporal dimension, as shown in Figure 5, which uses parallel temporal convolution [11] to extract temporal features at different scales, and dynamically adjusts the weights of each branch by combining with the attention mechanism, and then superimposes them in a nonlinear way to enhance the expressive ability of the feature map.

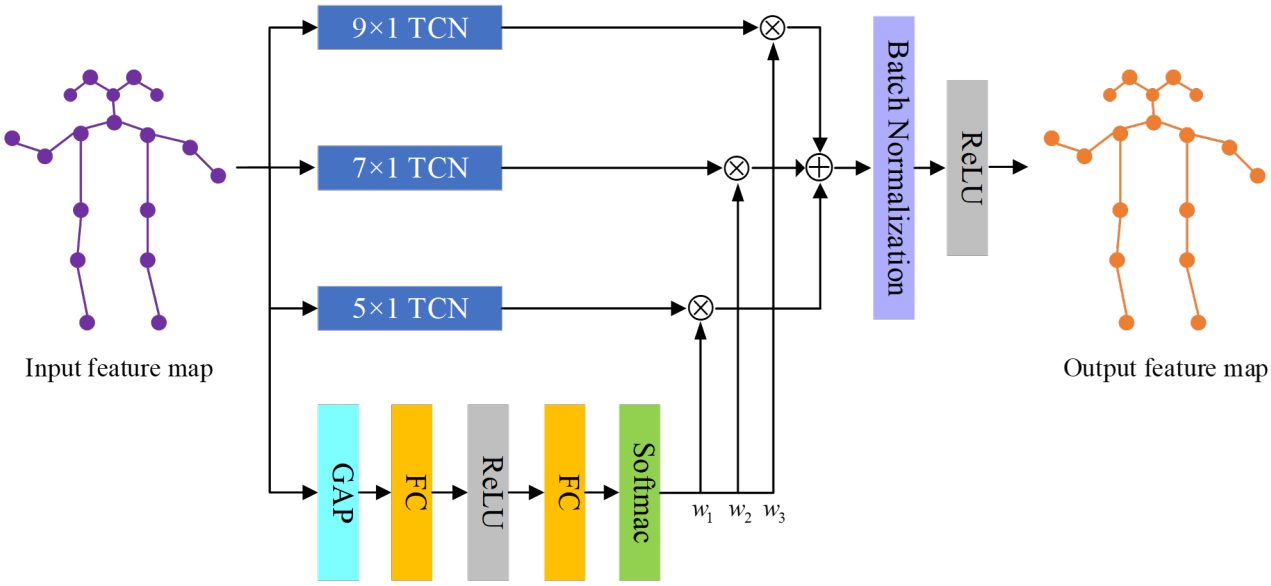


Figure 5. Multi-scale time convolution module based on channel attention mechanism.

The module uses three temporal convolutional branches with convolutional kernel sizes $(5 \times 1, 7 \times 1, 9 \times 1)$ in parallel to capture the dependencies between skeleton frame sequences of different durations. The channel attention [12] is modified in the attention branch by converting the dimensions of the RGB image feature maps (H, W, C) to the dimensions of the human skeleton frame sequence (T, V, C), i.e., replacing the height of the image, H, and the width of the image, W, by the time step, T, in the sequence of skeleton frames and the number of joints in the individual skeletons, V. The input feature maps are convolved with the multiscale time convolution, and then multiplied by the attention branch to adjust the branch weights, and finally the input feature maps are multiplied with the attention branch to adjust the branch weights. The input feature maps are multiscale time-convolved, then multiplied with the attention branches to adjust the branch weights, and finally the outputs of each time-convolved branch are summed to aggregate the outputs of multiple parallel branches to obtain the fused multiscale spatio-temporal feature map, and the outputs can be defined as follows:

$$f_{out} = R\left(B\left(W_5 Conv_5(x) + W_7 Conv_7(x) + W_9 Conv_9(x)\right)\right) \tag{6}$$

Where x is the input, $Conv_n$, $n = 5, 7, 9$ is the temporal convolution at three different scales, $W_5 + W_7 + W_9 = 1$ is the assigned attention weights, B, R are the normalized Bath Normalization and ReLU activation functions, respectively.

