electronics

MDPI

Article

Sequence Segmentation Attention Network for Skeleton-Based Action Recognition

Yujie Zhang and Haibin Cai *

Software Engineering Institute, East China Normal University, Shanghai 200062, China
* Correspondence: hbcai@sei.ecnu.edu.cn

Abstract: With skeleton-based action recognition, it is crucial to recognize the dependencies among joints. However, the current methods are not able to capture the relativity of the various joints among the frames, which is extremely helpful because various parts of the body are moving at the same time. In order to solve this problem, a new sequence segmentation attention network (SSAN) is presented. The successive frames are encoded in each of the segments that make up the skeleton sequence. Then, we provide a self-attention block that may record the associated information among various joints in successive frames. In order to better recognize comparable behavior, a model of external segment action attention is employed to acquire the deep interrelation information among parts. Compared with the most advanced approaches, we have shown that the proposed method performs better on NTU RGB+D and NTU RGB+D 120.

Keywords: human action recognition; skeleton data; self-attention; attention mechanism

1. Introduction

Human motion recognition can play a great role in a number of industries, including intelligent video surveillance, virtual reality, and human–computer interaction. In contrast to other methods, a lot of interest has been shown in skeleton-based action recognition in recent years due to its strong robustness to complex environments and camera views. When describing how the body moves, position information is usually used to help explain the process of body movement. According to the actual situation, you can use 2D or 3D coordinates to mark the joint position. It is simple to obtain the skeletal data using pose estimating methods or depth cameras [1].

Most of the methods used in the past [2–4] are based on artificial features that do not gain the temporal and spatial features of the skeletal structure. In recent years, deep learning has been widely used in the field of convolutional neural networks (CNNs) [5–7] and recurrent neural networks (RNNs) [8–10]. Despite their many successes over the past few years, they cannot be used to uncover important connections between joints. Rather, they can only be used to process routine data in Euclidean space. The joints in the human skeleton act as vertices, and the body's bones act as edges in a graph that is formed according to natural rules. In recent years, graph convolutional networks (GCNs) have been widely used in machine learning. A novel method named spatio-temporal graph convolutional networks (ST-GCN) is suggested by Yan et al. [11]. The GCN and one-dimensional temporal convolution are the two essential parts of the ST-GCN. Through these two parts, we can obtain space and time information, respectively. GCN can extract the position information and bone structure information of bones by using the graph structure. One-dimensional time convolution mainly acts on adjacent frames and can obtain joint motion information. Shi et al. [12] propose the two-stream adaptive graph convolution network (2s-AGCN) that utilizes an adaptive adjacency matrix to capture the mutual relationships of the nodes. Furthermore, by modeling the joint and bone data, the 2s-AGCN can improve performance. Subsequently, many methods based on AGCN

Citation: Zhang, Y.; Cai, H. Sequence Segmentation Attention Network for Skeleton-Based Action Recognition. *Electronics* **2023**, *12*, 1549. https://doi.org/10.3390/electronics12071549

Academic Editor: Donghyeon Cho

Received: 22 February 2023
Revised: 16 March 2023
Accepted: 23 March 2023
Published: 25 March 2023

and ST-GCN have since been developed in order to obtain an action recognition level of performance that has never before been attained.

While ST-GCN and its successors have demonstrated the capability of extracting spatial and temporal features, the majority of GCN-based models are deficient. It is observed that there is a correlation between the different joints in a number of successive frames. In the "Wear a shoe" action, for example, the hand will drive the body downward and close to the foot, and the joint will be displaced as a result of the overall motion. As a result of the motion, joints that were not associated in the previous frame will be connected in the next frame. As a result, it is helpful to obtain the relevant characteristics of the neighboring frames from different joints. However, these approaches are not able to capture this association efficiently. In the GCN-based approach, for example, Yan et al. [11] construct a spatio-temporal graph that ignores the relationships between different joints within the same frame and only propagates associated information between the same joints in different frames.

We propose a new sequence segmentation attention network (SSAN) for skeleton-based action recognition according to the foregoing. In particular, a skeleton sequence is split into a number of mutually non-overlapping segments, each consisting of a number of consecutive frames. Due to the correlation between the various joints in successive frames, we propose an internal self-attention (ISA) block. Using this block, it is possible to easily and efficiently extract the correlative information among the joints in every segment. Due to the reduction in the temporal dimension by using segmentation, the computation cost of the block is small. Moreover, a segment can be thought of as a decomposed action, whereas a whole action is made up of several decomposed actions. On the basis of this, a module called the external segment attention (ESA) block is proposed to group these sub-actions and record important temporal and spatial information that occurs between segments. This module improves the recognition accuracy of similar actions.

The contributions of this paper can be summarized as follows:

- In this paper, we propose a segmentation and encoding strategy for a skeleton sequence, which links joints in consecutive frames into a series of sequences and inserts a position-coding module to obtain spatio-temporal features.
- We suggest using an internal self-attention block to record pertinent information between various joints in adjacent frames and an external segment attention block to merge all actions.
- The validity of each module was established by ablation experiments. On two large datasets, NTU RGB+D and NTU RGB+D 120, our model performs superbly.

