applied sciences

MDPI

Article

Key Frame Selection for Temporal Graph Optimization of Skeleton-Based Action Recognition

Jingyi Hou [1,2,3,*], Lei Su [1,2,3] and Yan Zhao [4]

[1] School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China; g20208723@xs.ustb.edu.cn
[2] Institute of Artificial Intelligence, University of Science and Technology Beijing, Beijing 100083, China
[3] Key Laboratory of Intelligent Bionic Unmanned Systems, Ministry of Education, University of Science and Technology Beijing, Beijing 100083, China
[4] School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China; yanzhao@ustb.edu.cn
* Correspondence: houjingyi@ustb.edu.cn

Abstract: Graph neural networks (GNNs) are extensively utilized to capture the spatial–temporal relationships among human body parts for skeleton-based action recognition. However, due to the inefficient information propagation caused by redundant sampling of video frames in the temporal domain, we focus on refining temporal graphs through key frame selection. To this end, we propose a multi-stage key frame selection (MSKFS) method, aiming to find the most representative frames as the graph nodes to learn compact temporal graph representations of human skeletons for action recognition. The MSKFS progressively selects key frames in two stages: (1) salient posture frame selection based on the global dynamics of body parts and (2) key frame refinement and alignment according to intra-frame correlations. The first stage captures the most salient information and aligns the corresponding information of skeleton sequences within the same category. The second stage enriches the subtle information for the integrity of the information derived by the salient frames. Moreover, variational inference is applied to differentiate the key frame refinement and alignment procedure, allowing the end-to-end optimization of arbitrary graph-based models to represent the obtained compact graph for skeleton-based action recognition. Our MSKFS method achieves state-of-the-art performances on two challenging action recognition datasets.

Keywords: action recognition; key frame selection; graph neural network; skeleton sequence; variational inference

Citation: Hou, J.; Su, L.; Zhao, Y. Key Frame Selection for Temporal Graph Optimization of Skeleton-Based Action Recognition. *Appl. Sci.* **2024**, *14*, 9947. https://doi.org/10.3390/app14219947

Academic Editor: João M. F. Rodrigues

Received: 10 September 2024
Revised: 18 October 2024
Accepted: 27 October 2024
Published: 31 October 2024

1. Introduction

Action recognition is a critical component in the field of computer vision, and skeleton-based methods have gained growing interest due to the compact structure and background interference resistance of the skeleton sequences. The goal of skeleton-based action recognition is to interpret human actions through the modeling of spatial and temporal patterns embedded in sequences of skeleton data.

Early deep models often transform a sequence of skeleton data into sequential vectors and complete the recognition task using CNNs or RNNs. These methods struggle to uncover the information of global relationships between human body parts either spatially or temporally. The relationships between body parts can be modeled by graph neural networks (GNNs). Yan et al. [1] infer the relationships between body joints via the presented spatial–temporal graph convolutional network (ST-GCN), which achieves superior performances on the task of skeleton-based action recognition. The ST-GCN learns frame-level features by using GNNs to reason body joints' spatial relationships and models the temporal relationships of the frame-level features via convolutional operations. Since then, there has been an increasing number of methods based on GNNs for skeleton-based action

recognition, such as refining graph topology structures [2–10], reasoning the relations between the partial and the whole graphs [11], improving the robustness of noise interference resistance [12,13] and efficient computing [14,15]. Among them, graph topology refinement methods mainly focus on revealing the latent spatial relationships between body parts beyond only reasoning the physical relations.

Differently, we mainly focus on optimizing the temporal structure of the spatial–temporal graph that represents the skeleton-based action. In this paper, we consider two main problems in the temporal modeling of skeleton-based actions with GNNs. One problem of modeling temporal relations in human actions stems from the inefficiency of information propagation through frame-level features that are highly redundant. The presence of numerous redundant and misleading frames within each skeleton sequence poses a challenge to GNNs that aim to learn discriminative representations of temporal relations, and these redundant frames might even lead to incorrect predictions [14,15]. We also consider another problem of temporal relation representation with GNNs, i.e., the performance might be affected by the dramatically variant lengths of different skeleton sequences. For example, the action "drink water" in the NTU RGB+D 60 dataset [16] takes 66 frames for Subject No. 2 and 150 frames for Subject No. 3. Generally, two main methods are employed to conquer this problem by fixing the lengths of different sequences: zero or loop padding and random sampling. Intuitively, padding increases the computation expense and introduces a lot of irrelevant frames. Random sampling makes the training procedure unstable for the robustness of GNNs [17]. We focus on graph optimization in the temporal domain and the uniformity of the length of the sequence by selecting a certain number of key frames.

To tackle the above challenges, we propose a multi-stage key frame selection (MSKFS) method, which optimizes the temporal graph by automatically discovering a few representative frames from the skeleton sequence to recognize the skeleton-based actions. For an arbitrary skeleton sequence, the key frame selection is accomplished efficiently in two stages based on the global dynamics of body parts and the relationships between frames. Firstly, each skeleton graph is separated into five human body parts, and the global dynamics are used for salient posture frame selection. Secondly, we dynamically supplement and refine the key frames by exploiting the relationships between the key frames selected in the first phase and all the frames in the video to estimate the posterior of the key frames via variational inference with the initially selected salient posture frames as the prior. Finally, a new spatial–temporal graph organized by the key frames is inputted into a GNN-based model to recognize human actions.

The primary contributions of our work are highlighted in the following points:

- We propose a new multi-stage method that conducts key frame selection to align action sequences within the same category as well as preserves the most representative frames including both salient and subtle dynamic information for skeleton-based action recognition.
- We propose a plug-and-play module based on variational inference to simultaneously refine temporal relationships for constructing the spatial–temporal graph and optimize the parameters of the GNNs.
- The proposed method achieves state-of-the-art performances on two large-scale datasets.

2. Related Work

This work refines the temporal graph topology structure using a multi-stage key frame selection method for skeleton-based action recognition. First of all, we discuss deep learning methods for recognizing skeleton-based human actions from the perspective of spaces where the skeleton sequences are processed. Then we describe the key frame selection and its applications for computer vision tasks, especially for action recognition.

2.1. Skeleton-Based Action Recognition

Recently, a variety of deep learning methods for skeleton-based action recognition have been developed. These methods are mainly categorized into two types: space–time-based methods and graph-based methods. CNNs and RNNs are typical models for the space–time-based methods, and a representative of the graph-based methods is the GNN.