3.4. **A model for identifying violation behaviors of electric power operators.** Using the adaptive graph convolution and multiscale temporal convolution mentioned above, a spatio-temporal graph convolution block was composed as shown in Figure 6, where AGCN denotes adaptive graph convolution and MSTCN denotes multiscale temporal convolution based on the channel attention mechanism, with an intermediate Dropout layer with a dropout rate of 0.5 added.

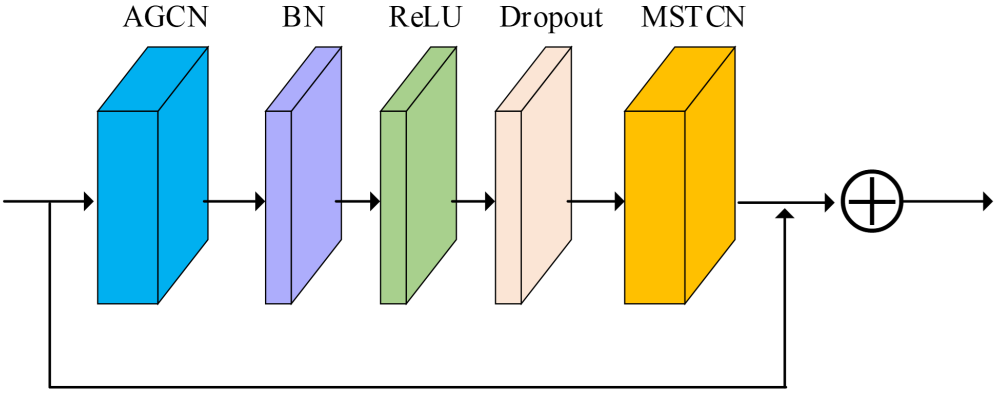


Figure 6. Spatio-temporal map convolution block

As shown in Figure 7, the spatio-temporal graph convolutional network consists of a stack of 9 spatio-temporal graph convolutional blocks as described above, and the number of output channels of each block is 64, 64, 64, 128, 128, 128, 256, 256 and 256. BN layer is added at the beginning to normalize the input data, and then global average pooling is per-formed to pool the feature maps of different samples to the same size. Then global average pooling is done to pool the feature maps of different samples to the same size and finally Softmax classifier is passed to get the prediction.

The coordinates of the bone joints are first-order information. Considering the second-order information of the bones, the length and direction also contain the behavioral

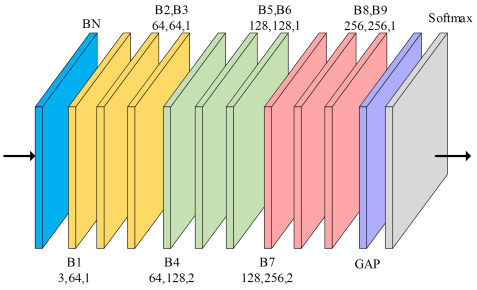


Figure 7. Convolutional network of spatio-temporal maps.

characteristics, a two-stream structure is designed to enhance the effect of behavioral recognition, as shown in Figure 8.

