electronics

MDPI

Article

Non-Uniform Motion Aggregation with Graph Convolutional Networks for Skeleton-Based Human Action Recognition

Chengwu Liang [1,2,*,†], Jie Yang [1,2,†], Ruolin Du [3], Wei Hu [1,2] and Yun Tie [4]

[1] School of Electrical and Control Engineering, Henan University of Urban Construction, Pingdingshan 467036, China; yangjie@ctgu.edu.cn (J.Y.); huwei1999@ctgu.edu.cn (W.H.)
[2] College of Electrical Engineering and New Energy, China Three Gorges University, Yichang 443002, China
[3] School of Transportation and Civil Engineering, Nantong University, Nantong 226019, China; 2233320008@stmail.ntu.edu.cn
[4] School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China; ieytie@zzu.edu.cn
* Correspondence: liangchengwu@huuc.edu.cn
† These authors contributed equally to this work.

Abstract: Skeleton-based human action recognition aims to recognize human actions from given skeleton sequences. The literature utilizes fixed-stride sampling and uniform aggregations, which are independent of the input data and do not focus on representative motion frames. In this paper, to overcome the challenge of the fixed uniform aggregation strategy being unable to focus on discriminative motion information, a novel non-uniform motion aggregation embedded with a graph convolutional network (NMA-GCN) is proposed for skeleton-based human action recognition. Based on the skeleton quality and motion-salient regions, NMA is able to focus on the discriminative motion information of human motion-salient regions. Finally, the aggregated skeleton sequences are embedded with the GCN backbone for skeleton-based human action recognition. Experiments were conducted on three large benchmarks: NTU RGB+D, NTU RGB+D 120, and FineGym. The results show that our method achieves 93.4% (Xsub) and 98.2% (Xview) on NTU RGB+D dataset, 87.0% (Xsub) and 90.0% (Xset) on the NTU RGB+D 120 dataset, and 90.3% on FineGym dataset. Ablation studies and evaluations across various GCN-based backbones further support the effectiveness and generalization of NMA-GCN.

Keywords: human action recognition; skeleton modality; frame sampling; motion salient region; motion aggregation

Citation: Liang, C.; Yang, J.; Du, R. ; Hu, W.; Tie, Y. Non-Uniform Motion Aggregation with Graph Convolutional Networks for Skeleton-Based Human Action Recognition. *Electronics* **2023**, *12*, 4466. https://doi.org/10.3390/electronics12214466

Academic Editors: Donghyeon Cho and George A. Tsihrintzis

Received: 5 September 2023
Revised: 18 October 2023
Accepted: 26 October 2023
Published: 30 October 2023

1. Introduction

Human action recognition is one of the most important tasks in the field of computer vision, and has a wide range of applications in fields such as video surveillance, industrial control, autonomous driving, and human–computer interaction [1–7]. Human actions can be represented through various data modalities, such as RGB, depth, skeleton, infrared, etc. Compared with traditional RGB data, skeleton data contains the positional information of human skeleton joints, which is unaffected by background and lighting conditions. This characteristic is better for constructing topological structures and dynamic spatiotemporal representations. In addition, based on RGB images or depth information, 2D or 3D skeleton data can be easily obtained through pose estimation algorithms or Microsoft Kinetics devices [1,2]. Thus, the skeleton-based action recognition task has attracted much attention.

Existing skeleton-based action recognition tasks are primarily addressed by deep neural network models, which include three dominant categories based on convolutional neural network (CNN) [8–10], graph convolutional network (GCN) [11–13], and recurrent neural network (RNN) [14–16] methods. GCN-based methods process skeleton data by constructing a spatiotemporal skeleton graph based on the natural connection relationship

of human joints, whereas CNN- and RNN-based methods respectively convert the skeleton sequence data into a pseudoimage or vector sequence for processing. As is widely acknowledged with the research community, informative frame sampling strategies are crucial for action recognition tasks [17–19]. However, existing skeleton-based action recognition methods have primarily focused on designing more advanced model architectures to improve recognition performance, overlooking the importance of frame sampling methods in action recognition tasks.

Current human action recognition methods generally adopt a fixed aggregation strategy for testing and training. Figure 1a,b shows two fixed temporal aggregation methods taken from [13] and [20], respectively. For frame sampling in a video, the dominant approaches [21–23] adopt the uniform sampling strategy [20], which splits the video into N segments and then randomly selects one frame from distinctive segments. For frame sampling in a skeleton sequence, the typical approach is to obtain input skeleton sequences by random cropping and interpolation [12,13]. However, human actions performed by people with private habits may have varied motion speeds or distinctive dynamics, even when they are performing the same actions. In addition, the information of each frame is not uniform in the temporal dimension, meaning that the fixed aggregation strategy is likely to ignore important information in the input data. Furthermore, the discriminative movement phases associated with action categories are different and should not be treated equally. Therefore, the purpose of this study is to design a novel aggregation method that can adaptively select representative frames according to the confidence information and motion distance, thereby improving the performance of skeleton-based action recognition. Figure 1c shows the core idea of our method, which is to select key frames based on motion-salient regions.

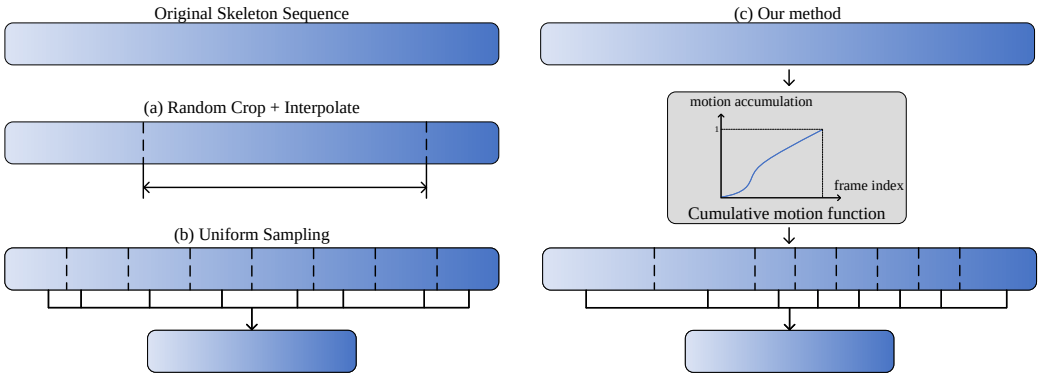

Figure 1. Comparison and visualization of distinctive temporal aggregation methods: (**a**) random cropping and interpolation method [13], where subsequences are cropped from the original sequence and interpolated to the given length (64); (**b**) uniform Sampling method [20], where each sequence is split into N non-overlapping segments of equal length, a frame is randomly sampled from each segment, and the frames are aggregated into a new subsequences; (**c**) our method, which splits the sequence into N non-uniform segments based on motion-salient regions, after which the representations of selected samples are learned and aggregated for effective human action recognition.

As mentioned above, selecting key frames is a critical issue in human action recognition. Recent studies have adopted reinforcement learning to train agents with policy gradient methods for selecting video frames [17,24–27]. AdaFrame [17] is a memory-augmented LSTM with the aim of searching for frames to use over time; the model is trained using the policy gradient method. Ocsampler [27] addresses the frame selection problem by processing a whole video sequence at once, leading to a significant improvement in efficiency while preserving accuracy. To avoid complicated training strategies, other studies have proposed a differentiable and lightweight policy network as a video frame sampler [18,19,28]. Scsampler [18] is a lightweight clip-sampling model capable of identifying the most salient temporal clips in a long video. Arnet [19] uses a lightweight

policy network to select optimal frame resolutions or even skip frames without losing any accuracy. These methods improve the recognition performance of video action recognition by the appropriate frame samplers.

Although the problem of video frame sampling has been studied extensively, limited attention has been paid to the sampling method for skeleton sequences. In order to make the model focus on more representative information in the skeleton sequence, Tang et al. [29] proposed a deep reinforcement learning method for selecting a fixed number of key frames; this approach gradually adjusts the selected frames according to the discriminability of the current frame for action recognition and its relationship in the entire action sequences. Shi et al. [30] designed a lightweight policy network for integration into a skeleton-based action recognition model. The policy network can adaptively select the optimal number of joints in the corresponding frame by calculating the features of a small number of joints or directly eliminating the corresponding frame. Compared with the fixed aggregation strategy, the aforementioned methods have achieved better results; however, they require a more complex model training strategy. These inflexible sampling approaches are limited by their fixed model architectures, and cannot be widely utilized in other skeleton-based action recognition models.

