applied sciences

MDPI

Article

STJA-GCN: A Multi-Branch Spatial–Temporal Joint Attention Graph Convolutional Network for Abnormal Gait Recognition

Ziming Yin, Yi Jiang, Jianli Zheng and Hongliu Yu *

School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
* Correspondence: yhl98@hotmail.com

Abstract: Early recognition of abnormal gait enables physicians to determine a prompt rehabilitation plan for patients for the most effective treatment and care. The Kinect depth sensor can easily collect skeleton data describing the position of joints in the human body. However, the default human skeleton model of Kinect includes an excessive number of many joints, which limits the accuracy of the gait recognition methods and increases the computational resources required. In this study, we propose an optimized human skeleton model for the Kinect system and streamline the joints using a center-of-mass calculation. We integrate several techniques to propose an end-to-end, spatial–temporal, joint attention graph convolutional network (STJA-GCN) architecture. We conducted experiments with a fivefold cross-validation on two common datasets of information on abnormal gaits to evaluate the performance of the proposed method. The results show that the STJA-GCN achieved 93.17 and 92.08% accuracy on the two datasets, and compared to the original spatial–temporal graph convolutional network (ST-GCN), the recognition accuracy increases by 9.22 and 20.65%, respectively. Overall, the results demonstrate that the STJA-GCN can accurately recognize abnormal gaits and, thus, can support low-cost rehabilitation assessments at community hospitals or in patients' homes.

Keywords: spatial temporal graph convolution; abnormal gait recognition; early multi-branch fusion; attention mechanism

Citation: Yin, Z.; Jiang, Y.; Zheng, J.; Yu, H. STJA-GCN: A Multi-Branch Spatial–Temporal Joint Attention Graph Convolutional Network for Abnormal Gait Recognition. *Appl. Sci.* **2023**, *13*, 4205. https://doi.org/10.3390/app13074205

Academic Editor: Luigi Portinale

Received: 28 January 2023
Revised: 24 March 2023
Accepted: 25 March 2023
Published: 26 March 2023

1. Introduction

Abnormal gaits can seriously affect patients' daily living and behavior. The most common causes of abnormal gaits include degenerative, neurological, and musculoskeletal diseases among older people [1]. The classification and diagnosis of abnormal gaits can help physicians identify underlying diseases and conditions to realize their early treatment, thus reducing their impacts on patients and their families. However, traditional gait analysis methods [2,3] tend to be highly subjective and inefficient because they mainly rely on clinicians' individual experience. Moreover, some gait analysis and measurement devices are relatively expensive and cumbersome to use. Such systems cannot be widely adopted in community hospitals or used by families at home, which limits the early detection of related diseases. This calls for a low-cost, easy-to-operate, and reliable method to enable community hospitals and families to accurately identify abnormal gaits.

Two types of sensors are commonly used for human gait analysis, including wearable [4,5] and non-wearable sensors [6,7]. Compared to wearable sensors, non-wearable sensors provide a convenient way to collect gait data from subjects in their most realistic state as a non-invasive measurement tool. Microsoft has launched two home-oriented physical gaming devices, called Kinect v2 and Azure Kinect DK. These typical non-wearable sensors are suitable for abnormal gait analysis. Using a depth camera, the Kinect can acquire both the RGB data of images and depth data of each pixel. The Kinect system can be used to easily capture three-dimensional (3D) data on human body poses without

attaching sensors to the body. The accuracy of data captured by the Kinect sensor has been shown in several studies [8–10]. The Azure Kinect DK is more accurate than the Kinect v2 and Kinect v1 [11]. A software development kit (SDK) is also provided for Kinect, which can be used to obtain a digital skeleton to represent information on human poses. In combination with the rapid development of artificial intelligence algorithms, this SDK provides a robust technical basis with suitable hardware and software for vision-based gait analysis and abnormal gait recognition.

The emergence of depth sensors like those used in the Kinect system considerably facilitates the acquisition of 3D skeleton data on the human body and has enabled the development of many gait analysis methods based on digital skeletons. In very early research, Bayesian classifiers [12,13] and artificial neural network (ANN) models [14] were used to extract skeleton data from Kinect V1 and compute gait features to identify Parkinson's disease. Subsequently, Guo et al. [15] compared the accuracy of support-vector machine (SVM) classifiers and a long short-term memory (LSTM) architecture in abnormal gait recognition. They found that gaits can be more accurately classified by directly importing skeleton data into the LSTM. Chen et al. [16] fused manual and depth features extracted by a convolutional neural network (CNN)-LSTM network, and classified gaits using an SVM model. Using an autoencoder based on a recurrent neural network (RNN), Jun et al. [17] extracted features from 3D skeleton data and identified abnormal gaits using a classifier.

Because the human skeleton can be modeled as a series of non-Euclidean graphs, these methods cannot effectively learn the underlying spatial relationships between skeleton joints. The advancement of graph convolutional networks (GCNs) allows for a more reasonable interpretation of the natural graph structure of the human skeleton. Consequently, skeleton-based action recognition algorithms have been studied more intensively from the perspective of spatial–temporal graph convolution. As shown in Table 1, the spatial–temporal graph convolutional network (ST-GCN) proposed by Yan et al. [18] provided the earliest feature extraction model for skeleton series in time and space, providing a skeleton-based action recognition model. ST-GCN models have inspired many spatial–temporal graph convolutional networks. Shi et al. [19] added an attention mechanism to the ST-GCN, and adopted a two-stream adaptive GCN (2s-AGCN) to capture richer action features. However, the two-stream framework inevitably increased the computing load of the network. Chen et al. [20] proposed a multiscale spatial–temporal GCN (MST-GCN) designed to capture the relationship between short- and long-range joints while enriching the perceptual field of the model in time and space. Similarly, Cheng et al. [21] proposed a Shift Graph Convolution Network (Shift-GCN) by adding the shift operation to graph convolution, which not only expanded the perceptual field of the model, but also greatly reduced the computational complexity of the algorithm. Liu et al. [22] combined multiscale graph deconvolution with G3D, a unified spatial–temporal graph convolutional operator, into a deconvolution-unification GCN. Their network was shown to promote the direct exchange of action information across space and time and to learn features effectively. Song et al. [23] created a residual GCN (ResGCN), which replaced the fusion of multiple models with early fusion between the inputs of multiple branches to minimize the computational resources required. All these studies considered the recognition of whole-body actions, without focusing on lower limb gait. Thus, these methods cannot be directly applied to recognize abnormal gaits. Moreover, most of the existing GCNs require a long period of training, because multiple streams are fused in the later stage of modeling, and the sub-models of different streams need to be trained separately before fusion is performed.