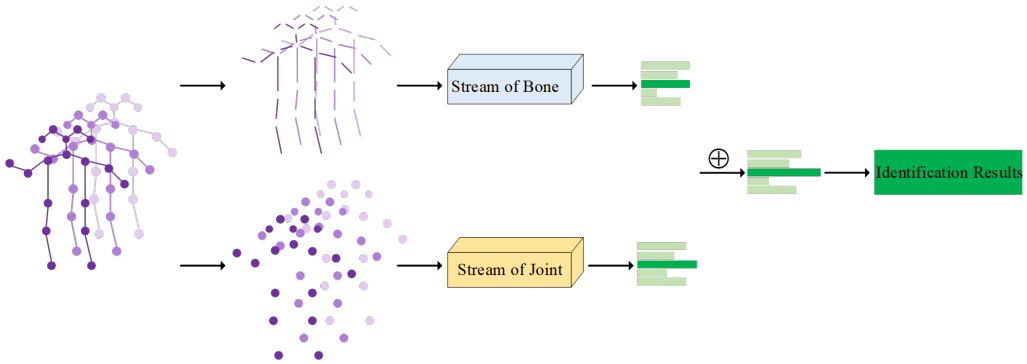


Figure 8. Model for recognizing violation behaviors of electric power operators with dual-flow structure.

Since a skeletal edge consists of two joints connected together, the joint near the center of gravity of the skeleton is defined as the source joint and the joint away from the center of gravity is defined as the target joint. Each skeletal edge is represented by a feature vector pointing from the source joint to the target joint, which contains both length and direction information. Assuming that the source joint $v_1 = (x_1, y_1)$ and the target joint $v_2 = (x_2, y_2)$, the vector of skeletal edges can be expressed as $b_{v_1 v_2} = (x_2 - x_1, y_2 - y_1)$. In the human skeleton graph, the number of joints is one more than the number of skeletal edges. A 0-valued bone vector is added so that each bone edge corresponds to a joint, and in this way the network of bone streams can be designed in the same way as the joint streams. Using the skeletal and joint streams to represent the network of input joints and skeletal edges, respectively, the Softmax scores of the two streams are added together to obtain a fusion score and predict the action labels.

4. **Experiments and Analysis.**

We conducted ablation experiments on a private dataset of electric power work sites and compared the experiments with other current state-of-the-art methods.

4.1. **Environment Configuration.**

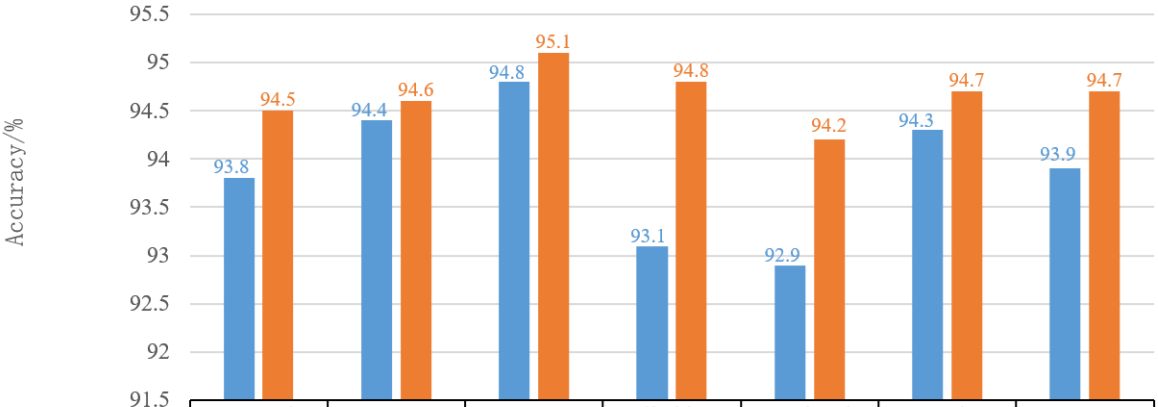
The experiments were performed on a server consisting of Ubuntu 18.04 with Linux kernel, python 3.7, pytorch 1.5.0+cu101, and an NVIDIA Tesla T4 GPU.

4.2. **Datasets.** The private dataset consists of 600 self-recorded videos of violation behaviors in the power operation scenarios, including 100 videos of smoking, talking on the phone, climbing over fences, climbing and stepping on equipment, and 100 videos of normal behaviors, captured by the deployment ball. Each behavior is captured from 8 angles, with 0° facing the human body and every 45° in clockwise direction. The dataset was divided into training and testing sets according to 7:3, with a training learning rate of 0.1 and an epoch of 100.