To address the aforementioned challenges of skeleton frame sampling, in this paper we propose a novel non-uniform motion aggregation embedded with graph convolutional network (NMA-GCN) for skeleton-based human action recognition. The proposed NMA-GCN consists of three components: a confidence-based refinement module (CRM), a non-uniform motion sampling module (NMS), and a graph convolutional network (GCN) backbone. Our key contribution is to design the non-uniform motion aggregation (NMA), which consists of CRM and NMS. First, in order to address the skeleton noise problem, the CRM is used to eliminate poor-quality frames according to the confidence information of the skeleton joints. Then, by calculating the motion distance between adjacent frames and constructing a cumulative motion distribution function, the NMS module is used for key frame sampling and aggregation. Finally, the aggregated skeleton sequences are embedded with the GCN backbone to predict the action categories. To verify the effectiveness of our proposed NMA-GCN, we report experimental results on three large action recognition datasets: NTU RGB+D [31], NTU RGB+D 120 [32], and FineGym [33]. Our experimental results show that NMA-GCN achieves improved performance compared to CNN-based, RNN-based, and GCN-based methods. Our main contributions are as follows:

(1) A novel non-uniform motion aggregation embedded with graph convolutional network (NMA-GCN) is proposed for skeleton-based human action recognition, overcoming the challenge of fixed uniform aggregation strategies being unable to focus on the discriminative motion information of human actions.
(2) A non-uniform motion aggregation (NMA) consisting of CRM and NMS is designed to discover the non-uniform importance of skeleton frames along the temporal dimension and aggregate key skeleton features based on motion-salient regions.
(3) The proposed NMA is embedded within a GCN backbone, providing a practical framework for skeleton-based human action recognition. An ablation study and extensive experiments are conducted on the NTU RGB+D, NTU RGB+D 120, and FineGym datasets, demonstrating the effectiveness and generalization ability of the proposed NMA-GCN method.

2. Related Works

2.1. Skeleton-Based Action Recognition

Skeleton data, which are more compact and lightweight, have recently become widely used in human action recognition tasks. The deep learning methods used for skeleton-based action recognition include three main streams: GCN-based methods, CNN-based methods, and RNN-based methods.

GCN-based methods construct a human skeleton graph according to the natural connection relationship of skeleton joints and bones, then use graph convolutional neural

networks to learn the action dynamics and features. The pioneering work in this field is the spatiotemporal graph convolutional neural network (ST-GCN) [11], which adopts spatial graph convolution and temporal convolution to model spatial and temporal features, respectively. Based on ST-GCN, other GCN-based methods such as two-stream adaptive GCN (2S-AGCN) [12], spatiotemporal graph routing GCN (STGR-GCN) [34], multi-scale unified spatiotemporal GCN (MS-G3D) [35], action–structure GCN (AS-GCN) [36], richly activated GCN (RA-GCN) [37], GCN–hidden conditional random field (GCN-HCRF) [38], feedback GCN (FGCN) [39], and channel-wise GCN (CTR-GCN) [13] have been proposed to solve skeleton-based action recognition tasks. A number of works have adopted multi-scale modeling to enhance spatiotemporal modeling capability. MS-G3D [35] is a decoupled multi-scale aggregation scheme that seeks to eliminate the redundant dependencies of node features between different neighborhoods, ensuring that the multi-scale feature aggregator can effectively capture graph-level node information on human skeletons. Other works have introduced attention mechanisms to enhance the spatiotemporal features. 2S-AGCN [12] is an adaptive graph convolutional network that adaptively learns the graph topology of different GCN layers and skeleton sequence samples in order to better adapt to the hierarchical structure of GCN. STGR-GCN [34] learns the importance of different spatiotemporal connection graphs for graph fusion via the frame attention mechanism.

CNN-based methods first convert skeleton sequence data into 2D or 3D images using computer graphic methods or into heatmaps using image rendering generation technology, then utilize these converted representations for feature learning. Potion [40] aggregates heatmap information from the same joint along the temporal dimension. DynaMotion [41] converts stacked pose heatmaps into 2D pseudoimages using a learnable encoder. Considering that the process of generating 2D pseudoimages causes a certain degree of information loss, PoseC3D [10] obtains 3D inputs by stacking heatmaps along the temporal dimension, thereby retaining all the input information during the transformation process while modeling with 3D CNN.

RNN-based methods are suitable for processing time series data due to their unique structure. Therefore, various methods have applied and adapted RNN and Long Short-Term Memory (LSTM) to model skeleton sequences, such as spatiotemporal LSTM network with trust gates (Trust Gate ST-LSTM) [42], spatiotemporal attention LSTM (STA-LSTM) [14], global context-aware attention LSTM (GCA-LSTM) [43], view adaptation LSTM (VA-LSTM) [44], spatial reasoning and temporal stack learning (SR-TSL) [15], and attention-enhanced graph convolutional LSTM (AGC-LSTM) [16].

The aforementioned methods are dedicated to designing more advanced model structures. In contrast, we aim to design a plug-and-play aggregation method in order to effectively select informative skeleton frames and aggregate the motion representation.

2.2. Frame Sampling and Representation Aggregation Methods

For human action recognition tasks, in addition to the network structure design, frame sampling and representation aggregation are crucial, especially when dealing with large-scale action datasets [17,24,29,45–47]. Temporal segment networks (TSN) [45] is a simple and efficient uniform sampling method that is currently widely used in various deep action recognition methods. However, the uniform sampling method considers each frame in the video sequence to be equally important, and does not pay attention to more representative input frames. Several works have adopted reinforcement learning to solve the problem of sampling key frames [17,24,25,29,47].

In the paper [17] and its journal version [47], a conditional computation framework (AdaFrame) was introduced for fast video recognition by selecting relevant frames with the aim of searching for which frames to use over time. Tang et al. [29] proposed a deep progressive reinforcement learning (DPRL) method for selecting key frames in skeleton sequences. DPRL gradually adjusts the selected frames based on two factors, namely, the quality of the selected frames and the relationship between the selected frames and the whole video. In [24], an end-to-end deep reinforcement approach was proposed for

classifying videos based on the fact that the information in video frames is not equally distributed over time. The reinforcement learning-based approach for efficient spatially adaptive video recognition (AdaFocus) [46] seeks to explore the spatial redundancy in video recognition; results have shown that AdaFocus can improve computational efficiency when used with lightweight CNNs. Instead of exploring network architectures, Wu et al. [25] focused on a developing a multi-agent reinforcement learning-based sampler that relies on a frame sampling strategy for effective untrimmed video recognition.

To avoid complex reinforcement learning training strategies, other works have considered a fully differentiable framework for frame sampling [18,19,30,48]. Scsampler [18] uses an extremely lightweight network as the sampler to sample clips based on salience scores. Meng et al. [19] proposed a novel and differentiable approach in which the optimal resolution is selected for each frame, with the goal of improving both accuracy and efficiency. Shi et al. [30] proposed a lightweight policy network to select the optimal number of skeleton joints, and solved non-differentiable problems by using the straight-through Gumbel estimator [49] algorithm. Although the above methods have lead to improvements in the action recognition task, they rely on complex training strategies and are difficult to generalize to other domains. Zhi et al. [48] proposed a simple and explainable motion-guided sampler (MGSampler) that does not depend on training data and is adaptable enough to process various video contents; however, it is only applicable to RGB data.

Considering the drawbacks of the aforementioned methods, in this paper we design a dynamic non-uniform motion aggregation method for skeleton data. By calculating the motion distance between adjacent frames and constructing the motion cumulative distribution function, our proposed non-uniform motion module is able to perform key frame sampling and aggregation.

3. Methods

In this section, we describe the proposed novel non-uniform motion aggregation embedded with graph convolutional network (NMA-GCN) for skeleton-based human action recognition. As shown in Figure 2, NMA-GCN consists of three components: a confidence-based refinement module (CRM), a non-uniform motion sampling module (NMS), and a GCN backbone. The combination of the CRM and NMS modules makes up the non-uniform motion aggregation (NMA), which can aggregate more representative skeleton frames from the original skeleton sequence. First, CRM refines the skeleton sequence by eliminating the poor quality skeleton frames. Then, NMS constructs the cumulative motion distribution function by calculating the motion distance of each frame to sample and aggregate the key skeleton frames. Finally, the aggregated skeleton sequences are fed into the GCN backbone to predict the class scores. In the following subsections, we explain these three components in detail.