2. Related Work

2.1. Skeleton-Based Action Recognition

The development of depth sensing techniques (such as Kinect [13]) and pose estimation algorithms [1,14] makes it possible to collect bone data in real-time by locating key joints. The skeleton data are robust to the changes in the lighting, the changes in the scene, and the complexity of the background. This robustness has been shown to be useful for data-driven tasks such as skeleton-based action recognition. In the past, it was difficult to extract useful information from skeleton sequences automatically in motion recognition. Therefore, people need to manually and painstakingly collate the required characteristics of the tag. Some conventional methods [3,15–17] rely on actions view-invariant characteristics. Some examples of these characteristics are skeletal quads [3,15], cluster sparse vocabulary encoding [16], and canonical representation of transform-based functions [17]. Several methods in the past have combined data from the various 3D action datasets modes. To increase performance, many works [18–20] have associated depth information with the skeleton. HOG features [18] and Fourier temporal pyramids [20] are used to represent the depth information, or they can use random decision forests for modeling [19]. Many machine learning applications have been made possible by recent developments in deep learning. RNN and CNN are the most popular models. RNN-based models [8,21–23]

generally associate the coordinates of all the joints in each frame as vectors, then the vector sequence is used as input to model the characteristics of the action. LSTM-IRN [24] provides the Interaction Relational Network to assure proper learning of the interaction mode. CNN-based models [6,25,26] generate a pseudo image by stacking a sequence of vectors, which is then converted into an image classification task. In order to perform better than a single network, the two-stream-based model [27] combines RNN and CNN, which work on skeleton and RGB images.

Nevertheless, because a skeleton is not a vector or a two-dimensional grid, joint vector sequence and semantic images are not suitable for representing the skeleton structure. Zhou et al. [28] propose a new bottom-up mechanism for learning category-level human semantic segmentation as well as multi-person pose estimation in a federated and end-to-end manner. With compact, efficient and powerful properties, it exploits structural information at different human granularities to alleviate the difficulties of human segmentation. In recent years, Yan et al. [11] proposed ST-GCN, which is used to model skeleton data directly as a graph and has improved performance. Obinata et al. [29] deal only with the relation of the local joints among the frames. The high power of the adjacency matrix can increase the perceptual field of the graph convolution, but it can result in node weight bias and inefficient remote modeling. In order to address this issue, Liu et al. [30] remove unnecessary dependencies between adjacent neighbors by using the proposed neighborhood de-entanglement approach. Li et al. [31] proposed a spatial temporal graph routing (STGR) network, which utilizes the framework of attention and the global self-attention mechanism. Asymmetric correlation measurement and high-level representation are used to calculate context information in the superior CA-GCN [32] approach proposed by Zhang and Ye et al. [33] proposed Dynamic GCN for the automatic learning of skeleton topology by using a new convolutional neural network. Cheng et al. [34] proposed a novel shift graph convolutional network (Shift-GCN). To improve performance and decrease the computational complexity of spatial graph convolution, this approach uses a unique shift graph and lightweight point-wise convolutions. Liu et al. [30] deployed more efficient spatio-temporal edges for 3D graph convolution.

2.2. Attention Mechanism

In recent years, attention mechanisms have shown great prospects in the wide application of machine learning in various fields. The attention mechanism is designed to concentrate on the key components or characteristics of the input that are important to decision making. Self-attention [35] refers to the attention mechanism with the same input and output dimensions, which can also be called intra-attention. The MATNet proposed by Zhou et al. [36] designs an asymmetric attention module in the dual-stream encoder, with the help of which the appearance features are transformed into a motion-focused representation at each res stage, allowing for a tight hierarchical interaction between object motion and appearance right during encoder encoding. There are many studies on the importance of attention to graph structure. For example, Lee et al. [37] proposed the graph attention model (GAM), which employs an RNN framework using a self-attention mechanism to deal with the graph structural data. The graph attention networks (GAT) framework was proposed by VeliBckovi et al. [38]. This method is equivalent to the combination of multiple self-attention modules, combining the characteristics of each head by using a function called concatenation or averaging. Multiple attention mechanisms can focus on a variety of different information. By paying attention to neighbors to calculate the set of hidden representations of each node in the network, we can obtain the degree of influence of neighbor nodes on themselves and assign different weights to distinguish their degree of concern to other nodes. As a result, the model's capacity is increased. Zhang et al. [39] proposed a self-attention mechanism based on graph structure, which computes an additional gating mechanism for each head, in contrast to GAT, which is based on the premise that the contributions of the attention heads are equal. Since it has been

shown that the self-attention mechanism can enhance the performance of classification, it is worth incorporating the self-attention mechanism in our strategy.

3. Methods

In this section, we begin by summarizing the overall architecture of this approach. In the next section, we propose a new method of spatial-temporal segment coding. Secondly, an internal self-attention block is presented, which is used to model the relationship between various joints. At last, it presents an external segment attention block, which can be used to combine the movement information among sequence segments.

3.1. Overall Architecture

Figure 1 illustrates the overall architecture of our model. A skeleton sequence composed of V0 node and T0 frame is input into the model, and it is transmitted to the segments encoding module to obtain the corresponding format of data. In the segments encoding module, the skeleton sequence is split into segments of the same number of frames, and each frame is linked in each segment. Then, we input the encoded data into SSAN and obtain the related information on spatial-temporal joints. The module consists of an ISA block and an ESA block. The ISA block is used to model the relationship among the links in consecutive frames, and the sub-action is assembled by using the ESA block. After compressing and reducing the amount of feature data in the global average pooling layer and the full connection layer, we achieve the goal of classification.