4.3. **Assessment of Indicators.** The evaluation index is the accuracy rate of behavior recognition, i.e., the percentage of correct classifications among all classifications, and the higher the accuracy rate is, the better the recognition performance of the method is. The accuracy rate is calculated as follows.

$$\text{Accuracy} = \frac{\text{Number of correctly categorized samples}}{\text{Full sample size}} \tag{7}$$

4.4. **Results.** In order to prove the effectiveness of each module in the model, ablation and comparison experiments were conducted. Firstly, an ablation experiment was carried out on the adaptive graph convolution in the model, as shown in Figure 9. Adaptive graph convolution improves the overall accuracy of the behavior recognition model by 0.7%, among which the recognition accuracy of climbing equipment is 1.7% higher, and the recognition accuracy of fence crossing is 1.3% higher, which indicates that adaptive graph convolution increases the flexibility of the model graph structure, strengthens the correlation of the human body's non-physical connections, and improves the recognition accuracy.



	Normal behavior	Smoke	Phone	Climbing equipment	Crossing the fence	Stepping onequipment	Total
Original graph convolution	93.8	94.4	94.8	93.1	92.9	94.3	93.9
Adaptive graph convolution	94.5	94.6	95.1	94.8	94.2	94.7	94.7

Figure 9. Comparison of adaptive graph convolution ablation experiments.

In order to illustrate the effectiveness of adaptive graph convolution, the connection strengths between (a) the right hand for climbing equipment and (b) the right foot for fence crossing at layer 9 of AGCN and other joints are visualized in Figure 10, where red is the result of the adaptive adjacency matrix and blue is the result of the original adjacency matrix. Each circle represents a joint, and its size represents the connection strength between a joint and other joints. For (a) the climbing behavior, it can be seen that the adaptive map convolution makes the right hand connected to other joints that are not physically connected, while the original map convolution makes the right hand

connected to the right elbow only. For (b), the behavior of crossing the fence, it can be seen that the adaptive graph convolution also connects the right foot to other joints that are not physically connected, whereas the original graph convolution connects the right foot to the right knee only. Therefore, the global adaptive graph convolution shows the potential dependence on the behavioral and kinematic relationships between the global joints of the skeleton. Compared with the predefined neighbor matrix based on the a priori knowledge of the human body, the non-physical structure of the joints in the neighbor matrix of the global adaptive graph convolution is strengthened, so that the strength of the global joints can be calculated automatically from the input skeleton data, and the strength of the global joints can be calculated in the network training, so that the joints of the human skeleton will realize the node weights of the human skeleton as the network level is deepened. In the network training, the joints of human skeleton are dynamically updated with the deepening of the network hierarchy to realize the node weights.

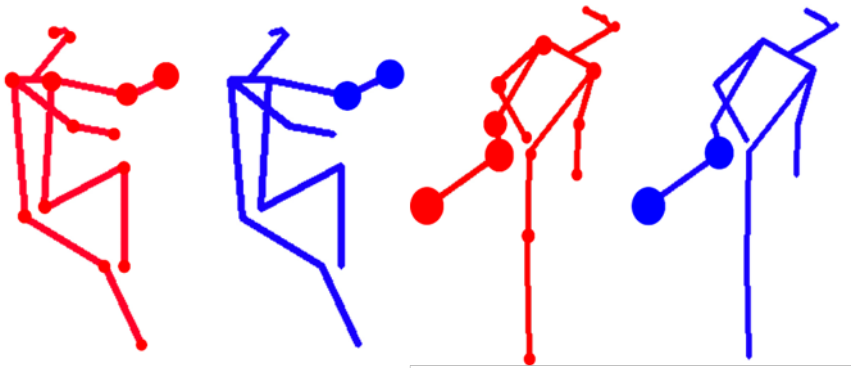


Figure 10. Visualization of joint strength.