Benefiting from the recurrent connections in hidden layers [18], RNNs are able to model temporal sequences. RNN-based methods first organize the skeleton joint coordinates of their corresponding frames into a vector sequence, which is then fed into an RNN model to learn the dynamic relationships. Meanwhile, inspired by the great success of CNNs in 2D/3D image tasks, CNNs are also used to extract features from skeleton-based action sequences. In CNN-based methods, the skeleton joints of a sequence are manually transferred into a pseudo image to learn spatial–temporal features. However, they are limited in modeling the structural relationships involved in the skeleton sequence due to the neglect of the natural structure information of human skeletons.

Numerous advanced methods model the human skeleton sequence as a graph, a representation that is widely used to describe relationships. The skeleton spatial–temporal graph model (ST-GCN) proposed by Yan et al. [1] is the first model discovering spatial structural information and temporal dynamic information from skeleton data. Owing to the representation ability of a spatial–temporal graph and learning ability of a GCN model, the ST-GCN has seen substantial enhancements in recognition performance. Based on the graph representation of human skeletons, a number of GNN-based methods focusing on optimizing the topology structure of the graph are proposed. Gao et al. [19] introduce the graph regression model to optimize the graph topology by observing multiple frames of skeleton data. Li et al. [2] extend a generalized spatial skeleton graph to capture higher-order associations between body parts by inferring the connections of body joints and the human body structure. Peng et al. [4] discover the latent relationships between joints by searching the network architecture with a multiple-hop strategy. Gao et al. [7] add a latent node to the skeleton spatial–temporal graph to discover the implicit connections within individual frames and across multiple frames. Yang et al. [20] reconstruct a centrality graph to model the topology structure of the skeleton sequence by investigating key information on the spatial graph for recognizing human actions. Zeng et al. [8] build skeleton graphs dynamically based on some input human poses instead of using a fixed pre-defined graph. Chen et al. [9] optimize the channel-wise topology structure to learn various topologies in different channels for recognizing human actions with skeleton data. Given the technical robustness of their method in representing spatial structural information, we adopt it as our backbone model. Our focus, however, is on capturing temporal information to further enhance performance, especially in handling complex input data. Chen et al. [21] present a dual-head graph network to learn a multi-granular spatio-temporal graph for tackling the problem of large variations in human actions. To learn coarse-level action presentation, they subsample features along the temporal dimension, while our method directly learns to select and align key frames within the same class to capture time-invariant representations. Liu et al. [22] propose a unified spatial–temporal graph representing skeleton sequences with a Transformer, where the temporal information can be aggregated for hierarchical modeling. Pang et al. [23] learn both local and global information for the graph representation via a GCN and a Transformer, which is trained by a contrastive learning strategy to integrate relations between human joints within and between frames. Different from previous methods, this work aims to optimize the temporal graph via key frame selection.

2.2. Key Frame Selection for Action Recognition

Compared to static images, videos containing human actions provide richer information, primarily including movement patterns such as amplitude and velocity. Additionally, the coordination among different body parts can be fully exhibited in videos.

However, videos composed of a large number of frames contain redundant information [24–26], so we only need a few key frames to represent the actions in videos. There exist some studies that already explore the issues in action recognition. Zhao et al. [27] select the discriminative frames by feature integration of the key frames and their neighbors. Ding et al. [28] split the sequence into segments to select the key frames automatically from each segment. Dong et al. [29] introduce a non-differentiable hard attention mechanism where the attention block is trained via a deep reinforcement learning method to select key frames. As for the skeleton-based action recognition task, only a few methods based on key frames are proposed. Tang et al. [26] distill the key frames with richer information by exploring the sequence redundancy on the basis of deep progressive reinforcement learning. We formulate the process for selecting key frames in two stages by mining the dynamics of human body parts and adaptively adding key frames from the sequence according to the spatial–temporal relationships between body parts.

3. Multi-Stage Key Frame Selection

Given a skeleton sequence $\mathcal{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N\}$ composed of N frames, where $\boldsymbol{X}_t \in \mathbb{R}^{C \times V}$ represents the initialized features (e.g., the spatial location information and the motion information) of the V body joints performed by all the subjects (usually 1 or 2 subjects) in the t-th frame, our goal is to optimize the temporal graph by selecting T key frames as the graph nodes in two stages to recognize human actions from skeleton data. We propose a novel method called multi-stage key frame selection (MSKFS), and Figure 1 depicts the pipeline of the proposed method. Firstly, we discover global dynamics of human body parts to find T_{init} key frames based on salient postures. Then, we supplement and refine the key frames from the number of T_{init} to T using variational inference according to the prior of the salient posture frames. Finally, a GNN-based model with a compact spatial–temporal graph as input is utilized for the task of skeleton-based action recognition.

3.1. Salient Posture Frame Selection

We conduct salient posture frame selection by choosing some exclusive positions by global dynamics of human body parts. An example of salient posture frame selection (action "hand waving") is illustrated in Figure 2. We first divide the skeleton graph in each frame into five body parts (i.e., "trunk", "left arm", "right arm", "left leg" and "right leg") and denote $\boldsymbol{X}_t^s \in \mathbb{R}^{C \times V_s}$ as the s-th body part feature, where $\sum_{s=1}^{5} V_s = V$. We conduct average pooling across all the axes of each body part feature $\boldsymbol{X}_t^s$ to obtain the part global pose dynamics $\boldsymbol{p}^s = [p_1^s, p_2^s, \cdots, p_N^s]^\top$, where

$$p_t^s = \text{AverPooling\_2D}(\boldsymbol{X}_t^s) \in \mathbb{R}^1. \tag{1}$$

The Savitzky–Golay (S-G) filter [30] is applied to remove the high-frequency noise from the $\boldsymbol{p}^s$ to obtain the representative dynamics:

$$\boldsymbol{p} = \text{S} - \text{G}(\boldsymbol{p}^s, \beta_l, \beta_o). \tag{2}$$

where $\boldsymbol{p} = [p_1, p_2, \cdots, p_N]^\top$. The β_l represents the length of the filter window, and β_o is the order of the fitted polynomial. Afterwards, we select the body part with the largest variance to remove the unnecessary low-frequency information,

$$\hat{s} = \underset{s}{\text{argmax}} \text{Var}(\boldsymbol{p}^s). \tag{3}$$

Figure 3 shows a pair of the representative dynamics of the body part "left leg" of an action "jump up" displayed by two subjects as an example. The action is composed of five phases: stand (before jump), squat (before jump), jump, squat (after jump) and stand (after jump). We find that the changes in actions conducted by the two subjects are consistent, thus extracting representative frames of the corresponding phases according to

the dynamics can help to align the inter-class actions and generate temporal scale-invariant representations. From the figure, we also observe that the phase of standing upright is dispensable for the "jump up" action recognition, and the other phases are more necessary and discriminative. Frames at time steps with high volatility in the dynamics can just satisfy the aforementioned characteristics.

