symmetry

MDPI

Article

Dual Attention-Guided Multiscale Dynamic Aggregate Graph Convolutional Networks for Skeleton-Based Human Action Recognition

Zeyuan Hu † and **Eung-Joo Lee** *,†

Department of Information Communication Engineering, Tongmyong University, Busan 48520, Korea; zeyuanhu410@126.com

* Correspondence: ejlee@tu.ac.kr

† These authors contributed equally to this work.

Received: 13 August 2020; Accepted: 21 September 2020; Published: 24 September 2020

Abstract: Traditional convolution neural networks have achieved great success in human action recognition. However, it is challenging to establish effective associations between different human bone nodes to capture detailed information. In this paper, we propose a dual attention-guided multiscale dynamic aggregate graph convolution neural network (DAG-GCN) for skeleton-based human action recognition. Our goal is to explore the best correlation and determine high-level semantic features. First, a multiscale dynamic aggregate GCN module is used to capture important semantic information and to establish dependence relationships for different bone nodes. Second, the higher level semantic feature is further refined, and the semantic relevance is emphasized through a dual attention guidance module. In addition, we exploit the relationship of joints hierarchically and the spatial temporal correlations through two modules. Experiments with the DAG-GCN method result in good performance on the NTU-60-RGB+D and NTU-120-RGB+D datasets. The accuracy is 95.76% and 90.01%, respectively, for the cross (X)-View and X-Subon the NTU60dataset.

Keywords: human action recognition; multiscale graph convolutional networks; dynamic aggregation; hierarchical level semantic information; spatial and temporal correlation

1. Introduction

Human action recognition is widely used in many scenarios, such as human-computer interaction [1], video retrieval [2], and medical treatment security [3]. In recent years, with the development of deep learning technology, human skeleton action recognition based on joint type, frame index, and 3D position identification has been widely studied. Moreover, the continuous progress of posture estimation technology [4] makes it easier to acquire bone posture and action data. Second, compared with RGB human action video, skeleton data are more robust and computationally efficient. In addition, skeletal action recognition is a supplement to RGB video action recognition [5] and is a higher level abstract representation of human posture and action.

To improve the recognition accuracy of skeleton movements, researchers need to use deep learning technology to simulate the spatial-temporal nature of bone sequences [5,6]. Examples include the recursive neural network (RNN) [7,8], deep convolutional neural network (CNN) [9–12], and attention [13,14] and graph convolutional network (GCN) [15–19]. In the early stage, RNN/LSTM uses short-term and long-term timing sequence dynamics to model the bone sequence, while CNN adjusts the bone data to the appropriate input (224×224) and learns the correlation.

Although these methods achieve good recognition effects, they cannot effectively recommend long-term dependencies and deep semantic information due to limited acceptance domains. Moreover, these methods increase the complexity of the model. In short, the key to human motion recognition

involves the key type and frame index, namely the semantic information. Semantic information can effectively reveal the temporal and spatial structure of human joints, that is the semantic meaning of two different joints is very different, e.g., sitting and standing can vary in the sequence of frames.

Most of the above research methods ignore the importance of semantic information (such as previous CNN-based works [12,20,21]). These methods typically overlook the semantics by implicitly hiding them in the 2D skeleton map (e.g., with rows corresponding to different types of joints and columns corresponding to frame indexes).

To address the limitations of these methods, we propose a dual attention-guided multiscale dynamic aggregate graph convolutional network (DAG-GCN), which makes full use of high-level semantics to realize skeleton-based human motion recognition. First, a hierarchical framework is constructed by exploring the association and motion frame correlation between bone nodes. To better model the skeleton node association, in addition to dynamics, a graph convolutional layer is superimposed to aggregate the deep semantics of key nodes. This approach realizes the adaptive updating of nodes and the information transmission between nodes. Second, in order to better model the correlation of action frames, we perform a space maximum pooling operation on all related node features in the same frame to obtain the frame-level feature representation. Finally, the frame indexing information is embedded in the dual attention-guiding mechanism; the frame sequence is indexed, and thus, the higher level semantics are determined.

To summarize, the main contributions of this paper are as follows:

(1) We propose a dual attention-guided multiscale dynamic aggregate graph convolutional network for skeleton-based human action recognition. We aim to explore the importance of joint semantics and enhance the dependencies between the semantics of different modules.
(2) The proposed DAG-GCN uses a node-level module and guided-level module to hierarchically mine the spatial-temporal correlation of frames and strengthen the semantic dependency between them.
(3) The node-level module performs multilayer graph convolution, which captures the position and velocity information of bone joints through the graph nodes. This information passes through the multilayer transmission to constitute the deep-layer semantics of bone joints. The guided-level module is composed of positional attention and channel attention, and this module refines the joint semantics step-by-step and establishes and strengthens the dependencies between frames.

The rest of this paper is organized as follows: In Section 2, we review related work on human action recognition using deep learning techniques. In Section 3, we introduce the dual attention-guided multiscale dynamic aggregate graph convolutional network (DAG-GCN). Section 4 gives the test results of the DAG-GCN human action recognition framework on datasets. A summary and future research plan are given in Section 5.

2. Related Work

The human skeleton action recognition method based on deep learning is obviously better than the traditional manual method. For example, Amir Shahroudy et al. [22] divided LSTM cells into five subcells corresponding to five parts of the human body: trunk, arms, and legs. The recognition effect was relatively good. Liu Jun et al. [23] designed a spatiotemporal LSTM network to capture the context dependence of joints in the spatiotemporal domain, in which the joints provide information for different types of joints at each step. To some extent, the researchers were effective at distinguishing between different types of joints. Majd M et al. [24] proposed a CLSTM framework for perceptive motion data, as well as spatial features and temporal dependencies. Gemmule H et al. [25] focused on learning salient spatial features using a CNN and then mapped their temporal relationship with the aid of LSTM networks. Zufan Zhang et al. [26] addressed the human action recognition issue by using a Conv-LSTM and fully connected LSTM with different attentions. However, convolutional neural networks demonstrated their superiority in both accuracy and parallelism [27–29]. These CNN-based

networks transform the skeleton sequence to a skeleton map of some target size and explore the spatial and temporal features. For example, Torpey D et al. [30] separately extracted local appearance and motion features using a 3D-CNN from sampled snippets of a video.