When using different sized convolutional kernels to capture contextual information at different time scales, the recognition results will be affected by the different scales of the contextual information. The larger the convolutional kernel, the larger the sensory field, and the more contextual information can be captured, but some irrelevant contextual information will also be captured. In order to find a suitable combination of convolution kernels, several combinations of different convolution kernel sizes (3×1,5×1,7×1), (5×1,7×1,9×1), (7×1,9×1,11×1), and (9×1,11×1,13×1) are set up in the multiscale temporal convolution module to address the above problems. The experimental results are plotted in Table 1.

Table 1. Comparison experiments of different convolutional kernel combinations.

Convolution kernel size	**Accuracy/%**
3×1, 5×1, 7×1	92.1
5×1, 7×1, 9×1	**94.7**
7×1, 9×1, 11×1	93.1
9×1, 11×1, 13×1	92.3

The experimental results show that when the convolutional kernel combination in this module is (5×1,7×1,9×1), the recognition accuracy is the highest on the private dataset of electric power operation. By analyzing the accuracy of various behaviors, the accuracy degradation for convolutional kernel combinations of (3×1,5×1,7×1) is due to the confusion of smoking, talking on the phone, and touching the face in normal behaviors. The subject defines the smoking behavior as the action of lowering the arm vertically

and then bringing the cigarette to the mouth to smoke twice; the phone call behavior is defined as lowering the arm vertically and then bringing the phone to the ear all the time. The consecutive skeleton frames contain relevant information in the time domain, and the first half of these three behaviors are the same, all of them are the arm is vertically lowered, and then the phone is brought to the face. If the sensory field is small, the context information of the skeleton frame sequences in the time domain can not be extract-ed sufficiently, and it is not possible to extract the whole process of the behavioral features. When the convolution kernel combinations are (7×1,9×1,11×1) and (9×1,11×1,13×1), the receptive field is larger, which ignores some details in the skeleton frame sequences and extracts some irrelevant information, resulting in a decrease in the overall accuracy of the recognition of various behaviors.

In summary, multiple TCN branching modules are used to extract the dependencies between different durations, and smaller convolutional kernels are used to extract features in shorter time periods, and larger convolutional kernels are used to extract features in longer time periods. The last step is to utilize the features in a more efficient way. In order to better utilize these features, the features of different durations are finally fused to solve the dependency problem between different durations and improve the performance of the model.

In order to verify whether the introduction of the channel attention mechanism can improve the recognition accuracy, ablation experiments were also conducted, and the experimental results are shown in Table 2.

Table 2. Comparison of recognition accuracy with and without attention branching.

Methodologies	**Accuracy/%**
Attention included	94.7
Attention not included	93.9

The experimental results show that the recognition accuracy is higher with the attention branch, and the recognition accuracy is improved by 0.8% on the private dataset of electric power operations, which proves that the addition of the attention mechanism is effective. The added attention branch can dynamically adjust the weights of each TCN branch according to the input feature maps, and sum up the output feature maps in a nonlinear way, so that the output feature maps have stronger expressive ability, effectively differentiate between different action categories, and improve the recognition accuracy.

In order to verify the effectiveness of the second-order skeletal information on behavior recognition, joint flow, skeletal flow and dual-flow methods are used for comparison experiments, as shown in Table 3, the dual-flow meth-od is better than the single-flow method.

Table 3. Comparison of recognition accuracy for different input methods.

Methodologies	**Accuracy/%**
Stream of bone	92.6
Stream of joint	93.2
Two-stream	**94.7**

In order to prove the overall performance of the model, a comparison experiment with other graph convolution-based behavior recognition methods was conducted on a private

Table 4. Comparison of recognition accuracy of different methods.

Methodologies	Accuracy/%
ST-GCN	86.7
2S-AGCN	92.8
AS-GCN	90.5
Ours	**94.7**

dataset of electric power operations, and the pro-posed method has the highest recognition accuracy, as shown in Table 4.