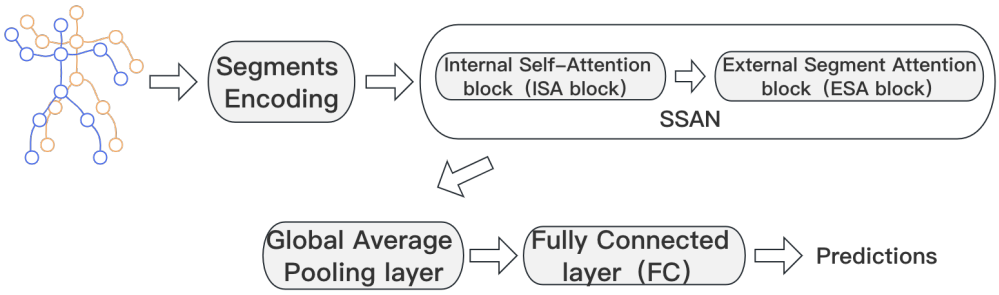

Figure 1. Illustration of the overall architecture of the proposed method. The segments encoding module rejoins the skeleton sequence segments and encodes them into a data format that has been added to the location ID. The ISA block is used to obtain the relationship of the node to different nodes within adjacent frames. The ESA block is used to integrate all the decomposed actions. Finally, the classification scores are obtained through the global average pooling layer and the full connection layer.

3.2. Segments Encoding

We describe an approach for encoding the joints in order to model the relation between different joints in successive frames. Figure 2 depicts the skeleton sequence segments encoding process.

First, the original skeleton sequence X is fed into the feature mapping layer, and the input channel is extended from C0 to the a set number C1. A single convolution layer using the Leaky ReLU and BatchNorm makes up the feature mapping layer. The main role of the feature mapping layer is to keep the input distribution of each layer of the neural network the same during training, speed up convergence and prevent overfitting. In each convolutional layer, the data are in three dimensions. It can be thought of as a number of two-dimensional images stacked on top of each other, each of which is called a feature map, and in the input layer, if it is a color image, there are typically three feature maps (red, green and blue). There are a number of filters between the layers, the size (width, height) and depth (length), which are set manually, commonly 3×3 or 5×5. In general, the number of the next layer of feature maps depends on the number of convolution kernel,

which results in a new channel number C1. The skeleton sequence is then separated into T contiguous segments that do not overlap. There are n frames in each sequence segment. Then, within each segment, the frame sequence is flattened. The spatio-temporal segment encoding layer consists of a convolutional layer with a Leaky ReLU function, into which X is fed to obtain the final segment encoding. Because the tensors generated by segment coding do not include the sequence of the joints, and the identification of the joints can not be differentiated so that the performance of motion recognition can be decreased [40]. In view of this problem, a location-coding module is adopted to mark every joint. It is important to note that in order to model all the joints in a segment, the same joints from different frames must be distinguished from each other in a section so that all the joints are labeled with different IDs.

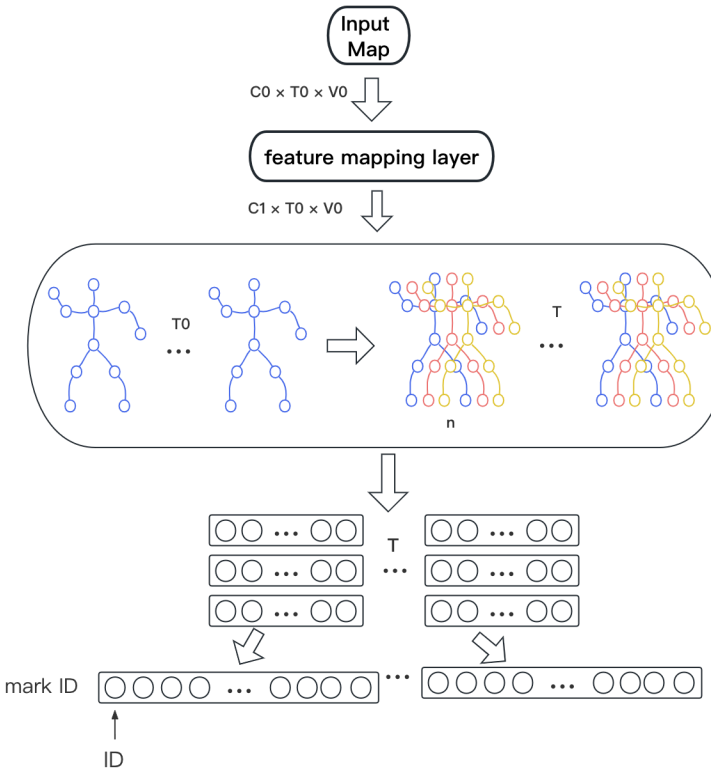

Figure 2. The process of segment encoding.