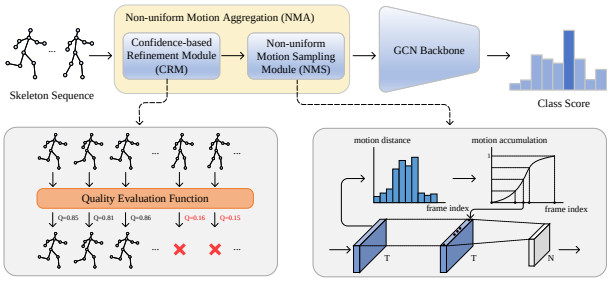

Figure 2. Overview of the proposed NMA-GCN. The key contribution is the design of a non-uniform motion aggregation (NMA), which consists of a confidence-based refinement module (CRM) and a non-uniform motion sampling module (NMS). The CRM refines the original skeleton sequence based on the quality of the skeleton frames, while the NMS module constructs cumulative motion distribution functions to select key frames.

3.1. Confidence-Based Refinement Module (CRM)

The skeleton-based human action recognition task takes human skeleton sequences as input data. Due to issues such as occlusion and estimation algorithm errors, skeleton data quality varies greatly between frames, whether 3D skeleton pose data captured by a Kinetics camera or 2D skeleton pose data from a pose estimation algorithm. Poor quality skeleton data not only affects the recognition performance of the model, it creates difficulty for skeleton frame sampling methods that are based on motion-salient regions. To address this problem, we propose the use of a confidence-based refinement module (CRM). The purpose of the CRM is to discover and remove a whole skeleton frame with poor quality in order to refine the original skeleton sequences for non-uniform motion representation aggregation. Specifically, the CRM calculates the quality score of each skeleton frame based on the confidence value of the skeleton joints, then eliminates the skeleton frames with lower quality scores.

For a given input skeleton sequence $\mathbf{X}_0 \in \mathbb{R}^{T_0 \times V \times C}$, where T_0, V, and C respectively denote the number of frames of the raw skeleton sequence, the number of joints, and the dimension of the joints, for the i-th joint in t-th frame, the 2D skeleton data obtained from the pose estimator can be represented as coordinate triplets (x_t^i, y_t^i, c_t^i), where (x_t^i, y_t^i) is the coordinate information of the joint and c_t^i is the confidence information of the joint. In this paper, 2D skeleton data are used as input and the confidence information of the 2D skeleton data is used to calculate the quality score of each frame to evaluate the quality of the skeleton frame. Then, the input skeleton data are refined by removing low quality frames. The quality score of each skeleton frame Q_t is obtained from the quality evaluation function f_q, which can be formulated as follows:

$$Q_t = f_q(\mathbf{c_t}) = 1/V \sum_{i=1}^{V} c_t^i \tag{1}$$

where $Q_t \in [0, 1]$ represents the quality score of the t-th frame and $\mathbf{c_t} = \{c_t^1, c_t^2, \ldots, c_t^i, \ldots, c_t^V\} \in \mathbb{R}^{V \times 1}$ represents the confidence vector of the t-th frame. Here, we use the mean function as the quality evaluation function. Specifically, the mean of the confidence value of each joint in a single skeleton frame is taken as the quality score of the corresponding frame.

The visualization of skeleton frames and their quality scores is shown in Figure 3. The skeleton frames with lower quality scores are highlighted in red and bolded. For the action category "Falling", the skeleton frames that have high quality scores are informative and representative for the intrinsic dynamic of action categories, which is useful for feature learning. It can be seen that skeleton frames with low quality scores, such as $Q_t < 0.2$, have large estimation errors. In addition, poor quality frames make it difficult to recognize the specific human shape, which may cause interference during model training.

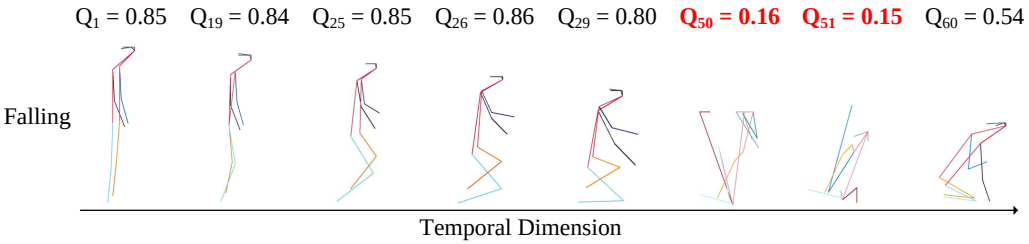

Figure 3. Visualization of skeleton frames and their quality scores. Lower quality scores ($Q_t < 0.2$) are indicated in red, and it can be observed that corresponding skeleton frames are difficult to recognize the specific human poses.

Based on the obtained quality scores, skeleton frames with quality scores greater than the threshold θ are retained:

$$M(Q_t) = \begin{cases} 1, & \text{if} \quad Q_t > \theta \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $M(Q_t)$ is a binary mask. Binary masks are commonly used to select the informative parts from the input features, as in [50,51]. Unlike these previous studies, however, here we adopt a threshold hyperparameter to generate a binary mask to ensure the selection of meaningful skeleton frames. Specifically, the t-th frames are pruned when $M(Q_t) = 0$ and retained when $M(Q_t) = 1$.

After processing by the CRM, the refined skeleton sequence $\mathbf{X} \in \mathbb{R}^{T \times V \times C}$ is obtained, where $T < T_0$ denotes the number of frames in the refined skeleton sequence. Then, the generated skeleton data are sent as new input to the subsequent sampling module for key frame sampling and aggregation.

3.2. Non-Uniform Motion Sampling (NMS) Module

Existing skeleton-based action recognition methods typically use random cropping and interpolation to obtain a fixed-length skeleton sequence as input to the model. However, because the importance of each frame of the skeleton sequence is not uniform along the temporal dimension, using input-independent sampling methods may lead to important motion information being ignored. Therefore, we propose the use of a non-uniform motion sampling (NMS) module to select informative frames by focusing on the more salient part of the motion information in the skeleton sequences. In this module, the cumulative motion distribution function of the skeleton sequences is designed for dynamic frame sampling and aggregation.

The NMS module aims to generate a final skeleton sequence of length N from the refined skeleton sequence $\mathbf{X} \in \mathbb{R}^{T \times V \times C}$ obtained by the CRM. To enable this module to pay more attention to the more salient part of the skeleton sequence, it is first necessary to calculate the motion distance S_t of each frame. The motion distance S_t is used to measure the magnitude of motion at the corresponding moment; more salient motion corresponds to larger S_t and vice versa.

In this paper, considering that the Euclidean distance between human joints has been widely used to analyze human poses and as the motion feature [52–54], we calculate the motion distance S_t based on the Euclidean distance. By first calculating the Euclidean distance between the same joint in adjacent frames, the motion distance of a single joint can be obtained. This process can be formulated as follows:

$$S_t^i = \sqrt{\left(x_t^i - x_{t-1}^i\right)^2 + \left(y_t^i - y_{t-1}^i\right)^2} \tag{3}$$

where S_t^i represents the motion distance of the i-th joint in the t-th frame and x_t^i and y_t^i represent the horizontal and vertical coordinate values of the i-th joint in the t-th frame, respectively. After obtaining the motion distance of a single joint, the motion distances of all joints in a frame are summed to generate the overall motion distance at the corresponding frame. In addition, considering that the estimation error of the skeleton data may have a negative effect on the motion distance calculation, we introduce the average confidence value of the adjacent frame joints based on Equation (3), formulated as follows:

$$S_t = \sum_{i=1}^{V} \left(S_t^i \times \frac{c_t^i + c_{t-1}^i}{2} \right) \tag{4}$$

where S_t represents the overall motion distance of the t-th frame.

Figure 4 shows the motion distance distribution of three actions. It can be seen that the distribution of the different actions has a large difference, resulting in discriminative information for action recognition. This shows that the calculation of motion distance based on the Euclidean distance is feasible. For the first action ("reading"), the whole movement is not obvious. Therefore, there is little difference in the motion distance at different frames. For the other two actions ("punching/slapping other person" and "hugging other person"), the motion distance varies greatly at different stages of the whole action. The regions with greater variability of motion distance are critical stages of the action and contain more discriminative information, which is beneficial for action recognition. As a result, more attention should be paid to those regions in the skeleton sequence where the motion is more salient.

