Action Recognition with Multi-stream Motion Modeling and Mutual Information Maximization

Yuheng Yang[1] , **Haipeng Chen**[1] , **Zhenguang Liu** [2*] , **Yingda Lyu**[3*] , **Beibei Zhang**[5] , **Shuang Wu**[4*] , **Zhibo Wang**[2] and **Kui Ren** [2]

[1]College of Computer Science and Technology, Jilin University
[2]School of Cyber Science and Technology, Zhejiang University
[3]Public Computer Education and Research Center, Jilin University
[4]Black Sesame Technologies
[5]Zhejiang Lab

yangyh20@mails.jlu.edu.cn, {chenhp, ydlv}@jlu.edu.cn, {liuzhenguang2008, bzeecs}@gmail.com, wushuang@outlook.sg, {zhibowang, kuiren}@zju.edu.cn

Abstract

Action recognition has long been a fundamental and intriguing problem in artificial intelligence. The task is challenging due to the high dimensionality nature of an action, as well as the subtle motion details to be considered. Current state-of-the-art approaches typically learn from articulated motion sequences in the straightforward 3D Euclidean space. However, the *vanilla* Euclidean space is not efficient for modeling important motion characteristics such as the joint-wise angular acceleration, which reveals the driving force behind the motion. Moreover, current methods typically attend to each channel equally and lack theoretical constrains on extracting task-relevant features from the input.

In this paper, we seek to tackle these challenges from three aspects: (1) We propose to incorporate an acceleration representation, explicitly modeling the higher-order variations in motion. (2) We introduce a novel Stream-GCN network equipped with multi-stream components and channel attention, where different representations (*i.e.*, streams) supplement each other towards a more precise action recognition while attention capitalizes on those important channels. (3) We explore feature-level supervision for maximizing the extraction of task-relevant information and formulate this into a mutual information loss. Empirically, our approach sets the new state-of-the-art performance on three benchmark datasets, NTU RGB+D, NTU RGB+D 120, and NW-UCLA.

1 Introduction

Having a mental representation of what actions other humans are performing is crucial for us to adjust our behaviors

---

* Corresponding author
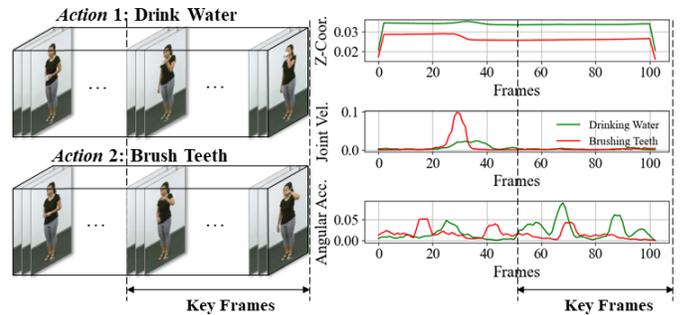https://github.com/ActionR-Group/Stream-GCN

Figure 1: The illustration of similar actions and their features at different orders. **Left**: Showcase two videos of similar actions, *"drinking water"* and *"brushing teeth"*, the key frames are video clips that are decisive for identifying the action. **Right**: Take the hand joint as an example, the curves indicate trajectories of motion characteristics for various orders on z-dimension.

and plan our course of actions [Liu *et al.*, 2021; Hao *et al.*, 2022a]. Similarly, the capacity for machines to model and recognize human actions is very much coveted. Now, action recognition spawns a wide spectrum of applications including public space surveillance [Wang *et al.*, 2016], human-robot interaction [Hao *et al.*, 2022b], violence detection [Singh *et al.*, 2018], and autonomous driving [Lu *et al.*, 2018; Hao *et al.*, 2022b]. As such, action recognition has attracted extensive attention in the past decade [Shi *et al.*, 2019b; Chen *et al.*, 2020b; Chen *et al.*, 2020a].

Fundamentally, action recognition amounts to learning a mapping from a pose sequence to an action class label. The task is inherently challenging due to the high dimensionality of the input pose sequence and the subtle motion details which differentiate different actions. In this paper, we focalize action recognition with 3D articulated pose sequences as inputs, given that 3D articulated pose can now be conveniently acquired through commodity motion capture systems or extracted from videos via off-the-shelf pose estimation methods like [Wang *et al.*, 2014].

Earlier approaches for action recognition [Liu *et al.*, 2016]

cast the human body as a set of 3D joint coordinates, which in fact treat different joints as independent entities and fail to capture connections between joints [Liu *et al.*, 2019b]. Recently, GCN (Graph Convolutional Network) based methods have become popular solutions as the human skeleton inherently possesses a tree-like graph structure. **A first class of work** puts the efforts in selecting and enriching features to better represent motion [Shi *et al.*, 2019a; Shi *et al.*, 2019b], facilitating action recognition from the input feature perspective. For instance, [Shi *et al.*, 2019b] advocates incorporating the lengths and directions of bones as second-order input features, which are informative for action recognition. [Shi *et al.*, 2019a] adds frame-wise bone and joint velocities as additional motion features to enhance the body representation. **Another line of work** resorts to devising sophisticated adjacency matrices, seeking to pursue better modeling of joint dependencies. For example, [Liu *et al.*, 2020] introduces a multi-scale aggregation scheme, leveraging distance adjacency matrices to measure the degree of dependency between two arbitrary skeleton joints. [Li *et al.*, 2019] presents a method to adaptively learn the non-physical dependencies among joints by an encoder-decoder structure, which automatically infers the link strengths between joints.