Position encoding is a secondary representation of each piece of sequence information in a sequence using positional information. The attention model itself does not have the ability to learn sequence positional information like an RNN, which has decided at the time of model definition that the order of the input-by-input information is equivalent to the order in which the information appears. The attention model needs to actively feed the sequential information of the sequence to the model. Each length of sequence should have a unique position code. For example, for sequences of lengths 600 and 6, the position encoding from 1 to 6 should be the same. For the length of the untrained sequence, the model should be easy to generalize. In addition, the encoding result should be bounded and have a relatively appropriate range because if it is too large, it will override the weight of the model. Therefore, the choice of the sine and cosine function for position encoding can well achieve the above requirements. According to [41], the relative position information of various joints is encoded utilizing periodic sine-cosine functions at various frequencies:

$$PE(np, 2i) = sin(np/10,000^{2i/C_{in}}), \quad (1)$$

$$PE(np, 2i+1) = cos(np/10,000^{2i/C_{in}}), \tag{2}$$

where n is the total of frames in each segment, p is the joint's location, and i is the vector's dimension for position encoding. With the beginning of the sequence as the reference position, p represents the position of the joint in the sequence. The value of p ranges from 0 to the length of the sequence.

3.3. Internal Self-Attention

In essence, self-attention is a mapping from a query to a collection of key-value pairs. Following the skeleton sequence's position-coding and spatial-temporal segment coding, we can construct the relations among the input tags by means of multi-headed self-attention. The ISA block is shown in Figure 3.

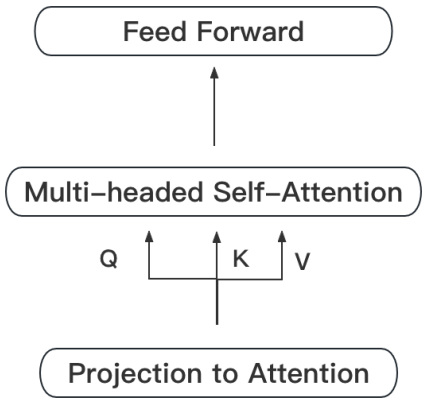

Figure 3. The structure of an internal self-attention block.

In particular, in the calculation of self attention, it is necessary to take into account not only the effect of all the others but also the effect on the others. Thus, the encoding sequence X is normally projected by means of a convolution layer to query Q, key K, and value V:

$$\hat{X} = Conv_{2D(1\times1)}(X) \in \mathbb{R}^{3C\times T\times V} \tag{3}$$

$$Q, K, V = Split(\hat{X}) \in \mathbb{R}^{C\times T\times V} \tag{4}$$

In this way, we can obtain Q, K, V by dividing the channel of X. The weight can then be determined by calculating the similarity between query Q and the key K's transposition. The dot-product is simply utilized as the similarity function, just like the ordinary Transformer. The Tanh function is then used to normalize the weights that were acquired. This research offers an optimal correlation matrix A to analyze the intrinsic topology of the human body while taking into account the fixed relationship of human joints. Moreover, the intensity of the attention map is further adjusted using an optimized parameter α. The final attention is then calculated by multiplying the final attention weight by the matching amount V:

$$X_{dot} = Tanh(\frac{QK^T}{\sqrt{C}}) \in \mathbb{R}^{T\times V\times V} \tag{5}$$

$$X_{attn} = V(\alpha \times X_{dot} + A) \in \mathbb{R}^{C\times T\times V} \tag{6}$$

where $\sqrt{C}$ is used to prevent too much inner product and to increase the gradient's stability during training.

A better performance can be attained by using a multiple-headed self-attention scheme, which enables the model to learn the relevant information from a variety of sub-spaces. In order to achieve a better performance, a multiple-headed self-attention scheme is used, which enables the model to learn the relevant information from a variety of sub-spaces. In particular, the self-attention calculation is carried out on a plurality of groups that are projected with different learning parameters. Then, a plurality of attention groups is connected.

A convolution layer with a 1×1 kernel size projects the learned attention into the output space. To enhance the network's performance, a feed-forward layer based on 1×1 2D convolutions is proposed:

$$X_{ISA} = Conv_{2D(1\times1)}(X_{attn}) \in \mathbb{R}^{C\times T\times V} \tag{7}$$

3.4. External Segment Attention

A movement may be considered as a series of successive sub-movements, for example, "Drinking water" consists of "holding the cup", "raising the arm", "opening the mouth" and other actions. In our approach, a decomposed action is included in each segment, and the ISA block is used to model a sequence of N frames. If the relationship between the decomposed actions is established, it will assist in recognizing actions and distinguishing similar actions. For this reason, it is proposed that an ESA block be used to bring together these decomposed actions. As illustrated in Figure 4, the inter-segment motion information is synthesized in this paper using a multi-scale convolution method.

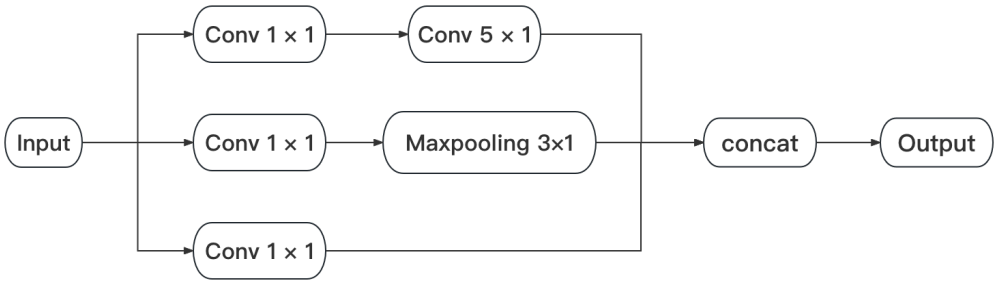

Figure 4. The structure of the external segment attention block.

