applied sciences

MDPI

Article

Chinese Traffic Police Gesture Recognition Based on Graph Convolutional Network in Natural Scene

Kang Liu [1,†], Ying Zheng [1,†], Junyi Yang [2], Hong Bao [3,*] and Haoming Zeng [1]

1 School of Mechanical Electronic & Information Engineering, China University of Mining & Technology-Beijing, Beijing 100083, China; kangliu@cumtb.edu.cn (K.L.); buu_zhengying@126.com (Y.Z.); haoming_zeng@126.com (H.Z.)
2 Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100049, China; yjy7894@126.com
3 College of Robotics, Beijing Union University, Beijing 100101, China
* Correspondence: baohong@buu.edu.cn
† These authors contributed equally to this work.

Abstract: For an automated driving system to be robust, it needs to recognize not only fixed signals such as traffic signs and traffic lights, but also gestures used by traffic police. With the aim to achieve this requirement, this paper proposes a new gesture recognition technology based on a graph convolutional network (GCN) according to an analysis of the characteristics of gestures used by Chinese traffic police. To begin, we used a spatial–temporal graph convolutional network (ST-GCN) as a base network while introducing the attention mechanism, which enhanced the effective features of gestures used by traffic police and balanced the information distribution of skeleton joints in the spatial dimension. Next, to solve the problem of the former graph structure only representing the physical structure of the human body, which cannot capture the potential effective features, this paper proposes an adaptive graph structure (AGS) model to explore the hidden feature between traffic police gesture nodes and a temporal attention mechanism (TAS) to extract features in the temporal dimension. In this paper, we established a traffic police gesture dataset, which contained 20,480 videos in total, and an ablation study was carried out to verify the effectiveness of the method we proposed. The experiment results show that the proposed method improves the accuracy of traffic police gesture recognition to a certain degree; the top-1 is 87.72%, and the top-3 is 95.26%. In addition, to validate the method's generalization ability, we also carried out an experiment on the Kinetics–Skeleton dataset in this paper; the results show that the proposed method is better than some of the existing action-recognition algorithms.

Keywords: graph convolution network; attention mechanism; traffic police gesture recognition

Citation: Liu, K.; Zheng, Y.; Yang, J.; Bao, H.; Zeng, H. Chinese Traffic Police Gesture Recognition Based on Graph Convolutional Network in Natural Scene. *Appl. Sci.* **2021**, *11*, 11951. https://doi.org/10.3390/app112411951

Academic Editor: Athanasios Nikolaidis

Received: 23 November 2021
Accepted: 12 December 2021
Published: 15 December 2021

1. Introduction

In just over a decade, automated driving technology [1,2] has achieved an impressive breakthrough in theoretical research and technology application, and automated driving vehicles are now regarded as a research hotspot by universities and research institutions worldwide. At present, unmanned vehicles with automated driving technology can achieve good autodriving functions in simple, closed-road environments. However, the existing technology does not have the ability to understand complex and uncertain road scenes as human drivers can; for example, the scenes with bad weather, such as heavy snow or fog; irregular road situations, such as ponding water and narrow paths; and special road scenes, such as emergencies and multivehicle confluence. The core problem is that human-like understanding and interaction cognition in complex environments are difficult to achieve, which seriously influences the safety and reliability of vehicles. However, long into the future, automated driving vehicles and human-driven vehicles will coexist. If the core problem we mentioned above cannot be solved, it will be very difficult for automated driving vehicles to reach level L4 and above; these levels were set by J3016 [3].

(J3016 is the taxonomy for driving automation proposed by the Society of Automobile Engineers (SAE) International. It defines six levels of driving automation, ranging from no driving automation (Level 0) to full driving automation (Level 5). Therefore, the key to solving this problem is the creation of interaction cognition technology, which should aim to understand traffic police body language in road driving.

It is hard to restore order with traffic signals and lights alone in emergency situations such as traffic jams and pedestrians' retrograde in junctions. These types of situations are dangerous to both humans and cars; therefore, flexible gestures used by traffic police are needed to help resume traffic flow. According to [4], under testing situations, driverless vehicles should have the ability to understand traffic command gestures. In other words, automated vehicles must recognize the gestures of traffic police correctly, and then, make a driving decision that fits the gesture given by police in real time. If the communication between automated driving vehicles and traffic police is unclear, or specifically, the vehicle cannot recognize the gestures of traffic police, this means the vehicle cannot make good decisions, which can cause accidents.

The gestures given by Chinese traffic police are made and governed by the Ministry of Public Security of the People's Republic of China. According to "Road Traffic Safety Law of the People's Republic of China" [5] and its enforcement regulations, there are eight kinds of new traffic police gesture signals, as shown in Figure 1, which are stop, go straight ahead, turn left, turn left waiting, turn right, lane change, slow down, and pull over. The gestures used by traffic police are a set of normative gesture actions that are mainly actions of the upper part of the body. The gestures used by traffic police consist of a series of continuous actions; if one was to only distinguish gestures using their spatial features, this may lead to an incorrect interpretation. Figure 2 displays two examples of different gestures (stop and pull over); however, these two gestures have similar spatial features. On this occasion, temporal consistency also aids effective recognition. Thus, spatial features and temporal features should both be considered in the recognition of gestures used by traffic police.

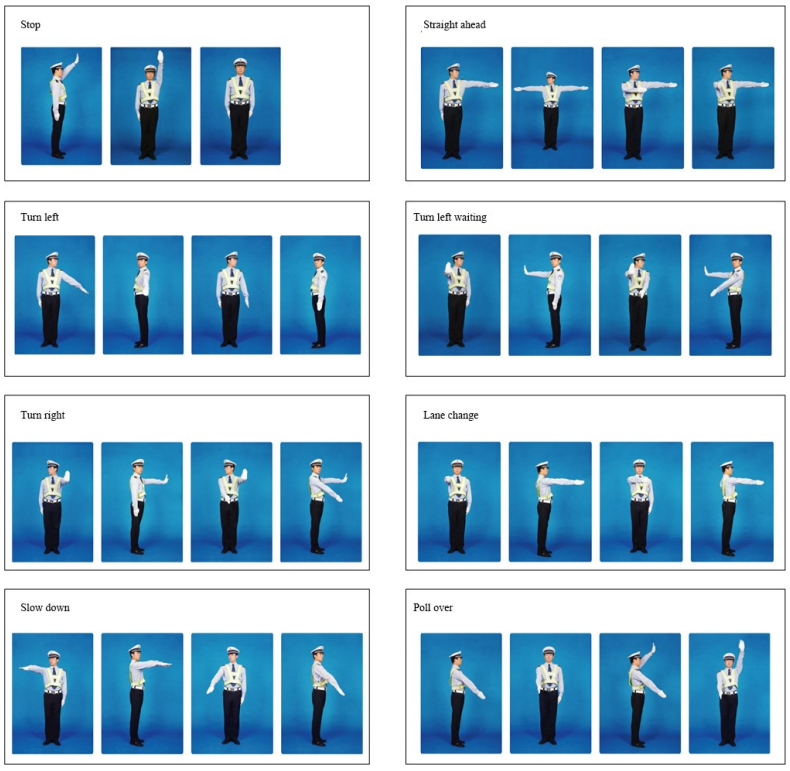

Figure 1. Actions used in gestures used by Chinese traffic police.