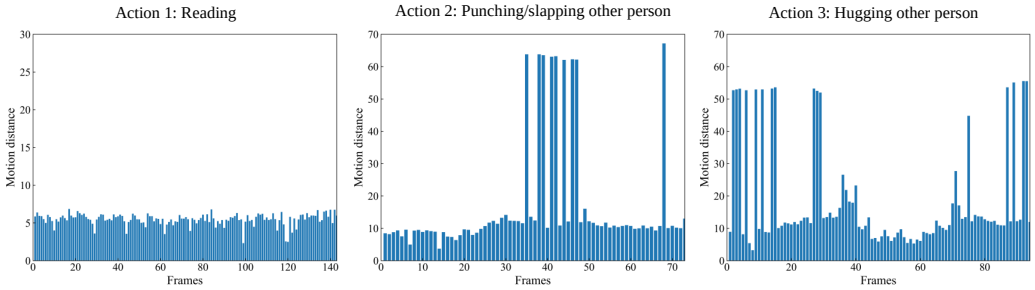

Figure 4. The motion distance distribution of three actions: "reading" (**left**), "punching/slapping other person" (**middle**), and "hugging other person" (**right**).

The cumulative motion distribution function can be constructed by normalizing the motion distance calculated from Equation (4) and then accumulating it along the temporal dimension, which can be formulated as follows:

$$\sum_{t}^{T} \tilde{S}_t = 1 \tag{5}$$

where $\tilde{S}_t$ represents the normalized motion distance. In order to further control the smoothness of the cumulative motion distribution function, we introduce the smoothness hyperparameter μ to adjust the original cumulative motion distribution, as in [48]. The smaller the value of μ, the smoother the cumulative motion distribution curve. The calculation of the normalized motion distance $\tilde{S}_t$ can be described as follows:

$$\tilde{S}_t = \frac{(S_t)^{\mu}}{\sum_{i=1}^{T}(S_i)^{\mu}} \tag{6}$$

The cumulative motion distribution function of two different actions are shown in Figure 5. Compared to Figure 4, it can be seen that a larger slope of the curve indicates more salient motion in the corresponding frame. It can be observed that the cumulative motion distribution curve of the two actions are quite different, which provides good discriminative information for action recognition. In addition, it can be seen that a lower value of μ results in a more uniform motion distribution. This allows the smoothness of the cumulative motion distribution curve to be controlled.

In addition, the final cumulative motion distribution function is obtained based on both joint and bone modalities. The method of obtaining the cumulative motion distribution function is approximately the same for both modalities; the only difference lies in the process of calculating the motion distance in Equation (4). For the joint modality, the motion distance is used to directly calculate the Euclidean distance of the joints in adjacent frames based on Equation (4). For the bone modality, the bone vector is first calculated based on the natural connectivity of the human body, then the Euclidean distance of the bone vector is calculated for the adjacent frames. Taking the motion of the human upper arm as an example, the upper arm is located between the elbow and the wrist; the

wrist joint point and the elbow joint point are defined as w_t and e_t, respectively, and the corresponding two-dimensional coordinates are (x_t^w, y_t^w) and (x_t^e, y_t^e). The arm vector $\boldsymbol{u_t}$ is first calculated according to the human connection relationship, then the motion distance of the bone modality is calculated based on the arm vector of adjacent frames. This can be formulated as follows:

$$\boldsymbol{u_t} = w_t - e_t = (x_t^w - x_t^e, y_t^w - y_t^e) \tag{7}$$

$$S_t = |\boldsymbol{u_t} - \boldsymbol{u}_{t-1}| \tag{8}$$

where $\boldsymbol{u_t}$ and $\boldsymbol{u_{t-1}}$ represent the arm vectors at frame t and frame $t-1$, respectively, while $||$ represents the modulo operation. It can be seen that the motion distance calculated based on the joint is the absolute motion distance, while the motion distance calculated based on the bone vector is the relative motion distance. Combining the cumulative motion distribution function of the two modalities can better reflect the corresponding motion characteristics compared to using either modality alone.

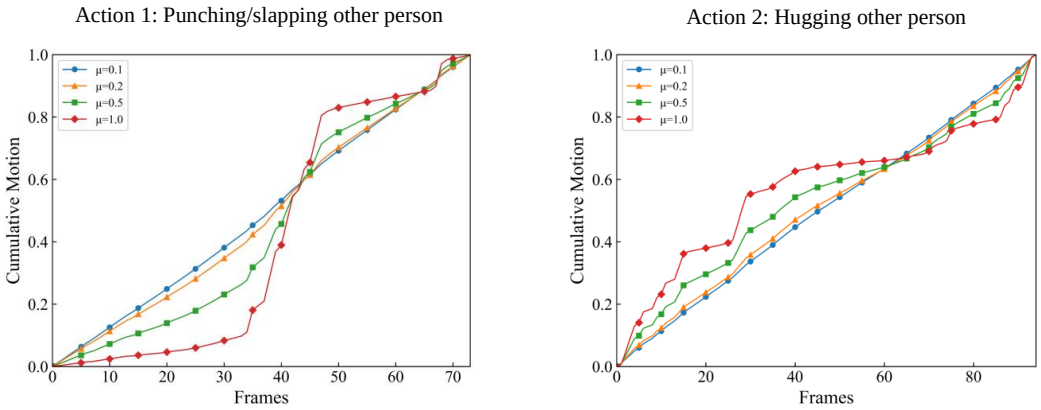

Figure 5. The cumulative motion distribution with different values of μ of two actions: "punching/slapping other person" (**left**) and "hugging other person" (**right**).

Finally, according to the cumulative motion distribution function, two strategies are proposed, namely, slope sampling and cumulative sampling. As shown in Figure 6, the slope sampling strategy directly samples the most salient region of the motion. More specifically, the slope sampling strategy generates N temporal segments based on the slope of the cumulative motion distribution function, then randomly samples frames from each segment. The cumulative sampling strategy generates N temporal segments by evenly dividing the cumulative motion distribution function and then randomly sampling one frame in each segment. The cumulative sampling strategy generates N temporal segments by evenly dividing the cumulative motion distribution function and then randomly sampling one frame in each segment. Afterwards, the selected frames are aggregated to produce the final skeleton subsequence.

The skeleton sequences obtained according to the NMS module are fed into the GCN backbone to predict the action categories.

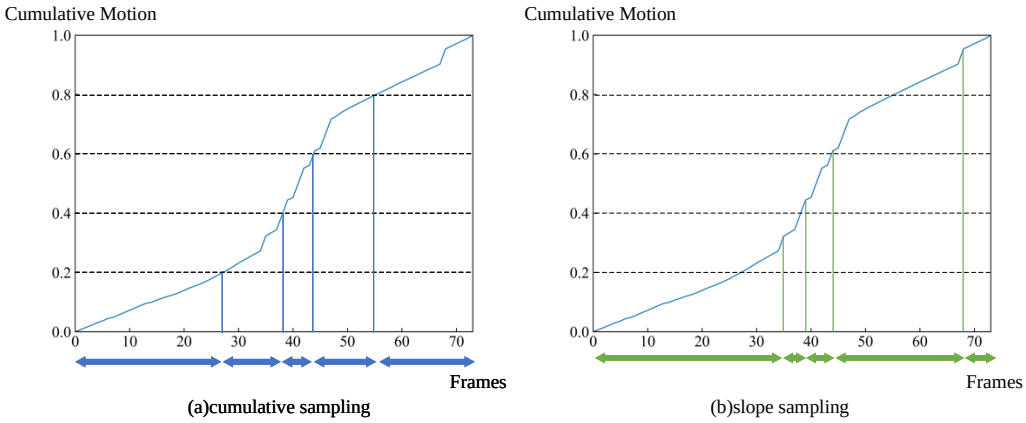

Figure 6. Two different sampling strategies based on the cumulative motion distribution function: (**a**) the cumulative sampling strategy divides the skeleton sequence into N segments based on the same cumulative motion of each segment, then randomly samples frames from each segment; (**b**) the slope sampling strategy divides the skeleton sequence into N segments based on the slope of the cumulative motion distribution function, then randomly samples frames from each segment. Blue and green arrows indicate the temporal segments divided by the two sampling strategies, respectively.

3.3. GCN Backbone

GCN-based models have been widely used in skeleton-based human action recognition, where they have seen significant progress. It is important to note that the architectural design of the backbone is not the main contribution of our method. The GCN backbone used in our method can be replaced by any other GCN-based network, such as [13,55]. In this section, we use the representative GCN backbone (ST-GCN [11]) to process the skeleton sequences aggregated by the proposed NMA.