Encouraged by the above observations, we select salient posture frames by simply finding the peaks and troughs of the representative dynamics calculated by the comparison of neighboring values, which is formulated as

$$\boldsymbol{t}_s = \text{FindingPT}(\boldsymbol{p}, \beta_d, \beta_p), \tag{4}$$

where $\boldsymbol{t}_s \in \mathbb{R}^{L_1}$ is the indices of the selected salient posture frames, and β_d, β_p are the hyperparameters of distance and prominence, respectively. Specifically, the distance β_d represents the minimum horizontal time steps between the sampled peaks or troughs. If there is distance between two neighboring peaks or troughs, the smaller one will be removed. The prominence β_p is the minimum vertical distance between the peak or trough and the surrounding baseline. The sequence of salient posture frames is generated by extracting the corresponding selected frames of the video according to the chronologically sorted indices.

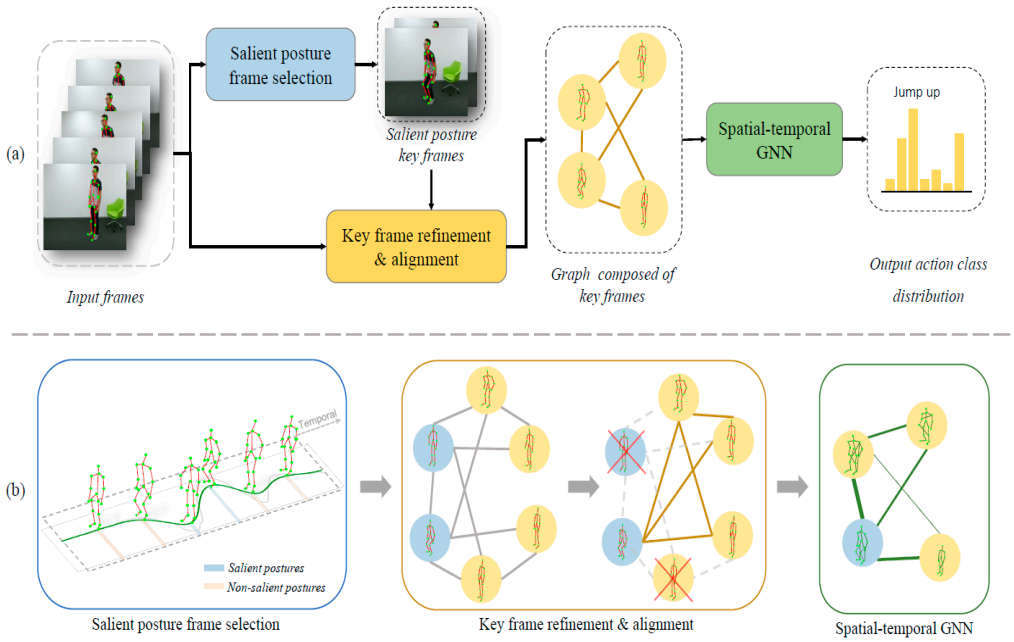


Figure 1. Overview of our MSKFS. (**a**) Architecture of the proposed method. The skeleton data involved in video frames are inputted into the modules of salient posture frame selection and key frame refinement and alignment, where the output salient posture frames of the former module are regarded as the prior for refining and aligning the selected frames of the latter module. The skeleton structures in the selected frames can thus be constructed as a spatial–temporal graph and fed into a GNN to generate the final predicted action category. (**b**) Data flow through three main modules of our method. In the salient posture frame selection, the salient postures are selected from the uniformly sampled frames from the input sequence. In the key frame refinement and alignment, we refine the selection of key frames as the graph nodes with the guidance of salient postures. In the spatial–temporal GNN, the edges of the graph can be further refined and the information of each graph can be propagated to represent the action.

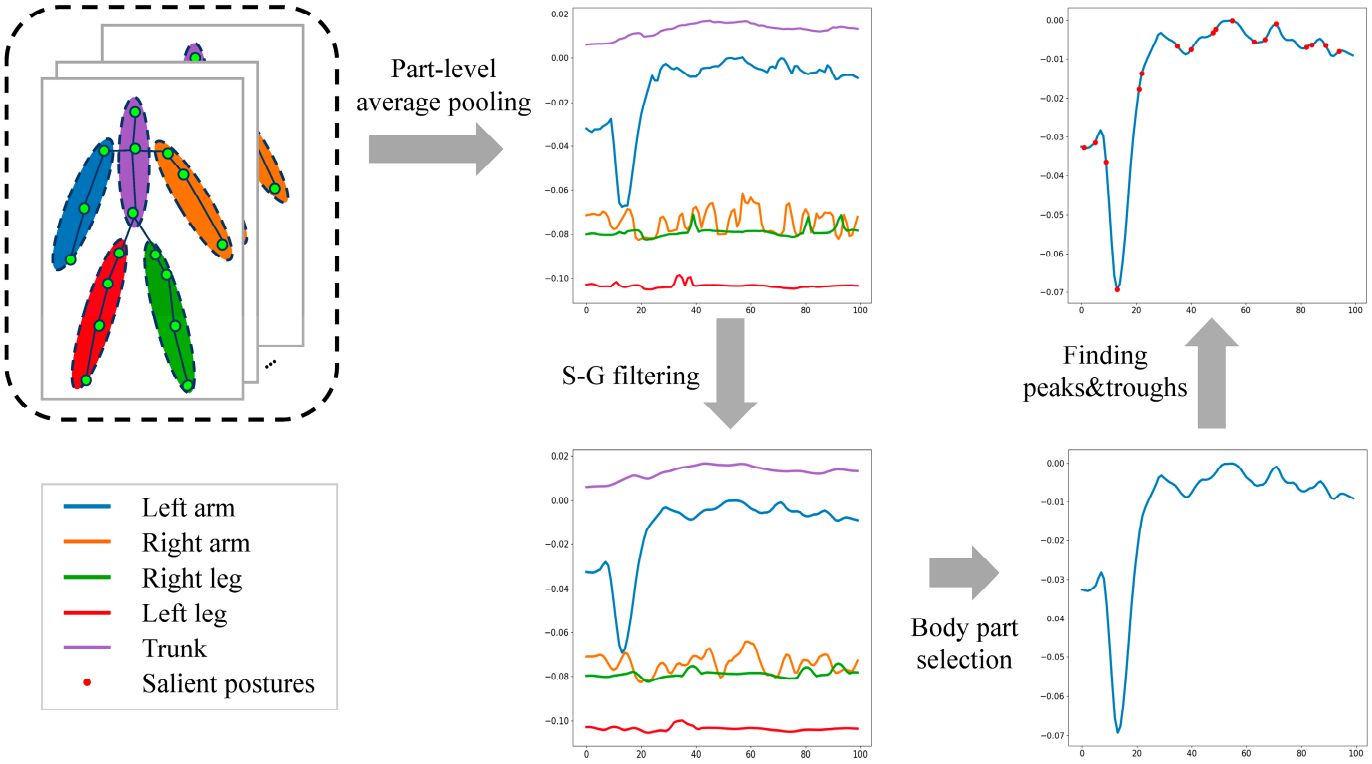
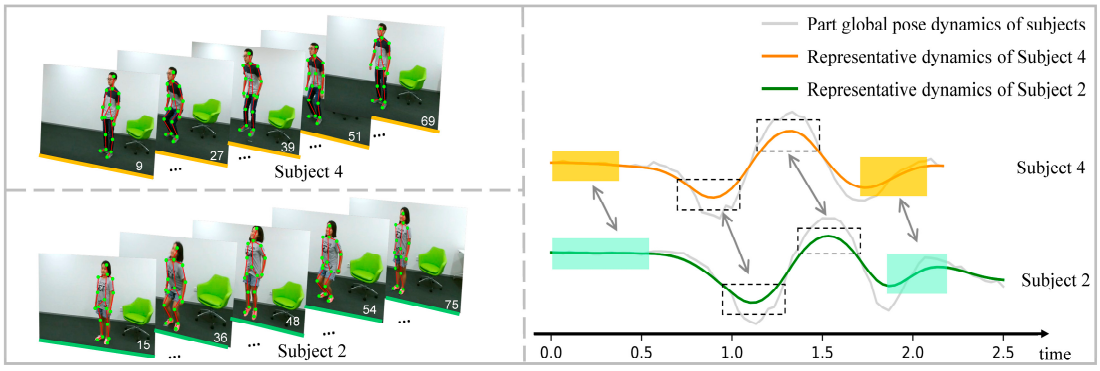