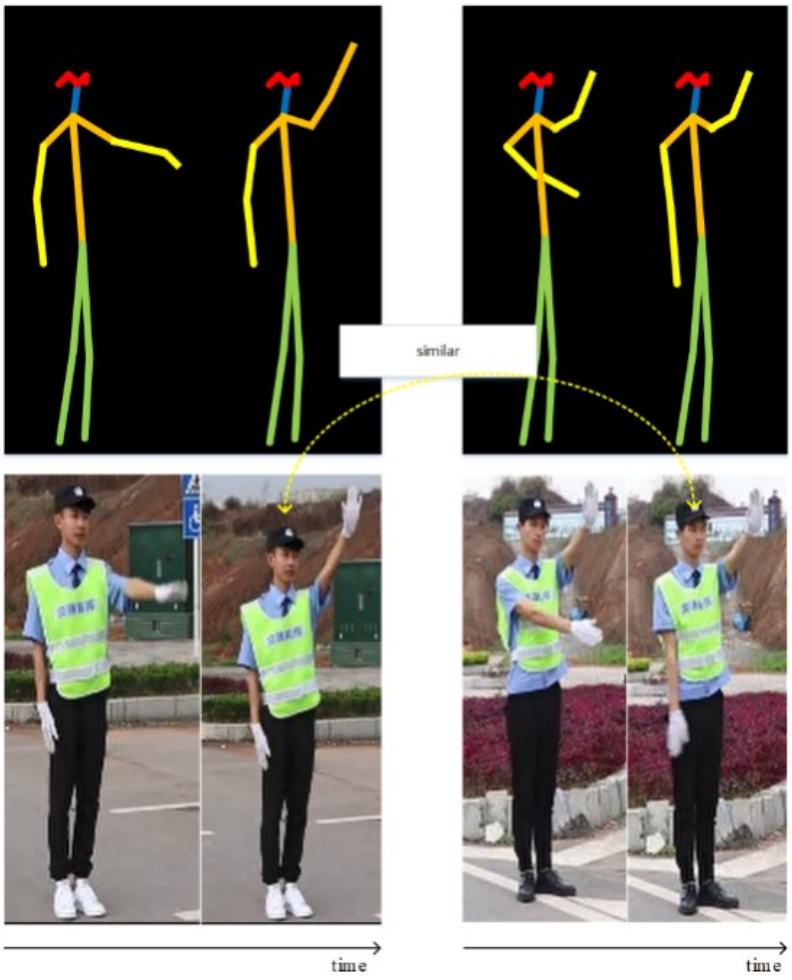

Figure 2. Two actions with similar spatial features belonging to different gestures.

Frequently used traffic police gesture recognition methods can roughly be divided into contact methods, based on external devices, and noncontact methods, based on computer vision. Contact methods based on external devices can recognize gestures used by traffic police with high speed and precision, but in this method, traffic police officers need to wear extra sensors, and the operation process is cumbersome, which increases the workload of traffic police. In addition, the equipment is expensive, which also limits the promotion of the application of this kind of method. Noncontact traffic police gesture recognition methods based on computer vision mainly capture picture data or video data using vision sensor equipment such as high-definition cameras to recognize gestures used by traffic police through digital image processing. These types of methods have an advantage in that the recognition process can be carried out without traffic police having to wear any extra equipment. However, the existing traffic police gesture recognition methods based on computer vision simply compile gestures into simple sequence data or pseudo-images, which leads to the problem of the original features of the image data being lost. To summarize, there are two main issues present in existing traffic police gesture recognition technologies that should be addressed: one challenge is the construction of a traffic police gesture dataset, as there has been no complete and public traffic police dataset until now. The other challenge is the question of how to extract information regarding gestures used by traffic police effectively. As the urban road transport environment is unpredictable, the extraction of traffic police gesture features in video images may be influenced by many factors such as light conditions, road conditions, etc., which directly affect the efficiency of traffic police gesture recognition.

To solve the problems outlined above, this paper designs a solution for use in the collection of a traffic police gesture dataset on real roads and established a dataset which is reasonable and valid. In addition, this paper proposes a traffic police gesture action technology based on the GCN. First, the skeleton joint data of traffic police images were obtained using the OpenPose algorithm. Then, an ST-GCN base network was used to collect the spatial–temporal features of gestures used by traffic police. As the skeleton map structure of the ST-GCN is predefined and only represents the physical structure of the human body, it cannot effectively characterize the logical association factors between gestures used by traffic police. Thus, this paper proposes the use of an AGS model to extract the associated feature information in the node data of the skeletons traffic police members. In addition, this paper proposes a TAS to extract the temporal sequence feature information regarding gestures used by traffic police in the temporal dimension. In addition, joints have different levels importance in different gesture actions. For gestures used by traffic police, the key joints of action mainly focus on the six joints of the arm, which leads to the problem of imbalanced distribution of information in the feature space. Thus, a spatial attention mechanism was introduced into the improved ST-GCN model, and the attention module was able to adjust the weight of key gesture joints dynamically, choose the effective gesture node feature information, and improve the accuracy of the recognition of gestures used by traffic police.

This paper extracted features of gestures used by traffic police with an improved ST-GCN model and recognized actions in the gestures of traffic police. The contributions of this paper are as follows:

1. We captured data regarding the gestures used by traffic police in real urban roads and established a complete traffic gesture dataset, including actions used in the gestures of traffic police on duty information regarding their labeling.
2. We addressed the problems that the skeleton graph of the ST-GCN basic model only represents the physical structure of the human body and that the global information regarding the gestures used by traffic police is lost; an AGS was proposed to extract the associated feature information regarding traffic police's skeleton joint data.
3. Features in the temporal dimension are crucial to the recognition of gestures used by traffic police; the TAS was proposed in the AGS to collect the temporal sequence feature information regarding gestures used by traffic police in a time series.

2. Related Work

2.1. Traffic Police Gesture Recognition

The worldwide use of traffic police gesture recognition technology can roughly be divided into two kinds of methods: methods based on external devices and methods based on computer vision. Methods based on external devices collect gesture features by placing sensor equipment on traffic police's bodies to collect traffic information regarding actions used in gestures. For example, Ref. [6,7] both collected important information such as the motion trail of traffic police's arms and the position of hands, etc., through accelerometers, then recognized the actions of gestures used by traffic police according to these. Traffic police gesture recognition methods based on computer vision are noncontact methods which do not require equipment to be worn and are convenient. Literature [8,9] has proposed a method that can be used to recognize gestures used by traffic police in complex environments. First, dark channel prior and kernel density estimation are used to extract the torso and arms of the police as the foreground region. Then, the coordinates of a pixel in the upper arms and forearms of traffic police are determined by the max-covering scheme method. Finally, the pose and action of the traffic police are recognized through the geometrical relationship of the rotation joint. Guo et al. [10] proposed a five-part body model to recognize gestures used by traffic police in complex scenes; it differed greatly from the former methods which required a pretraining process or 3D measuring equipment to construct a human body model. They used the max-covering scheme to learn the five-part body model automatically. Literature [11,12] has proposed a method based on cumulative

block intensity vector (CBIV) using the n-frame cumulative difference to collect the features of traffic police.

Compared with the action recognition methods based on video image data, skeleton-data-based action recognition technology has strong robustness when used in complex, dynamic scenes. Using sensors such as Kinect (2012) [13], ASUS Xtion PRO LIVE (2011) RGB-D cameras [14], etc., skeleton data can be collected easily. Ma et al. [15] used Kinect2.0 to collect data from 10 volunteers and constructed a traffic police command gesture skeleton joint dataset. They used convolutional operation to analyze the location change of skeleton joints to extract features in the temporal dimension and extract spatial features by analyzing the relative position of skeleton nodes at the same time. The extracted spatial–temporal features of gestures used by traffic police were used to train the ST-CNN model and achieve the recognition of gestures used by traffic police. Zhang C. et al. [16] constructed a traffic police gesture model based on joints and skeletons after analyzing the hinged feature of gestures used by traffic police. Furthermore, a convolutional pose machine (CPM) was introduced to extract the key joints of gestures used by traffic police to collect the relative length of the skeleton in gestures used by traffic police and its angle between acceleration of gravity as the spatial context feature. Meanwhile, long short-term memory (LSTM) was used to extract the features of gestures used by traffic police in temporal series. At last, the spatial features and temporal features were integrated to recognize gestures used by traffic police.

As the skeleton joint graph has the characteristic of graph structure data, the ST-GCN [17] was the first to use the GCN model to collect skeleton data concerning spatial–temporal features in a human skeleton topology graph, which achieved the aim of recognizing human actions. Most of the following recognition models were designed and modified based on the GCN model, for example, actional–structural graph convolutional networks (AS-GCNs) [18], two-stream adaptive graph convolutional networks (2S-AGCNs) [19], and channelwise topology refinement graph convolution (CTR-GCN) [20].