Aiming to address these problems, this paper proposes a method of abnormal gait recognition based on a multi-branch, spatial–temporal, joint attention graph convolution network (STJA-GCN). In order to enhance the extraction of key features in walking gait, the connection between human skeleton joints is simplified, and the multi-branch joint motion features are fused in the early stage to form an end-to-end single-stream model

architecture, which not only preserves the multiple motion characteristics of the human skeleton, but also avoids the complex calculation of multi-stream model fusion.

Table 1. Summary of Existing GCN Models.

Model	Year	Number of Streams	Description
ST-GCN [18]	2018	1 s	Firstly constructed a spatial–temporal graph. The parameters of each layer are fixed.
2s-AGCN [19]	2019	2 s	Added adaptive graph convolutional layers. The network has a large amount of calculation.
Shift-GCN [21]	2020	4 s	Proposed non-local shift graph convolution. High computational complexity due to shift operations.
MS-G3D [22]	2020	2 s	Redefined the k-order adjacency matrix. Designed a unified spatio-temporal graph convolution operator.
Res-GCN [23]	2020	1 s	Introduced a residual GCN with bottleneck structure and part-wise attention module.
MST-GCN [20]	2021	4 s	Proposed a multi-scale spatial and temporal graph convolution module.

2. Materials and Methods

2.1. Human Skeleton Model for Abnormal Gait Recognition

The human skeleton reflects the important features of body motions. The coordinates of the main joints can be effectively extracted with a depth sensor, as in the Kinect system. The classification accuracy of human behaviors improves as more joints are recognized. As for the abnormal gait recognition, more attention should be paid to the motion of lower limbs. Since pathological gaits may often manifest with abnormal bending of the legs, the body joints extracted by the depth sensor must be winnowed to some extent to improve the feature extraction of limb joints and enhance the efficiency of the training process.

As shown in Figure 1, the Kinect V2 can acquire the coordinates of 25 body joints, including 8 joints of the hands; Azure Kinect DK can acquire the coordinates of 32 body joints, including 8 joints of the hands and 6 of the head. In abnormal gait recognition, adding this many joints on the hands and head does not substantially help. Thus, the 8 hand joints were simplified as a single center-of-mass key point for each of the left and right hands, and the 6 head joints were simplified to a single key point. The coordinates of each key point are the mean coordinates of the joints on that side of the body.

$$\begin{aligned} ComHand_{x,y,z} &= \frac{\sum I_{x,y,z}}{4}, \\ ComHead_{x,y,z} &= \frac{\sum I_{x,y,z}}{5}, \end{aligned} \tag{1}$$

where $ComHand_{x,y,z}$ is the x, y, and z three-dimensional space coordinates of the center-of-mass key point for the hand, and $ComHead_{x,y,z}$ is the x, y, and z three-dimensional space coordinates of the center-of-mass key point for the head; these new keypoints are shown as red dots in Figure 1. $I_{x,y,z}$ is the x, y, and z three-dimensional space coordinates of the joint in the corresponding part, as shown by the yellow dots in Figure 1.

Thus, any human skeleton graph can be represented as $G = \{V, E\}$, where V is the set of joints in the streamlined skeleton, and E is the set of edges corresponding to the physical connections between the joints. For any spatial–temporal series $T = \{1, 2, \ldots M\}$, the corresponding spatial–temporal graph of the skeleton can be represented as $STG = \left\{G^t | t \in [1, 2, \ldots M]\right\}$, where M is the number of consecutive frames of the skeleton series, and G^t is the corresponding skeleton graph at time t.

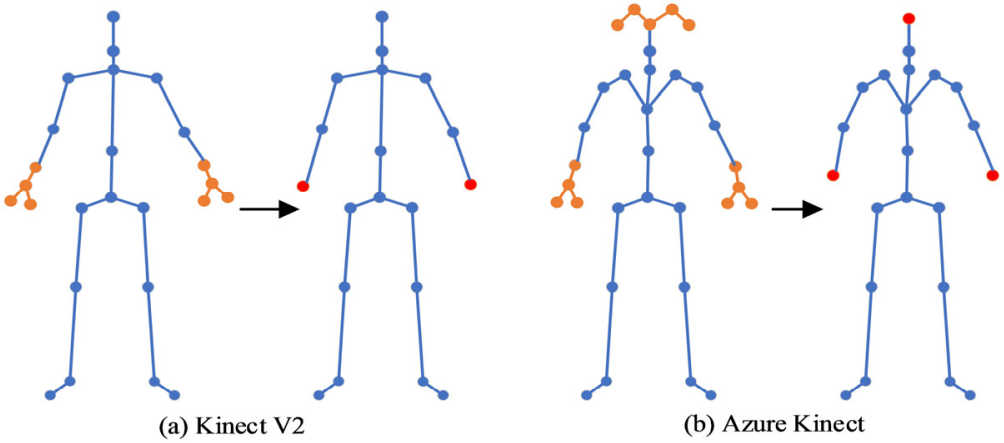

Figure 1. Schematic of the joint reconstruction. (**a**) Kinect V2 Joint reconstruction, (**b**) Azure Kinect Joint reconstruction. Both of these simplifications are based on decreasing the number of joints on the head and the hands.