ST-GCN utilizes stacked N blocks to process the skeleton data, while each ST-GCN block consists of a spatial graph convolution (GCN) module and a temporal convolution (TCN) module. The GCN module adopts a learnable topological graph defined on the spatial dimension for spatial feature fusion. The TCN module extracts the temporal features using 1D convolution (kernel size 9) on the temporal dimension. For a given input $\mathbf{X}_{\text{in}} \in \mathbb{R}^{V \times D}$, where D denotes the feature dimension, the GCN modulecan be formulated as follows:

$$\text{GCN}(\mathbf{X}_{\text{in}}) = \sum_{k=1}^{K} (\mathbf{A}_k \mathbf{X}_{\text{in}}) \mathbf{W}_k \tag{9}$$

where $\mathbf{A}_k \in \mathbb{R}^{V \times V}$ denotes the coefficient matrix derived from a predefined joint topology, K denotes the number of matrices, and $\mathbf{W}_k$ is a learnable weight matrix. In addition, the residual connection is applied to each ST-GCN block. The computations of the ST-GCN block can be summarized as follows:

$$\mathbf{X}_{\text{out}} = \text{TCN}(\text{GCN}(\mathbf{X}_{\text{in}})) + \mathbf{X}_{\text{in}} \tag{10}$$

where $\mathbf{X}_{\text{out}}$ denotes the output feature of the ST-GCN block. Afterwards, a global pooling layer is applied to the resulting tensor to obtain the high-level feature vectors. Finally, the SoftMax classifier maps the feature vectors to the probability scores of K candidate action categories.

4. Experiments

In this section, we describe the extensive experiments conducted to evaluate the effectiveness of the proposed NMA-GCN. First, the datasets and implementation details are introduced. Then, we describe the exhaustive ablation studies performed on the NTU RGB+D dataset [31]. We further demonstrate the generalization of the proposed NMA with

distinctive backbones. Finally, we compare our NMA-GCN with state-of-the-art methods on three widely used datasets.

4.1. Datasets

We conducted experiments on three public action recognition datasets: NTU RGB+D [31], NTU RGB+D 120 [32], and FineGym [33]. Following convention, we report the Top-1 accuracy for NTU RGB+D and NTU RGB+D 120 and the mean Top-1 accuracy for FineGym.

The NTU RGB+D dataset was captured by Microsoft's second-generation Kinect camera. The content includes three types of actions: daily, health-related, and two-person interactions. Each sample contains four modalities: RGB, depth map, 3D skeleton, and infrared sequence. The NTU RGB+D dataset was collected in 2016; it contains 60 action categories and 57,000 video samples. The NTU RGB+D dataset has two recommended settings. (1) The cross-subject (Xsub) benchmark includes 40,320 samples for training and 16,560 for evaluation; in this setting, as introduced in reference [13,31], the dataset is split into training and testing sets according to 40 subjects, with half of the subjects used for training and the rest for testing. (2) The cross-view (Xview) benchmark contains 37,920 samples for training and 18,960 videos for evaluation; as introduced in reference [55,56], the training samples in this set come from camera views 2 and 3, while the evaluation samples are all from camera view 1. We report the top-1 accuracy on both benchmarks.

The NTU RGB+D 120 dataset is an extended version of NTU RGB+D dataset which was proposed in 2019. The NTU RGB+D 120 dataset contains 120 action classes and 114,000 video samples. The NTU RGB+D 120 dataset has two recommended settings. (1) In the cross-subject (Xsub) benchmark setting, as introduced in reference [32,35], 106 subjects are split into training and testing groups, with each group consisting of 53 subjects. (2) In the cross-setup (Xset) benchmark setting, as introduced in [6,56], the dataset is split into training and testing sets based on the camera setup ID, with the even-numbered IDs used for training and the odd-numbered IDs for testing. We report the Top-1 accuracy for both benchmarks.

The FineGym dataset is a fine-grained action recognition dataset containing 29,000 video samples of 99 gymnastics action classes, 23,000 videos for training, and 6000 videos for testing. The gymnastics video samples collected in FineGym dataset have strong similarity in terms of background information, which can help to prevent the model solving classification tasks by learning information that is unrelated to actions.

4.2. Implementation Details

We implemented the proposed NMA-GCN using the PYSKL [57] toolbox. All experiments were performed on a server with an Intel Xeon Platinum 8160Ts CPU and NVIDIA GeForce RTX 3090Ti GPUs. In our experiments, multiple GCN-based backbones were adopted to verify the effectiveness of our method. For all datasets, $N = 100$ frames were sampled from each skeleton sequence and then used as input. The networks were trained for 80 epochs using a stochastic gradient descent (SGD) optimizer with a Nesterov momentum of 0.9 and weight decay of 0.0005. The batch size and initial learning rate were set to 32 and 0.1, respectively.

In terms of data preprocessing, we obtained the skeleton data following [10]. HRNet [58], a 2D pose estimator pretrained on the COCO [59] dataset, was adopted to extract 2D skeleton data. Using HRNet, 2D pose information (x, y, c) was generated from a skeleton frame, where (x, y) represents the two-dimensional coordinate information of the joints and c is the confidence information of the corresponding joints.

4.3. Ablation Study

To verify the effectiveness of each module of the proposed NMA-GCN and evaluate the hyperparameters, ablation experiments were conducted using the ST-GCN [11] backbone on NTU RGB+D [31] dataset with the Xsub setting.

4.3.1. Effectiveness of the Threshold Hyperparameter θ

The proposed confidence-based refinement module (CRM) addresses the skeleton noise problem by eliminating poor quality frames; the threshold hyperparameter θ described in Equation (2) is used to control the quality of the retained skeleton frames. The skeleton frames are removed when their quality score is smaller than θ, with $\theta = 0$ meaning that no frames are deleted.

We performed ablation experiments on the CRM module with different values of θ; the results are shown in Figure 7. It can be seen that $\theta = 0.2$ achieves the best result, and is able to improve performance from 88.7% to 89.9%, an improvement of 1.2%, which confirms the effectiveness of CRM for skeleton-based action recognition. Furthermore, the CRM module yields much worse performance when θ is greater than 0.4. The main reason for this is that a larger value of θ causes a large amount of frames to be removed, resulting in the destruction of the complete motion process. Unless otherwise specified, we use $\theta = 0.2$ in all of the following experiments.

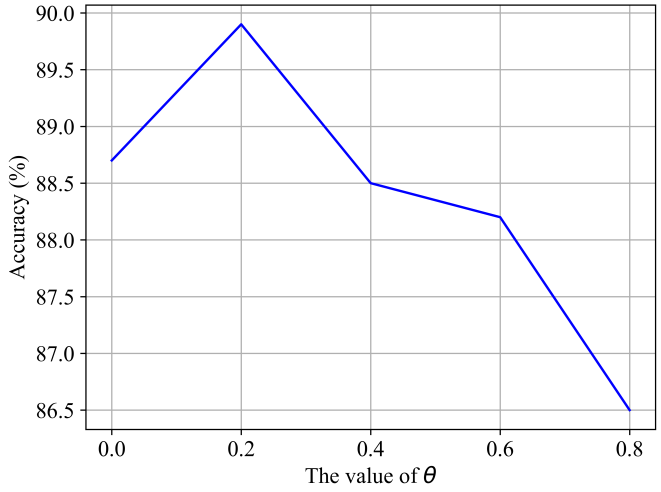

Figure 7. Performance comparison of different values of θ on the NTU RGB+D (Xsub) benchmark.

4.3.2. Effectiveness of the Smoothness Hyperparameter μ

The non-uniform motion sampling (NMS) module is the core component of the proposed NMA. It selects the frames which contain more discriminative motion information. As shown in Figure 5, the smoothness hyperparameter μ described in Equation (6) is used to control the smoothness degree of the cumulative motion distribution function.

To investigate the influence of the hyperparameter μ, we performed ablation experiments on different μ values; the results are reported in Table 1. When $\mu = 1.0$, the cumulative motion distribution function remains the same as in the original distribution. It can be seen that the original NMS module ($\mu = 1.0$) improves performance from 88.7% to 89.7%. In addition, $\mu = 0.5$ achieves the best result by a margin of 2.8%, indicating the effectiveness of NMS. Unless otherwise mentioned, we set $\mu = 0.5$ in all of the following experiments.

Table 1. Performance comparison of different values of μ on the NTU RGB+D (Xsub) benchmark.