2.2. Graph Convolutional Neural Networks

With regard to graph structure data, the present model of transferring the graph into a set of vectors in the data preprocessing stage cannot guarantee the integrity of graph structure information, and the obtained results heavily depend on the results of graph preprocessing. Marco Gori [21] et al. first proposed the graph neural network (GNN) model, constructed the learning process on graph data directly, and mapped nodes and edges of a graph to a low-dimensional space through the model. The spectral network [22] was the first method used to introduce convolution into the GNN, which defined convolution operation in the Fourier domain, but this method leads to potential intensive operation, and the convolution kernel has no characteristic of locality. Mikael Henaff [23] et al. introduced parameters with smoothing factors, which afforded locality to the convolution kernel of the spectral network. The graph convolution kernel defined by spectral network depends on the Laplacian matrix of the graph; thus, parameters cannot be shared in other graphs, and the complexity of network calculation is high. Michael Defferrard [24] et al. proposed ChebNet based on the polynomial convolution kernel, which greatly improved computational efficiency. Adaptive graph convolutional neural networks (AGCNs) [25] learn not only the relationships between the nodes of the original graph structure but also the possible potential relationships between the nodes. However, the convolution kernels of the abovementioned models depend on the basis vector of the Laplacian matrix's features, which depend on the structure of the graph. That means that this kind of model is trained for specific structures, cannot be applied to different structure graphs directly, and has poor generalization capability. At the same time, methods based on spectral decomposition need to process the complete graph in the calculation. There is a high time complexity in matrix decomposition, which makes it difficult to extend to large-scale graph data to learn. David Duvenaud [26] et al. proposed neural FP, which uses different weight matrices for nodes with different amounts of degree. However, when

the scales of nodes are large and the degrees of nodes are various, this model cannot be applied because there are too many parameters. Mathias Niepert [27] et al. tried to sort the nodes, selected a fixed amount of neighbor nodes, and convolved them, imitating the method of the convolutional network, which transferred the learning problem of graph structure into the traditional Euclidean data learning problem. Federico Monti [28] proposed MoNet, which used pseudo-coordinates to transform GNN models into Gaussian kernel hybrid models. Geodesic convolutional neural networks (GCNN) [29], anisotropic CNN (ACNN) [30], GCN [31], diffusion-convolutional neural networks (DCNN) [32], etc., can be regarded as MoNet models. William Hamilton [33] et al. proposed a universal inductive reasoning framework that sampled and aggregated neighbor node features to generate a representation of a node.

3. Methods

3.1. Traffic Police Gesture Recognition Network

The traffic police gesture recognition method proposed in this paper was based on the ST-GCN. Figure 3 shows the complete process of traffic police gesture recognition:

1. Use openpose algorithm to process collected videos containing gestures used by traffic police; extract skeleton data.
2. Construct the AGS on the collected skeleton data, extract the spatial–temporal features of gestures used by traffic police with the improved ST-GCN, and recognize gestures used by traffic police.

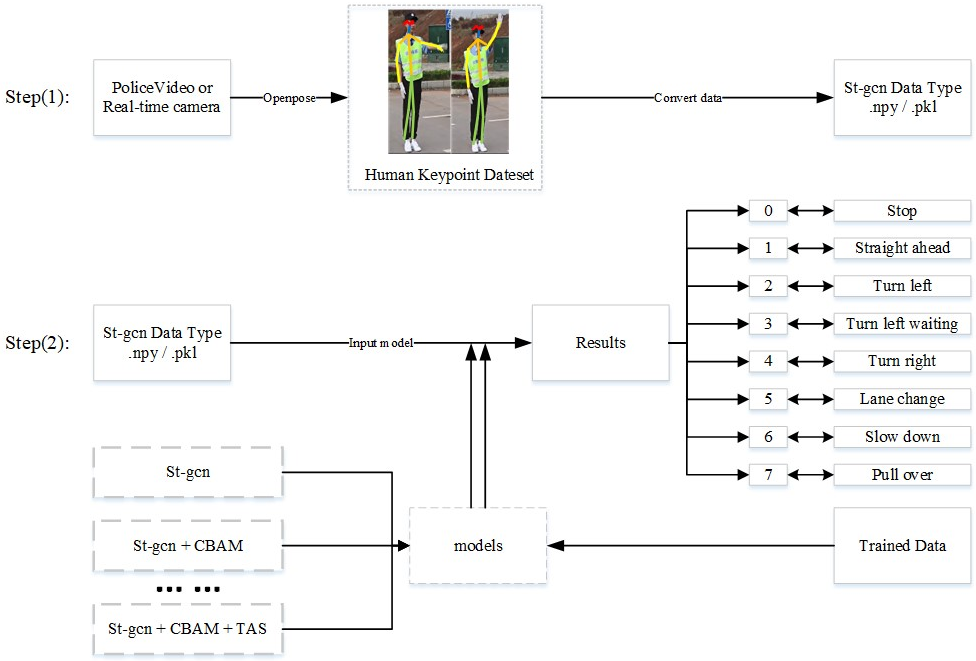

Figure 3. The complete process of traffic police gesture recognition.

The traditional ST-GCN model designed a universal skeleton representation for action recognition: a spatial–temporal skeleton graph. The nodes of the skeleton graph are human joints, and edges in the graph have spatial–temporal features. Edges can be divided into two categories: the connection between the human joints using bones is the first kind of edge, and the connection of the same node in different frames is the second kind of edge, as shown in Figure 4. In the spatial–temporal skeleton graph of the ST-GCN, the second kind of edge connects the same joints, which have a linear relationship, meaning it overcomes the requirement for the manually designed extraction of joint features or the design of traversal rules of joints in traditional methods.

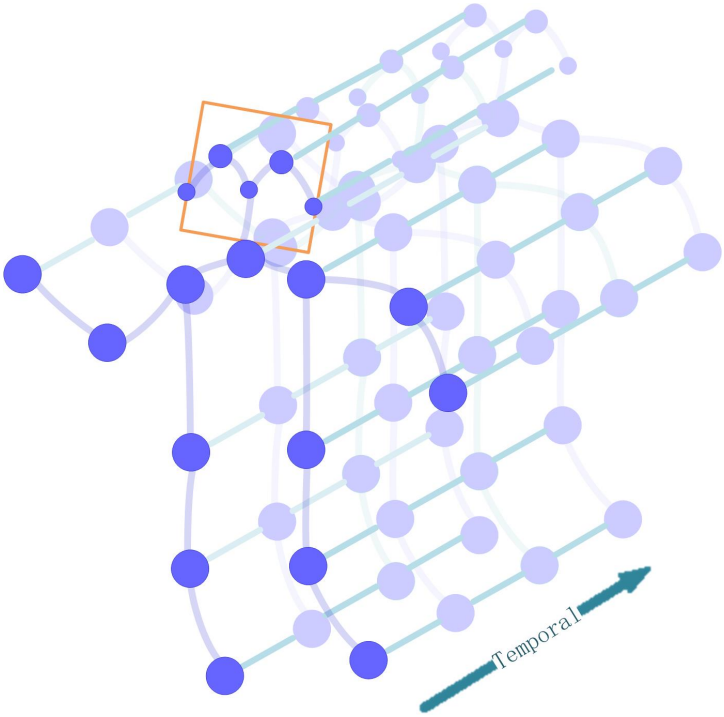

Figure 4. The spatial–temporal graph of skeleton sequence.

However, there are some drawbacks to the ST-GCN method, which are as follows:

1. The skeleton graph, which is the physical structure of the human body, is predefined. It cannot ensure that it is the best skeleton graph for traffic police gesture recognition; for example, the dependency between the two hands of a traffic police member cannot be captured.
2. There are layers in the structure of GCN; different semantics are contained in different layers. However, the graph topology structure of the ST-GCN model is fixed in all the network layers and extracts the feature of the same topology, which causes them to lose the flexibility of information modeling.
3. A fixed graph structure cannot be the optimal representative method for different gesture actions; joints have different levels of importance in different gestures.

Based on the analysis above, this paper extended the topology structure of the traffic police skeleton graph and learned features of gestures used by traffic police from the spatial dimension and temporal dimension. The attention mechanism was also introduced to learn the importance of each joint in different traffic gestures. The network model of traffic police gesture recognition proposed in this paper is shown in Figure 5.