Figure 2. Example data flow of the proposed salient posture frame selection based on the global dynamics of body parts. The global dynamics of different body parts are plotted with curves in different colors. The salient postures can be reflected by dynamic changes.



(a) Observed Skeleton Sequence("Jump up") (b) Global Dynamic Information

Figure 3. Example of the dynamics of body part "left leg" in action "jump up". The left part of the figure is a pair of two skeleton sequences that are performed by Subject No. 2 and Subject No. 4. The right part represents a pair of two dynamics in the body part "left leg".

3.2. Key Frame Refinement and Alignment

In the second stage of our MSKFS, we adaptively augment and refine the key frames based on the selected salient posture frames in the first stage. In addition to the discriminative information carried by the salient posture frames, we still need more detailed information to represent some local continuous variants (i.e., high-order features) of the motion and periodic trends that are important features of time signals. We believe that frames containing the aforementioned characteristics share some similarities with the salient posture frames in certain spaces. For example, frames representing local continuous variants are similar in appearance, and the most representative variants are the neighborhoods of the salient posture frames. Accordingly, we can learn to map the joint-level features of

the frames into a common feature space and select the final key frames according to the similarity measurement in this space.

In this paper, we use the variational inference to estimate the posterior of the selected key frames conditioned on the features that carry information related to the similarity measurement. There are two benefits as follows. First, the reparameterization trick enables the key frame selection operation differentiable for optimization with gradient descent. Second, the variational inference allows the knowledge to constrain the distribution of the temporal graph node, which improves the generalization and robustness of the proposed model.

To be specific, a spatial graph convolution layer is first utilized to learn the frame-level representation $\boldsymbol{g}_t$. At time step t, the human bodies are regarded as a graph where vertices represent body joints and edges characterize their connections. The graph feature of the v-th vertex at time t is calculated by

$$\boldsymbol{g}_t = \text{AverPooling\_1D}(\text{ReLU}(\boldsymbol{A}\boldsymbol{X}_{\text{t}}^{\top}\boldsymbol{W})), \tag{5}$$

where $\boldsymbol{W}$ is a learnable weight matrix, $\boldsymbol{A} \in \{0,1\}^{V\times V}$ represents the adjacency matrix and AverPooling_1D$(\cdot)$ means the average pooling operation along the axis of body joints.

Given the frame-level graph features, we then use a multi-head attention operation to learn the similarity-measurement-based feature $\boldsymbol{s}_t$ for the key frame refinement and alignment. For the t-th frame, i-th head attention operation is formulated as

$$\boldsymbol{s}_t[i] = f_{v_i}([\boldsymbol{g}_{t_s}, \boldsymbol{g}_t]) \cdot \text{Softmax}(\frac{f_{k_i}([\boldsymbol{g}_{t_s}, \boldsymbol{g}_{\text{t}}])^{\top} f_{q_i}(\boldsymbol{g}_t)}{n_{\text{head}}}), \tag{6}$$

where $f_{q_i}(\cdot)$, $f_{k_i}(\cdot)$ and $f_{v_i}(\cdot)$ are linear functions, and n_{head} denotes the number of attention heads. Thus, we have the similarity-measurement-based feature $\boldsymbol{s}_t$ by concatenating the features calculated by different heads of attention operations:

$$\boldsymbol{s}_t = [\boldsymbol{s}_t[1]; \boldsymbol{s}_t[2]; \ldots; \boldsymbol{s}_t[n_{\text{head}}]]. \tag{7}$$

We learn the posterior of each key frame conditioned on the input human skeleton sequence:

$$q(w_t|\mathcal{X}) = \text{Bernoulli}(\pi_t), \tag{8}$$

where

$$w_t = \begin{cases} 0, & \text{if } L + \pi_t < 0, \\ 1, & \text{otherwise,} \end{cases} \tag{9}$$

is an indicator variable representing whether to select the frame at t as the key frame, and $L \sim$ Logistic, i.e.,

$$\begin{aligned} &L = \log U - \log(1-U), \\ &U \sim \text{Uniform}(0,1). \end{aligned} \tag{10}$$

Bernoulli(π_t) denotes the Bernoulli distribution with parameter π_t that is calculated by a linear function given $\boldsymbol{s}_t$. For variational inference, we consider the form of the prior as

$$p(w_t) = \text{Bernoulli}(\pi_t^{\text{prior}}). \tag{11}$$

However, even though we apply the reparameterization trick to the Bernoulli distribution, the parameter of the distribution still cannot be optimized during the gradient descent. To ensure the feasibility of optimization, we relax the discrete distribution to a continuous distribution [31] as:

$$q(w_t|\mathcal{X}) = \text{BinConcrete}(\pi_t, \lambda),\ p(w_t) = \text{BinConcrete}(\pi_t^{\text{prior}}\lambda_{\text{prior}}), \tag{12}$$

where λ_{prior} and λ are the scale parameters of the continuing distribution. The BinConcrete distribution is defined as

$$w_t = \sigma(\frac{L + \pi_t}{\lambda}), \tag{13}$$

where $\sigma(\cdot)$ is the sigmoid function. For the posterior distribution, the parameter $\pi_t = f_\pi(\boldsymbol{s}_t)$ where $f_\pi(\cdot)$ is a linear function. For the prior distribution, the parameter π_t^{prior} is calculated by

$$\pi_t^{\text{prior}} = \log(Pr_t) - \log(1 - Pr_t), \tag{14}$$

and Pr_t is the frequency of the selected salient posture key frame at time step t.