2.2. Early Multi-Branch Fusion of Spatial–Temporal Joint Attention GCN

The proposed STJA-GCN mainly combines an early multi-branch fusion strategy with multiple spatial–temporal joint attention graph convolutional modules. As shown in Figure 2, three types of motion features are obtained by processing the data of the real skeleton series, including joint positions, motion velocities, and skeleton features. Thereafter, the three types of features are fused with the corresponding implicit features of the lower limbs, and then imported through the spatial–temporal joint attention graph convolutional modules. The spatial–temporal joint attention graph convolutional layer consists of multiple temporal and spatial convolutions. In addition, we also introduce a spatial–temporal joint attention module here, and the classification results are output through the fully connected layer.

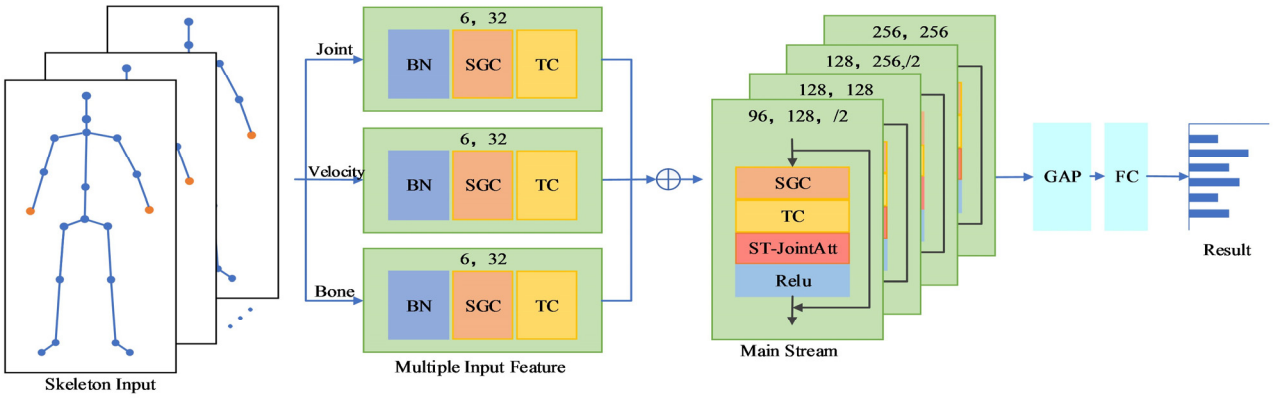

Figure 2. STJA-GCN model structure (The two numbers in each block denote input and output channels, and /2 represents a stride of 2. BN: BatchNorm, SGC: Spatial Graph Convolution, TC: Temporal Convolution, GAP: Global Average Pool, FC: Fully Connected Layer).

2.2.1. Multi-Branch Input Features and Fusion

The 2 s-AGCN model was the first model to fuse multiple streams at the decision layer. Subsequently, many multi-stream models have been developed for skeleton-based action recognition. However, the decision-layer fusion approach needs to handle various data inputs by independently training multiple models, which significantly increases the complexity of the model. Consequently, this approach cannot be applied to end-to-end training processes. In this study, we developed an architecture for early multi-branch fusion, in which temporal and spatial convolution are performed on each branch to extract

the respective features, which are then fused and imported to the spatial–temporal graph convolutional layer. This process both diversifies input features and simplifies the required computations. Therefore, the proposed architecture can process more input data and exhibits improved performance.

Let $X = \left\{ x \in \mathbb{R}^{C_{in} \times T_{in} \times V_{in}} \right\}$ be the set of the 3D body joint coordinate series corresponding to the input spatial–temporal graph of the skeleton STG, where Cin, Tin, and Vin are the input coordinates, number of frames, and number of joints, respectively. The corresponding set of joint positions can be expressed as $R = \{r_i | i = 1, 2, \ldots V_{in}\}$, where

$$r_i = x[:, :, i] - x[:, :, c], \tag{2}$$

with c being the index of the central spinal joint. Thus, the input of the joint position feature consists of both X and R.

The set of motion velocity features is composed of fast motions $F = \{f_t | t = 1, 2, \ldots T_{in}\}$ and slow motions $S = \{s_t | t = 1, 2, \ldots T_{in}\}$, where

$$\begin{aligned} f_t &= x[:, t+2, :] - x[:, t, :], \\ s_t &= x[:, t+1, :] - x[:, t, :] \end{aligned} \tag{3}$$

The set of skeleton features contains the joint length $L = \{l_i | i = 1, 2, \ldots V_{in}\}$ and joint angle $A = \{a_i | i = 1, 2, \ldots V_{in}\}$. The length ($l_i$) and angle ($a_{i,w}$) of each joint can be calculated as

$$\begin{aligned} l_i &= x[:, :, i] - x\left[:, :, i_{adj}\right], \\ a_{i,w} &= arccos\left(\frac{l_{i,w}}{\sqrt{l_{i,x}^2 + l_{i,y}^2 + l_{i,z}^2}} \right) \end{aligned} \tag{4}$$

where i_{adj} is the joint adjacent to the i-th joint, and $w \in \{x, y, z\}$ is the 3D coordinates of the joint.

2.2.2. Spatial–Temporal Graph Convolutional Attention Module

According to the definition of ST-GCN, the convolution of the graph with any frame t can be written as

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in N(v_{ti})} \frac{f_{in}(v_{tj}) \cdot w(l_{ti}(v_{tj}))}{Z_{ti}(v_{tj})}, \tag{5}$$

where v_{ti} is the i-th joint of the t-th frame; f_{in} and f_{out} are the input and output features of the corresponding joint, respectively; $N(v_{ti})$ is the set of joints adjacent to joint v_{ti}; Z_{ti} is the normalization term to balance the contributions of different adjacency sets; and $w(\cdot)$ is a weight function that assigns weights indexed by the label function $l_{ti}(\cdot)$, which constructs multiple adjacency sets $N(v_{ti})$ by assigning different labels to different graph nodes.