Upon investigating and experimenting on the released code of state-of-the-art methods [Chen *et al.*, 2021a; Yan *et al.*, 2018; Chen *et al.*, 2021b], we empirically observe that: **First**, current approaches typically revolve around employing lower-order features such as joint coordinates or velocities in the Euclidean space. Despite their simplicity, these features might not be favourable for describing subtle motion details such as joint angular acceleration and motion trends. Figure 1 illustrates two similar actions, "drinking water" and "brushing teeth". It is evident to see that the joint coordinates of the two actions are very close to each other, while the velocity features of the two actions are also very similar. However, the joint angular accelerations of the two actions are quite different. **Second**, most existing approaches attend to each channel *equally*, dissevering the fact that different channels contribute unequally in recognizing an action. For instance, in identifying the action "drinking water", vertical movements (z-dimension) of the hand joint play a more vital role than horizontal movements (xy-dimensions). Apparently, it is not a good idea to assign equal weight to each dimension. **Third**, current approaches typically employ the conventional CE (Cross-Entropy) loss to supervise the learning of motion information, lacking an effective constraint on guaranteeing that task-relevant features are extracted.

In this paper, we embrace three key designs to tackle the challenges. Technically, **(1)** We consider incorporating angular accelerations as complementary information and extract them within the framework of rigid body kinematics. These higher-order acceleration features provide an additional input stream, on top of the original Euclidean space motion features, potentially improving the completeness and robustness of the model. **(2)** We introduce a novel GCN network equipped with multiple input streams and channel-wise attention, where different input streams are assembled while attention weights capitalize on those important channels. **(3)** We further engage an information-theoretic objective between the extracted deep features and action labels. Maximizing this mutual information objective drives our model to fully mine task-relevant information while reducing task-irrelevant nuisances.

Thereafter, we conduct extensive experiments on three large benchmark datasets including NTU RGB+D, NTU RGB+D 120, and NW-UCLA. Empirically, our approach consistently and significantly outperforms state-of-the-art methods. Interestingly, we observe that existing methods indeed have difficulties in distinguishing between similar actions.

To summarize, our **key contributions** are as follows: 1) A new motion representation is proposed, which enriches the lower-order motion representations with higher-order motion representations. 2) A multi-stream graph convolution network is presented to comprehensively extract knowledge from multiple representations. Within the framework, the distinct weights of different channels are considered. To explicitly supervise the knowledge extraction from the input motion sequence, we theoretically analyzed the mutual information between the extracted deep feature and the label, arriving at a loss that maximizes the task-relevant information. 3) Our method achieves the new state-of-the-art performance on three benchmark datasets and overall provides interesting insights. The implementations are released, hoping to facilitate future research.

2 Method

Broadly, a human skeleton can be cast as a set of joints and bones. We refer to the set of joints as a set of nodes $\mathcal{V} = \{v_i\}_{i=1}^n$. Further, we model the set of bones, each of which connects a pair of joints, as a set of edges $\mathcal{E} = \{e_i\}_{i=1}^{n \times n}$. Thus, we conveniently depict a human skeleton pose as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $n = |\mathcal{V}|$ and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.

Problem. Presented with an observed 3D skeleton motion sequence $\mathcal{S} = \langle \mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_m \rangle$, where $\mathcal{G}_i$ is the skeleton pose of the human in the ith frame, we are interested in predicting its action class label $y \in \mathbb{R}$. Put differently, we seek to learn the non-linear mapping from a motion sequence $\mathcal{S}$ to its action class label y.

Method overview. Overall, our method consists of three key components: a) motion representations engaging both lower-order and higher-order motion features, b) a novel Stream-GCN network that is equipped with channel-wise attention, and is able to integrate higher-order and lower-order representations, and c) a mutual information objective that effectuates feature-level supervision for maximizing the extraction of task-relevant information. Specifically, we first leverage multiple motion representations to frame the motion context. Then, the motion sequences are fed into the proposed Stream-GCN network to predict the action label. Within the network, we introduce a mutual information objective to supervise the prediction. We would like to highlight that our approach has an edge in engaging acceleration to capture subtle motion details, while our network games in maximizing the extraction of task-relevant information and getting rid of task-irrelevant nuisances. In what follows, we elaborate on the three key components in detail.

2.1 Motion Representation

To represent human motion, there have been efforts to outline a pose as 3D coordinates of all skeletal joints [Liu *et al.*, 2019b]. The 3D joint coordinates offer an intuitive and direct way to precisely sketch the motion. Unfortunately, such a representation regards different joints as independent entities and is unfavourable for describing higher-order subtle motion details such as bone rotation trends [Shi *et al.*, 2019a].