In recent years, graph convolutional networks (GCNs) have proven effective in structured and unstructured Euclidean data processing and have been widely used to model structured human skeleton action data. Yan et al. [15] proposed a spatial-temporal graph convolutional network and treated each bone joint as a node of the graph. Tang et al. [31] enhanced a predefined graph by defining the edges to better construct the graph. To consider these interactions of different human parts, Liu R et al. [18] proposed a novel structure-induced graph convolutional network (Si-GCN) framework to enhance the performance of the skeleton-based action recognition task.

Li F et al. [32] stated that traditional graph convolution cannot completely cover each action execution and proposed a novel multi-stream and enhanced spatial-temporal graph convolutional network to aggregate more temporal features. Shiraki K et al. [33] proposed an acquisition of optimal connection patterns for skeleton-based action recognition with graph convolutional networks. To capture more semantics, Xiaolu Ding et al. [19] proposed a semantics-guided graph convolutional network with three types of semantic graph modules to capture action-specific latent dependencies.

Although the above GCN methods [34–37] effectively capture the semantic information of the skeleton joints, they do not refine and fuse the semantic information, thus reusing redundant information, increasing the operating efficiency of the network and reducing the accuracy of the network's recognition of human skeleton movements. For the proposed DAG-GCN, we design two modules and a boot module (node) to better capture the skeleton's key points of semantic information, pass this information through the multilayer figure convolution superposition, and realize the node semantic information between the layers. Second, a boot module is used to analyze the frame index and to gradually refine the bone joint points of semantic filtered redundant information. This approach preserves the important spatial and temporal structural information and builds dependencies.

3. Dual Attention-Guided Dynamic Aggregate Graph Convolution

Proof of overview. In this section, we consider the correlation between the joints and the dependence between the motion frames. We propose a dual attention-guided multiscale dynamic aggregate graph convolution model (DAG-GCN) that is composed of bone joint-level modules and guided-level modules. The goal is to exploit the semantics of bone joints and movement frames for skeleton-based human action recognition. This approach removes redundant dependencies between node features from different neighborhoods, which allows for powerful multiscale aggregators to effectively capture graph-wide joint relationships on human skeletons. A guided-level module operator facilitates direct information flow across spacetime for effective feature learning. Integrating the dynamic aggregation scheme with a guided-level module provides a powerful feature extractor (dual attention-guided multiscale dynamic aggregate graph convolutional networks (DAG-GCNs)) with multiscale receptive fields across both spatial and temporal dimensions. The direct multiscale aggregation of features in spacetime further enhances the model performance. Figure 1 shows the DAG-GCN model recognition framework.

The skeleton-based human recognition framework of the proposed dual attention-guided graph convolutional network consists of a bone joint-level module and a frame guided-level module. In "concatenation", we learn the two stream representation of a bone joint by concatenating the spatial and temporal information of a skeleton joint. In the bone joint-level module, we use three multiscale graph convolution layers (MS-GCNs) to build the dependencies of different bone joints. To gradually refine the spatial and temporal two stream semantic information of bone joints and strengthen the dependence between motion frames, we use dual attention for guidance. □

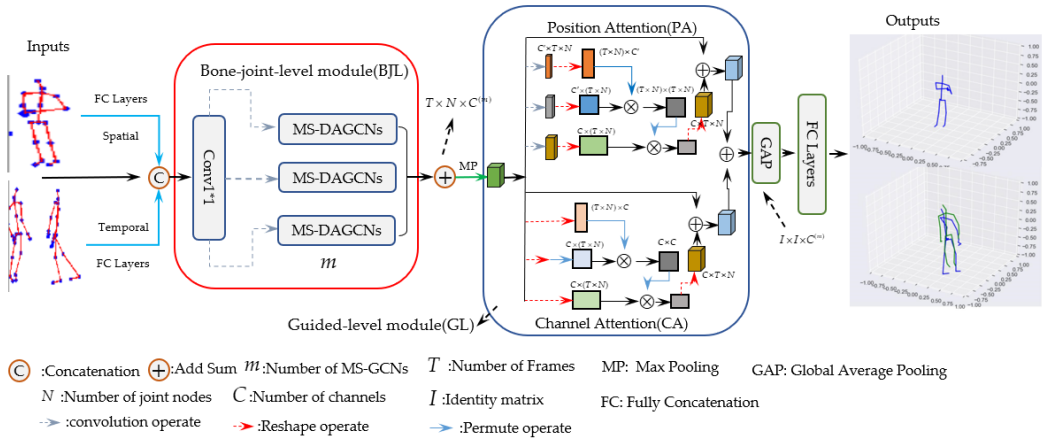

Figure 1. In the recognition framework of the DAG-GCN model, BJL is a bone joint-level module based on multiscale dynamic aggregate graph convolutional networks that describes and aggregates the bone joint semantic information. GL is a guided-level module based on dual attention that describes and gradually refines the semantic information of the bone joint and motion frames. "CA" is a channel attention module; "PA" is a position attention module. MS, multiscale.