In the proposed approach, we adopt a distance-based division method to define the label assignment. Specifically, we define that $l_{ti}(v_{tj}) = d(v_{ti}, v_{tj})$, where $d(v_{ti}, v_{tj})$ denotes the graph distance between v_{ti} and v_{tj}. The joints with the same distance comprise a subset and share a learnable weight function $w(\cdot)$. By the adjacency matrix A, Equation (5) above can be transformed into the form given below:

$$f_{out} = \sum_{d=0}^{D} W_d f_{in} \left(\Lambda_d^{-\frac{1}{2}} A_d \Lambda_d^{-\frac{1}{2}} \odot M_d \right), \tag{6}$$

where D is the preset maximum graph distance; f_{in} and f_{out} are the input and output feature maps, respectively; $\odot$ is a point-by-point convolution; A_d is a d-order adjacency matrix that labels joint pairs with graph distance d; Λ_d is the normalization operator of A_d, W_d, and M_d, which are learnable parameters for implementing the convolution operation and tuning the importance of each edge.

In this study, we introduce a spatial–temporal joint attention mechanism. The proposed approach includes an added attention module after the spatial–temporal graph convolution to form a new spatial–temporal convolutional layer. In the field of skeleton-based action recognition, previous attention mechanisms were mainly realized by multilayer perceptron (MLP) models, in which each channel or spatial dimension was independently processed, while the other dimensions were equally distributed to individual units. Given that temporal and spatial information may be correlated with each other in a spatial–temporal skeleton series, in the proposed model, we included a spatial–temporal joint attention mechanism (ST-JointAtt) to distinguish between the most informative joints in a particular frame and the whole skeleton series together. The mechanism helps to identify the most critical gait regions and joints in the entire walking process.

Figure 3 illustrates the ST-JointAtt module. First, the skeleton series is temporally and spatially pooled for frames and joints, respectively. Next, the two eigenvectors are stitched together, and the information is compressed by a fully connected layer. Subsequently, two independent fully connected layers are used to obtain the inter-frame and inter-joint attention scores. Finally, the frame score and joint score are multiplied by the outer product of channels. The result can be regarded as the attention score of the entire skeleton series. This process can be expressed as follows.

$$\begin{aligned} f_{inner} &= \theta((\,pool_t(f_{in}) \oplus pool_v(f_{in}))\cdot W), \\ f_{out} &= f_{in} \odot (\sigma(f_{inner}\cdot W_t) \otimes \sigma(f_{inner}\cdot W_v)), \end{aligned} \tag{7}$$

where f_{in} and f_{out} are the input and output feature maps, respectively; $\oplus$ is the splicing operation; $\otimes$ and $\odot$ are the out-of-frequency-domain product operation and element product operation, respectively; $pool_t(\cdot)$ and $pool_v(\cdot)$ are the average pooling operations in the frame dimension and joint dimension, respectively; $\theta(\cdot)$ and $\sigma(\cdot)$ are the Sigmoid and Hard Swish activation functions, respectively; and W is the training parameters.

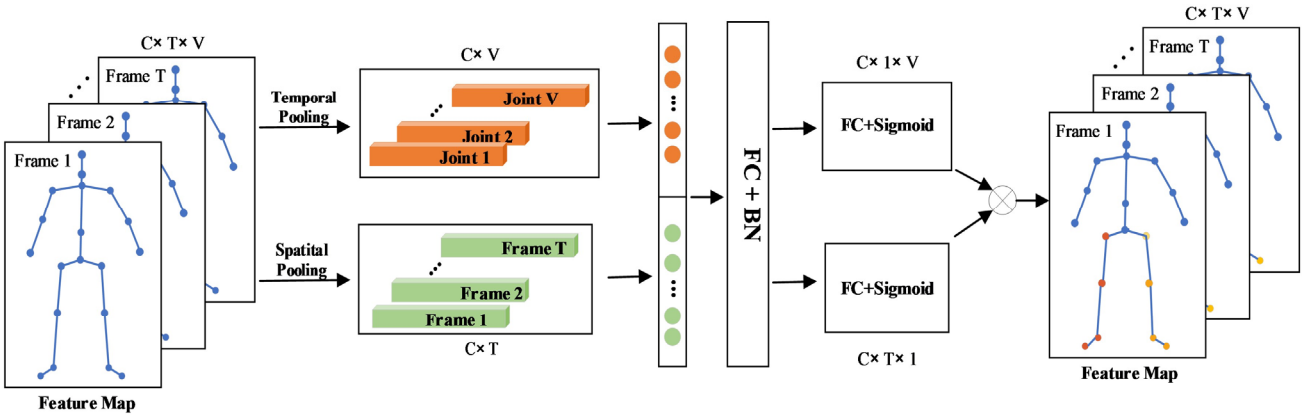

Figure 3. ST-JointAtt model structure.

3. Results

Herein, we conducted experiments to evaluate the performance of the proposed model on two common datasets with the same abnormal gait type and data format [24]. The two datasets were collected by different versions of depth sensors, namely, the Kinect v2 and Azure Kinect DK. The body joints captured by the two depth sensors are shown in Figure 4. For convenience, the two datasets are abbreviated as Kinect25 and Azure32, indicating the number of joints.

In this study, we first conducted several ablation experiments using the Kinect25 data to verify the gait classification effectiveness of each operation in the STJA-GCN model. Then, the performance of the STJA-GCN on each of the two datasets was compared with that of typical existing graph convolution methods.