Motivated by this, we incorporate both higher-order and lower-order motion representations. For lower-order representations, we include joint coordinates, bone lengths, joint velocity, and bone velocity. For higher-order representations, we engage in joint and bone angular acceleration representations. These different representations (*i.e.*, streams) supplement each other towards a more complete and precise motion characterization. Below, we formulate the joint (or bone) angular acceleration representation.

The articulate human skeleton can be considered as rigid bodies (*i.e.* bones) connected by joints. The motion of a joint in relation to its neighbouring joints can therefore be characterized in the framework of rigid body kinematics. Within this framework, the angular acceleration of a joint has a significant role. Since the acceleration of an object depends directly on the force acting on it, the acceleration change actually characterizes the changes of forces exerted upon joints, which serve as the driven source for motion. Take drinking water as an example, conventional methods merely consider the position displacement of the raised arm. They omit the acceleration change of the arm in the process of lifting the hand, moving upwards with a constant speed, and decelerating when approaching the mouth.

Instantaneous angle. We denote v_i^p as the coordinate of the i-th joint at the p-th frame. Supposing we are to calculate the acceleration feature of v_i^p. We set $\delta_i^p = v_i^p - v_{i-1}^p$ as an axis vector from joint v_{i-1}^p to joint v_i^p. Upon this, we define the instantaneous absolute angle as follows [Aslanov and Ledkov, 2012]:

$$\theta_i^p = \overbrace{\theta_{i-1}^p}^{convected} + \underbrace{\widetilde{\theta}_i^p}_{relative} = \theta_{i-1}^p + \arccos\left(\frac{\delta_i^{p+1} \cdot \delta_i^p}{|\delta_i^{p+1}| \cdot |\delta_i^p|}\right), \tag{1}$$

where θ_{i-1}^p is the instantaneous convected rotation angle of joint v_{i-1}^p w.r.t. the fixed coordinate system, and $\widetilde{\theta}_i^p$ is the instantaneous relative rotation angle as opposed to joint v_{i-1}^p.

Instantaneous angular velocity. Further, we obtain the instantaneous angular velocity ω_i^p of joint v_i^p by differentiating Equation (1) w.r.t. t:

$$\omega_i^p = \frac{\mathrm{d}\theta_i^p}{\mathrm{d}t} = \overbrace{\omega_{i-1}^p}^{\mathrm{d}\theta_{i-1}^p/\mathrm{d}t} + \underbrace{\widetilde{\omega}_i^p}_{\mathrm{d}\widetilde{\theta}_i^p/\mathrm{d}t}. \tag{2}$$

Instantaneous angular acceleration. To facilitate understanding, we visualize *instantaneous angle* θ_i^p and *instantaneous angular velocity* ω_i^p. Subsequently, we acquire the

angular acceleration from the instantaneous angular velocity. Mathematically, we formulate the angular acceleration of joint v_i^p as ε_i^p. Naively, we could obtain the acceleration by differentiating Equation (2) w.r.t. t, yielding:

$$\varepsilon_i^p = \frac{\mathrm{d}\omega_i^p}{\mathrm{d}t} = \underbrace{\varepsilon_{i-1}^p}_{\mathrm{d}\omega_{i-1}^p/\mathrm{d}t} + \overbrace{\widetilde{\varepsilon}_i^p + \omega_{i-1}^p \times \omega_i^p}^{\mathrm{d}\widetilde{\omega}_i^p/\mathrm{d}t}, \tag{3}$$

where ε_i^p measures the angular acceleration of joint v_i^p. It is evident to see that the derivation of the above process also applies to computing the angular acceleration of bones.

2.2 Stream-GCN with Multi-stream Representation Fusion and Channel Attention

Up to this point, we have presented how to obtain higher-order motion representations. Next, we introduce our Stream-GCN (GCN with multi-stream representation fusion and channel attention) network that extracts cues from higher-order and lower-order motion representations to recognize an action. The overall architecture of the network is shown in Figure 2. The network absorbs six input streams: four lower-order motion representations (joint coordinates, bone lengths, joint velocity, and bone velocity) and two higher-order motion representations (joint and bone angular accelerations). Each input stream is processed by a Stream-GCN network to predict an action class distribution. Their class distributions are then ensembled to give the final class distribution.

Stream-GCN network architecture. Typically, the operation of a GCN layer is given by:

$$X^{r+1} = \sigma(\hat{A}X^r M^r), \tag{4}$$

where $X^r \in \mathbb{R}^{n \times l_r}$ and $X^{r+1} \in \mathbb{R}^{n \times l_r}$ are the features of rth and $r + 1^{th}$ layers, respectively. n is the number of nodes in the graph, which translates to the number of joints in this task. l_r is the length of joint features at the rth layer. $\sigma(\cdot)$ is an activation function, *e.g.*, ReLU. Matrix $M^r \in \mathbb{R}^{l_r \times l_{r+1}}$ is a network parameter (transformation matrix). Filter matrix $\hat{A}$ is computed based on $\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$, where $\tilde{A} = A + I$ and adjacency matrix $A \in \mathbb{R}^{n \times n}$ characterizes the connections between joints. In the adjacency matrix A, $A_{i,j} = 1$ if there exist a bone connects the ith and j^{th} joints, and $A_{i,j} = 0$ otherwise. $\tilde{D} \in \mathbb{R}^{n \times n}$ is a degree matrix and $\tilde{D}_{i,i} = \sum_j \tilde{A}_{i,j}$.