Method	μ	Top1 (%)
ST-GCN [11]	–	88.7
ST-GCN+NMS (Ours)	0.2	89.1
	0.5	91.5
	1.0	89.7

4.3.3. Ablation Experiment on the CRM and NMS Modules

In this section, we describe the ablation experiments conducted to separately investigate the effectiveness of each component of the proposed NMA. As shown in Table 2, the recognition accuracy of the ST-GCN model without CRM or NMS is 88.7%. We separately added CRM and NMS to the baseline, obtaining recognition accuracies of 89.9% (1.2% improvement) and 89.3% (0.6% improvement), respectively. The low improvement with the NMS module may be due to the fact that poor quality frames generate incorrect cumulative distribution functions for key frame sampling. Finally, the combination of CRM and NMS boosts recognition accuracy to 91.5%, demonstrating the effectiveness and complementarity of the two modules.

Table 2. The respective impacts of the two modules on the NTU RGB+D (Xsub) benchmark.

Method	CRM	NMS	Top1 (%)
ST-GCN [11]			88.7
	✓		89.9 (+1.2)
		✓	89.3 (+0.6)
	✓	✓	91.5 (+2.8)

4.3.4. Evaluation of NMA and Comparison

The proposed non-uniform motion aggregation (NMA) is the main contribution of our work, consisting of the CRM and NMS modules together. Here, we use "NMA-GCN" to denote NMA embedded within the ST-GCN backbone. To verify the effectiveness of the proposed NMA, we conducted a comparison and evaluation of different sampling strategies. We compared our NMA with two sampling strategies: (1) fixed stride sampling: a subsequence of N frames with a fixed stride is randomly selected from the original skeleton sequence; (2) uniform sampling: a subsequence of N frames is sampled uniformly along the temporal dimension.

As shown in Table 3, the results show that NMA with the cumulative sampling strategy outperforms fixed stride sampling and uniform sampling by 3.3% and 2.8%, respectively, indicating the effectiveness of selecting frames with more discriminative motion information. It can be observed that the proposed slope sampling strategy results in worse performance compared to the cumulative sampling strategy. This experimental result demonstrates that the proposed cumulative sampling strategy is more capable of handling the motion variations and capturing the complete motion process. Thus, the proposed NMA with the cumulative sampling strategy was adopted in all subsequent experiments.

Table 3. Performance comparison of different sampling strategies on the NTU RGB+D (Xsub) benchmark.

Method	Sampling Strategy	Top 1 (%)
ST-GCN [11]	fixed stride sampling	88.2
	uniform sampling	88.7
NMA-GCN (Ours)	slope sampling (Ours)	88.9
	cumulative sampling (Ours)	91.5

4.4. Recognition Accuracy of Different Action Classes

To further study the performance improvement of different action categories contributed by the proposed NMA-GCN as compared to ST-GCN [11], Table 4 shows the recognition accuracy of the top ten action categories with the most significant performance improvement compared to the ST-GCN baseline on the NTU RGB+D (Xsub) benchmark. As shown in Table 4, as compared with the ST-GCN baseline, our NMA-GCN achieves better recognition performance in all ten action categories. The accuracy improvements of actions such as "clapping", "saluting", and "pointing to something with finger" all exceed

10%. However, the recognition accuracy of actions such as "writing" and "reading" is not as high as expected. The reason for this may be that the amplitude of these actions is relatively small, preventing the proposed NMA based on the salient motion region from performing optimally.

Table 4. Performance improvements on the top ten action categories with the most significant performance improvement when using the proposed NMA-GCN on the NTU RGB+D dataset.

Top Ten Improved Action Classes	ST-GCN (Baseline)	NMA-GCN (Ours)
1. clapping	65.9%	87.9% (+22.0%)
2. salute	84.1%	97.5% (+13.4%)
3. pointing to something with finger	68.8%	79.7% (+10.9%)
4. taking a selfie	77.5%	87.3% (+9.8%)
5. rub two hands together	80.4%	89.5% (+9.1%)
6. put the palms together	87.0%	95.7% (+8.7%)
7. use a fan/feeling warm	81.8%	90.2% (+8.4%)
8. reading	48.0%	54.9% (+6.9%)
9. check time	85.5%	92.0% (+6.5%)
10. writing	51.8%	57.4% (+5.6%)

Furthermore, we present the normalized confusion matrix for the proposed method on the NTU RGB+D (Xsub) benchmark in Figure 8. It can be seen that most of the actions are accurately recognized, with confusion occurring mainly among two ambiguous actions, namely, "reading" and "writing". Due to the similarity between ambiguous actions and spatiotemporal representations, recognition is challenging and results in confused misclassification. The reason for this is that these inconspicuous actions lack body movement and motion information, meaning that the advantages of our method are not applicable.

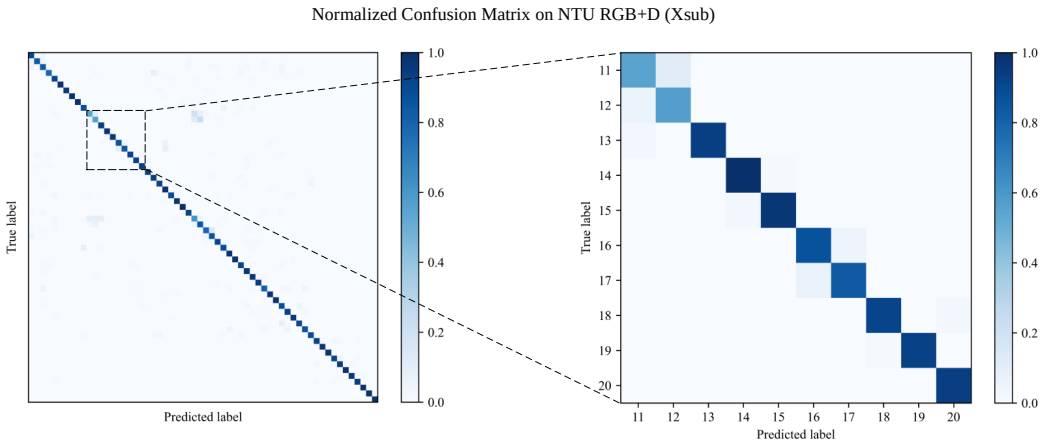

Figure 8. Normalized confusion matrix for the proposed method on the NTU RGB+D (Xsub) benchmark.

4.5. Ablation Study and Performance Comparison of Different Backbones

To further examine and demonstrate the generalization ability of our proposed NMA-GCN, we applied NMA to different GCN-based backbones and performed comparative experiments on the NTU RGB+D [31], NTU RGB+D 120 [32], and FineGym [33] datasets; for the FineGym dataset, we report the mean Top-1 accuracy, while for the others we report the Top-1 accuracy.

In this experiment, we chose ST-GCN [11], MS-AAGCN [55], and CTR-GCN [13] as the backbones for comparison. As shown in Table 5, when employing our NMA, MS-AAGCN achieves a 2.5% performance improvement on the FineGym dataset. On the Xsub setting of the NTU RGB+D 120 dataset, the proposed NMA yields a 2.6% performance improvement for ST-GCN. On the Xsub setting of the NTU RGB+D dataset, the proposed NMA yields a 2.8% performance improvement for ST-GCN. In addition, NMA is able to achieve higher

performance improvements (at least 1.5%) on the FineGym and NTU RGB+D 120 datasets. The main reason for this is that the action categories in these two datasets are more rich and diverse. They have a large motion magnitude, making the strengths of our proposed NMA method more advantageous.

Our ablation studies show that the proposed NMA can consistently improve action recognition performance across different GCN-based backbones, demonstrating the generalization ability of the proposed method.

Table 5. Performance improvements for different GCN-based backbones on the NTU RGB+D (Xsub), NTU RGB+D 120 (Xsub), and FineGym datasets. In keeping with common practice, the Top-1 accuracy (%) is reported for the NTU RGB+D and NTU RGB+D 120 datasets, while the mean Top-1 accuracy (%) is reported for the FineGym dataset.

Methods	Backbone	FineGym (%)	NTU RGB+D 120 (%)	NTU RGB+D (%)
Baseline	ST-GCN [11]	85.1	81.8	88.7
NMA-GCN (Ours)	ST-GCN [11]	86.7	84.4	91.5
Baseline	MS-AAGCN [55]	86.7	81.9	89.4
NMA-GCN (Ours)	MS-AAGCN [55]	89.2	83.4	90.0
Baseline	CTR-GCN [13]	88.5	82.2	90.4
NMA-GCN (Ours)	CTR-GCN [13]	90.3	84.4	90.7

4.6. Comparison with State-of-the-Art Methods

The multi-stream modality fusion strategy has commonly been employed in previous state-of-the-art methods [11–13,20,55,60]. In order to conduct a fair comparison, we followed the same multi-stream fusion strategy as in [13,55]. Our model was trained on four streams (joint, bone, joint motion, and bone motion) on the NTU RGB+D and NTU RGB+D 120 datasets. The joint stream used the original skeleton coordinates as input. The bone stream used the differential of spatial coordinates as input. The joint motion stream and bone motion stream used the differential of the temporal dimension of the corresponding data as input. The score-level fusion strategy was adopted to obtain the fused score, which was then used for prediction.