3.1. Multi-Scale Dynamic Aggregates

Proof of traditional GCNs. Graph convolutional networks [36], which have been proven to be effective for processing structured data, have also been used to model the structured skeleton data. The human skeleton graph is denoted as $\zeta = (\vartheta,\varepsilon)$, where $\vartheta = (\vartheta_1,\vartheta_2,\cdots,\vartheta_N)$ is the set of N bone joint nodes and ε is the edge set of bone joints. $A \in R^{N\times N}$ is an adjacency matrix from graph ζ, where $A_{i,j} = 0$ has no edges from ϑ_i to ϑ_j; otherwise, $A_{i,j} = 1$ means there are edges, and this node is self-connecting. For a human action skeleton sequence, we denote all bone joint node feature sets as $\chi = \{\chi_{t,k} \in R^C | t = 1,\cdots,T; k = 1,\cdots N\}$, where C is the dimensional feature vector for bone joint node ϑ_n at time t and T is the total of movement frames. Thus, the input skeleton action sequence is adequately expressed structurally by the graph; the layer-wise update rule of traditional GCN layers [34–37] can describe features at time t as:

$$\chi_t^{(l)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}\chi_t^{(l-1)}W^{(l-1)}) \quad (1)$$

where $\widetilde{D}$ is the diagonal degree matrix, $\widetilde{A}$ is a Laplacian normalized self-connecting adjacency matrix, $\sigma(\bullet)$ is an activation function of graph convolution layers, and W is a weight matrix. The traditional GCN captures semantics information of bone joint nodes by aggregating and transmitting the features of neighbor nodes. □

Proof of dynamic aggregate GCNs. Although this method effectively aggregates the structural information of neighbor nodes, this leads to the captured semantic information being transmitted only in local regions, making it impossible to capture the long-distance dependence relationship and multiscale structure information of moving joints. In addition, in traditional GCNs, the graph is fixed throughout the convolution process, which reduces the performance of features if the input graph structure is not suitable. Thus, we propose a multiscale dynamic aggregate GCN [36–38] in which the graph structure can be gradually refined during the process. The ability of the graph structure to capture multiscale features is improved by fusing the currently embedded feature and the graph information used in the previous layer. The dynamic aggregate process is as follows:

Step 1. We denote the adjacency matrix A as:

$$A_{i,j} = \begin{cases} 1, & \vartheta_i \in N(\vartheta_j); \vartheta_j \in N(\vartheta_i); \vartheta_i = \vartheta_j \\ 0, & otherwise \end{cases} \tag{2}$$

where ϑ_i represents a bone joint node and $N(\vartheta_i)$ represents the set of neighbors of the node ϑ_i.

Step 2. Based on the definition, we denote the aggregate kernel ψ of the lth layer as:

$$\psi^{(l)} = A^{(l)} + \varpi K^E \tag{3}$$

where $A^{(l)}$ is the lth layer of the adjacency matrix. ϖ is the weight of the kernel $K^E = \chi_t^{(l)}(\chi_t^{(l)})^T$.

Step 3. We can see the advantages of the aggregate strategy. The introduction of embedded information aims to explore more accurate graph structures, and the detailed information contained in the graph structure is fully utilized. Thus, the graph structure can be dynamically updated as:

$$A^{(l+1)} \leftarrow A(A^{(l)} + \varpi K^E)A^T + \theta^{(l)} I \tag{4}$$

where A is the adjacency matrix of the initial graph convolution layer. According to Equation (1), we can denote the adjacency matrix $A^{(l+1)}$ as:

$$\begin{cases} A^{(l+1)} = \sum_{\vartheta_z \in N(\vartheta_i)} \sum_{\vartheta_u \in N(\vartheta_j)} A_i A_{j,u} \Psi_{z,u} + \theta^{(l)} I_{i,j} \\ \Psi_{z,u} = A_{z,u}^{(l)} + \varpi < \chi_z^{(l)}; \chi_u^{(l)} > \end{cases} \tag{5}$$

where $< \bullet >$ is the inner product vectors. $\theta^{(l)}$, $I_{i,j}$ is the relative importance and identity matrix from bone joint nodes of ϑ_i and ϑ_j.

Step 4. The contextual information revealed by different scales is helpful to describe the rich local features of the skeleton action region from different levels. We capture multiscale spatial-temporal information by constructing the graph structure of different neighborhood scales. Specifically, at scale S, each bone joint node ϑ_i is connected to its $S - hop$ neighbor. Then, the bone joint nodes on scale S can be denote as:

$$V_S(\vartheta_i) = V_{S-1}(\vartheta_i) \cup V_1(V_{S-1}(\vartheta_i)) \tag{6}$$

where $V_S(\vartheta_i)$ is the set of $S - hop$ neighbor bone joint nodes, and the values of S are 1, 2, and 3. □

Proof of the MS-GCN module aggregate: To obtain more effective multiscale spatial and temporal semantic information of skeleton-based actions, the dependence between different scales of semantic information and the ability to represent relevant information are strengthened. This reduces the use of redundant information. We reaggregate multiple multi-scale graph convolutional network (MS-GCN) modules at time t as:

$$\chi_t = \sigma(\sum_{m=1}^{M} MGs_t^m), 1 \le M \le 3 \tag{7}$$

where MGs is the multiscale graph convolution module. According to Equation (1), the multiscale graph convolutional network (MS-GCN) can be denoted as:

$$\chi_t^{(l)} = \sigma(\sum_{s=0}^{S} \widetilde{D}_s^{-\frac{1}{2}} \widetilde{A}_s \widetilde{D}_s^{-\frac{1}{2}} \chi_t^{(l-1)} W^{(l-1)}) \tag{8}$$

where $\widetilde{D}_{i,j} = \sum_j \widetilde{A}_{i,j}$ and $\chi_t^{(l)} \in R^{T \times N \times C^{(l)}}$.

To summarize, the association between joints is captured by the joint node-level module, and the context multiscale structure semantics of different levels are captured. In addition, the aggregated multiscale graph convolutional network (MS-GCN) is used to explore the correlation of structural

skeleton actions and the dependence between joints. We use the semantics of the joint type and the dynamics to learn the graph connections among the nodes (different joints) within a frame. The joint type information is helpful for learning the suitable adjacency matrix (i.e., relations between joints in terms of connecting weights). Take two source joints, foot and hand, and a target joint, head, as an example: intuitively, the connection weight value from foot to head should be different from the value from hand to head even when the dynamics of foot and hand are the same. Second, as part of the information of a joint, the semantics of joint types takes part in the message passing process in MS-GCN layers. □

3.2. Guided-Level Module

When the node-level module encodes skeleton data, it will ignore the correlation between moving frames, weaken the representation effect of related information, and strengthen the representation ability of unrelated features. In addition, to further refine the multiscale semantic information and strengthen the dependency between frames, we introduced self-guided attention to gradually refine the multiscale semantic information of joint nodes, which helps to encode local and global semantics, establish long-term dependency relationships between moving frames adaptively, eliminate redundant information, and highlight the representation of relevant information on joint nodes.