Conventional methods such as [Shi *et al.*, 2019a; Chen *et al.*, 2021a] directly adopt equal channel weights for all joints and bones in each frame. This scheme fails to attend to the distinct roles of different channels in recognizing an action. Motivated by this, we propose a cross-channel attention module to adaptively assign weights to different channels. Formally, in our settings, the input feature is $X \in \mathbb{R}^{n \times c \times m}$, where c is the number of channels and m is the number of frames in the given motion sequence. Put differently, the initial feature length (at the 0th layer) $l_0 = c \times m$.

Different from Equation (4), upon the rth layer, we apply one layer graph convolution as below:

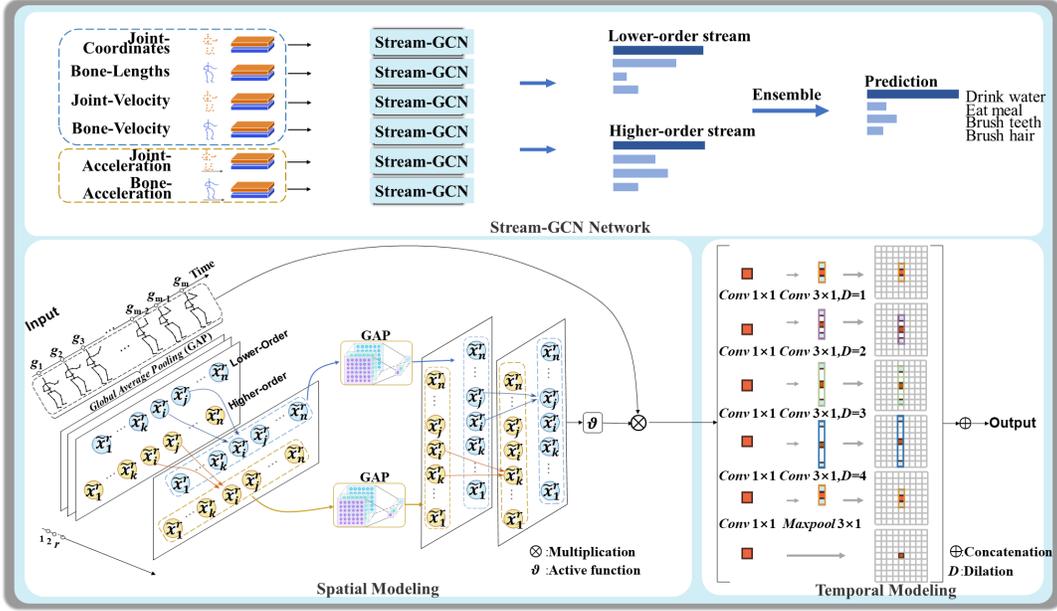$$X^{r+1} = \sigma\left(\hat{A}\tau(X^r)M^r\right), \tag{5}$$

Figure 2: The overall pipeline of our Stream-GCN Network. The goal is to identify the action class by virtue of the motion sequence. For clarity of illustration, we only show one-layer spatial and temporal modeling in this figure. We first absorb input streams from the lower-order and high-order representations. For each stream, the spatial modeling conducts a channel attention module that consists of a set of pooling and convolution operations, yielding the features with channel weights. The temporal modeling adopts multi-scale convolutions, capturing long-range temporal dependencies within a motion sequence. For multi-stream fusion, each stream predicts an action class distribution, which is ensembled to approach the final class distribution.

where $\tau(\cdot)$ is a cross-channel attention function that boosts the synergy of different channels. More specifically, to compute the weights of different channels, $\tau(\cdot)$ consists of a set of pooling, convolution, and subtraction operations, formulated in Equations (6)–(9).

Formally, we first apply linear transformation $\mathcal{L}$, which is a 1×1 convolution, to X^r, transforming X^r from $\mathbb{R}^{n \times c_r \times m}$ to $\mathbb{R}^{n \times c'_r \times m}$. In favor of better aggregating spatial features, we apply a global pooling operation $\Psi(\cdot)$ in the temporal domain on features X^r to sketch its spatial features:

$$\tilde{X}^r = \Psi\big(\mathcal{L}(X^r)\big), \tag{6}$$

where $\tilde{X}^r \in \mathbb{R}^{n \times c'_r}$ is a set of joint spatial features, namely $\tilde{X}^r = \{\tilde{x}^r_1, \tilde{x}^r_2, \ldots, \tilde{x}^r_n\}$ and $\tilde{x}^r_i \in \mathbb{R}^{c'_r}$.