3.3. Skeleton-Based Action Recognition

Using the obtained posterior distribution, we can select the key frames from all the frames of the entire video. Multiplied with $\sigma(w_t)$, the graph features $\boldsymbol{g}_t$ of the selected key frames with top T posterior probabilities encoded with a Transformer encoding layer and inputted into a GNN-based model for the final prediction. We apply the CTR-GCN [9] here as the base GNN-based model to obtain the predicted action label y.

Consequently, the entire model can be optimized by maximizing the variational lower bound:

$$V_{LB} = \mathbf{E}_{q(w_t|\mathcal{X})}[\log p(y|\mathcal{X}, w_{1:N})] - \sum_{t=1}^{N} \text{KL}[q(w_t|\mathcal{X})||p(w_t)], \tag{15}$$

where $p(y|\mathcal{X}, w_{1:N})$ is obtained by the output of the GNN-based action classification model in our method, and KL divergence is calculated as

$$\begin{aligned} &\text{KL}[q_\theta(w_t|\mathcal{X})||p(w_t)] \\ &= \log \tfrac{\lambda}{\lambda_{\text{prior}}} + \pi_t - \pi_t^{\text{prior}} + \tfrac{1}{K} \sum_{k=1}^{K} [(\lambda_{\text{prior}} - \lambda) \log(w_t^{(k)}(1 - w_t^{(k)})) + \Delta], \end{aligned} \tag{16}$$

where

$$\Delta = 2 \log \frac{\exp(\pi_t^{\text{prior}})(w_t^{(k)})^{-\lambda^{\text{prior}}} + (1 - w_t^{(k)})^{-\lambda^{\text{prior}}}}{\exp(\pi_t)(w_t^{(k)})^{-\lambda} + (1 - w_t^{(k)})^{-\lambda}}. \tag{17}$$

During the inference procedure, we can directly select the key frames with top T posterior of $q(w_t|\mathcal{X})$.

4. Experiments

4.1. Datasets

Our method is evaluated on the NTU RGB+D 60 dataset [16] and the NTU RGB+D 120 dataset [32].

The NTU RGB+D 60 dataset has a total of 56,880 samples executed by 40 volunteers, with ages spanning from 10 to 35 years. All data are captured using the Microsoft Kinect v2 sensor, capturing actions from 3 different camera positions. Within the dataset, the action samples are classified into 60 classes: 40 of these cover common daily behaviors, 9 focus on health-related actions and 11 involve mutual actions performed by 2 players. There are two distinct criteria used to partition the dataset into training and test sets: (1) cross-subject (X-sub), where the training set and the test set are split according to the character IDs; and (2) cross-view (X-view), where the training set comprises samples from camera views 2 and 3, and the test set contains samples from camera view 1.

The NTU RGB+D 120 dataset is an extension of NTU RGB+D 60 by adding an additional 60 classes, resulting in a total of 114,480 samples across 120 classes, all performed by 106 volunteers. Like its predecessor, this dataset is split into training and test sets using two criteria: (1) cross-subject (X-sub), where the training and test sets are derived from

53 different subjects each, and (2) cross-setup (X-setup), where the training set consists of samples from setups with even IDs, and the test set comprises those with odd IDs.

4.2. Implementation Details

Network setting. The numbers of channels of the backbone model are set to 64–64–64–128–128–128–256–256–256, where each module comprises a GCN and a TCN. In the salient posture frame selection set, the filter window is configured with a length of 8, and the polynomial order is set to 2. The hyperparameters defined in Equation (4), i.e., distance and prominence, are set empirically to values of 7 and 0.0001, respectively.

Training. The SGD optimizer is employed with a momentum of 0.9 and a warm-up phase is introduced for 5 epochs at a learning rate of 0.1 to enhance training stability. The learning rate is decreased by a factor of 0.1 at epochs 20, 35, 45 and 55. We also implement a weight decay of 0.0004. For two large datasets, the batch size is 64.

Data processing. Following [26], we first interpolate each sequence into f = 100 frames in which the first and last frames remain the same. This strategy reduces the use of a lot of computing resources. To augment the data during training, we utilize random rotation following the method in [33].

Evaluation metric. Following the instructions of [16,32,34], we use the accuracy metric to evaluate the performances of different methods on NTU RGB+D 60 and NTU RGB+D 120. The metric is calculated by

$$\text{Accuracy} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}(\hat{y}_i = y_i) \tag{18}$$

where M is the total number of samples, $\hat{y}_i$ represents the prediction of the i-th sample, y_i is the ground-truth of the i-th sample and $\mathbf{1}(\cdot)$ denotes the indicator function.

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

In Tables 1 and 2, our proposed method is benchmarked against state-of-the-art methods on the NTU RGB+D 60 and NTU RGB+D 120 datasets, respectively. In order to ensure comparability, we employ the multi-stream fusion strategy at the score level, as is common among most compared methods that use various types of features to boost action recognition performance. Specifically, we first conduct experiments using two single-stream strategies employing two different input features: joint coordinates (joint) and the differential values of joints in the same frames (bone), respectively. Then we perform experiments using a 4-stream strategy that leverages 4 kinds of features as input, including the aforementioned joint and bone features, along with two motion features of the joint and bone that are calculated by computing the differential of these features along the time dimension (fusion-4s). To show the compatibility of our MSKFS method, in addition to using CTR-GCN as the backbone of our approach (ours (CTR-GC [9])), we also adopted BlockGCN as the backbone (ours (Block-GC [10])), which learns topological knowledge from the physical connections.

Table 1. Comparison with the state-of-the-art action recognition methods (accuracy, %) on the NTU RGB+D 60 dataset. The best results for each feature setting and split are indicated in bold.