First, we transfer the multiscale semantics captured by the node-level module to the maximum pooling layer. Second, we input the pooling feature map into the self-guided attention module to generate detailed attention features. We denote the attention feature $GAtt_t$ at time t as:

$$GAtt_t = GAttModule(MP(\chi_t)) \tag{9}$$

where $GAttModule(\bullet)$ is the attention-guided module. $MP(\bullet)$ is the maximum pooling layers.

The attention-guided module is composed of positional attention (PA) and channel attention (CA) [39,40]. Positional attention can further obtain rich context representation, while channel attention can be regarded as a response of specific classes to strengthen the correlation and dependence between different semantic information and motion frames. Suppose the feature graph of attention input is $F = MP(\chi_t) \in R^{T\times N\times C}$, where T is the frame dimension, N is the bone joint dimension, and C is the channel dimension.

Proof of Positional Attention. The input feature map F is reconstructed by convolution block transfer to generate a new feature map $F_0^{PA} \in R^{(T\times N)\times C'}$. In the other branch, F follows the same operations and is then reconstructed, generating another new feature map $F_1^{PA} \in R^{C'\times(T\times N)}$. The two new feature maps are multiplied and generate the positional attention coefficient $\alpha_{i,j}^{PA}$ as:

$$\alpha_{i,j}^{PA} = \frac{exp(F_{0,i}^{PA}, F_{1,j}^{PA})}{\sum_{i=1}^{T\times N} exp(F_{0,i}^{PA}, F_{1,j}^{PA})} \tag{10}$$

where $\alpha_{i,j}^{PA}$ is the effect of the ith position on the jth position. In the third branch, we again reconstruct the input feature F and generate new feature map $F_2^{PA} \in R^{C\times(T\times N)}$. Thus, the positional attention feature map F^{PA} is:

$$F^{PA,j} = \lambda \sum_{i=1}^{T\times N} \alpha_{i,j}^{PA} F_{2,j}^{PA} + F_j \tag{11}$$

where the value of λ is gradually increased by learning.

In summary, the positional attention further selectively aggregates joints and frames the global context information to the captured features. □

Proof of channel attention. Channel attention aims to reconstruct the channel of input feature map F. The channel attention coefficient $\alpha_{i,j}^{CA}$ is:

$$\alpha_{i,j}^{CA} = \frac{exp(F_{0,i}^{CA}, F_{1,j}^{CA})}{\sum_{i=1}^{C\times C} exp(F_{0,i}^{CA}, F_{1,j}^{CA})} \tag{12}$$

where $F_0^{CA} \in R^{(T\times N)\times C}$ and $F_1^{CA} \in R^{C\times(T\times N)}$ are the reconstructed feature maps by different branches. Thus, the channel attention feature map $F^{CA,j}$ is:

$$F^{CA,j} = \pi \sum_{i=1}^{C} \alpha_{i,j}^{CA} F_{2,j}^{CA} + F_j \tag{13}$$

where the value of π is gradually increased by learning. The feature map $F_{2,j}^{CA} \in R^{C\times(T\times N)}$ is reconstructed by other branches.

In summary, channel attention reaggregates captured semantic information into original features, highlights dependencies between same-type nodes and motion frames, and increases the ability to distinguish properties between classes.

Finally, we transfer and transform the captured multiscale spatial and temporal semantic information through the global average pooling layer (GAP), and we input the fully connected layer (FC layer) to complete the skeleton action recognition. □

4. Experimental Results and Analysis

In this section, we conduct correlation experiments to verify the effectiveness of the proposed DAG-GCN framework for skeleton-based human action recognition, and we provide the experimental results and analysis. More precisely, we compare the DAG-GCN method with other state-of-the-art approaches on NTU60-RGB-D, NTU120-RGB-D, and other public datasets. Then, we present the results of ablation experiments on multiscale dynamic aggregate operations and show the efficiency of the DAG-GCN recognition framework.

4.1. Datasets

Proof of the NTU60-RGB+D (NTU60) dataset [22,31]: This dataset contains 60 preformed action classes with 56,880 skeleton sequences from 40 different subjects. Each human skeleton graph is represented by 25 (in this paper, N = 25) body bone joint nodes. Each movement frame contains one to two action subjects. In the cross-view (X-View), the 40 action subjects are randomly divided into 60% and 40% for the training and test sequences, respectively. For the cross-Sub(X-Sub), the action subjects sequence contains 40,091 training and 16,487 testing examples. □

Proof of the NTU120-RGB+D (NTU120) dataset [23,32]: This dataset was collected using the Kinect camera and is an extension of NTU60-RGB+D with 113,945 examples and 120 action classes. For the cross-view (X-View) and cross-Sub (X-Sub), half of the 106 human action subjects are used for training, and the remaining parts are used for testing. □

4.2. Training and Implementation Details

Proof of the framework setting. To gather as much multiscale and context semantic information as possible, in the bone joint node-level module, the number of MS-GCN layers was set to two, and the scale (S) was set to three ($1 \leq S \leq 3$). Before each MS-GCN and input CNN layer, the number of CNN neurons was set to 64, and the number of MS-GCN blocks was set to three (r = 3). The guided-level module contains position attention and channel attention. Finally, we used the Adam optimizer. The initial learning rate was set to 0.005, and the batch size was 64. The weight decay was set to 0.0005, and the epochs were set to 400. □

Proof of the training environment. All the experiments were processed on Pytorch Version 1.3.0, and the Python Version was 3.6, built on two NVidia Tesla P100 GPUs. To improve the accuracy, we used the process mode of label smoothing, and the smoothing factor was set to 0.1. Then, we used the cross-entropy loss to train the DAG-GCN model for human skeleton action recognition. □

4.3. Ablation Experiments

To verify the influence of each component in the DAG-GCN model with regard to recognition accuracy, we present and analyze the experimental results of each component on the NTU60 and NTU120 public datasets.