Next, we explore to capture the channel interaction across joint spatial features. As shown in Figure 2, given a pair of joints (v_i, v_j) and their corresponding joint spatial features $(\tilde{x}^r_i, \tilde{x}^r_j)$. We design a correlation modeling function $\mathcal{Z}(\cdot)$ that aggregates channel-wise features between v_i and v_j [Chen *et al.*, 2021a]:

$$\mathcal{Z}(\tilde{x}^r_i, \tilde{x}^r_j) = \tilde{x}^r_i - \tilde{x}^r_j. \tag{7}$$

To better exploit the contribution of each channel, we leverage a global pooling operation $\Phi(\cdot)$ on $\mathcal{Z}(\tilde{x}^r_i, \tilde{x}^r_j)$ to obtain a channel feature $\Phi(\mathcal{Z}(\tilde{x}^r_i, \tilde{x}^r_j)) \in \mathbb{R}^{c'_r}$. Thereafter, we employ *attention mechanism* [Wang *et al.*, 2019] on the channel feature to quantify the contributions of different channels in

action recognition:

$$\vartheta\bigg(\mathcal{F}\big(\Phi\big(\mathcal{Z}(\tilde{x}^r_i, \tilde{x}^r_j)\big), k\big)\bigg), \tag{8}$$

where ϑ is an activation function, and $\mathcal{F}$ is a $1D$ convolution with kernel size of k. In summary, cross-channel attention function $\tau(\cdot)$ has the form:

$$\tau(X^r) = X^r \vartheta\bigg(\mathcal{F}\big(\Phi\big(\mathcal{Z}(\tilde{x}^r_i, \tilde{x}^r_j)\big), k\big)\bigg). \tag{9}$$

To model the action in the temporal domain, we design a multi-scale temporal modeling that follows [Liu *et al.*, 2020]. As shown in Figure 2, the module consists of six parallel temporal convolutional branches. Each branch starts with a 1×1 convolution to aggregate features between different channels. The first four branches contain a 3×1 temporal convolution with four different dilations to obtain multi-scale temporal fields. The Maxpool of the fifth branch is applied to emphasize the most significant feature information among consecutive frames. The last 1×1 convolution is added to the residual-preserving gradient during backpropagation.

2.3 Mutual Information Objective

We can certainly train the Stream-GCN with CE loss, as is done in most previous methods [Ye *et al.*, 2020]. Given our systematic examination of extracting motion features for action recognition, it would be fruitful to investigate whether introducing a novel supervision method at the feature level would facilitate the task.

Let Z denote the extracted deep features from input X. Inspired by [Liu *et al.*, 2022; Tian *et al.*, 2021], we try to enforce the correlation between extracted deep features Z and action label Y, seeking to maximize the extraction of task-relevant information and get rid of task-irrelevant nuisances.

Mutual information (MI). MI measures the amount of information shared between two random variables. Formally, MI between Z and Y quantifies the statistical dependency of encoding variables Z and action label Y:

$$\mathcal{I}(Y;Z) = \mathbb{E}_{p(Y,Z)}[log \frac{p(Y,Z)}{p(Y)p(Z)}], \qquad (10)$$

where $p(Y, Z)$ is the probability distribution between Z and Y, while $p(Z)$ and $p(Y)$ are their marginals [Liu *et al.*, 2022].

Mutual information loss. In order to ensure that the maximum amount of information about Y is extracted to the deep feature Z, while reducing the information that is not relevant to Y in Z, our primary objective can be formulated as:

$$IB(Y, Z) = \max \mathcal{I}(Y;Z|X). \qquad (11)$$

Due to the notorious difficulty of the conditional MI computations especially in neural networks [Tian *et al.*, 2021], we perform a simplification. We factorize Equation (11) as follows:

$$IB(Y, Z) = \overbrace{\mathcal{I}(Y;Z)}^{relevancy} - \underbrace{[\mathcal{I}(Z;X)]}_{compression} + \overbrace{[\mathcal{I}(Z;Y|X)]}^{redundancy}. \qquad (12)$$

The first term forces features to be informative about the label Y. The second term compresses Z to be concise. The third term constraints Z to include task-relevant information in X. These MI terms are depicted in Figure 3 and can be estimated by existing MI estimators [Tian *et al.*, 2021]. In our experiments, we employ the Variational Information Bottleneck (VIB) [Alemi *et al.*, 2016] and Variational Self-Distillation (VSD) [Tian *et al.*, 2021] to estimate each MI term.

3 Experiments

In this section, we conduct extensive experiments to empirically evaluate our method on three benchmark action recognition datasets. Broadly, we intend to answer the following research questions:

- **RQ1:** How does the proposed method compare against the state-of-the-art approaches for skeleton-based action recognition?

- **RQ2:** How much do different components of Stream-GCN contribute to its performance?

- **RQ3:** What interesting insights and findings can we obtain from the empirical results?

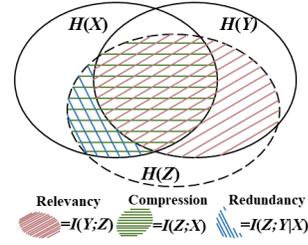Next, we first present the experimental settings, followed by answering the above research questions one by one.

Figure 3: Our mutual information maximizes $I(Y;Z)$, compresses $I(Z;X)$, and preserves $I(Z;Y|X)$. $H(\cdot)$ denotes entropy. The visualization is inspired by [Tian *et al.*, 2021].

3.1 Experimental Settings

Datasets. We adopt three widely used action recognition datasets, namely NTU-RGB+D, NTU-RGB+D 120, and Northwestern-UCLA, to evaluate the proposed method.