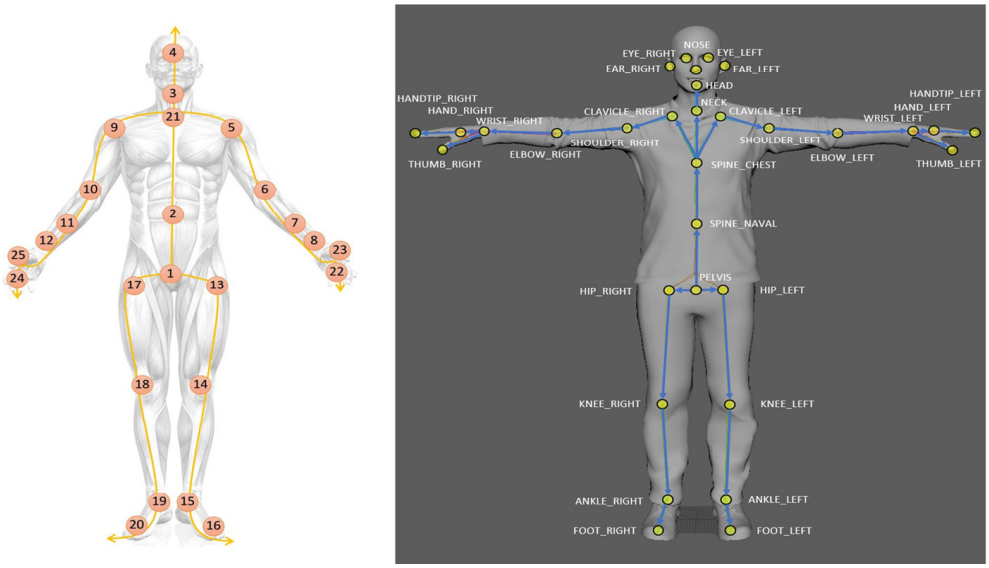

Figure 4. Joints acquired by Kinect v2 and Azure Kinect.

3.1. Datasets

(1) Kinect25

A total of 6 sensors were placed on each side of a 10 m by 3 m walkway. The sensors on the same side were separated by a distance of 2.2 m. The skeleton data were generated by each sensor individually. The data acquisition was terminated when the subject was less than 1 m from the sensor. Ten healthy subjects participated in the experiment. According to the guidelines, each subject was asked to simulate five pathological gaits, including analgesic, stiff-legged, staggering, striding, and Trendelenburg gaits. In total, each subject walked in each gait 20 times. Thus, the final dataset contained 7200 samples of gait data.

(2) Azure32

A 4 m-long walkway was set up with an Azure Kinect DK sensor placed at the end. Twelve healthy males participated in the data acquisition. After watching a video of pathological gaits, each subject was trained and then simulated five pathological gaits, including analgesic, stiff-legged, staggering, striding, and Trendelenburg gaits. In total, each subject walked in each gait 20 times. Thus, the final dataset contained 1440 samples of gait data.

Figure 5 shows the types of pathological gaits in the datasets. For the division method of the dataset, we randomly select the data of seven subjects in the kinect25 dataset as the training sets, and the data of the remaining three subjects as the testing sets. In the azure32 dataset, the data of eight subjects are randomly selected as the training sets, and the data of the remaining three subjects are used as the testing sets.

3.2. Experimental Setup

All experiments were conducted on a 64-bit Windows operating system with an NVIDIA GeForce RTX 3090 (24 G) graphics card, using the Python 3.7 programming language and the PyTorch 1.12.1 library. The stochastic gradient descent (SGD) optimizer was called to train the model from end to end. Each experiment was run for 80 epochs. The learning rate was initialized as 0.1 and reduced by 0.1 times at the 10th and 50th epochs. The batch size was set to 32 and the weight decay to 0.001. Each input consisted of a continuous skeleton of 200 frames, and less than 200 frames of the skeleton series were padded by zeros.

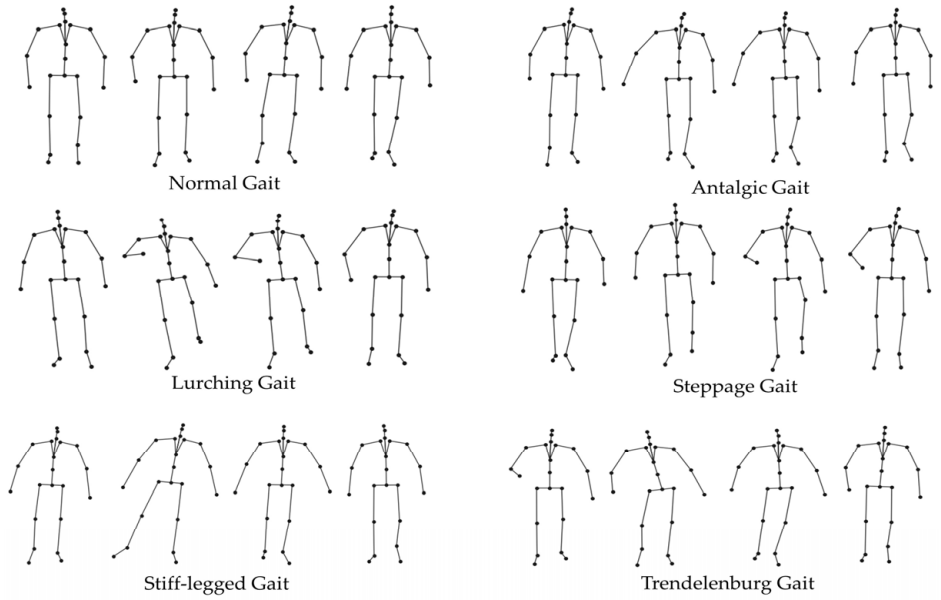

Figure 5. Normal and pathological gaits.

The main stream of the proposed STJA-GCN comprised four stacked spatial–temporal graph convolutional layers. The ResNet mechanism was introduced to each layer. Each layer included 9 spatial–temporal convolutional kernels with a dropout of 0.5. The output size of the first 2 layers was set to 128, and that of the last 2 layers to 256. In the first and third layers, the step size was set to 2 to reduce the time dimension by half of the number of frames and improve the computational efficiency.

To obtain more convincing results, a fivefold cross-validation was performed on each dataset. The performance of the models were measured by the cross-validated average top-1 accuracy, FLOPs, and the number of model parameters (#Params). FLOPs represents the number of operations performed by the model and indicates its computational complexity.

3.3. Ablation Experiments

To verify the effectiveness of the proposed method in joint streamlining and the individual components of the proposed model, we conducted separate contrastive technical ablation experiments using the Kinect25 system with ST-GCN as a baseline model. Each experiment was performed 10 times, and the contrastive models were compared in terms of the average accuracy.