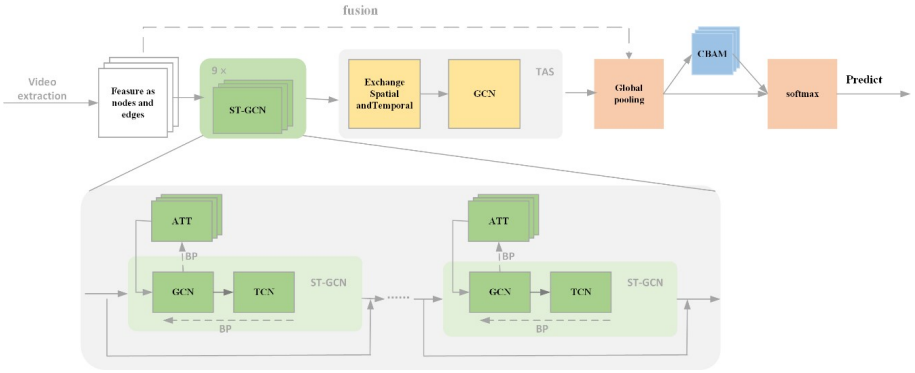

Figure 5. The architecture of traffic police gesture recognition network.

3.2. Spatial Graph Convolution

To better explain the module, here, we describe it from skeleton data in a single frame. The skeleton graph of the ST-GCN in the same frame only represents the physical structure of the human body, which is not the optimal representation with regard to gestures used by traffic police. For example, the spatial locations of hands are crucial in traffic police gesture semantic representation, which have potential contact. Thus, in this paper, an AGS was constructed to find the potential features in human joints. The left figure in Figure 6 shows the connection method of the ST-GCN; the right figure of Figure 6 is the spatial–temporal graph extended by this paper, which connects several neighbor joints. The skeleton graph constructed in this paper is a nondirectional spatial–temporal graph in the same frame, $G =< V, E >$, which contains skeleton sequence data of N joints and T frames. The set for nodes is shown in Equation (1):

$$V = \{v_{ti} | t = 1, \cdots, T, i = 1, \cdots, N\} \tag{1}$$

In Equation (1), v_{ti} represents the i joints in t frame, t means the serial number of frame, and i means the number of joints. There are two kinds of edges in the skeleton graph, and no manual distribution is added in the process of connecting edges, which means that the model can deal with more datasets and improves its universality.

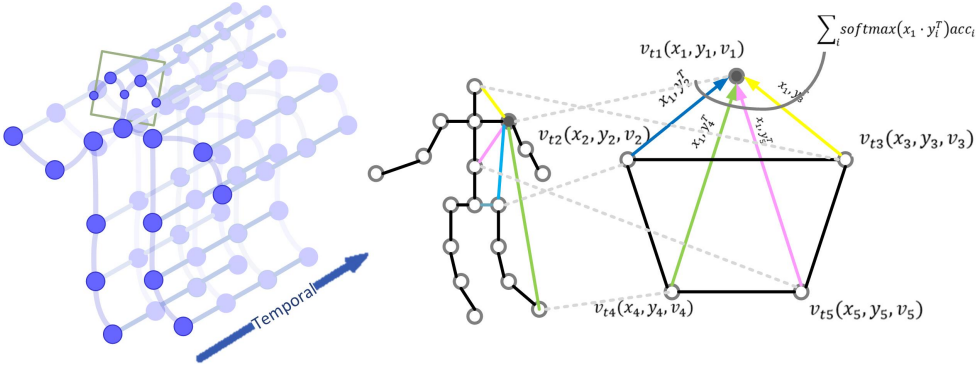

Figure 6. The proposed skeleton graph on spatial dimension.

In the traditional convolution method, the output of a single channel in the spatial position x can be represented as Equation (2):

$$f_{out}(X) = \sum_{h=1}^{K} \sum_{w=1}^{K} f_{in}(P(x, h, w)) \cdot W(h, w) \tag{2}$$

The size of convolution kernel is $K \times K$, f_{in} is input feature graph, $P(\cdot)$ is the sampling function, and $W(\cdot)$ is the weight function. For the sampling function, images have the same number of neighbor pixels. Sampling function $P(\cdot)$ means the selection of the surrounding neighbor pixels centered on the X pixel. However, in skeleton graphs, the number of neighbor nodes is different: the neighbor nodes set is defined by Equation (3):

$$B(v_{ti}) = \{v_{ti} | d(v_{tj}, v_{ti}) \leq D\} \tag{3}$$

In Equation (3), $d(v_{tj}, v_{ti})$ represents the shortest path from skeleton joints v_{tj} to v_{ti}; take $D = 1$. Sampling function $P : B(v_{ti}) \rightarrow V$ is shown in Equation (4):

$$p(v_{ti}, v_{tj}) = v_{tj} \tag{4}$$

The sampling function in this paper takes the self-learning method to avoid loss in long-distance cross-information from partial convolution, for example, cooperating information between hands.

Compared with the sampling function, the weight function is more complicated. A rigid mesh exists around the center of the image, and neighbor pixels have a fixed spatial sequence. However, there is no potential sequence in skeleton graphs. The skeleton graph in one frame is shown as in Figure 7a, in which root joints were painted in red. The ST-GCN proposed three partitioning strategies, which were Uni-labeling (Figure 7b), Distance partitioning (Figure 7c), Spatial configuration partitioning (Figure 7d), respectively. Here, we use the spatial configuration partitioning strategies from the the ST-GCN model, as shown in Figure 7d. This strategy divides the set into three subsets: root joints themselves (green nodes in Figure 7d), joints closer to the skeleton center of gravity than root joints (blue nodes in Figure 7d), and joints further from the skeleton center of gravity than root joints (yellow joints in Figure 7d). The specific indication is shown in Equation (5).

$$l_{ti}(v_{tj}) \begin{cases} 0 & if \quad rj = ri \\ 1 & if \quad rj < ri \\ 2 & if \quad rj > ri \end{cases} \tag{5}$$

Weight function is shown in Equation (6):

$$w(v_{ti}, v_{tj}) = w'(l_{ti}(v_{tj})) \tag{6}$$

Taking the redefined sampling function and weight function in Equation (2), graph convolution in the frame can be obtained as in Equation (7):

$$f_{out}(v_i) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(P(v_{ti}, v_{tj})) \cdot W(v_{ti}, v_{tj}) \tag{7}$$

In Equation (7), $Z(\cdot)$is the regular term and the cardinal number of the relative subset.

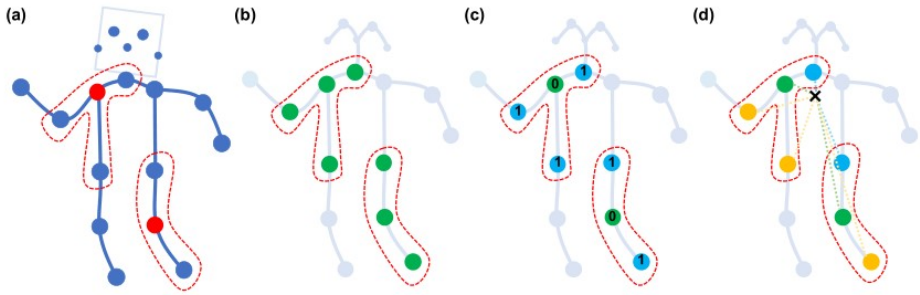

Figure 7. Partitioning strategies for constructing convolution operations.

The movement trail of traffic police gestures mainly occurs on the upper body of the human. Different joints have different levels of influence in each gesture. For example, the joints of the arm are more important because the range of motion is larger. This leads to the number of less-effective or noneffective features being far larger than the number of effective features, affecting the recognition of gestures used by traffic police. Based on this, inspired by the process of attention mechanisms, this paper introduces the convolutional block attention module (CBAM) [34]. The CBAM module enhances the effective features in the channel domain and spatial domain at the same time, while the attention mechanism in the channel domain adds weight to features in each channel; the value of weight means the importance of the channel feature, that is, which channels in the input are valid. For simplicity's sake, the operation in the spatial domain can be understood as finding the key joint of the graph and extracting effective feature information.

As shown in Figure 8, in the channel domain, to model the dependency between channels and aggregate spatial features, we used average pooling and max pooling to obtain two different channel background descriptions, F^c_{avg} andF^c_{max}. Then, we used the

multilayer perceptron to calculate the two different channels above and summed them element-by-element. Finally, we generated the final channel attention traffic police gesture feature graph M_C through the activation function, as shown in Equation (8):

$$\begin{aligned} M_C(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1(W_0(F^c_{max}))) \end{aligned} \tag{8}$$

In the equation above, σ is ReLU activation function, W_0 is the first layer of the fully connected layer, and the length of the output vector is $r \times C$. W_1 is the second layer of the fully connected layer, and the length of the output vector is C.