In particular, the X_{ISA} channel dimension is reduced by 1×1 convolutions to concentrate on more efficient features:

$$X_1 = \phi_1(X_{ISA}) \in \mathbb{R}^{\frac{C}{2}\times T\times V} \tag{8}$$

$$X_2 = \phi_2(X_{ISA}) \in \mathbb{R}^{\frac{C}{4}\times T\times V} \tag{9}$$

$$X_3 = \phi_3(X_{ISA}) \in \mathbb{R}^{\frac{C}{4}\times T\times V} \tag{10}$$

After that, a 5×1 convolution layer with a dilation rate of 1 and a 3×1 maximum pooling layer, respectively, receive the produced features X_1 and X_2. Time information $X_1^{'}$ is extracted by the convolution layer, and key features $X_2^{'}$ are acquired by the maximum pooling layer. It is important to note that even if this convolution layer's expansion rate is 1, there are n frames per segment, making the implementation of the expansion convolution over n frames equal. This saves on computing costs because the high expansion rate is avoided. Finally, the branches are aggregated and connected to obtain the aggregation feature X_{ESA} between segments:

$$X_{ESA} = Concat(X_1^{'} + X_2^{'} + X_3) \in \mathbb{R}^{C\times T\times V} \tag{11}$$

The remaining connections are subsequently employed to stabilize network training. It is important to note that all outputs associated with the remainder must be standardized.

4. Experimental Results and Discussion

In this section, we have performed a wide range of comparative experiments to evaluate the effectiveness of the proposed method. First of all, the experiment setup and datasets are introduced. The contribution of each component is assessed using SSAN based on NTU RGB+D skeleton data in the section that follows. Finally, by contrasting SSAN with the most advanced techniques on two datasets, NTU RGB+D and NTU RGB+D 120, the superiority of SSAN is demonstrated.

4.1. Datasets

NTU RGB+D is one of the most popular data sets in this field. It consists of 56,880 video-clips, with one action in each, and there are 60 classes in all. The Microsoft Kinect Depth Sensor recorded the dataset. There were three cameras for each action, each positioned at the same height but with a different horizontal angle: -45° , 0° , 45°. A total of 40 participants, ranging in age from 10 to 35, carried out the tasks. Each subject contains 25 3D coordinates, yet there are only two subjects in a single video clip. Two criteria are suggested in the data set's founding document [9]: (1) Cross-subject (X-sub): The data set consists of a training set and a validation set. The training set consists of 40,320 video clips, and the validation set contains 16,560 video clips, and the participants are not the same. Of the 40 subjects, half will be used for testing and the other half for training. (2) Cross-view (X-view): The training package consists of 7920 videos taken by the camera at 0°, 45°, and the verification package includes 18,960 frames that are taken by the camera at -45°.

NTU RGB+D 120 is an extension of NTU RGB+D. The data set consists of 114,480 actions, which are implemented by 106 different subjects. The dataset consists of 32 settings, with each setting indicating a particular position and background. The original method [42] suggested two criteria: (1) Cross-subject (X-sub120): 106 subjects are divided into training and examination groups, with half of the 106 being used for training and the other half for examination. (2) Cross-setup (X-set120) benchmark: Even setup IDs are used for the training data, whereas odd setup IDs are used for the test data.

4.2. Experimental Setting

By playing back the actions, all skeletal sequences were padded to 60 frames. With a weight decline of 0.0004 and a Nesterov momentum of 0.9, we train our model using an SGD optimizer. The training time is 120 s, and the initial learning rate is 0.1. The batch size is 64. Each segment consists of 3 successive frames. The output channels are 64, 64, 128, 128, 256, 256 and 256, and the spatial temporal segment attention block is set to 7.

4.3. Ablation Study

4.3.1. Ablation Study for SSAN

The validity of SSAN was examined on the NTU RGB+D dataset. With the exception of the object being compared, all other conditions are the same to allow for a fair comparison. The comparison tests in Table 1 show that SSAN is effective. In the table, SA means that the self-attention is computed only in and among the frames. The parameter n is the number of consecutive frames and acts on the internal self-attention block. Option $n = 1$ means that only the relationship of joints within frames is modeled. Option $n = 3$ means that the relationship of each joint between three consecutive frames is modeled at the same time. PE means position encoding. Its function is to distinguish the joints in different parts of the same frame and the joints in the same part of adjacent frames so as to obtain more spatio-temporal information. ESA stands for the external segment attention (ESA) block, and the purpose of this option is to explore the impact of the association information between the segmented sub-actions on the model.

Table 1. Using the NTU RGB+D dataset, the SSAN ablation research was conducted.

SA	$n = 1$	$n = 3$	PE	ESA	X-Sub (%)	X-View (%)
yes	yes	-	-	-	89.7	94.6
-	yes	-	yes	yes	90.1	94.9
-	-	yes	-	yes	90.0	95.1
-	-	yes	yes	-	90.1	95.2
-	-	yes	yes	yes	92.9	96.7

From the above, we can find that the model using only the self-attention mechanism obtains the worst result. The model choosing three frames as segmentation conditions, adding the position encoding module and ESA block achieves the best results. Selecting three successive frames as a sequence segmentation is more accurate than a single frame because of the correlation between the different joints of adjacent frames. It confirms that our ISA module is valid. Moreover, the precision of the model with no position coding is inferior to that of the whole model. It is mainly due to the difference in the function of space and time, and the rational utilization of the sequence information can enhance the performance of the system. In addition, not using ESA blocks degrades model accuracy. The reason is that this model can be used to build the relationship of each decomposition action and catch the key parts of the motion, which is helpful to differentiate similar behavior and increase the performance.