NTU-RGB+D. NTU-RGB+D [Shahroudy *et al.*, 2016] is tailored for the skeleton-based action recognition task. It contains 56,880 video samples from 60 action classes performed by 40 volunteers. Each sample contains one action and is guaranteed to have a maximum of two subjects, which is captured by three Microsoft Kinect v2 cameras. This dataset provides two sub-benchmarks: (1) Cross-Subject (X-Sub): data for 20 subjects is used as the training data, while the rest is used as test data. (2) Cross-View (X-View) divides the training and test sets according to different camera views.

NTU-RGB+D 120. NTU-RGB+D 120 [Liu *et al.*, 2019a] is the current largest human action recognition dataset, which extends NTU-RGB+D with an additional 60 action classes and 57,600 video samples. It contains a total of 114,480 samples from 120 action classes, which are executed by 106 volunteers and captured by three Kinect cameras. Within the dataset, two benchmarks are maintained: (1) Cross-Subject (X-Sub), which categorizes 53 subjects into the training class and the other 53 subjects into the test class. (2) Cross-Setup (X-Set), which arranges data items with even IDs into the training group and odd IDs into the test group.

Northwestern-UCLA. The Northwestern-UCLA dataset contains 1,494 video samples in 10 classes [Wang *et al.*, 2014]. These videos were collected by filming ten actors with three Kinect cameras. We follow the evaluation protocol mentioned in [Wang *et al.*, 2014], where videos collected by the first two cameras serve as the training samples and the rest serve as test samples.

Implementation details. We conduct experiments on a computer equipped with an Intel Xeon E5 CPU at 2.1GHz, three NVIDIA GeForce GTX 1080 Ti GPUs, and the RAM of 64GB. We leverage PyTorch 1.1 to implement our model. We apply pre-processing method of [Zhang *et al.*, 2020] to the data in NTU-RGB+D and NTU-RGB+D 120. Moreover, we initialize the data in Northwestern-UCLA according to the approach of [Cheng *et al.*, 2020b]. We apply stochastic gradient descent (SGD) with 0.9 Nesterov momentum to train the Stream-GCN model. For NTU-RGB+D and NTU-RGB+D 120 datasets, the number of training epochs is set to 65 with the first 5 epochs being warm-up epochs, which help stabilize the training process. For NTU-RGB+D and NTU-RGB+D

Methods	X-Sub(%)	X-Set(%)
ST-LSTM [Liu *et al.*, 2016]	55.7	57.9
2s-AGCN [Shi *et al.*, 2019b]	82.9	84.9
4s Shift-GCN [Cheng *et al.*, 2020b]	85.9	87.6
DC-GCN+ADG [Cheng *et al.*, 2020a]	86.5	88.1
MS-GCN [Liu *et al.*, 2020]	86.9	88.4
Dynamic GCN [Ye *et al.*, 2020]	87.3	88.6
MST-GCN [Feng *et al.*, 2021]	87.5	88.8
Ta-CNN [Xu *et al.*, 2022]	85.4	86.8
EfficientGCN-B4 [Song *et al.*, 2022]	88.3	89.1
Stream-GCN(Joint stream)	85.8	86.7
Stream-GCN(Bone stream)	86.5	88.3
Stream-GCN(4 streams)	89.2	90.4
Stream-GCN(6 streams)	**89.7**	**91.0**

Table 1: Comparisons of the Top-1 accuracy(%) with the state-of-the-art methods on the NTU RGB+D 120 dataset.

Methods	X-Sub(%)	X-View(%)
IndRNN [Liu *et al.*, 2016]	81.8	88.0
2s-AGCN [Shi *et al.*, 2019b]	88.5	95.1
SGN [Zhang *et al.*, 2020]	89.0	94.5
AGC-LSTM [Si *et al.*, 2019]	89.2	95.0
DGNN [Shi *et al.*, 2019a]	89.9	96.1
4s Shift-GCN [Cheng *et al.*, 2020b]	90.7	96.5
DC-GCN+ADG [Cheng *et al.*, 2020a]	90.8	96.6
Dynamic GCN [Ye *et al.*, 2020]	91.5	96.0
MS-GCN [Liu *et al.*, 2020]	91.5	96.2
MST-GCN [Feng *et al.*, 2021]	91.5	96.6
Ta-CNN [Xu *et al.*, 2022]	90.4	94.8
EfficientGCN-B4 [Song *et al.*, 2022]	91.7	95.7
Stream-GCN	**92.9**	**96.9**

Table 2: Comparisons of the Top-1 accuracy(%) with the state-of-the-art methods on the NTU RGB+D dataset.

Methods	Top-1(%)
Lie Group [Veeriah *et al.*, 2015]	74.2
Ensemble TS-LSTM [Lee *et al.*, 2017]	89.2
4s Shift-GCN [Cheng *et al.*, 2020b]	94.6
DC-GCN+ADG [Cheng *et al.*, 2020a]	95.3
Ta-CNN [Xu *et al.*, 2022]	96.1
Stream-GCN	**96.8**

Table 3: Comparisons of the Top-1 accuracy(%) with the state-of-the-art methods on the Northwestern-UCLA dataset.