To verify the effectiveness of the proposed NMA-GCN, we conducted a comparison with state-of-the-art alternatives on three large datasets. We choose GCN [13] as the neural network backbone. Table 6 presents the experimental results on the NTU RGB+D and NTU RGB+D 120 datasets. Because the literature contains many different neural architectures, we compared CNN-based, RNN-based, and GCN-based methods on the NTU RGB+D and NTU RGB+D 120 datasets. We followed the acknowledged experimental settings to demonstrate the effectiveness of the proposed NMA-GCN. CNN-based methods transform the skeleton sequence into pseudoimages. For instance, Caetano et al. [61] proposed tree structure reference joints image (TSRJI), a skeleton image representation that combines the use of reference joints and a tree structure skeleton. Ke et al. [62] proposed a method they termed "Clips+CNN+MTLN". This method transforms a skeleton sequence into clips, then uses a CNN to extract the frame-level feature and a multi-task learning network to process all frames jointly. RNN-based methods have the capacity to capture the dynamic dependencies in sequential data [2]. Because the human skeleton is a natural graph structure, GCN-based methods are widely used to process skeleton data; thus, we choose representative GCN-based methods for comparison. In addition, more GCN-based methods were chosen for comparison, as our proposed method falls under this category. The great ability of GCN backbones and the effectiveness of our method are demonstrated by this comparison with different neural network architectures.

Table 6. Performance comparison, showing the Top-1 accuracy (%) of our proposed method and existing state-of-the-art methods on the NTU RGB+D and NTU RGB+D 120 datasets.

Type	Methods	NTU RGB+D		NTU RGB+D 120	
		Xsub (%)	Xview (%)	Xsub (%)	Xset (%)
CNN	TSRJI [61]	73.3	80.0	65.5	59.7
	SkeleMotion [63]	76.5	84.7	67.7	66.9
	Clips + CNN + MTLN [62]	79.6	84.8	58.4	57.9
	RotClips + MTCNN [64]	81.1	87.4	62.2	61.8
	Banerjee et al. [65]	84.2	89.7	74.8	76.9
	3SCNN [66]	88.6	93.7	-	-
RNN	Trust Gate ST-LSTM [42]	69.2	77.7	58.2	60.9
	STA-LSTM [14]	73.4	81.4	-	-
	GCA-LSTM [43]	74.4	82.8	58.3	59.2
	VA-LSTM [44]	79.4	87.6	-	-
	SR-TSL [15]	84.8	92.4	-	-
	AGC-LSTM [16]	89.2	95.0	-	-
GCN	ST-GCN [11]	81.5	88.3	70.7	73.2
	AS-GCN [36]	86.8	94.2	78.3	79.8
	RA-GCN [37]	87.3	93.6	81.1	82.7
	2s-AGCN [12]	88.5	95.1	79.2	81.5
	GCN-HCRF [38]	90.0	95.5	-	-
	MS-AAGCN [55]	90.0	86.2	-	-
	FGCN [39]	90.2	96.3	85.4	87.4
	AdaSCN [30]	90.5	95.3	85.9	86.8
	Shift-GCN [56]	90.7	96.5	85.9	87.6
	NMA-GCN (Ours)	93.4	98.2	87.0	90.0

As shown in Table 6, our proposed NMA-GCN achieves competitive results and is able to generalize well across datasets. For instance, on both settings of the NTU RGB+D dataset, the best recognition accuracy of our NMA-GCN is 93.4% and 98.2%, which is significantly higher than CNN-based methods [64,66], RNN-based methods [16,43], and other GCN-based methods [12,30,56]. On the NTU-RGB+D 120 dataset, our proposed NMA-GCN outperforms the approach from [56] by 1.1% and 2.4% for the Xsub and Xset settings, respectively. Furthermore, Table 7 illustrates the experimental results on the FineGym dataset. Because the official FineGym dataset is only available in RGB video, the methods using FineGym for experimental validation are mainly RGB-based methods. Therefore, we chose skeleton-based, RGB-based, and multi-modality methods for comparison. It can be seen that our method achieves competitive performance on the FineGym dataset.

Based on the results of this comparison with existing state-of-the-art methods, our proposed NMA-GCN can effectively improve skeleton-based human action recognition performance and has good generalization ability.

Table 7. Performance comparison, showing the mean Top-1 accuracy (%) of our proposed method and existing state-of-the-art methods on the FineGym dataset.

Methods	Modality	Mean Top-1 Accuracy (%)
ActionVLAD [67]	RGB	50.1
I3D [68]	RGB	63.2
TSN [20]	RGB, Flow	76.4
TRN [69]	RGB, Flow	79.8
TRNms [69]	RGB, Flow	80.2
TSM [21]	RGB, Flow	81.2
ST-GCN [11]	Skeleton	85.1
RSANet [70]	RGB	86.4
RGBSformer [71]	RGB, Skeleton	86.7
MS-AAGCN [55]	Skeleton	86.7
CTR-GCN [13]	Skeleton	88.5
NMA-GCN (Ours)	Skeleton	90.3

5. Conclusions

In this work, we propose a non-uniform motion aggregation embedded within a graph convolutional network (NMA-GCN) for skeleton-based human action recognition. Compared with existing state-of-the-art methods, our proposed NMA-GCN improves accuracy by 2.7% (Xsub) and 1.8% (Xview) on the NTU RGB+D dataset, 1.1% (Xsub) and 2.4% (Xset) on the NTU RGB+D 120 dataset, and 1.8% on the FineGym dataset. Based on the results of this study, the proposed non-uniform motion aggregation can effectively learn discriminative human motion representations from the salient information of skeleton sequences, achieving improved recognition performance compared to fixed uniform aggregation methods from the literature. Our ablation study demonstrates that the proposed NMA-GCN is able to generalize well across various GCN-based backbones.

Furthermore, we believe that our NMA-GCN can be further improved through a variety of means. Potential future work could involve extending the proposed method from selecting informative skeleton frames to selecting informative skeleton joints. Another potential future improvement could be to explore the information of the RGB modality in order to compensate for the skeleton's lack of fine-grained human–object interaction recognition. In addition, we hope to improve our proposed NMA-GCN to reduce its computational consumption for use in practical scenarios.

Author Contributions: Conceptualization, C.L. and J.Y.; Data curation, J.Y.; Formal analysis, J.Y.; Funding acquisition, C.L.; Investigation, J.Y., R.D., and W.H.; Methodology, J.Y.; Project administration, C.L.; Resources, C.L.; Software, J.Y. and W.H.; Supervision, C.L. and Y.T.; Validation, C.L. and J.Y.; Visualization, J.Y., R.D., and W.H.; Writing—original draft preparation, J.Y.; Writing—review and editing, C.L., R.D., W.H., and Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 62176086.

Data Availability Statement: We used the processed skeleton data for all datasets provided by the PYSKL toolbox [57]. The NTU RGB+D dataset can be found at https://download.openmmlab.com/mmaction/pyskl/data/nturgbd/ntu60_hrnet.pkl, accessed on 16 March 2023. The NTU RGB+D 120 dataset can be found at https://download.openmmlab.com/mmaction/pyskl/data/nturgbd/ntu120_hrnet.pkl, accessed on 16 March 2023. The FineGym dataset can be found at https://download.openmmlab.com/mmaction/pyskl/data/gym/gym_hrnet.pkl, accessed on 15 March 2023.

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

References

1. Zhang, J.; Li, W.; Ogunbona, P.O.; Wang, P.; Tang, C. RGB-D-based action recognition datasets: A survey. *Pattern Recognit.* **2016**, *60*, 86–105. [CrossRef]
2. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human Action Recognition From Various Data Modalities: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3200–3225. [CrossRef] [PubMed]
3. Rodomagoulakis, I.; Kardaris, N.; Pitsikalis, V.; Mavroudi, E.; Katsamanis, A.; Tsiami, A.; Maragos, P. Multimodal human action recognition in assistive human-robot interaction. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China , 20–25 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2702–2706.
4. Liang, C.; Qi, L.; He, Y.; Guan, L. 3D human action recognition using a single depth feature and locality-constrained affine subspace coding. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2920–2932. [CrossRef]
5. Gedamu, K.; Ji, Y.; Gao, L.; Yang, Y.; Shen, H.T. Relation-mining self-attention network for skeleton-based human action recognition. *Pattern Recognit.* **2023**, *139*, 109455. [CrossRef]
6. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1474–1488. [CrossRef]
7. Yu, B.X.; Liu, Y.; Zhang, X.; Zhong, S.h.; Chan, K.C. MMNet: A Model-Based Multimodal Network for Human Action Recognition in RGB-D Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3522–3538. [CrossRef]
8. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Skeleton-based action recognition with convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong , 10–14 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 597–600.