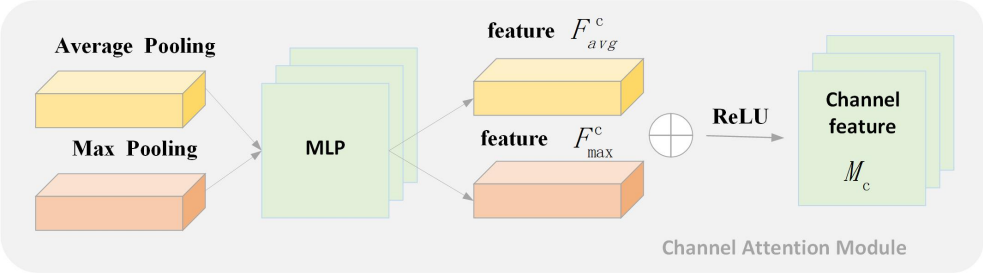

Figure 8. Channel attention module.

As shown in Figure 9, to extract feature information regarding key nodes from the spatial domain, first average pooling and max pooling must be taken with regard to the input traffic police gesture features. Then, the feature extracted must be spliced and the convolution taken once to ensure the output traffic police gesture features are in the same dimension as the input traffic police gesture features. The calculation of spatial attention mechanism (M_s) is shown in Equation (9):

$$M_S(F) = \sigma\left(f^{7*7}([AvgPool(F); MaxPool(F)])\right) = \sigma(f^{7*7}(\left[F^c_{avg}; F^c_{max}\right])) \tag{9}$$

σ represents Sigmoid function, and f^{7*7} means that a convolution kernel with a size of 7×7 is used in the convolution layer.

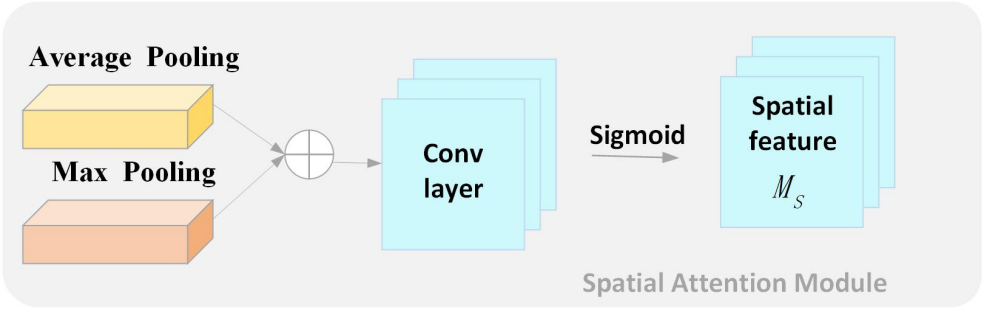

Figure 9. Spatial attention module.

3.3. Temporal Graph Convolution

ST-GCN methods only connect the same joints in different frames in the temporal dimension, which can only extract the trace relationship of the same joints in neighbor frames. However, gestures used by traffic police consist of a set of different actions, so simply extracting the relative action features in neighbor frames is not enough to describe the semantics of whole gestures. Here, we connect the same joints in several frames to a joint in one frame to better extract rich traffic police gesture features, as shown in Figure 10. The left of Figure 10 is the temporal graph of the ST-GCN; the right of Figure 10 is the extended skeleton graph proposed by this paper. To better explain, the right of

Figure 10 only illustrates the connection situation of the adjacent five frames; all the joints are connected similarly in real operation.

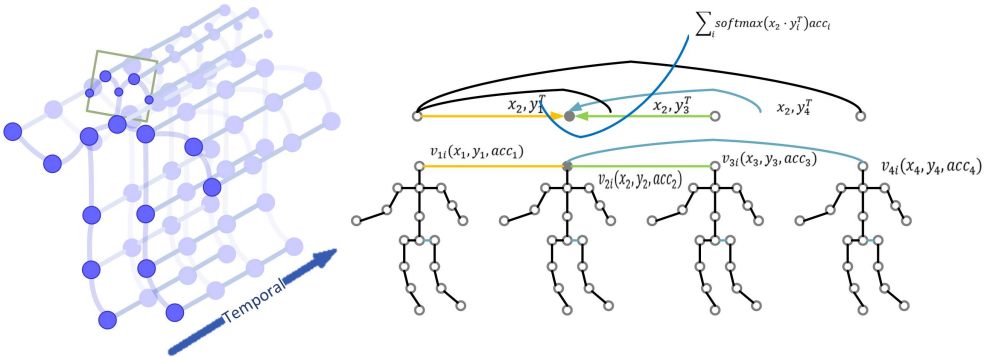

Figure 10. The proposed skeleton graph on temporal dimension.

Here, we designed a TAS module; its structure is shown in Figure 11. It was used to extract traffic police gesture features in the temporal dimension. The TAS studies the dynamic features of each joint through every frame separately in the temporal dimension. Every joint is recognized as independent from one another; the TAS collects the relevance between frames by comparing the dynamic change of the same joint in the temporal dimension and mines the features ignored in the traditional ST-GCN model.

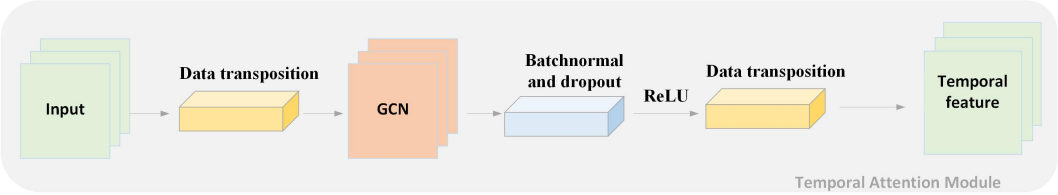

Figure 11. Temporal attention module.

This paper borrows the ideas used in the temporal convolution network (TCN) [35] and applies its mode of processing spatial features to the temporal domain. A TAS module can capture the dependency relationship in temporal sequence. It can extract partial features in temporal sequence by forming a sequence of feature vectors with feature vectors captured in each time frame. The convolution calculation is layerwise; it calculates the frames from every moment at the same time. From Equation (3), we can see that the selection of the convolution domain relies on the joints' distribution in space, and division is based on the distance of joints; convolution domain is determined by the distance between joints.

Based on these considerations, we directionally transposed the input data and exchanged the feature data in the spatial dimension and temporal dimension, which transformed (N, M, C, T, V) into (N, M, C, V, T). N is batch_size, and C is the number of channels, taking $C = 3$. T is the figure for frames, and V is the number of skeleton joints. M is the amount of people that appeared in each frame of the video, which was $M = 1$ here because there was only one traffic police member in the video of traffic police on duty. In sequence data frames, we used the downscaling operation through convolution summation, which achieved temporal graph convolution, that is

$$M_T(F)_{NCTW} = \sum_K \sum_V F_{NKCTV} \cdot A_{KVW} \tag{10}$$

N, K, C, T, V, W represent Batch_Size, KernelSize, channels, the number of frames, the number of skeleton joints, and the number of people, respectively.

Finally, the batch normal operation and dropout operation were performed on the features extracted, the amount of midfeature and data redundancy were reduced, the process of training was sped up, and the performance was improved. In the temporal dimension, the characteristics of the connection between skeleton joints are listed below:

- Unlike the connection of skeleton joints in a frame, the connection of skeleton joints in the temporal sequence itself is used as a parameter to train. Furthermore, it can self-learn through a fully connected matrix and learnable weight hyperparameters.
- The size of the convolution kernel determines the connection effect between frames and enriches the contact between frames. It can be regarded as a dynamic recognition process from a certain point of view.