Ensemble Methods	Acceleration Streams		Acc(%)
	Joint	Bone	
4 Streams			89.2
5 Streams		✓	89.4
5 Streams	✓		89.3
6 Streams	✓	✓	**89.7**

Table 4: Study on removing acceleration streams.

Ye *et al.*, 2020] achieve much better results than single stream methods [Zhang *et al.*, 2020; Si *et al.*, 2019]. This reveals the importance of incorporating different representations to approach more precise action recognition.

3.3 Ablation Study (RQ2)

We perform ablation experiments to examine the influence of each component in our method (*i.e.*, Motion Representation, Cross-channel attention Module, and MI Loss). All the ablation studies are conducted on the NTU-RGB+D 120 cross-subject benchmark dataset. We also investigate the impact of modifying the parameter settings of the cross-channel attention module.

Motion representation. To verify the impact of the proposed acceleration representations in modeling higher-order motion features. We tried removing the joint angular acceleration stream and bone angular acceleration stream respectively. Specifically, we *first* tried removing all acceleration streams. *Then*, we studied removing only the acceleration stream of the joint. *Finally*, we also tried removing only the acceleration stream of bone. The quantitative results are demonstrated in Table 4, empirical results reveal that using the high-order representation indeed boosts the action recognition accuracy.

Cross-channel attention module. Next, we are interested in the effect of removing the cross-channel attention module. The empirical results are demonstrated in Table 5. Remarkably, the result in the second line of Table 5 shows that when the cross-channel attention module is removed, the accuracy is decreased (89.7% vs 89.4%).

Mutual information loss. We validate the effect of the MI Loss in Section 2.3. As illustrated in the third and the fourth lines of Table 5, the 0.4% accuracy improvement provides empirical evidence that our proposed MI loss is effective. This reveals that the additional feature-level supervision facilitates the extraction of task-specific information.

Parameter settings. By default, we use cross-channel attention to assign different weights to different channels in

120 datasets, the initial learning rate is set to 0.1 and decays by 0.1 every 35 epochs, the batch size is selected as 64. For the Northwestern-UCLA dataset, the initial learning rate is set to 0.01 and decays by 0.0001 every 50 epochs, the batch size is set to 16.

3.2 Comparison with Existing Methods (RQ1)

We first empirically compare the proposed model with the state-of-the-art methods. The experimental results are summarized in Tables 1–3. Table 1 and Table 2 present results on NTU RGB+D 120 and NTU RGB+D datasets, while Table 3 illustrates results on the Northwestern-UCLA dataset. From the tables, we have the following observations. **(a)** The proposed method consistently outperforms state-of-the-art approaches on all three benchmark datasets. For instance, on the NTU120 RGB+D dataset, state-of-the-art method [Song *et al.*, 2022] achieves a 88.3% action recognition accuracy. In contrast, Stream-GCN obtains a 89.7% accuracy. **(b)** Generally, GCN based approaches (such as [Shi *et al.*, 2019a; Shi *et al.*, 2019b; Liu *et al.*, 2020]) perform better than RNN based methods (such as [Liu *et al.*, 2016]). The empirical evidence suggests that GCN models are more suitable for action recognition tasks as they may conveniently model the graph structure of the human body. **(c)** Another finding is that multi-stream models usually outperform single-stream ones. For example, on the NTU RGB+D 120 dataset, multi-stream models [Cheng *et al.*, 2020b; Song *et al.*, 2022;

Attention Module	Mutual Information	Params	Acc(%)
		1.21M	88.9
	✓	1.23M	89.4
✓		1.64M	89.3
✓	✓	1.65M	**89.7**

Table 5: The impact of each component.

Methods	ϑ	Kernel size				
		$K = 1$	$K = 3$	$K = 5$	$K = 7$	$K = 9$
Stream-GCN	Sig(·)	**85.8**	84.0	84.3	84.0	79.9
Stream-GCN + Mask	Sig(·)	84.4	84.1	84.4	84.3	83.9
Stream-GCN	ReLU(·)	84.1	84.1	84.2	84.0	84.2
Stream-GCN + Mask	ReLU(·)	84.0	83.9	84.3	84.0	84.3
Stream-GCN	Tanh(·)	83.7	84.3	84.0	84.1	84.0
Stream-GCN + Mask	Tanh(·)	83.9	84.4	84.2	83.9	83.9

Table 6: Accuracy (%) with different parameter settings on the cross-channel attention module.

recognizing an action. We explore the hyper-parameter settings in the cross-channel attention module with joint stream (*i.e.*, joint coordinates). The key hyper-parameters include the activation function ϑ, the kernel size of the $1D$ convolution, and the mask on the adjacency matrix. As shown in Table 6, we observe that Stream-GCN obtains the best performance when ϑ is a Sigmoid function, the kernel size $K = 1$, and the adjacency matrix is without a mask.