9. Yan, A.; Wang, Y.; Li, Z.; Qiao, Y. PA3D: Pose-action 3D machine for video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7922–7931.
10. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978.
11. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
13. 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 IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13359–13368.
14. 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 AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
15. 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.
16. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2019; pp. 1227–1236.
17. Wu, Z.; Xiong, C.; Ma, C.Y.; Socher, R.; Davis, L.S. Adaframe: Adaptive frame selection for fast video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1278–1287.
18. Korbar, B.; Tran, D.; Torresani, L. Scsampler: Sampling salient clips from video for efficient action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 6232–6242.
19. Meng, Y.; Lin, C.C.; Panda, R.; Sattigeri, P.; Karlinsky, L.; Oliva, A.; Saenko, K.; Feris, R. Ar-net: Adaptive frame resolution for efficient action recognition. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; Part VII, pp. 86–104.
20. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 20–36.
21. Lin, J.; Gan, C.; Han, S. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE/CVF international Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7083–7093.
22. Li, Y.; Ji, B.; Shi, X.; Zhang, J.; Kang, B.; Wang, L. Tea: Temporal excitation and aggregation for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 909–918.
23. Wang, L.; Tong, Z.; Ji, B.; Wu, G. Tdn: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1895–1904.
24. Fan, H.; Xu, Z.; Zhu, L.; Yan, C.; Ge, J.; Yang, Y. Watching a small portion could be as good as watching all: Towards efficient video classification. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018.
25. Wu, W.; He, D.; Tan, X.; Chen, S.; Wen, S. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6222–6231.
26. Zheng, Y.D.; Liu, Z.; Lu, T.; Wang, L. Dynamic sampling networks for efficient action recognition in videos. *IEEE Trans. Image Process.* **2020**, *29*, 7970–7983. [CrossRef]
27. Lin, J.; Duan, H.; Chen, K.; Lin, D.; Wang, L. Ocsampler: Compressing videos to one clip with single-step sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13894–13903.
28. Wang, J.; Yang, X.; Li, H.; Liu, L.; Wu, Z.; Jiang, Y.G. Efficient video transformers with spatial-temporal token selection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 69–86.
29. Tang, Y.; Tian, Y.; Lu, J.; Li, P.; Zhou, J. Deep progressive reinforcement learning for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5323–5332.
30. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13413–13422.
31. 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.

32. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701. [CrossRef]
33. Shao, D.; Zhao, Y.; Dai, B.; Lin, D. Finegym: A hierarchical video dataset for fine-grained action understanding. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2616–2625.
34. Li, B.; Li, X.; Zhang, Z.; Wu, F. Spatio-temporal graph routing for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8561–8568.
35. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 143–152.
36. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3595–3603.
37. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 1915–1925. [CrossRef]
38. Liu, K.; Gao, L.; Khan, N.M.; Qi, L.; Guan, L. A multi-stream graph convolutional networks-hidden conditional random field model for skeleton-based action recognition. *IEEE Trans. Multimed.* **2020**, *23*, 64–76. [CrossRef]
39. Yang, H.; Yan, D.; Zhang, L.; Sun, Y.; Li, D.; Maybank, S.J. Feedback graph convolutional network for skeleton-based action recognition. *IEEE Trans. Image Process.* **2021**, *31*, 164–175. [CrossRef] [PubMed]
40. Choutas, V.; Weinzaepfel, P.; Revaud, J.; Schmid, C. Potion: Pose motion representation for action recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7024–7033.
41. Asghari-Esfeden, S.; Sznaier, M.; Camps, O. Dynamic motion representation for human action recognition. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 557–566.
42. Liu, J.; Shahroudy, A.; Xu, D.; Kot, A.C.; Wang, G. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 3007–3021. [CrossRef] [PubMed]
43. Liu, J.; Wang, G.; Hu, P.; Duan, L.Y.; Kot, A.C. Global context-aware attention lstm networks for 3d action recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1647–1656.
44. 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 Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2117–2126.
45. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2740–2755. [CrossRef] [PubMed]
46. Wang, Y.; Chen, Z.; Jiang, H.; Song, S.; Han, Y.; Huang, G. Adaptive focus for efficient video recognition. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 16249–16258.
47. Wu, Z.; Li, H.; Xiong, C.; Jiang, Y.G.; Davis, L.S. A Dynamic Frame Selection Framework for Fast Video Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1699–1711. [CrossRef]
48. Zhi, Y.; Tong, Z.; Wang, L.; Wu, G. Mgsampler: An explainable sampling strategy for video action recognition. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1513–1522.
49. Jang, E.; Gu, S.; Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv* **2016**, arXiv:1611.01144.
50. Yang, W.; Zhang, J.; Cai, J.; Xu, Z. Relation selective graph convolutional network for skeleton-based action recognition. *Symmetry* **2021**, *13*, 2275. [CrossRef]
51. Liu, N.; Zhao, Q.; Zhang, N.; Cheng, X.; Zhu, J. Pose-guided complementary features learning for amur tiger re-identification. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
52. Nie, Q.; Wang, J.; Wang, X.; Liu, Y. View-invariant human action recognition based on a 3D bio-constrained skeleton model. *IEEE Trans. Image Process.* **2019**, *28*, 3959–3972. [CrossRef]
53. Bai, Z.; Ding, Q.; Xu, H.; Chi, J.; Zhang, X.; Sun, T. Skeleton-based similar action recognition through integrating the salient image feature into a center-connected graph convolutional network. *Neurocomputing* **2022**, *507*, 40–53. [CrossRef]
54. Gao, Y.; Liu, Z.; Wu, X.; Wu, G.; Zhao, J.; Zhao, X. Skeleton-based human action recognition by the integration of euclidean distance. In Proceedings of the 2021 9th International Conference on Information Technology: IoT and Smart City, New York, NY, USA, 22–25 December 2021; pp. 47–51.
55. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Trans. Image Process.* **2020**, *29*, 9532–9545. [CrossRef]

56. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 183–192.
57. Duan, H.; Wang, J.; Chen, K.; Lin, D. Pyskl: Towards good practices for skeleton action recognition. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 7351–7354.
58. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef]
59. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; Part V, pp. 740–755.
60. Zhou, H.; Liu, Q.; Wang, Y. Learning discriminative representations for skeleton based action recognition. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2023; pp. 10608–10617.
61. Caetano, C.; Brémond, F.; Schwartz, W.R. Skeleton image representation for 3d action recognition based on tree structure and reference joints. In Proceedings of the 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Rio de Janeiro, Brazil, 28–31 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 16–23.
62. 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.
63. Caetano, C.; Sena, J.; Brémond, F.; Dos Santos, J.A.; Schwartz, W.R. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–8.
64. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. Learning clip representations for skeleton-based 3d action recognition. *IEEE Trans. Image Process.* **2018**, *27*, 2842–2855. [CrossRef] [PubMed]
65. Banerjee, A.; Singh, P.K.; Sarkar, R. Fuzzy integral-based CNN classifier fusion for 3D skeleton action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2206–2216. [CrossRef]
66. Liang, D.; Fan, G.; Lin, G.; Chen, W.; Pan, X.; Zhu, H. Three-stream convolutional neural network with multi-task and ensemble learning for 3d action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019.
67. Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; Russell, B. Actionvlad: Learning spatio-temporal aggregation for action classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 971–980.
68. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
69. Zhou, B.; Andonian, A.; Oliva, A.; Torralba, A. Temporal relational reasoning in videos. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 803–818.
70. Kim, M.; Kwon, H.; Wang, C.; Kwak, S.; Cho, M. Relational self-attention: What's missing in attention for video understanding. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8046–8059.
71. Shi, J.; Zhang, Y.; Wang, W.; Xing, B.; Hu, D.; Chen, L. A Novel Two-Stream Transformer-Based Framework for Multi-Modality Human Action Recognition. *Appl. Sci.* **2023**, *13*, 2058. [CrossRef]