4.3.1. Ablation Study on the Proposed Node Module

Multiscale structure features contain important semantic information of a skeleton action sequence. To further verify the effectiveness of different numbers of scale dynamic graph structures, we built multiple graph convolution models and discuss the experimental results for the NTU60 and NTU120 datasets. Table 1 lists the comparison results for different scales of graph models, where the evaluation index is accuracy (%).

Table 1. The comparison results on NTU60/120 with the cross (X)-View and X-Sub(%).

Model	X-View (NTU60)	X-Sub (NTU60)	X-View (NTU120)	X-Sub (NTU120)
$(m = 1, S = 1)/GL$	85.19	69.57	56.57	39.76
$(m = 1, S = 2)/GL$	87.43	76.72	70.63	50.46
$(m = 1, S = 3)/GL$	89.08	78.91	71.16	63.74
$(m = 3, S = 1)/GL$	87.12	84.99	74.09	71.46
$(m = 3, S = 2)/GL$	92.88	86.51	74.09	71.46
$(m = 3, S = 3)/GL$	95.76	90.01	82.44	79.03
$(m = 6, S = 3)/GL$	90.53	84.81	74.23	71.39
$(m = 6, S = 6)/GL$	91.80	83.46	72.44	69.88

In Table 1, $S = 1, 2, 3$ is the scale of the multiscale graph convolution layers (MS-GCN layers). $m = 1, 3, 6$ is the number of MS-GCN blocks. GL is a guided-level module that contains position attention and channel attention.

According to the experimental results in Table 1, we draw multiple conclusions as follows:

(1) To learn the multiscale structure information of a skeleton sequence, the number of scales and blocks of the bone joint node module are changed. To verify the results, we find that $(m = 1, S = 3)/GL$ outperforms $(m = 1, S = 1)/GL$ by 3.89% and 9.34%, respectively, for the X-View and X-Sub accuracy on the MTU60dataset. However, the accuracy for the NTU120 was improved by 14.59% and 23.98%, respectively. The reason is that more scale structural information can be captured, and the structure information of different scales is complementary. This alleviates the deficiency of the single scale information in describing the joints and motion frames of bones.

(2) Multiscale local and global contest semantic information is beneficial for message passing in MS-GCN blocks. This information also improves the ability of semantic information to represent bone joints and motion frames. For example, the experiment results of $(m = 3, S = 3)/GL$ are superior to those of $(m = 1, S = 1)/GL$. The accuracy was improved by 11.28% and 15.29%, respectively, for the X-View and X-Sub accuracy on the MTU120 dataset.

The main reason is that it is difficult for single MS-GCN blocks to explore semantic features of the skeleton actions with high-order structural information. For example, in the messaging process, the semantic information expressed should be different even if the 3D coordinates corresponding to different joints are the same. Introducing a multiscale graph convolution (MS-GCN) block allows the semantic information to represent bone joint and movement frames.

(3) Using the large-scale ($(m = 6, S = 6)/GL$), two stream joint semantics for the expressions of bone joint and movement frames at the same time, neither $(m = 6, S = 3)/GL$, nor $(m = 6, S = 6)/GL$

produce further benefits in comparison with $(m = 3, S = 3)/GL$. The main reason is that the large-scale structure contains too much redundant information, and the detailed information of the graph structure is lost, although multiple scales capture bone joints and motion frames from more levels.

4.3.2. Comparison to the State-of-the-Art

To further verify that the proposed model can better capture the multiscale and global contextual semantic information of bone nodes and motion frames compared with the state-of-the-art skeleton recognition model, we used the NTU60 and NTU120 public datasets in a comparison study. The experimental results obtained by many state-of-the-art human skeleton action recognition models are listed in Tables 2 and 3. At present, the state-of-the-art recognition methods contain graph and other (non-graph) models.

Table 2. Comparison of the results on NTU60/120 with X-View and X-Sub (%).

Models	X-View (NTU60)	X-Sub (NTU60)	X-View (NTU120)	X-Sub (NTU120)
$Part-AwareLSTM$ [22]	70.3	62.9	26.3	25.5
$STA-LSTM$ [5]	81.2	73.4	57.9	55.7
$CNN-MTLN$ [20]	84.8	79.6	57.9	58.4
$VA-RNN$ [41]	87.6	79.4	–	–
$ELATT-GRU$ [8]	88.4	80.7	–	–
$VA-CNN$ [12]	94.3	88.7	–	–
$ST-GCN$ [15]	88.3	81.5	–	–
$AS-GCN$ [42]	94.2	86.8	–	–
$GR-GCN$ [43]	94.3	87.5	–	–
$2S-GCN$ [17]	95.1	88.5	**84.9**	**82.9**
SGN [19]	94.5	89.0	81.5	79.2
$DAG-GCN$	**95.76**	**90.01**	82.44	79.03

In Table 2, a dash "–" indicates that these methods have not been experimentally tested on the dataset. A bold value indicates the highest value on the dataset.

According to the experimental results in Table 2, we draw the following conclusions:

(1) For the NTU60 dataset, the proposed DAG-GCN recognition method obtained the optimal results. DAG-GCN outperformed 2S-GCN and SGNby 0.66% and 1.26%, respectively, in the X-View accuracy. However, 2S-GCN outperformed DAG-GCN by 0.46% and 3.87%, respectively, for the X-View and X-Sub accuracy on the NTU120 dataset. This result indicates that the adaptive graph structure (2S-GCN) has a better embedding effect on large-scale input data. Moreover, the second-order information of the skeleton data is explicitly formulated and combined with the first-order information using a two stream framework, which brings a notable improvement in the recognition performance. The topology of the graph is adaptively learned for different GCN layers and skeleton samples in an end-to-end manner, which can better suit the action recognition task and the hierarchical structure of the GCNs.

(2) For the overall experimental results, we observe that the graph structure methods outperform non-graph methods with regard to accuracy on the NTU120 and NTU60 datasets. The main reason is that the graph-based structure methods can better obtain the deep semantic information of bone joints by aggregating and transferring neighbor node information. In addition, the correlation between adjacent nodes is used to strengthen the dependence between moving frames.