3.3.1. Effectiveness of Joint Streamlining

Previous studies based on the human skeleton have usually been performed on the original skeleton connections. To verify the effectiveness of joint streamlining for gait recognition, we conducted experiments using the Kinect25 data. Out of the original 25 joints, the hand joints were averaged to obtain 19 body joints, which were then trained on the ST-GCN. According to the experimental results in Table 2, the accuracy of the model is not considerably affected and even slightly improved after the redundant hand joints are reduced from the original Kinect dataset. Hence, the proposed joint streamlining operation does not reduce the accuracy of the model for abnormal gait recognition. Therefore, the removal of the joints does not hinder the ability of the model to recognize abnormal gaits. Moreover, the number of floating-point operations per second (FLOPs) during model training was reduced by 24.08%. Thus, joint streamlining simplified the model and sped up the training process.

Therefore, streamlining the hand joints is computationally helpful for abnormal gait recognition. In the subsequent experiments, the models were trained based on the streamlined data.

Table 2. Comparison of different numbers of joints.

Number of Joints	Top-1 (%)	FLOPs (×109)
25	83.2 ± 7.88	5.44
19	**83.95 ± 4.05**	**4.13**

3.3.2. Effectiveness of Attention Mechanisms

To enhance the ability to extract skeleton features, the attention module was embedded in the spatial–temporal convolutional layers of the STJA-GCN. The effectiveness of the module was verified by embedding the ST-JointAtt module in each convolutional layer of the ST-GCN and comparing the embedded module with other graph convolution-based attention modules. According to the experimental results shown in Table 3, the accuracy of the model is improved compared with ST-GCN, and the FLOPs and model parameters are increased by the inclusion of the attention mechanism. Compared with STCAtt [25] and PartAtt [23], the addition of the ST-JointAtt module yielded the largest improvement in the model accuracy. This demonstrates the effectiveness of ST-JointAtt in capturing key features in the spatial–temporal skeleton series.

Table 3. Comparison of different attention modules.

Model	Top-1 (%)	FLOPs (×109)	#Params (×106)
STGCN	83.95 ± 4.05	**4.13**	**3.07**
+STCAtt	86.42 ± 2.69	4.14	3.37
+PartAtt	84.47 ± 10.87	4.15	3.47
+ST-JointAtt	**88.81 ± 6.55**	4.16	3.47

3.4. Comparison of Accuracy between Diffent Models

Many studies [18–20] have shown that graph neural networks are more effective than traditional deep learning methods in the study of the human skeleton. Thus, we chose to compare the ST-GCN with several typical ST-GCN-based multi-stream graph convolutional networks, such as RA-GCNv1 [26], RA-GCNv2 [27], and 2s-AGCN on Kinect25 and Azure32. The results of the comparative analysis are shown in Tables 4 and 5.

Table 4. Comparison of results on Kinect25.

Model	Top-1 (%)	FLOPs (×109)	#Params (×106)
ST-GCN	83.95 ± 4.05	5.44	3.07
RA-GCNv1(3s)	91.21 ± 5.97	8.29	6.05
RA-GCNv2(3s)	91.23 ± 3.78	8.29	6.05
2s-AGCN	91.06 ± 3.93	9.46	6.94
STJA-GCN	**93.17 ± 2.76**	**2.44**	**2.18**

Table 5. Comparison of results on Azure32.

Model	Top-1 (%)	FLOPs (×109)	#Params (×106)
ST-GCN	71.43 ± 4.11	4.79	3.07
RA-GCNv1(3s)	73.38 ± 4.85	9.60	6.05
RA-GCNv2(3s)	35.41 ± 4.51	9.60	6.05
2s-AGCN(3s)	77.06 ± 6.97	10.94	6.94
STJA-GCN	**92.08 ± 2.92**	**2.82**	**2.18**

As shown in Tables 4 and 5, the STJA-GCN model outperforms the models used for comparison in terms of accuracy, FLOPs, and number of parameters on both datasets. The STJA-GCN is 9.22 and 20.65% more accurate than the ST-GCN on the Kinect25 and Azure32 datasets, respectively. Owing to the early fusion between branches and the reduction of spatial–temporal graph convolutional layers, the STJA-GCN achieves better FLOPs and number of parameters than RA-GCNv1, RA-GCNv2, and 2s-AGCN, as evidenced by its lower complexity and substantially fewer parameters. Moreover, the final accuracy of the model is improved across the board. However, RA-GCN v2 and other models with excessive numbers of convolutional layers fail to achieve desirable results on Azure32 owing to the small sample size of the dataset. In contrast, the STJA-GCN model achieves good recognition accuracy on the dataset.

Furthermore, the performance of the model was evaluated with a confusion matrix to reflect the difference between the actual and predicted values. Figure 6 shows the confusion matrices for maximum and minimum accuracies on the two datasets.