4. Experiments

4.1. Dataset

4.1.1. Traffic Police Gestures Data

In a real pathway, five cameras from five directions were used to collect traffic police gesture video data, whose video window size was 840 × 480, and which was stored in the format of AVI. In total, 20,480 videos were captured; each camera filmed 512 videos for each gesture among them. Eight kinds of gestures were used by traffic police: stop, straight ahead, turn left, turn left waiting, turn right, lane change, slow down, and pull over. Eight volunteers were invited to participate during data collection, and their age, gender, and height were different; each of the participants were labeled separately in the dataset. Specifically, we used five cameras set in the same horizontal height to capture five different horizontal views for one gesture action, which were 90 degrees left, 45 degrees left, 0 degrees front, 45 degrees right, and 90 degrees right. Furthermore, each of the volunteers were asked to only complete each traffic police gesture action once. In the dataset, each camera was assigned a consistent number, and each number had its own filming angle. To further increase the diversity of the data, different conditions with regard to distances between cameras and participants, lighting conditions, junction scene, and clothes of the participants were added. In the dataset, the gestures used by traffic police were divided into a training set, validation set, and a test set in the proportion of approximately 7:2:1 for each gesture. Since there are only eight kinds of gestures in Chinese traffic police gesture, taking top-1 and top-3 to validate our method is better (top-1 means to take the largest probability vector as the predicted result. If the classification result is correct, then the prediction is correct; otherwise, the result is wrong. Top-3 means if the correct result appears in the top 3 largest probability vectors, then the prediction is correct, otherwise the result is wrong). Figure 12 shows some data samples in the dataset; Table 1 shows exact condition settings in data collection.

Figure 12. Part of the traffic police gestures dataset.

Table 1. The condition settings of data collection.

Condition	Settings
People	Eight volunteers differing in age, gender, and height
Light conditions	Four light conditions: in the morning, noon, late afternoon, and evening
Scene	Four scenes: with pedestrians, no pedestrian, T-junction, and criss-cross crossing
Clothes	Two kinds of clothes: summer clothes and winter clothes
Cameras	Five cameras from different angles
Gesture	Eight kinds of standard traffic police gesture as the national standard
Distance	5 m, 70 m
Total amount	20,480

4.1.2. Kinetics–Skeleton Dataset

Kinetics–Skeleton dataset [36–38]: Kinetics–Skeleton is divided into three parts: Kinetics-400 [36] contains 400 action categories, Kinetics-600 [37] contains 600 action categories, and Kinetics-700 [38] contains 700 action categories. Each of the actions has 400–1150 video clips, and the length of each video clip is around 10 seconds. Action types include single-person actions, for example, painting, drinking, laughing, and punching; two-person interaction actions, such as hugging, kissing, and shaking hands; and normal actions such as opening gifts, mowing the lawn, and washing dishes. There are 306,245 videos in Kinetics-400. The training set has 246,245 videos in total, 250–1000 video clips are contained for each category. There are 495,547 videos in Kinetics-600. The training set has 392,622 videos in total, and each category contains 450–1000 video clips. There are 650,317 videos in Kinetics-700. The training set has 545,317 videos in total, and each category contains 450–1000 video clips. In every part of the Kinetics–Skeleton dataset, each action category has 50 videos in the validation set, and the testing set has 100 videos for each action category. Each action category in the Kinetics–Skeleton dataset contains a kind of behavior, but a particular video may contain several kinds of actions. For example, "texting" while driving or "hula dancing" while "playing ukulele". In these cases, the video will only be labeled with one tag, thus it will not exist in two action classes at the same time. Therefore, the accuracy of the top-5 is more appropriate (top-5 means that if the correct result appears in the top 5 largest probability vectors, then the prediction is correct, otherwise the result is wrong). Some of the data in Kinetics–Skeleton are shown in Figure 13.

Figure 13. Part of the Kinetics–Skeleton dataset.

4.2. Evaluation Result

4.2.1. Implement Detail

This paper used two Nvidia GeForce GTX 1080Ti graphics cards to parallel train the model during the training process; the deep learning framework used was PyTorch1.6. The original learning rate was 0.1, the CosineAnnealing method was used to adjust the learning rate, and batch_size was set as 256. Finally, we used the information cross-entropy loss function to perform deviation iterative calculation specifically to instantiate loss function in each module and summarize all the loss. In the positive iteration process, to avoid contingency, we chose training samples randomly and made the skeleton frame data of

each video 150 frames by cutting the redundant frames or playing it repeatedly if the number of frames was not enough.

4.2.2. Ablation Study

To prove the effectiveness of the improved modules in the proposed method, we performed an ablation study on the traffic police gesture dataset. This paper improved the ST-GCN base network from the spatial dimension and temporal dimension.

This paper introduced an attention mechanism and the AGS in the spatial dimension. In Figure 14, it is shown that the model-introduced attention mechanism and the model-introduced AGS both possessed lower loss than the base ST-GCN network in the training set and validation set, and the training speed was faster. It can also be seen from Figure 15 that there was an enhancement of accuracy in every improved model.

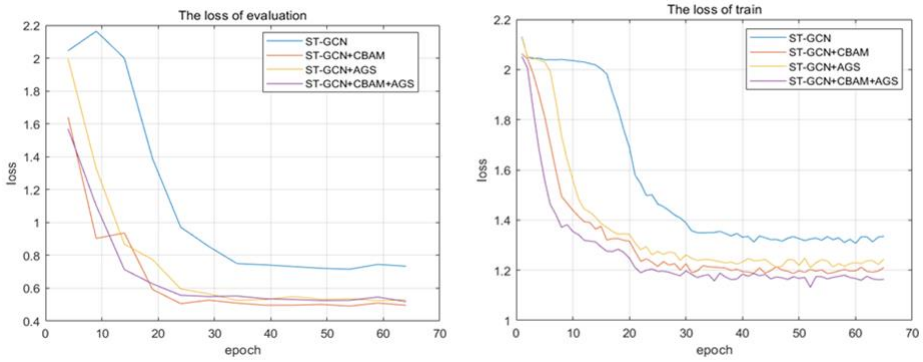

Figure 14. The trend of loss on spatial dimension.

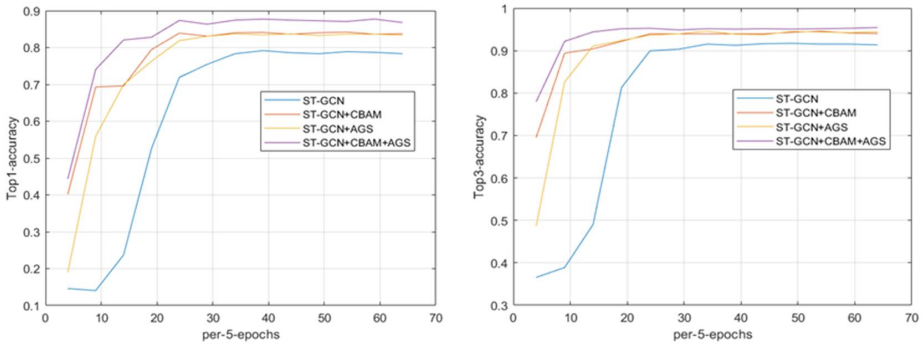

Figure 15. The trend of accuracy on spatial dimension.

Gestures used by traffic police consist of several gestures in order; their features in the temporal dimension are critical to traffic police gesture recognition. In the temporal dimension, we designed a TAS module to capture logical association feature information in the temporal dimension. As shown in Figure 16, the loss of the model that integrated the TAS module fell quicker and was more likely to reach a plateau. Figure 17 also shows that its accuracy was also much higher than the base ST-GCN network.

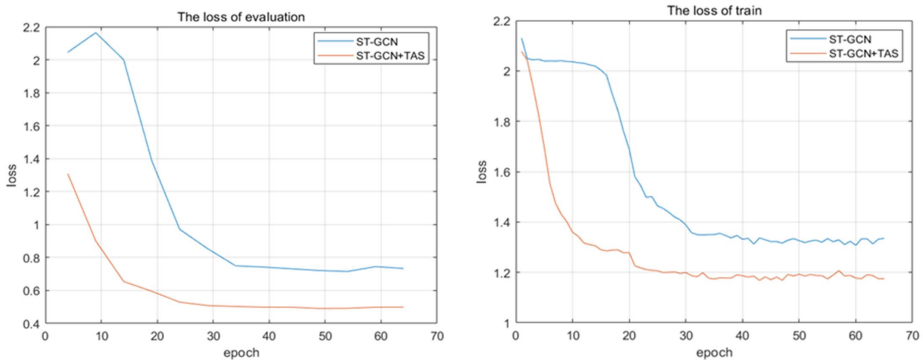

Figure 16. The trend of loss on temporal dimension.