Methods	Features	X-Sub	X-View
ST-GCN [1]	Joint	81.5	88.3
SR-TSL [35]	Joint	78.8	88.2
AS-GCN [2]	Joint	86.8	94.2
AGCN [3]	Joint	-	93.7
MS-G3D [36]	Joint	89.4	95.0
PL-GCN [37]	Joint	84.0	90.5
MST-GCN [38]	Joint	89.0	95.1
Skeletal [8]	Joint	89.0	95.3

Table 1. *Cont.*

Methods	Features	X-Sub	X-View
DualHead-Net [21]	Joint	90.3	**96.1**
TranSkeleton [22]	Joint	90.1	95.4
FR Head [39]	Joint	90.3	95.3
BlockGCN [10]	Joint	90.9	95.4
Ours (CTR-GC [9])	Joint	90.2	95.6
Ours (Block-GC [10])	Joint	**91.5**	**96.1**
DGNN [40]	Fusion-4s	89.9	96.1
Shift-GCN [14]	Fusion-4s	90.7	96.5
Dynamic GCN [5]	Fusion-4s	91.5	96.0
MST-GCN [38]	Fusion-4s	91.5	96.6
Skeletal [8]	Fusion-4s	91.6	96.7
CTR-GCN [9]	Fusion-4s	92.4	96.8
DualHead-Net [21]	Fusion-4s	92.0	96.6
Ta-CNN+ [41]	Fusion-4s	90.7	95.1
MCTM-Net [42]	Fusion-4s	92.8	96.8
FR Head [39]	Fusion-4s	92.8	96.8
BlockGCN [10]	Fusion-4s	**93.1**	97.0
Ours (CTR-GC [9])	Fusion-4s	92.7	96.9
Ours (Block-GC [10])	Fusion-4s	**93.1**	**97.1**

Table 2. Comparison with state-of-the-art for action recognition (accuracy, %) on the NTU RGB+D 120 dataset. The best results for each feature setting and split are indicated in bold.

Methods	Features	X-Sub	X-Setup
MST-GCN [38]	Joint	82.8	84.5
Skeletal [8]	Joint	83.5	85.7
DualHead-Net [21]	Joint	84.6	85.9
TranSkeleton [22]	Joint	84.9	86.3
HDGCN [43]	Joint	85.7	87.3
FR Head [39]	Joint	85.5	87.3
BlockGCN [10]	Joint	86.9	88.2
Ours (CTR-GC [9])	Joint	86.2	**88.7**
Ours (Block-GC [10])	Joint	**86.7**	88.6
Shift-GCN [14]	Fusion-4s	85.9	87.6
Dynamic GCN [5]	Fusion-4s	87.3	88.6
MST-GCN [38]	Fusion-4s	87.5	88.8
Skeletal [8]	Fusion-4s	87.5	89.2
CTR-GCN [9]	Fusion-4s	88.9	90.6
DualHead-Net [21]	Fusion-4s	88.2	89.3
Ta-CNN+ [41]	Fusion-4s	85.7	87.3
MCTM-Net [42]	Fusion-4s	89.3	91.0
FR Head [39]	Fusion-4s	89.5	87.3
BlockGCN [10]	Fusion-4s	90.3	91.5
Shift-GCN [14]	Fusion-4s	85.9	87.6
Ours (CTR-GC [9])	Fusion-4s	89.1	90.9
Ours (Block-GC [10])	Fusion-4s	**90.4**	**91.7**

As shown in the tables, our method performs better than the backbone methods on both datasets under all the data split settings, indicating the effectiveness of our method. Note that [9] also conduct data sampling instead of using all frames as input in their implementation. To be specific, this method randomly crops the center portion of input frames to a target size ranging from 50% to 100% of the original length for data augmentation during training and crops the 90% center part of the entire sequence at the test time. For a detailed understanding of the backbone model, please consult Section 4.5. Actually, our method is complementary to most skeleton-based action recognition methods that mainly consider encoding the spatial relationships of action, because we concentrate on the optimization of

temporal graph nodes at the early stage of the entire data flow, while other methods focus on either spatial graph representation or the design of GNNs. Comparison results with methods that process frame-level representations, such as DualHead-Net [21], show the substantial competitiveness of our proposed MSKFS method.

4.4. Analysis on the Length of Key Frame Sequence

In this part, we analyze the effects of different lengths of selected key frame sequences. How key frame length affects the action recognition performance is explored within the context of the NTU RGB+D 60 dataset. According to the key frame selection and refinement method in this paper, this section validates the effect of key frame length on classification performance using the joint modality under the X-sub and X-view split criteria. The accuracy of the action recognition task as it relates to key frame length is shown in Figure 4. For the X-sub setting, when the number of frames is less than 70, the recognition accuracy is highly sensitive to sequence length. Between 50 and 70 frames, the model's recognition accuracy significantly increases from 89.2% to 90.2%. This phenomenon may be due to an insufficient number of key frames where the model cannot capture enough motion information, leading to improved recognition accuracy as the number of key frames increases. When the frame length reaches 70, the accuracy does not rise significantly anymore but instead shows a certain decrease (at 80 frames), which could be mainly attributed to the inclusion of too many key frames, leading to redundant or noisy frames in the optimized skeleton sequence, thereby hindering the performance of the model. As the key frame length increases to 90, the task's accuracy experiences a slight increase but is almost indistinguishable from when the length is 70 frames. Therefore, for the X-sub setting, we select a key frame length of 70 to maximally include useful motion information while ensuring recognition accuracy and considering the computational efficiency of the model. Similarly, for the X-view setting, we choose a key frame length of 80. Considering that the average length of salient pose frames across all samples is approximately equal to 12, we set $T_{\text{init}} = 20$.

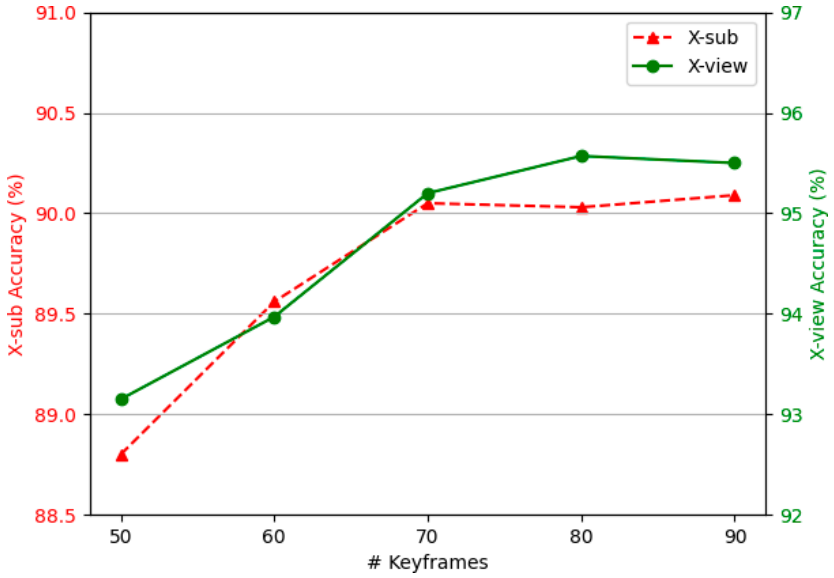


Figure 4. Action prediction performance of different lengths of the sequences.