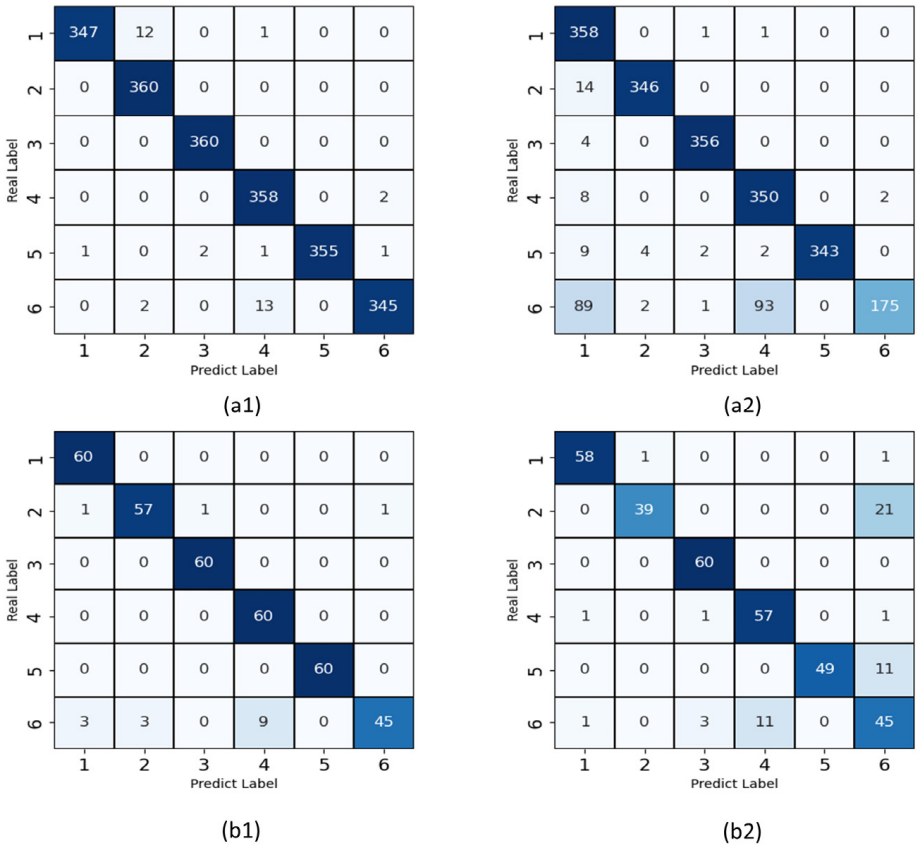

Figure 6. Comparison of confusion matrices. (**a1**,**a2**) are the training models with maximum and minimum accuracies on Kinect25, respectively, (**b1**,**b2**) are the training models with maximum and minimum accuracies on Azure32. The gait labels: 1-normal gait; 2-antalgic gait; 3-lurching gait; 4-steppage gate; 5-stiff-legged gate; 6-Trendelenburg gait.

As shown in Figure 6, the results of each gait type achieved good results, but the classification accuracy of the Trendelenburg gait in these four experiments was not as good as that of the other gaits; in the (a2) 360 Trendelenburg gait samples, 89 cases were misclassified as normal gait, and 93 cases were misclassified as steppage gait. Misclassification of the Trendelenburg gait as the steppage gait was also present in the other three experiments.

4. Discussion

In this study, we propose a novel method for abnormal gait recognition based on graph neural networks. After extracting the skeleton data of walking humans with depth

sensors, the body joints are streamlined by replacing unimportant joints in hands and head with fusion averages. This streamlining reduces the redundant joints and enhances the extraction of key walking features while reducing the computational load and training speed of the model. Then, a multi-branch model structure is constructed for feature fusion in the early stage. The original joint connections are divided into three feature branches, including those for joint positions, motion velocities, and skeleton features, to enhance the extraction of information between joints during walking. Finally, we also introduced the ST-JointAtt module, which is combined with spatial–temporal graph convolution to form a spatial–temporal joint attention graph convolution module. The ST-JointAtt can jointly process temporal attention and spatial attention. Thus, the most important joints can be identified in the whole skeleton series to focus more attention on the main features of abnormal gaits. In addition, the results of a numerical evaluation show that the proposed spatial–temporal joint attention graph convolution method, which fuses multi-branch features early, performed accurately on two public Kinect gait classification datasets. The experimental results show that the proposed method achieved improved efficiency and accuracy on an abnormal gait recognition task compared to existing methods.

The effectiveness of the proposed method was thoroughly verified through comparative experiments. The results show that joint streamlining does not affect the accuracy of the model for gait recognition, but does reduce the number of parameters required and the computational load. Compared with other attention mechanisms, the ST-jointAtt can significantly improve the accuracy of the model, indicating that it can help to discriminate abnormal gaits in both time and space.

It may also be observed from the confusion matrices that the STJA-GCN model is not sufficiently accurate to classify some gaits on the two datasets, but the model is generally able to distinguish abnormal gaits well on both datasets. Furthermore, the Trendelenburg gait was sometimes incorrectly classified as the steppage gate owing to the clinical similarity between the two gaits. The Trendelenburg gait is caused by an abnormal hip abduction: during walking, the weak gluteal muscle causes the pelvis to sag to the opposite side, shifting the center of gravity to the supporting leg. The same is true for the steppage gait: due to the paralysis of the calf extensor group, the patient needs to raise their leg while walking to keep their center of gravity balanced. Therefore, the center of gravity shifts to one side during the walking cycle in both gait patterns, leading to classification errors. Future research may consider capturing walking gait features from other angles or exploring more suitable attention modules for abnormal gaits. Due to data limitations, in this study, we only identified a few classes of abnormal gaits. In the future, we plan to consider further collaboration with clinical hospitals and rehabilitation centers to collect gait data from real patients and train a model to identify more classes of abnormal gaits.

5. Conclusions

Early identification of abnormal gaits is very important, especially at home and in community hospitals. Detecting abnormal gaits as soon as possible helps to provide appropriate treatment and care. Based on deep sensors, such as the Kinect system, in this study, we have proposed a multi-branch, spatial–temporal, joint attention graph convolutional network (STJA-GCN) for abnormal gait recognition. The STJA-GCN optimizes the human skeleton joint model, simplifies the structure of multi-stream models, and captures important joints in walking motion from multiple dimensions. The experimental results show that the proposed method outperformed the benchmark ST-GCN and exhibited major improvements in terms of accuracy, FLOPs, and number of parameters. The proposed method is expected to be used for the initial rehabilitation assessment of abnormal gaits for home use and in community hospitals.

Author Contributions: Conceptualization, Z.Y.; methodology, Z.Y. and Y.J.; software, Y.J.; writing—original draft preparation, Z.Y. and Y.J.; writing—review and editing, J.Z. and H.Y.; funding acquisition, H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China, grant number 2020YFC2005800 and 2020YFC2005801.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experimental data came from an online dataport (IEEEDataPort, url: https://ieee-dataport.org/documents/kinect-v2-skeleton-datasets-simulated-abnormal-gaits, accessed on 18 October 2021 and https://ieee-dataport.org/documents/azure-kinect-3d-skeleton-and-foot-pressure-data-pathological-gaits, accessed on 17 May 2022).

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments on our paper.