4.3.3. Performance of the Guided-Level Module

In this section, we discuss the different attention module performances of the skeleton action recognition framework. The performance results are listed in Table 3.

Table 3. Performance results of guided-level modules on NTU60/120 with the X-view and X-Sub (%).

Models	X-View (NTU60)	X-Sub (NTU60)	X-View (NTU120)	X-Sub (NTU120)
$BJL/(NG)$	90.69	78.88	65.49	62.57
$BJL/(PA)$	92.08	82.44	79.66	71.80
$BJL/(CA)$	93.75	85.94	81.58	76.66
$BJL/(GL)$	95.76	90.01	82.44	79.03

In Table 3, BJL indicates the bone joint-level module, NG the non-guided module, and PA and CA indicate position attention and channel attention, respectively.

According to the performance results in Table 3, we draw the following conclusions:

(1) For the NTU60 and NTU120 datasets, the proposed BJL/GL (DAG-GCN) recognition method obtained the optimal results, and compared to the non-guided module methods, we observe that by integrating either a position attention (PA) or channel attention (CA) module in the multiscale aggregate graph convolution (BJL), the performance improves by 11.13% and 16.46%, respectively, for the X-Sub accuracy, and by 5.07% and 16.95%, respectively, for the X-View accuracy. This result occurs because features generated by the proposed multiscale aggregate graph convolution are refined in the guided-level module frameworks in the proposed DAG-GCN, and the dependencies between motion frames are highlighted.

(2) In both datasets, compared to position attention (BJL/PA), the channel attention (BJL/CA) module has better performance in recognition accuracy. This finding demonstrates that even though both the position attention and channel attention modules exhibit a benign performance in recognition accuracy, the channel attention performance is better than that of position attention. In other words, the channel attention (CA) outperforms the position attention (PA) module when they are combined.

4.3.4. Visualization of the Recognition Results

Showing that the performance results are different through a correlation ablation experiment may not be enough to fully understand the advantage of the proposed DAG-GCN recognition framework. Although this framework improves the performance of skeleton action recognition (the results are shown in Tables 1–3), the training process and the impacts of different components are ignored. To this end, we visualize the impact of different components, and the training and testing process on the NUT60 datasets is shown below.

According to the training and testing process in Figure 2, we observe that the value of loss for the X-View on NTU60 dataset tends to be stable after being reduced by a certain degree.

The visualization recognition results of the consecutive frames from the NTU120 dataset are shown in Figure 3. According to the visualization results in Figure 3, we observe that both the "check time" and "clapping" actions are accurately recognized. Similarly, the semantics of continuous frames (in red) are better captured by the proposed recognition framework (DAG-GCN); that is, we can more effectively explore the detailed changes of the bone joints, especially when the region has a complex or subtle change (in green). These visualization results indicate that the DAG-GCNs can capture the finer semantic details of skeleton action while avoiding semantic loss and enhanced dependencies of consecutive frames. The selective aggregates of temporal and spatial information from a dual attention guided-level module help to capture contextual information, and the semantic information of the joints is gradually refined.

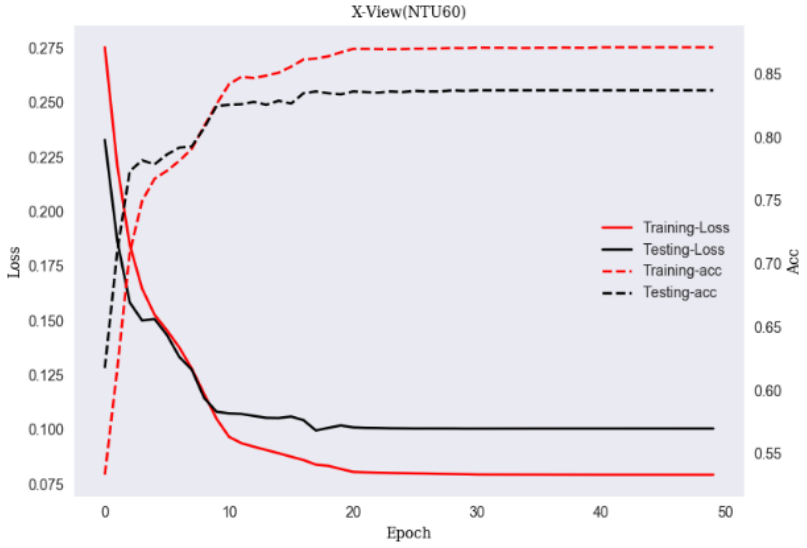

Figure 2. Training and testing process of accuracy and loss for X-view on the NTU60 dataset.

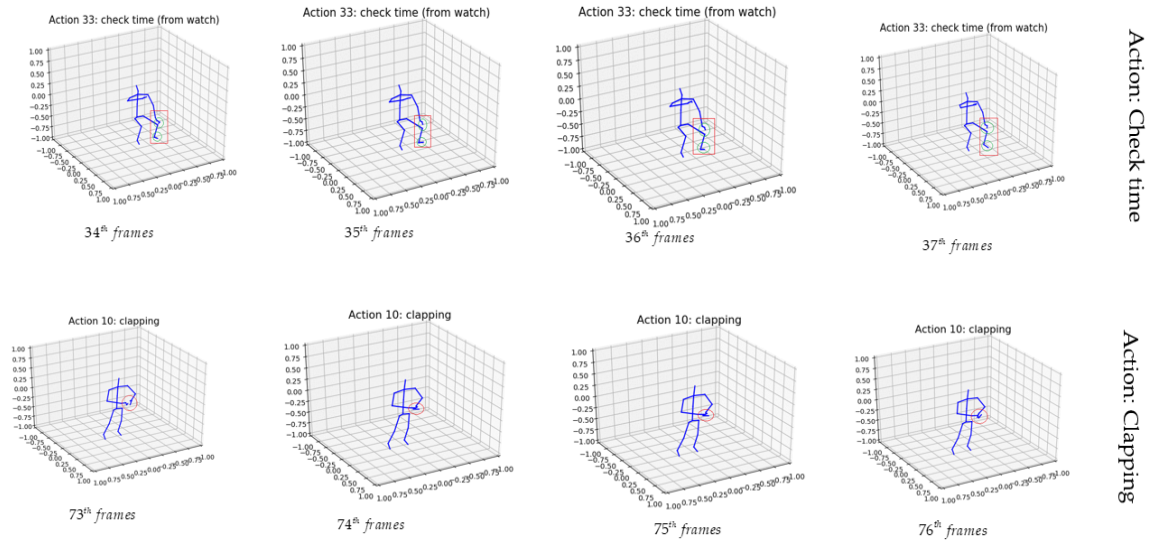

Figure 3. Visualization recognition results of the check time and clapping actions. Red indicates continuous action.