Figure 11 shows the visualization results of behavior recognition, and three representative frames of various behaviors are selected for display. It can be seen that five kinds of illegal behaviors, such as climbing equipment, climbing over fences, smoking, stepping on equipment and making phone calls, can be accurately recognized by operators at the electric power operation site.

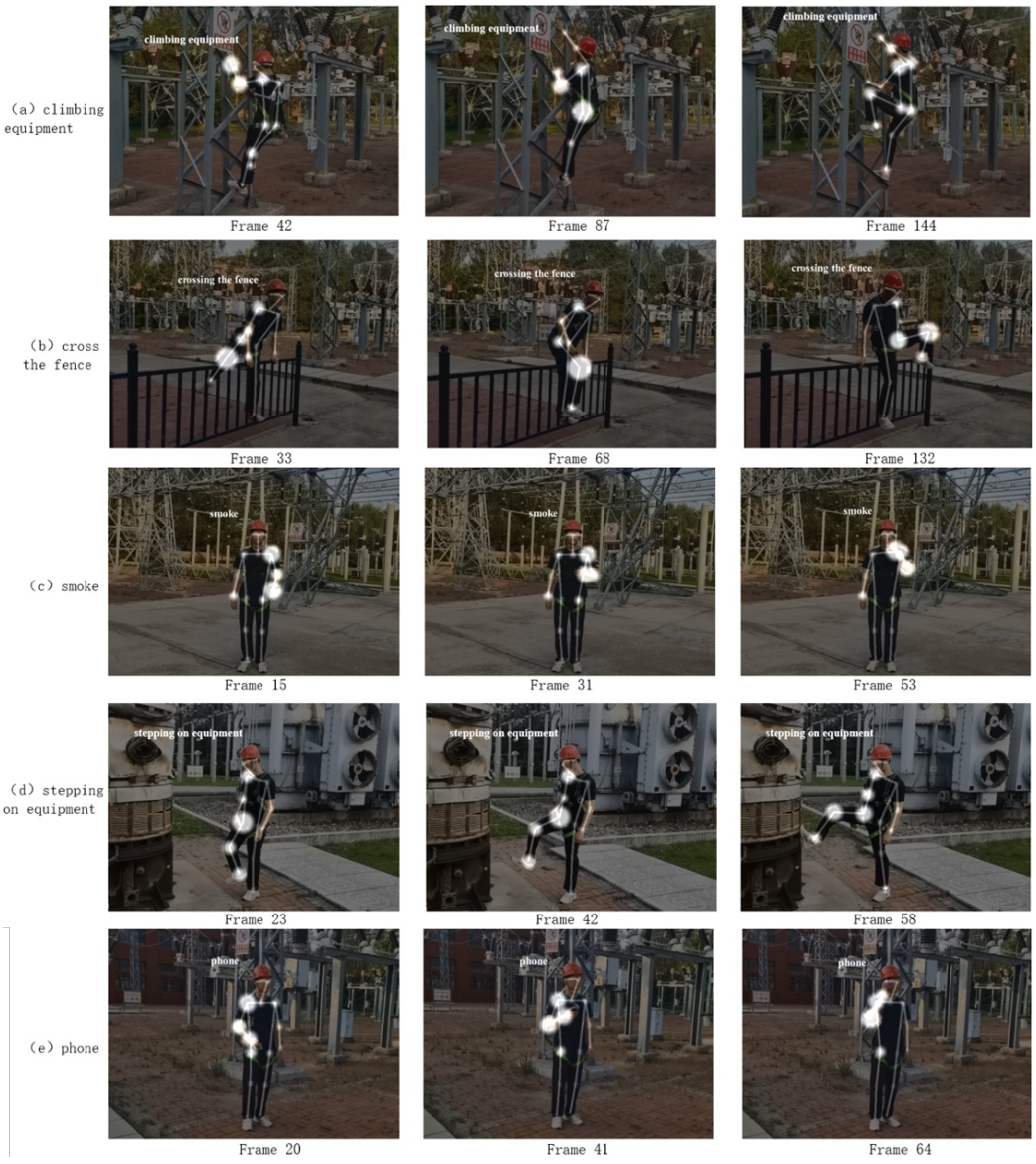


Figure 11. Visualization results of power operator violation identification

5. **Conclusion.** Electricity is the lifeblood of the country, and it is of great significance to standardize the operation, and the safety of the project is a matter of life and death. We propose a recognition method based on enhanced spatio-temporal graph convolution with dual-flow structure for the violation behaviors of electric power operators. Firstly, we use YOLOv5 to detect electric power operators, and then we construct a violation behavior recognition model with dual-flow structure by using adaptive graph convolution and designing a multi-scale temporal convolution module. The method can accurately identify the violation behaviors of electric power work personnel climbing equipment, climbing over fences, smoking, stepping on equipment, and making phone calls, which effectively reduces the probability of accidents at the work site. At present, the regulation of personal protective equipment for electric power operators has tended to be perfect, and this paper also proposes a new regulatory approach from the perspective of identification of violations by operators, but with the continuous complexity of the electric power system, the standardization of the operation process of the operators has also put forward higher requirements, the next step is to prepare for the combination of electric power operation tools for the installation of insulating brackets, hanging high-voltage grounding wires and other complex operational processes to design identification and scoring methods.

Acknowledgment. This work was supported by a grant from the Jilin Science and Technology Development Program Project (Grant No. 20240103018).

REFERENCES

[1] H.-H. Chen and J.-S. Guan, "Teacher–student behavior recognition in classroom teaching based on improved YOLO-v4 and Internet of Things technology," *Electronics*, vol. 11, no. 23, pp. 3998, 2022.