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

References

1. Chen, S.; Lach, J.; Lo, B.; Yang, G.Z. Toward pervasive gait analysis with wearable sensors: A systematic review. *IEEE J. Biomed. Health Inform.* **2016**, *20*, 1521–1537. [CrossRef] [PubMed]
2. Martínez-Martín, P.; Gil-Nagel, A.; Gracia, L.M.; Gómez, J.B.; Martínez-Sarriés, J.; Bermejo, F. Unified Parkinsons-Disease Rating-Scale Characteristics and Structure. *Mov. Disord.* **1994**, *9*, 76–83. [CrossRef] [PubMed]
3. Dunsky, A.; Dickstein, R.; Marcovitz, E.; Levy, S.; Deutsch, J.E. Home-Based Motor Imagery Training for Gait Rehabilitation of People with Chronic Poststroke Hemiparesis. *Arch. Phys. Med. Rehabil.* **2008**, *89*, 1580–1588. [CrossRef] [PubMed]
4. Tao, W.; Liu, T.; Zheng, R.; Feng, H. Gait analysis using wearable sensors. *Sensors* **2012**, *12*, 2255–2283. [CrossRef] [PubMed]
5. Suppa, A.; Kita, A.; Leodori, G.; Zampogna, A.; Nicolini, E.; Lorenzi, P.; Rao, R.; Irrera, F. l-DOPA and Freezing of Gait in Parkinson's Disease: Objective Assessment through a Wearable Wireless System. *Front. Neurol.* **2017**, *8*, 406. [CrossRef] [PubMed]
6. Paul, S.; Banerjee, A.; Ghoshal, R.; Tibarewala, D.N. Development of ultrasonic tachography system for gait analysis. *Int. J. Biomed. Eng. Technol.* **2016**, *20*, 66–95. [CrossRef]
7. Sun, B.; Zuo, Z.; Zeng, X.; Liu, T.; Lu, Q. Movement disorder detection via adaptively fused gait analysis based on kinect sensors. *IEEE Sens. J.* **2018**, *18*, 7305–7314.
8. Clark, R.A.; Mentiplay, B.F.; Hough, E.; Pua, Y.H. Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives. *Gait Posture* **2019**, *68*, 193–200. [CrossRef] [PubMed]
9. Galna, B.; Barry, G.; Jackson, D.; Mhiripiri, D.; Olivier, P.; Rochester, L. Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease. *Gait Posture* **2014**, *39*, 1062–1068. [CrossRef] [PubMed]
10. Tölgyessy, M.; Dekan, M.; Chovanec, Ľ. Skeleton Tracking Accuracy and Precision Evaluation of Kinect V1, Kinect V2, and the Azure Kinect. *Appl. Sci.* **2021**, *11*, 5756. [CrossRef]
11. Albert, J.A.; Owolabi, V.; Gebel, A.; Brahms, C.M.; Granacher, U.; Arnrich, B. Evaluation of the Pose Tracking Performance of the Azure Kinect and Kinect v2 for Gait Analysis in Comparison with a Gold Standard: A Pilot Study. *Sensors* **2020**, *20*, 5104. [CrossRef] [PubMed]
12. Procházka, A.; Vyšata, O.; Vališ, M.; Ťupa, O.; Schätz, M.; Mařík, V. Bayesian classification and analysis of gait disorders using image and depth sensors of Microsoft Kinect. *Digit. Signal. Process.* **2015**, *47*, 169–177. [CrossRef]
13. Dranca, L.; de Abetxuko Ruiz de Mendarozketa, L.; Goñi, A.; Illarramendi, A.; Navalpotro Gomez, I.; Delgado Alvarado, M.; Rodríguez-Oroz, M.C. Using Kinect to classify Parkinson's disease stages related to severity of gait impairment. *BMC Bioinform.* **2018**, *19*, 471. [CrossRef]
14. Ťupa, O.; Procházka, A.; Vyšata, O.; Schätz, M.; Mareš, J.; Vališ, M.; Mařík, V. Motion tracking and gait feature estimation for recognising Parkinson's disease using MS Kinect. *Biomed. Eng. Online* **2015**, *14*, 97. [CrossRef]
15. Guo, Y.; Deligianni, F.; Gu, X.; Yang, G.-Z. 3-D canonical pose estimation and abnormal gait recognition with a single RGB-D camera. *IEEE Robot. Autom. Lett.* **2019**, *4*, 3617–3624. [CrossRef]
16. Chen, F.; Cui, X.; Zhao, Z.; Zhang, D.; Ma, C.; Zhang, X.; Liao, H. Gait acquisition and analysis system for osteoarthritis based on hybrid prediction model. *Comput. Med. Imaging Graph.* **2020**, *85*, 101782. [CrossRef] [PubMed]
17. Jun, K.; Lee, D.-W.; Lee, K.; Lee, S.; Kim, M.S. Feature extraction using an RNN autoencoder for skeleton-based abnormal gait recognition. *IEEE Access.* **2020**, *8*, 19196–19207. [CrossRef]
18. 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, LO, USA, 2–7 February 2018.
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, Long Beach, CA, USA, 15–20 June 2019.
20. Chen, Z.; Li, S.; Yang, B.; Li, Q.; Liu, H. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021.

21. Cheng, K.; Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-based action recognition with shift graph convolutional network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
22. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
23. Song, Y.-F.; Zhang, Z.; Shan, C.; Wang, L. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In Proceedings of the 28th ACM international conference on multimedia, Seattle, WA, USA, 12–16 October 2020.
24. Jun, K.; Lee, Y.; Lee, S.; Lee, D.-W.; Kim, M. Pathological Gait Classification Using Kinect v2 and Gated Recurrent Neural Networks. *IEEE Access.* **2020**, *8*, 139881–139891. [CrossRef]
25. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks. *IEEE Trans. Image Process.* **2020**, *29*, 9532–9545. [CrossRef] [PubMed]
26. Song, Y.F.; Zhang, Z.; Wang, L. Richly Activated Graph. Convolutional Network for Action. Recognition with Incomplete Skeletons. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019.
27. Song, Y.-F.; Zhang, Z.; Shan, C.; Wang, L. Richly Activated Graph Convolutional Network for Robust Skeleton-Based Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 1915–1925. [CrossRef]