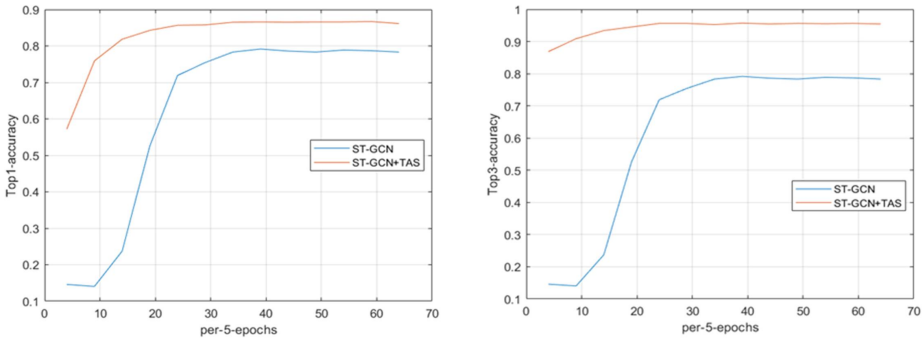

Figure 17. The trend of accuracy on temporal dimension.

As shown in Table 2, the first set of data are related data of the base network, which this paper is based on. The second set of data verify the recognition efforts of different modules related to gestures used by traffic police. The third set of data combine different modules, and the fourth set of data show the effort of the final solution proposed in this paper.

Table 2. The accuracy of traffic police gesture test set.

ID	ST-GCN	CBAM	AGS	TAS	Top-1 (%)	Top-3 (%)
1	√				79.16	91.72
2	√	√			84.19	94.6
3	√		√		83.72	94.6
4	√			√	86.7	95.72
5	√	√	√		86.05	94.6
6	√		√	√	86.5	95.44
7	√	√		√	86.7	95.72
8	√	√	√	√	87.72	95.26

Table 2 shows that networks that combine the CBAM, AGS, and TAS modules separately are better than the base network; the top-1 improved by 5.03%, 4.56%, and 7.54%, respectively. The experiment shows that the introduction of the CBAM can overcome the data imbalance distribution problem on the feature level to a certain extent, which readjusts the weight of different joints, enhances the effective features of gestures used by traffic police, and inhibits the less-effective or non-effective features. The AGS is not limited to the physical structure of humans; it can self-learn the potential relationship between skeleton joints, which improves the accuracy of traffic police gesture recognition effectively. The TAS module shows outstanding performance among all modules, because gestures used by traffic police consist of a series of actions and it can extract the trajectory features between actions, enriching the features of traffic police gesture. The experiment

results also verify the fact that information in the temporal dimension is critical in traffic police gesture recognition; while performing an ablation study on different modules, we found the accuracy of all module combinations to be improved, which further proves the effectiveness of every module. The method proposed in this paper had the highest accuracy among all, improving the top-1 by 8.56%.

Eight kinds of gestures used by traffic police consist of different actions, including repetitive actions and one-time actions. To analyze the difficulty level of each traffic police gesture recognition, experiments were performed on eight kinds of gestures used by traffic police, as shown in Table 3.

Table 3. The accuracy for every traffic police gesture.

Traffic Police Gesture	Stop	Straight Ahead	Turn Left	Turn Left Waiting
Top-1 (%)	97.96	89.68	83.09	86.90
Top-3 (%)	99.32	92.06	94.12	97.24
Traffic Police Gesture	**Turn Right**	**Lane Change**	**Slow Down**	**Pull Over**
Top-1 (%)	86.73	75.86	90.58	83.20
Top-3 (%)	97.35	95.86	98.55	92.00

In Table 3, it is shown that the recognition accuracy of lane change is relatively low because in the key gesture of the lane change gesture, arms are perpendicular to the torso, which makes the key joints coincide in the skeleton graph, and the spatial–temporal features of the gesture action are hard to extract. Figure 18 shows the similarity that exists between gestures used by traffic police. Gestures with deeper colors depict higher similarity, which makes confusion more likely.

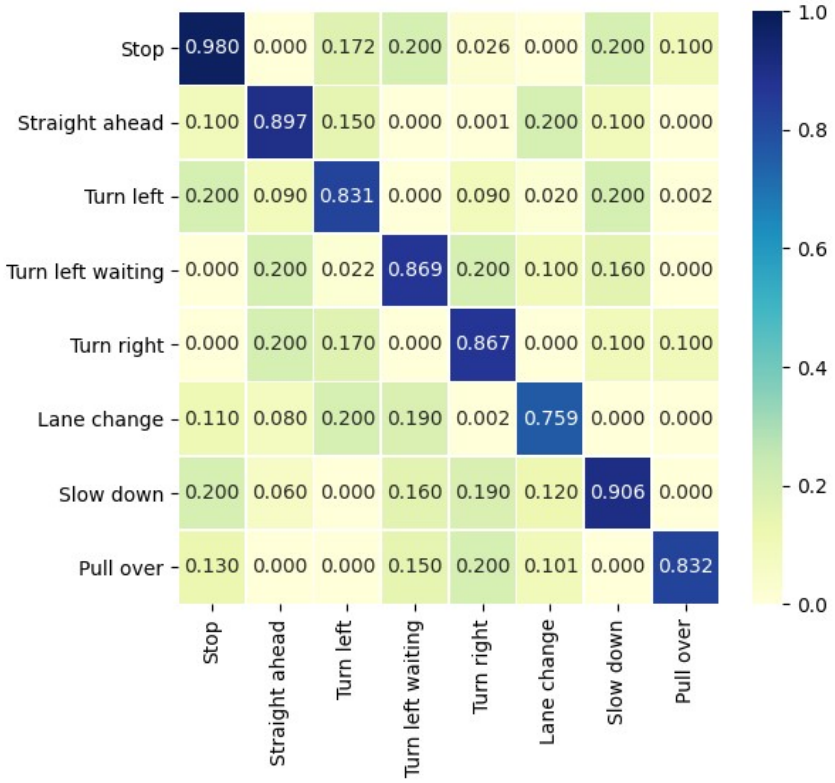

Figure 18. The traffic police gesture dataset for traffic police gesture recognition network sorted by classwise accuracy.

4.2.3. Comparison with the State-of-Art Result

To more effectively verify the effectiveness and generalization capability of our model, this paper used the Kinetics–Skeleton dataset to perform experimental verification, as shown in Table 4.

Table 4. Benchmark comparison for Kinetics–Skeleton model.

Methods	Top-1 (%)	Top-5 (%)
Feature Encoding [39]	14.9	25.8
Deep LSTM [40]	16.4	35.3
Temporal Conv [41]	20.3	40.0
ST-GCN [17]	30.7	52.8
Ours-Conv-Chiral [42]	30.9	52.6
Ours	32.11	54.31

"Feature Encoding" [39], which is shown in the table, is a method based on manually designed features that sort every frame of a video in temporal order and extract the feature of actions. Action recognition methods based on manually designed features are designed for specific actions, cannot be applied to other actions, and have poor generalization capability. They can only be applied to situations such as simple actions or a single scene. The deep LSTM [40] proposed by Shahroudy et al. divided the human body into five parts: the torso, two arms, and two legs. The LSTM is used to extract the context features of each part and merge them to a whole action feature to recognize actions. The temporal ConvNet [41] proposed by Kim and Reiter adjusted the TCN network encoder used in temporal action localization, introduced methods used in residual models, extracted the spatial–temporal features of actions, and enhanced the interpretability of the model's parameters and features. Inspired by the methods used in parameter-shared and parity symmetry, Raymond A [42] et al. designed equivalent transformation layers of frequently used layers in the deep network and proposed Ours-Conv-Chiral, which reduced the calculation of the model effectively. Deep LSTM, Temporal ConvNet, and Ours-Conv-Chiral are all deep learning methods which encode the action sequence or map it to pseudo-images to extract the feature of action, leading to the original feature of data being lost and affecting the accuracy of action recognition. Table 4 compares the recognition performance of top-1 and top-5; it can be seen that methods based on the graph network are generally better than the former methods based on manually designed features and deep learning. This paper improved the graph network model; the accuracy of top-1 and the accuracy of top-5 in the proposed model are 32.11% and 54.31%, respectively, which are superior values compared to other advanced models.