4.3.2. Impact of the Variable n

The influence of successive frames n is investigated, as illustrated in Figure 5. The NTU RGB+D skeleton data set has an average frame length of 83 frames, and we take 60 frames per frame. The results show that when $n = 3$, our model accuracy achieves the best value. It is impossible to effectively record the relationship between the successive frames if n is too small. If n is too big, the coupling relation of the successive frames becomes more complicated, and the correlation between the first frame and the last frame of every section is very low.

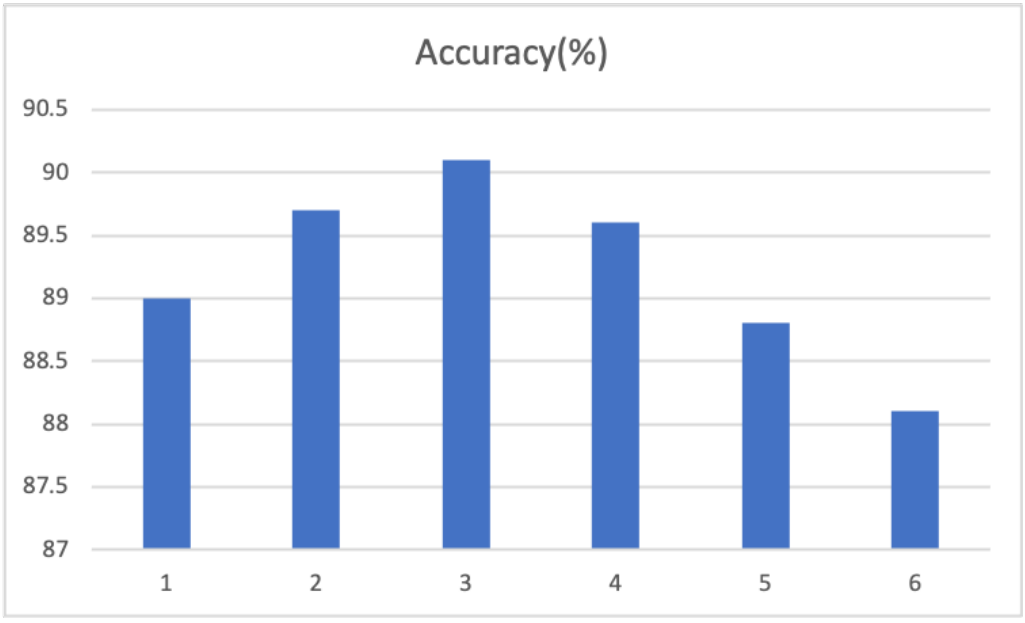

Figure 5. The influence of the parameter n on the model based on the NTU RGB+D skeleton data set.

4.3.3. Accuracy of Each Class

Figure 6 shows the accuracy of every category on the NTU RGB+D dataset. There are 60 different activity classifications represented on the horizontal axis. Relative to the ST-GCN baseline, we have made significant improvements in the majority of subjects, especially the "clapping", and "pointing to sth with finger" and "rub two hands together". Moreover, we have discovered that the precision of certain movements, such as "writing", "reading", is still very low. The main cause of this problem is that the motion only happens in a limited number of joints, and the effect magnitude is low. Therefore, this method can not be used to model the key joints.

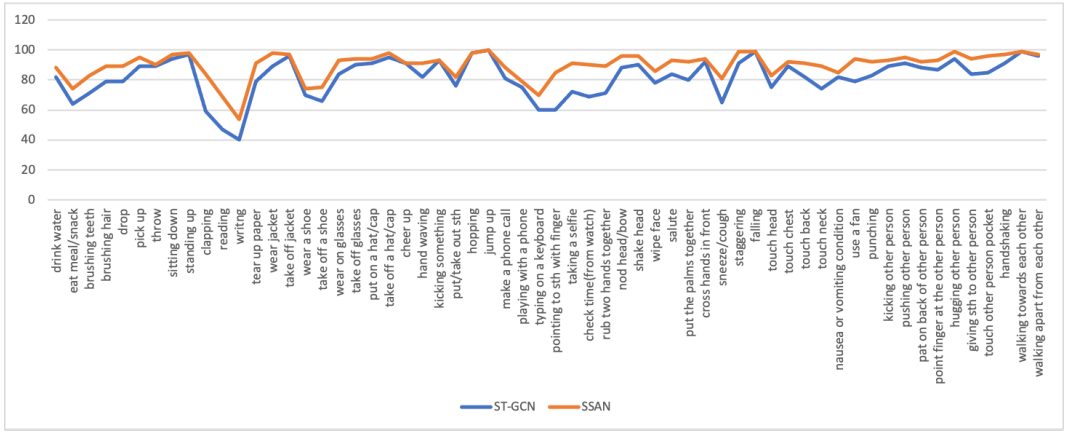

Figure 6. The influence of the number of consecutive frames n on the model based on the NTU RGB+D skeleton data set.