The visualization results of different attention modules (*PA* and *CA*) on the NTU60 datasets are shown in Figure 4. We observe that the response of semantic classes is more noticeable than that of channel attention (CA). Although position attention can effectively enhance specific semantics and locate key bone joint points, some non-keybone joint regions are still highlighted on the semantic maps. By contrast, we propose a guided-level (GL) module to capture semantic information that better focuses on the specific bone joints of the key regions of interest. In particular, we observe that captured semantic maps whose highlighted areas focus on a few key continuous bone joints can avoid similar continuous movements that might result in misrecognition of some actions.

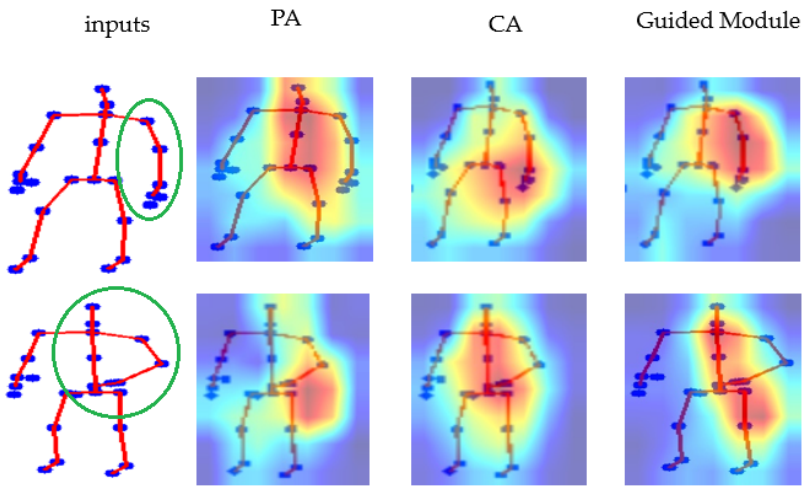

Figure 4. Visualization results of different attention modules. *PA* and *CA* indicate position attention and channel attention, respectively. The main parts of body movement (in green) are indicated as "*inputs*".

5. Conclusions and Next Research Works

We propose a dual attention-guided graph convolutional network for the task of human skeleton action recognition. The recognition framework incorporates a multiscale aggregate graph convolution strategy to combine multiscale semantic information of bone joints and consecutive frames at different levels and use guided attention to gradually aggregate and refine relevant local and global context semantic information. Then, the guided-level module filters irrelevant noisy information and improves the frameworks to focus on relevant class-specific bone joints in the skeleton. To validate our proposed DAG-GCN method, we conducted experiments on the NTU60 and NTU120 datasets in an action recognition task, and we presented the results of extensive ablation experiments to evaluate the effectiveness of the proposed methods.

The experimental results showed that the proposed recognition frameworks outperformed other models on the NTU60 datasets, mainly due to the rich multiscale contextual dependencies of local and global information. In addition, we enhanced the correlation between frames and bone joints to be more suitable for small-scale data. However, this approach was not superior to the 2S-GCN method on the NTU120 data, mainly because the three 2S-GCN types of semantic graph modules (structural graph extraction module, actional graph inference module, and attention graph iteration module) were employed in Sem-GCN to aggregate L-hop joint neighbor information in order to capture action-specific latent dependencies and to distribute importance levels. Compared with DAG-GCN, this approach is more suitable for large-scale data. In the future, we will explore a faster and more powerful semantic network for human skeleton action recognition.

Author Contributions: Z.H. conceived of and designed the experiments and wrote the paper. E.-J.L. supervised the study and reviewed the paper. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

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

Abbreviations

The following abbreviations are used in this manuscript:

DAG-GCNs	dual attention-guided multi-scale dynamic aggregate graph convolutional networks
GL	guided-level module
BJL	bone joint-level module
NG	non-graph