5. Conclusions

In this paper, we used skeleton data of traffic police to construct graph structure data as an input source of the model; using the ST-GCN as a base network, the spatial–temporal features of gestures used by traffic police were extracted, and finally, gestures used by traffic police were recognized. As the ST-GCN has a disadvantage in that it only learns the physical structure of the human body but ignores the potential relationship between joints in frames, we proposed the AGS to learn the potential relationship between nodes of traffic police skeletons. Based on the fact that the importance levels of nodes are different in different traffic police gestures, the attention mechanism was introduced to strengthen the weight of key nodes, to inhibit the nodes with little information volume or no information volume, and to improve the accuracy of traffic police gesture recognition. Another disadvantage is that joint feature information in the temporal dimension of the same joint in frames is too simple. This paper proposed the TAS to solve this, in which features in the temporal dimension are enriched by connecting the same joints in different frames to the same joints in one frame, which improves the accuracy of recognition.

The method proposed in this paper can only recognize gestures used by traffic police in simple scenes; in the future, we will continue to study traffic police gesture recognition methods that can be applied in complex scenes, which means we should extend the existing traffic police gesture dataset we established and locate and detect traffic police in complex scenes.

Author Contributions: Conceptualization, K.L. and Y.Z.; methodology, K.L. and Y.Z.; software, J.Y.; validation, Y.Z. and J.Y.; formal analysis, K.L.; investigation, Y.Z. and H.Z.; resources, Y.Z.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, K.L. and H.B.; visualization, H.Z.; supervision, H.B.; project administration, K.L.; funding acquisition, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Key Project of National Nature Science Foundation of China under grant 61932012 and in part by Natural Science Foundation of Shanxi under 201901D111467.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon request. The data are not publicly available due to copyright. Please submit a formal application to the corresponding author for any citations.

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

Abbreviations

The following abbreviations are used in this manuscript:

CBAM	Convolutional block attention module
GCN	Graph convolutional networks
GNN	Graph neural network
LSTM	Long short-term memory
ST-GCN	Spatial–temporal graph convolutional network
TAS	Temporal attentions mechanism
AGS	Adaptive graph structure

References

1. Ma, N.; Li, D.Y.; He, W.; Deng, Y.; Li, J.H.; Gao, Y.; Bao, H.; Zhang, H.; Xu, X.K.; Liu, Y.S.; et al. Future vehicles: Interactive wheeled robots. *Sci. China Inf. Sci.* **2021**, *64*, 1–3. [CrossRef]
2. Li, D.Y.; Ma, N.; Gao, Y. Future Vehicle: Learnable Wheeled Robot. *Sci. China Inf. Sci.* **2020**, *63*, 24–54. [CrossRef]
3. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Available online: https://www.sae.org/standards/content/j3016/ (accessed on 30 April 2021).
4. Contents and Methods of Field Test Capability Assessment for Automated Vehicle. Available online: http://jtgl.beijing.gov.cn/jgj/jgxx/gsgg/jttg/588465/683743/2020011514485067399.pdf (accessed on 11 February 2018).
5. Road Traffic Safety Law of the People's Republic of China. Available online: http://www.gov.cn/banshi/2005-08/23/content_25575.htm (accessed on 28 October 2003).
6. Wang, B.; Yuan, T. Traffic police gesture recognition using accelerometers. *J. Hainan Norm. Univ.* **2008**, 1080–1083.
7. Tao, Y.; Ben, W. Accelerometer-based Chinese Traffic Police Gesture Recognition System. *Chin. J. Electron.* **2010**, *2*, 270–274.
8. Fan, G.; Cai, Z.; Jin, T. Chinese Traffic Police Gesture Recognition in Complex Scene. In Proceedings of the IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications, Changsha, China, 16–18 November 2011.
9. Cai, Z.; Fan, G. Max-covering scheme for gesture recognition of Chinese traffic police. *Pattern Anal. Appl.* **2015**, *18*, 403–418. [CrossRef]
10. Fan, G.; Jin, T.; Zhu, C. Gesture Recognition for Chinese Traffic Police. In Proceedings of the International Conference on Virtual Reality and Visualization (ICVRV), Hangzhou, China, 25–26 September 2016.
11. Sathyaa, R.; Geethaa, M.K. Automation of Traffic Personnel Gesture Recognition. *Int. J. Inf. Process.* **2015**, *9*, 67–76.
12. Sathyaa, R.; Geethaa, M.K. Vision Based Traffic Personnel Hand Gesture Recognition Using Tree Based Classifiers. *Comput. Intell. Data Min.* **2015**, *2*, 187–200.
13. Microsoft Kinect. Available online: https://dev.windows.com/en-us/kinect (accessed on 15 December 2012).
14. ASUS Xtion RPO LIVE. Available online: https://www.asus.com/3D-Sensor/Xtion_PRO (accessed on 15 December 2011).

15. Ma, C.; Zhang, Y.; Wang, A.; Wang, Y.; Chen, G. Traffic Command Gesture Recognition for Virtual Urban Scenes Based on a Spatiotemporal Convolution Neural Network. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 37. [CrossRef]
16. He, J.; Zhang, C.; He, X.L.; Dong, R.H. Visual Recognition of Traffic Police Gestures with Convolutional Pose Machine and Handcrafted Features. *Neurocomputing* **2020**, *390*, 248–259. [CrossRef]
17. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
18. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
19. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
20. Chen, Y.X.; Zhang, Z.Q.; Yuan, C.F.; Li, B.; Deng, Y.; Hu, W.M. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021.
21. Gori, M.; Monfardini, G.; Scarselli, F. A new model for learning in graph domains. In Proceedings of the IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005.
22. Bruna, J.; Zaremba, W.; Szlam, A.; Lecun, Y. Spectral Networks and Locally Connected Networks on Graphs. *arXiv* **2013**, arXiv:1312.6203.
23. Henaff, M.; Bruna, J.; Lecun, Y. Deep Convolutional Networks on Graph-Structured Data. *arXiv* **2015**, arXiv:1506.05163.
24. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In Proceedings of the Thirtieth Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
25. Li, R.; Sheng, W.; Zhu, F.; Huang, J. Adaptive Graph Convolutional Neural Networks. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
26. Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv* **2015**, arXiv:1509.09292.
27. Niepert, M.; Ahmed, M.; Kutzkov, K. Learning Convolutional Neural Networks for Graphs. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016.
28. Monti, F.; Boscaini, D.; Masci, J.; Rodola, E.; Bronstein, M.M. Geometric Deep Learning on Graphs and Manifolds Using Mixture Model CNNs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
29. Masci, J.; Boscaini, D.; Bronstein, M.M.; Vandergheynst, P. Geodesic convolutional neural networks on Riemannian manifolds. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015.
30. Bronstein, M.M.; Bruna, J.; LeCun, Y.; Szlam, A.; Vandergheynst, P. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Process. Mag.* **2017**, *34*, 18–42. [CrossRef]
31. Kipf, T.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
32. Atwood, J.; Pal, S.; Towsley, D.; Swami, A. Sparse Diffusion-Convolutional Neural Networks. In Proceedings of the Thirty-First Conference on Neural Information Processing Systems, Long Beach, CA, USA, 16–20 June 2017.
33. Hamilton, W. L.; Ying, R.; Leskovec, J. Inductive Representation Learning on Large Graphs. In Proceedings of the Thirty-First Conference on Neural Information Processing Systems, Long Beach, CA, USA, 16–20 June 2017.
34. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
35. Bai, S.; Kolter, J.Z.; Koltun, V. Trellis Networks for Sequence Modeling. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
36. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Zisserman, A. The Kinetics Human Action Video Dataset. *arXiv* **2017**, arXiv:1705.06950.
37. Carreira, J.; Noland, E.; Banki-Horvath, A.; Hillier, C.; Zisserman, A. A Short Note about Kinetics-600. *arXiv* **2018**, arXiv:1808.01340.
38. Smaira, L.; Carreira, J.; Noland, E.; Clancy, E.; Zisserman, A. A Short Note on the Kinetics-700-2020 Human Action Dataset. *arXiv* **2020**, arXiv:2010.10864.
39. Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
40. Shahroudy, A.; Liu, J.; Ng, T. T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *IEEE Comput. Soc.* **2016**, *1*, 1010–1019.
41. Kim, T. S.; Reiter, A. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
42. Yeh, R.A.; Hu, Y.T.; Schwing, A.G. Chirality Nets for Human Pose Regression. In Proceedings of the Thirty-third Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.