4.4. Comparison with the Most Advanced Methods

The SSAN algorithm is compared with the most advanced methods in NTU RGB+D and NTU RGB+D 120 datasets. The accuracy comparison is shown in Table 2. There are approaches based on RNNs, CNNs, GCNs, and transformers. In contrast to RNN-based and CNN-based approaches, we have a notable advantage in SSAN. The main cause of the bad performance of CNN and RNN is that they cannot fully utilize skeleton data. On the other hand, GCN-based approaches are able to utilize skeleton data in an efficient way and achieve better recognition performance. Furthermore, the 3s-Cros SCLR is a non-supervised approach, which can exploit cross-view consistency and obtain more advantages than other supervisory approaches.

Our approach performs better than the majority of current approaches in both data sets. The ISA block is able to benefit from the relative knowledge of various joints among successive frames, and the ESA block is able to model the movement information among segments and capture the sensitive critical sections. Compared with similar approaches, such as ST-TR, we have a much smaller number of parameters, but we perform much better.

Table 2. On NTU RGB+D and NTU RGB+D120 skeleton datasets, the recognition accuracy and parameter size are compared with the most advanced methods.

Methods	Param ($\times 10^6$)	X-Sub(%)	X-View(%)	X-Sub120(%)	X-Set120(%)
3s-CrossSCLR [43]	-	86.2	92.5	80.5	80.4
ST-LSTM [10]	-	69.2	77.7	-	-
ST-GCN [11]	3.1	81.5	88.3	-	-
2s-AGCN [12]	6.9	88.5	95.1	82.9	84.9
Shift-GCN [34]	-	90.7	96.5	85.9	87.6
Dynamic-GCN [33]	14.4	91.5	96.0	85.9	87.6
CTR-GCN [44]	5.8	92.4	96.8	88.9	90.6
SSAN(Ours)	5.7	92.9	96.7	88.9	90.8

5. Conclusions

In this paper, a new sequence segmentation attention network approach is presented in this paper. The algorithm includes three modules: segment coding, ISA block and ESA block, in which a number of successive frames are coded into a sequence, the ISA block is adopted to efficiently capture the relation between the successive frames, and the ESA block is applied to sum up the movement between them. Ablative research has demonstrated the validity of this approach. Based on the NTU RGB+D and NTU RGB+D 120 datasets, the proposed SSAN is superior to existing state-of-the-art techniques. With the increase in video dimension, it will limit the attention mechanism. The skeleton data itself contains human body structure features. How to make better use of high-dimensional information and human body structure features is the direction of future improvement.

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

Funding: This research received no external funding.

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

References

1. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef] [PubMed]
2. Hussein, M.E.; Torki, M.; Gowayyed, M.A.; El-Saban, M. Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations. In Proceedings of the IJCAI 2013, 23rd International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013; Rossi, F., Ed.; IJCAI/AAAI: Palo Alto, CA, USA, 2013; pp. 2466–2472.
3. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014; IEEE Computer Society: Washington, DC, USA, 2014; pp. 588–595. [CrossRef]
4. Fernando, B.; Gavves, E.; Oramas, J.M.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 5378–5387. [CrossRef]
5. Du, Y.; Fu, Y.; Wang, L. Skeleton based action recognition with convolutional neural network. In Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015, Kuala Lumpur, Malaysia, 3–6 November 2015; IEEE: New York, NY, USA, 2015; pp. 579–583. [CrossRef]
6. Liu, H.; Tu, J.; Liu, M. Two-Stream 3D Convolutional Neural Network for Skeleton-Based Action Recognition. *arXiv* **2017**, arXiv:1705.08106 .
7. Li, B.; Dai, Y.; Cheng, X.; Chen, H.; Lin, Y.; He, M. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops, Hong Kong, China, 10–14 July 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 601–604. [CrossRef]
8. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 1110–1118. [CrossRef]
9. Shahroudy, A.; Liu, J.; Ng, T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 1010–1019. [CrossRef]
10. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. In Proceedings of the Computer Vision—ECCV 2016—14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part III; Lecture Notes in Computer Science; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9907, pp. 816–833. [CrossRef]
11. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, 2–7 February 2018; McIlraith, S.A., Weinberger, K.Q., Eds.; AAAI Press: Menlo Park, CA, USA, 2018; pp. 7444–7452.