4.5. Ablation Study

To further show that our method can exploit more informative representations for action recognition contributed by every module of MSKFS, we conduct an ablation study on the NTU RGB+D 60 dataset. Firstly, we directly apply all the frames as input and some commonly used key frame selection methods combined with our backbone model, CTR-GCN, to study the effectiveness of the proposed key frame selection module. These baselines are provided as follows:

All. All skeleton sequences are interpolated to 70 frames into the backbone action recognition model.

Uniform sampling [33]. This method segments the complete skeleton sequence into 20 equal segments and randomly selects a single frame from each segment to organize a 20-frame sequence as the input of the backbone model for training. Different from [33], we directly sample the middle frame in each segment to avoid introducing any randomness into the final results. In the test time, the target size is set to 20% of the original length.

Salient posture (ours). This method only selects the salient posture frames, i.e., the frames selected by the first stage of our method, as the input of the backbone model.

In addition, we validate the effectiveness of the second stage of our method, i.e., the key frame refinement and alignment (KFRA), by providing results of replacing the salient posture frame selection with the uniform-sampling strategy. Table 3 shows the comparison results. With the "all" strategy, classification accuracies of 92.3% for Top 1 and 96.3% for Top 5 are achieved, indicating that, without any manipulation, direct interpolation can reduce the frame length to some extent. However, compared to our strategy, i.e., salient posture, the all strategy yields lower accuracy, suggesting that selecting salient pose frames is necessary for action recognition tasks. The "uniform-sampling strategy" involves selecting 20 frames at equidistant intervals from the skeleton sequence, achieving classification accuracies of 92.1% and 95.9% on Top 1. Compared to the previous strategy, this strategy shows a decrease in classification accuracy and a noticeable gap compared to the proposed method. This suggests that, for different action sequences, frames selected based on intervals cannot directly be considered key frames, and uniformly selecting salient frames cannot serve as prior knowledge to guide key frame refinement. Our MSKFS method achieves better action recognition accuracy under both settings, proving that the proposed salient pose selection method aids in completing action recognition tasks.

Table 3. Accuracies (%) of different baseline methods. The best results for each split are indicated in bold.

Methods		w/o KFRA (%)		w/KFRA (%)	
		Top 1	Top 5	Top 1	Top 5
All	X-sub	92.3	98.2	-	-
	X-view	96.3	99.4	-	-
Uniform sampling	X-sub	92.1	98.3	92.1	97.9
	X-view	95.6	98.6	95.8	98.7
Salient posture (ours)	X-sub	92.2	98.1	**92.5**	**98.9**
	X-view	96.1	98.8	**96.9**	**99.7**

4.6. Visualization Result

We provide a visualization of some key frames finally selected by the proposed method. Figure 5a shows the images corresponding to key frames A to H of the action "hand waving". In frames A and B, the subject gradually raises her left hand. By frame C, the subject's left hand has reached the starting position for waving and begins to wave. Additionally, the curve shows that the subject completed four cycles of waving, with the amplitude decreasing gradually, which aligns with the key frames selected by the model. After key frame refinement and alignment, more frames unrelated to the action type were removed, and more detailed information can be added to minimize information loss. We also show the relationships between these frames by presenting the attention maps learned in our proposed model in Figure 5b. From the figure, we can observe that the relationship weights between these manually selected key frames, which we consider representative, are generally large, indicating the effectiveness of the temporal graph optimization. Another notable observation is frame C, the starting position of the action, showing strong relationships with many other key frames, underscoring its significance.

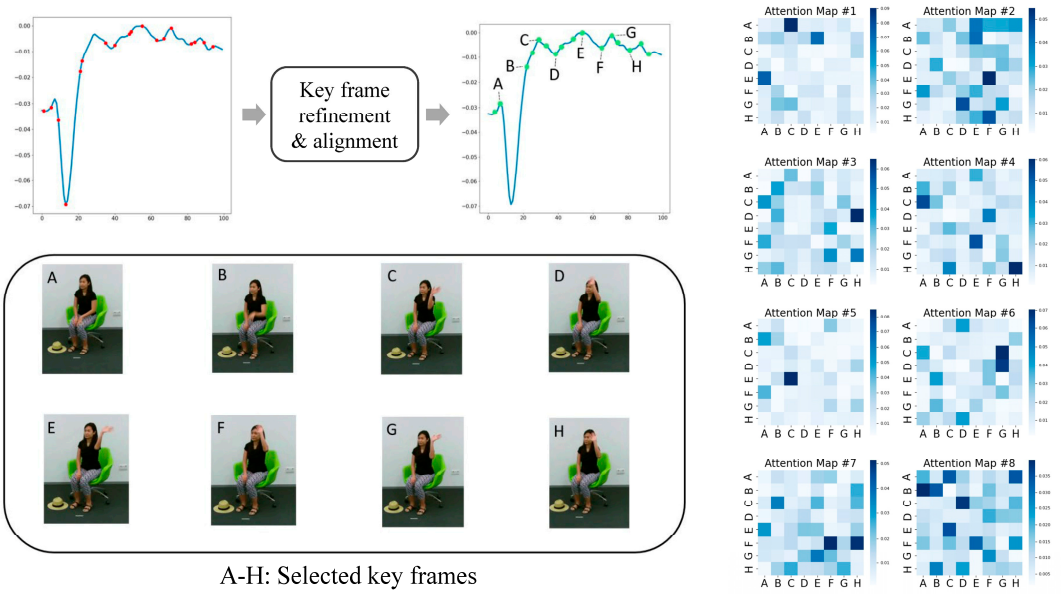


Figure 5. Visualization of (**a**) the key frame selection and alignment procedure and (**b**) the attention maps of the selected key frames. A–H are selected key frames.