[2] H. Wang and C. Schmid, "Action recognition with improved trajectories," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551-3558, 2013.
[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 27, 2014.
[4] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1933-1941, 2016.
[5] S.-J. Berlin and M. John, "Spiking neural network based on joint entropy of optical flow features for human action recognition," *The Visual Computer*, vol. 38, no. 1, pp. 223-237, 2022.
[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489-4497, 2015.
[7] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5533-5541, 2017.
[8] C. Feichtenhofer, H.-Q. Fan, J. Malik, and K.-M. He, "Slowfast networks for video recognition," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202-6211, 2019.
[9] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[10] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026-12035, 2019.
[11] S. Bai, J.-Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132-7141, 2018.
[13] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "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, 2019.

[14] W. Yan, X. Wang, and S. Tan, "YOLO-DFAN: Effective high-altitude safety belt detection network," *Future Internet*, vol. 14, no. 12, pp. 349, 2022.

[15] N.-K. Anushkannan, V.-R. Kumbhar, S.-K. Maddila, C.-S. Kolli, B. Vidhya, and R.-G. Vidhya, "YOLO algorithm for helmet detection in industries for safety purpose," *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, IEEE, pp. 225-230, 2022.

[16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291-7299, 2017.

[17] J.-D. Wang, K. Sun, T.-H. Cheng, B.-R. Jiang, C.-R. Deng, Y. Zhao, D. Liu, Y.-D. Mu, M.-K. Tan, X.-G. Wang, W.-Y. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349-3364, 2020.

[18] K. Zhou, T. Wu, C. Wang, J. Wang, and C. Li, "Skeleton-based abnormal behavior recognition using spatio-temporal convolution and attention-based LSTM," *Procedia Computer Science*, vol. 174, pp. 424-432, 2020.

[19] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2990-3001, 2020.

[20] F. Zhang, T.-Y. Wu, J.-S. Pan, G.-Y. Ding, and Z.-Y. Li, "Human motion recognition based on SVM in VR art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, no. 40, pp. 1-15, 2019.