12. 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, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; Computer Vision Foundation/IEEE: New York, NY, USA, 2019; pp. 12026–12035. [CrossRef]
13. Zhang, Z. Microsoft Kinect Sensor and Its Effect. *IEEE Multim.* **2012**, *19*, 4–10. [CrossRef]
14. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014; IEEE Computer Society: Washington, DC, USA, 2014; pp. 1653–1660. [CrossRef]
15. Evangelidis, G.D.; Singh, G.; Horaud, R. Skeletal Quads: Human Action Recognition Using Joint Quadruples. In Proceedings of the 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, 24–28 August 2014; IEEE Computer Society: Washington, DC, USA, 2014; pp. 4513–4518. [CrossRef]
16. Luo, J.; Wang, W.; Qi, H. Group Sparsity and Geometry Constrained Dictionary Learning for Action Recognition from Depth Maps. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, 1–8 December 2013; IEEE Computer Society: Washington, DC, USA, 2013; pp. 1809–1816. [CrossRef]
17. Rahmani, H.; Mian, A.S. Learning a non-linear knowledge transfer model for cross-view action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 2458–2466. [CrossRef]
18. Hu, J.; Zheng, W.; Lai, J.; Zhang, J. Jointly Learning Heterogeneous Features for RGB-D Activity Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2186–2200. [CrossRef] [PubMed]
19. Rahmani, H.; Mahmood, A.; Huynh, D.Q.; Mian, A.S. Real time action recognition using histograms of depth gradients and random decision forests. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, 24–26 March 2014; IEEE Computer Society: Washington, DC, USA, 2014; pp. 626–633. [CrossRef]
20. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Learning Actionlet Ensemble for 3D Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 914–927. [CrossRef] [PubMed]
21. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. In Proceedings of the Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Singh, S., Markovitch, S., Eds.; AAAI Press: Menlo Park, CA, USA, 2017; pp. 4263–4270.
22. 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, ICCV 2017, Venice, Italy, 22–29 October 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 2136–2145. [CrossRef]
23. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently Recurrent Neural Network (IndRNN): Building a Longer and Deeper RNN. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Washington, DC, USA, 2018; pp. 5457–5466. [CrossRef]
24. Perez, M.; Liu, J.; Kot, A.C. Interaction Relational Network for Mutual Action Recognition. *IEEE Trans. Multim.* **2022**, *24*, 366–376. [CrossRef]
25. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.A.; Boussaïd, F. A New Representation of Skeleton Sequences for 3D Action Recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2017, pp. 4570–4579. [CrossRef]
26. Liu, M.; Liu, H.; Chen, C. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognit.* **2017**, *68*, 346–362. [CrossRef]
27. Zhao, R.; Ali, H.; van der Smagt, P. Two-stream RNN/CNN for action recognition in 3D videos. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, 24–28 September 2017; IEEE: New York, NY, USA, 2017; pp. 4260–4267. [CrossRef]
28. Zhou, T.; Wang, W.; Liu, S.; Yang, Y.; Gool, L.V. Differentiable Multi-Granularity Human Representation Learning for Instance-Aware Human Semantic Parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; Computer Vision Foundation/IEEE: New York, NY, USA, 2021; pp. 1622–1631. [CrossRef]
29. Obinata, Y.; Yamamoto, T. Temporal Extension Module for Skeleton-Based Action Recognition. In Proceedings of the 25th International Conference on Pattern Recognition, ICPR 2020, Milan, Italy, 10–15 January 2021; IEEE: New York, NY, USA, 2020; pp. 534–540. [CrossRef]
30. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: New York, NY, USA, 2020; pp. 140–149. [CrossRef]
31. Li, B.; Li, X.; Zhang, Z.; Wu, F. Spatio-Temporal Graph Routing for Skeleton-Based Action Recognition. In Proceedings of the The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, HI, USA, 27 January–1 February, 2019; AAAI Press: Menlo Park, CA, USA; pp. 8561–8568. [CrossRef]
32. Zhang, X.; Xu, C.; Tao, D. Context Aware Graph Convolution for Skeleton-Based Action Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: New York, NY, USA, 2020; pp. 14321–14330. [CrossRef]

33. Ye, F.; Pu, S.; Zhong, Q.; Li, C.; Xie, D.; Tang, H. Dynamic GCN: Context-enriched Topology Learning for Skeleton-based Action Recognition. In Proceedings of the MM '20: The 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; Chen, C.W., Cucchiara, R., Hua, X., Qi, G., Ricci, E., Zhang, Z., Zimmermann, R., Eds.; ACM: New York, NY, USA, 2020; pp. 55–63. [CrossRef]
34. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: New York, NY, USA, 2020; pp. 180–189. [CrossRef]
35. Lin, Z.; Feng, M.; dos Santos, C.N.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A Structured Self-Attentive Sentence Embedding. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
36. Zhou, T.; Li, J.; Wang, S.; Tao, R.; Shen, J. MATNet: Motion-Attentive Transition Network for Zero-Shot Video Object Segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 8326–8338. [CrossRef] [PubMed]
37. Lee, J.B.; Rossi, R.A.; Kong, X. Graph Classification using Structural Attention. In Proceedings of the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, 19–23 August 2018; Guo, Y., Farooq, F., Eds.; ACM: New York, NY, USA, 2018; pp. 1666–1674. [CrossRef]
38. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
39. Zhang, J.; Shi, X.; Xie, J.; Ma, H.; King, I.; Yeung, D. GaAN: Gated Attention Networks for Learning on Large and Spatiotemporal Graphs. In Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, CA, USA, 6–10 August 2018; Globerson, A., Silva, R., Eds.; AUAI Press: Arlington, VA, USA, 2018; pp. 339–349.
40. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action-Gesture Recognition. In Proceedings of the Computer Vision—ACCV 2020—15th Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020; Revised Selected Papers, Part V; Ishikawa, H., Liu, C., Pajdla, T., Shi, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; Lecture Notes in Computer Science; Volume 12626, pp. 38–53. [CrossRef]
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
42. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.; Kot, A.C. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2684–2701. [CrossRef] [PubMed]
43. Li, L.; Wang, M.; Ni, B.; Wang, H.; Yang, J.; Zhang, W. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, lNashville, TN, USA, 19–25 June 2021; Computer Vision Foundation/IEEE: New York, NY, USA, 2021; pp. 4741–4750. [CrossRef]
44. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 13339–13348. [CrossRef]