5. Conclusions

In this paper, we have presented a temporal graph refinement method via multi-stage key frame selection to recognize skeleton-based human actions. Our method is capable of reducing the scale of the temporal graph and improving the efficiency of the information propagation between graph vertices. We have introduced a multi-stage key frame selection (MSKFS) method to select representative frames and discard redundant and misleading frames. The proposed MSKFS includes two stages: salient posture frame selection, key frame refinement and alignment. The MSKFS learns the salient features and subtle dynamic features that are consistent within inter-class for better spatial–temporal representation of skeleton-based human action sequences. Guided by the action recognition task, an end-to-end variational inference approach is employed to estimate the parameters of the posterior distribution for key frame selection. Experiments on several challenging datasets on skeleton-based action prediction demonstrate the superior performances of the proposed method.

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

Funding: This research was funded by Natural Science Foundation of China, grant numbers 62106021 and U20A20225.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The NTU RGB+D 60 dataset [16] and the NTU RGB+D 120 dataset [31] were used. This data can be found here: https://rose1.ntu.edu.sg/dataset/actionRecognition (accessed on 5 September 2024).

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

References

1. 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; pp. 7444–7452.
2. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; IEEE: New York, NY, USA, 2019; pp. 3595–3603.
3. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
4. Peng, W.; Hong, X.; Chen, H.; Zhao, G. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 2669–2676.
5. 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 ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 55–63.
6. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Trans. Image Process.* **2020**, *29*, 9532–9545. [CrossRef] [PubMed]
7. Gao, J.; He, T.; Zhou, X.; Ge, S. Focusing and diffusion: Bidirectional attentive graph convolutional networks for skeleton-based action recognition. *arXiv* **2019**, arXiv:1912.11521.
8. Zeng, A.; Sun, X.; Yang, L.; Zhao, N.; Liu, M.; Xu, Q. Learning skeletal graph neural networks for hard 3d pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11416–11425.
9. 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 International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 13339–13348.
10. Zhou, Y.; Yan, X.; Cheng, Z.-Q.; Yan, Y.; Dai, Q.; Hua, X.-S. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 2049–2058.
11. Song, Y.; Zhang, Z.; Wang, L. Richly activated graph convolutional network for action recognition with incomplete skeletons. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1–5.
12. Song, Y.; Zhang, Z.; Shan, C.; Wang, L. Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Trans. Circuit Syst. Video Technol.* **2021**, *31*, 1915–1925. [CrossRef]
13. Li, S.; Yi, J.; Farha, Y.A.; Gall, J. Pose refinement graph convolutional network for skeleton-based action recognition. *IEEE Robot. Autom. Lett.* **2021**, *6*, 1028–1035. [CrossRef]
14. Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-based action recognition with shift graph convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 180–189.
15. Song, Y.; Zhang, Z.; Shan, C.; Wang, L. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1474–1488. [CrossRef] [PubMed]
16. Shahroudy, A.; Liu, J.; Ng, T.; Wang, G. NTU RGB+D: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019. [CrossRef]
17. Duan, H.; Zhao, Y.; Chen, K.; Shao, D.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978.
18. 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.* **2022**, *45*, 3200–3225. [CrossRef] [PubMed]
19. Gao, X.; Hu, W.; Tang, J.; Liu, J.; Guo, Z. Optimized skeleton-based action recognition via sparsified graph regression. In Proceedings of the ACM International Conference on Multimedia, New York, NY, USA, 21–25 October 2019; pp. 601–610.
20. Yang, D.; Li, M.M.; Fu, H.; Fan, J.; Leung, H. Centrality graph convolutional networks for skeleton-based action recognition. *arXiv* **2020**, arXiv:2003.03007.
21. Chen, T.; Zhou, D.; Wang, J.; Wang, S.; Guan, Y.; He, X.; Ding, E. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In Proceedings of the ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 4334–4342.
22. Liu, H.; Liu, Y.; Chen, Y.; Yuan, C.; Li, B.; Hu, W. Transkeleton: Hierarchical spatial-temporal transformer for skeleton-based action recognition. *IEEE Trans. Circuit Syst. Video Technol.* **2023**, *33*, 4137–4148. [CrossRef]
23. Pang, C.; Lu, X.; Lyu, L. Skeleton-based action recognition through contrasting two-stream spatial-temporal networks. *IEEE Trans. Multimed.* **2023**, *25*, 8699–8711. [CrossRef]
24. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.

25. Tong, Z.; Song, Y.; Wang, J.; Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv* **2022**, arXiv:2203.12602.
26. 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.
27. Zhao, Z.; Elgammal, A.M. Information theoretic key frame selection for action recognition. In *Proceedings of the British Machine Vision Conference (BMVC), Leeds, UK, 1–4 September 2008*; Everingham, M., Needham, C.J., Fraile, R., Eds.; British Machine Vision Association: Durham, UK, 2008; pp. 1–10.
28. Ding, C.; Wen, S.; Ding, W.; Liu, K.; Belyaev, E. Temporal segment graph convolutional networks for skeleton-based action recognition. *Eng. Appl. Artif. Intell.* **2022**, *110*, 104675. [CrossRef]
29. Dong, W.; Zhang, Z.; Song, C.; Tan, T. Identifying the key frames: An attention-aware sampling method for action recognition. *Pattern Recognit.* **2022**, *130*, 108797. [CrossRef]
30. 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, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
31. Maddison, C.J.; Mnih, A.; Teh, Y.W. The concrete distribution: A continuous relaxation of discrete random variables. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
32. 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]
33. Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; Zheng, N. Semantics-guided neural networks for efficient skeleton-based human action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1112–1121.
34. Lan, G.; Wu, Y.; Hu, F.; Hao, Q. Vision-based human pose estimation via deep learning: A survey. *IEEE Trans. Hum.-Mach. Syst.* **2023**, *53*, 253–268. [CrossRef]
35. 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 September 2018; pp. 106–121.
36. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 140–149.
37. Huang, L.; Huang, Y.; Ouyang, W.; Wang, L. Part-level graph convolutional network for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11045–11052.
38. Chen, Z.; Li, S.; Yang, B.; Li, Q.; Liu, H. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 1113–1122.
39. Zhou, H.; Liu, Q.; Wang, Y. Learning discriminative representations for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10608–10617.
40. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with directed graph neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7912–7921.
41. Xu, K.; Ye, F.; Zhong, Q.; Xie, D. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; pp. 2866–2874.
42. Wu, C.; Wu, X.-J.; Xu, T.; Shen, Z.; Kittler, J. Motion complement and temporal multifocusing for skeleton-based action recognition. *IEEE Trans. Circuit Syst. Video Technol.* **2023**, *34*, 34–45. [CrossRef]
43. Lee, J.; Lee, M.; Lee, D.; Lee, S. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 10410–10419.