References

1. Kerdvibulvech, C. A Review of Augmented Reality-Based Human-Computer Interaction Applications of Gesture-Based Interaction. In Proceedings of the International Conference on Human-Computer Interaction, Orlando, FL, USA, 26–31 July 2019; Springer: Cham, Switzerland, 2019; pp. 233–242.
2. Zhang, K.; Sun, H.; Shi, W.; Feng, Y.; Jiang, Z.; Zhao, J. A Video Representation Method Based on Multi-view Structure Preserving Embedding for Action Retrieval. *IEEE Access* **2019**, *7*, 50400–50411. [CrossRef]
3. Hassan, M.M.; Ullah, S.; Hossain, M.S.; Alelaiwi, A. An end-to-end deep learning model for human activity recognition from highly sparse body sensor data in Internet of Medical Things environment. *J. Supercomput.* **2008**, *10*, 142–149.
4. Cao, Z.; Martinez, G.H.; Simon, T.; Wei, S.E.; Sheikh, Y.A. Open Pose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *99*, 1.
5. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. Skeleton-Indexed Deep Multi-Modal Feature Learning for High Performance Human Action Recognition. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6
6. Han, F.; Reily, B.; Hoff, W.; Zhang, H. Space-time representation of people based on 3D skeletal data. *Comput. Vis. Image Underst.* **2017**, *158*, 85–105. [CrossRef]
7. Si, C.; Jing, Y.; Wang, W.; Wang, L.; Tan, T. Skeleton-Based Action Recognition with Spatial Reasoning and Temporal Stack Learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 103–118.
8. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2117–2126.
9. Huynh-The, T.; Hua, C.H.; Kim, D.S. Learning Action Images Using Deep Convolutional Neural Networks For 3D Action Recognition. In Proceedings of the IEEE Sensors Applications Symposium (SAS), Sophia Antipolis, France, 11–13 March 2019; pp. 1–6.
10. Fan, H.; Luo, C.; Zeng, C.; Ferianc, M.; Que, Z.; Liu, S.; Niu, X.; Luk, W. F-E3D: FPGA-based Acceleration of an Efficient 3D Convolutional Neural Network for Human Action Recognition. In Proceedings of the IEEE 30th International Conference on Application-Specific Systems, Architectures and Processors (ASAP), New York, NY, USA,15–17 July 2019; pp. 1–8.
11. Wu, H.; Ma, X.; Li, Y. Hierarchical dynamic depth projected difference images–based action recognition in videos with convolutional neural networks. *Int. J. Adv. Robot. Syst.* **2019**, *16*. [CrossRef]
12. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1963–1978. [CrossRef]
13. Cho, S.; Maqbool, M.; Liu, F.; Foroosh, H. Self-Attention Network for Skeleton-based Human Action Recognition. *arXiv* **2019**, arXiv:1912.08435.
14. Liu, S.; Ma, X.; Wu, H.; Li, Y. An End to End Framework with Adaptive Spatio-Temporal Attention Module for Human Action Recognition. *IEEE Access* **2020**, *8*, 47220–47231. [CrossRef]
15. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv* **2018**, arXiv:1801.07455.
16. Kong, Y.; Li, L.; Zhang, K.; Ni, Q.; Han, J. Attention module-based spatial-temporal graph convolutional networks for skeleton-based action recognition. *J. Electron. Imaging* **2019**, *28*, 043032. [CrossRef]
17. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 12026–12035.

18. Liu, R.; Xu, C.; Zhang, T.; Zhao, W.; Cui, Z.; Yang, J. Si-GCN: Structure-induced Graph Convolution Network for Skeleton-based Action Recognition. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
19. Ding, X.; Yang, K.; Chen, W. A Semantics-Guided Graph Convolutional Network for Skeleton-Based Action Recognition. In Proceedings of the 2020 the 4th International Conference on Innovation in Artificial Intelligence, Xiamen China, 6–9 May 2020; pp. 130–136.
20. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A new representation of skeleton sequences for 3d action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3288–3297.
21. Du, Y.; Fu, Y.; Wang, L. Skeleton based action recognition with convolutional neural network. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 579–583.
22. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
23. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 816–833.
24. Majd, M.; Safabakhsh, R. Correlational Convolutional LSTM for Human Action Recognition. *Neurocomputing* **2019**, *396*, 224–229. [CrossRef]
25. Gammulle, H.; Denman, S.; Sridharan, S.; Fookes, C. Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 177–186.
26. Zhang, Z.; Lv, Z.; Gan, C.; Zhu, Q. Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions. *Neurocomputing* **2020**, *410*, 304–316.
27. Xiong, W.; Wu, L.; Alleva, F.; Droppo, J.; Huang, X.; Stolcke, A. The Microsoft 2017 Conversational Speech Recognition System. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5934–5938.
28. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
29. Liu, C.; Ying, J.; Yang, H.; Hu, X.; Liu, J. Improved human action recognition approach based on two-stream convolutional neural network model. *Vis. Comput.* **2020**, *6*, 28.
30. Torpey, D.; Celik, T. Human Action Recognition using Local Two-Stream Convolution Neural Network Features and Support Vector Machines. *arXiv* **2020**, arXiv:2002.09423.
31. Tang, Y.; Tian, Y.; Lu, J.; Li, P.; Zhou, J. Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5323–5332.
32. Li, F.; Zhu, A.; Xu, Y.; Cui, R.; Hua, G. Multi-stream and Enhanced Spatial-temporal Graph Convolution Network for Skeleton-based Action Recognition. *IEEE Access* **2020**, *8*, 97757–97770. [CrossRef]
33. Shiraki, K.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Acquisition of Optimal Connection Patterns for Skeleton-based Action Recognition with Graph Convolutional Networks. In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications–Volume 5: VISAPP, Valletta, Malta, 27–29 February 2020.
34. Wang, J.; Hu, J.; Qian, S.; Fang, Q.; Xu, C. Multimodal Graph Convolutional Networks for High Quality Content Recognition. *Neurocomputing* **2020**, *412*, 42–51. [CrossRef]
35. Qin, Y.; Mo, L.; Li, C.; Luo, J. Skeleton-based action recognition by part-aware graph convolutional networks. *Vis. Comput.* **2019**, *36*, 621–631.
36. Yang, D.; Li, M.M.; Fu, H.; Fan, J.; Leung, H. Centrality Graph Convolutional Networks for Skeleton-based Action Recognition. *Sensors* **2020**, *20*, 3499.
37. Yang, K.; Ding, X.; Chen, W. A Graph-Enhanced Convolution Network with Attention Gate for Skeleton Based Action Recognition. In Proceedings of the ICCPR '19: 2019 8th International Conference on Computing and Pattern Recognition, Beijing, China, 23–25 October 2018; Association for Computing Machinery: New York, NY, USA, 2019; pp. 342–347.

38. Rashid, M.; Kjellstrm, H.; Lee, Y.J. Action Graphs: Weakly-supervised Action Localization with Graph Convolution Networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 615–624.
39. Lee, J.; Jung, Y.; Kim, H. Dual Attention in Time and Frequency Domain for Voice Activity Detection. *arXiv* **2020**, arXiv:2003.12266.
40. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 3146–3154.
41. Zhang, P.; Xue, J.; Lan, C.; Zeng, W.; Gao, Z.; Zheng, N. Adding attentiveness to the neurons in recurrent neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 135–151.
42. 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 Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3595–3603..
43. Gao, X.; Hu, W.; Tang, J.; Liu, J.; Guo, Z. Optimized skeleton-based action recognition via sparsified graph regression. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 601–610.