3.4 Discussions (RQ3)

Visualize results. We visualize the learned attention maps of different layers (*layer* 1, *layer* 5, and *layer* 9) for the "drinking water" action in Figure 4. From the figure, we see that different layers of the learned attention map contain distinct semantics. Specifically, in the lower layers, it seems that the model focuses on the relations between hands, shoulders, and wrists. This is intuitive since the model will initially focus on the body parts (joints) where the movement takes place. In the intermediate layers, the arm is more concerned by the model, with attention to the other parts gradually diminishing. This demonstrates that the model is analyzing what exactly the arm is doing. In the higher layers, the information is highly aggregated. Meanwhile, the model will combine other joints that are helpful to identify the motion. As an example, for recognizing the action of "drinking water", the correlation between the hands and the head is detected and leveraged by the model. Experiments on three different benchmark datasets suggest that incorporating higher-order motion sequences achieves higher accuracy. Interestingly, a point that attracts our attention is: we empirically observe that as the number of action classes goes higher, although incorporating high-order representation still improves the accuracy, the advantage of incorporating high-order representation decreases. We plan to dive deeper into this phenomenon and come up with new solutions.

4 Related Work

Earlier skeleton-based action recognition methods model the motion sequences with convolutional neural networks or recurrent neural networks [Si *et al.*, 2018]. As these conven-
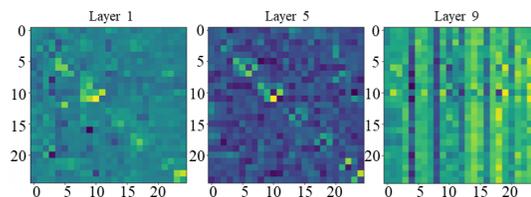

Figure 4: Examples of the learned attention maps at different layers for the drinking water action. The numbers denote different joints, e.g., *"number* 4" denotes *"head"* and *"number* 8" denotes *"left hand"*. The brighter area indicates that the weight of the correlation matrix is larger there, which means the correlation strength between the joints is stronger.

tional methods have inherent difficulties in capturing the connections between joints, they tend to suffer from unsatisfactory results. Recent efforts have embraced GCNs to model the human motion sequences as spatio-temporal graphs [Bian *et al.*, 2021]. [Yan *et al.*, 2018] is a pioneering work in this line of efforts, proposing a ST-GCN model to characterize the skeleton motion as a GCN. [Shi *et al.*, 2019b] proposes an adaptive graph convolution network based on ST-GCN, incorporating skeletal bone features as an additional input stream and adopting self-attention to learn the graph parameters. [Shi *et al.*, 2019a] introduces frame-wise bone and joint velocities as additional input streams for action recognition.

Existing methods tend to have difficulties in discriminating between similar actions. Meanwhile, to our best knowledge, the higher-order motion features and mutual information are rarely explored in skeleton-based action recognition [Tan *et al.*, 2021; Chen *et al.*, 2022]. This inspires us to seek methods to facilitate the model recognizing the action class precisely.

5 Conclusion

In this paper, we have proposed a motion representation derived from rigid body kinematics. The new representation captures the higher-order motion features, complementing conventional motion representations. We also present a Stream-GCN network equipped with multiple input streams and channel-wise attention, where different streams are ensembled while attention weights capitalize on those important channels. Finally, we introduce a mutual information objective for theoretic supervision on extracted features. Extensive experiments show that our method consistently surpasses state-of-the-art methods on three different benchmark datasets. Our code is released to facilitate researchers.

References

[Alemi *et al.*, 2016] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[Aslanov and Ledkov, 2012] Vladimir Aslanov and Alexander Ledkov. *Dynamics of tethered satellite systems*. Elsevier, 2012.

[Bian *et al.*, 2021] Cunling Bian, Wei Feng, Liang Wan, and Song Wang. Structural knowledge distillation for efficient skeleton-based action recognition. *IEEE Transactions on Image Processing*, 30:2963–2976, 2021.

[Chen *et al.*, 2020a] Jingjing Chen, Liangming Pan, Zhipeng Wei, Xiang Wang, Chong-Wah Ngo, and Tat-Seng Chua. Zero-shot ingredient recognition by multi-relational graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10542–10550, 2020.

[Chen *et al.*, 2020b] Jingjing Chen, Bin Zhu, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. A study of multi-task and region-wise deep learning for food ingredient recognition. *IEEE Transactions on Image Processing*, 30:1514–1526, 2020.

[Chen *et al.*, 2021a] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021.

[Chen *et al.*, 2021b] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1113–1122, 2021.

[Chen *et al.*, 2022] Meng Chen, Li Lu, Zhongjie Ba, and Kui Ren. Phoneytalker: An out-of-the-box toolkit for adversarial example attack on speaker recognition. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1419–1428. IEEE, 2022.

[Cheng *et al.*, 2020a] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *European Conference on Computer Vision*, pages 536–553. Springer, 2020.

[Cheng *et al.*, 2020b] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020.

[Feng *et al.*, 2021] Dong Feng, ZhongCheng Wu, Jun Zhang, and Tingting Ren. Multi-scale spatial temporal graph neural network for skeleton-based action recognition. *IEEE Access*, 9:58256–58265, 2021.

[Hao *et al.*, 2022a] Yanbin Hao, Shuo Wang, Pei Cao, Xinjian Gao, Tong Xu, Jinmeng Wu, and Xiangnan He. Attention in attention: Modeling context correlation for efficient video classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7120–7132, 2022.

[Hao *et al.*, 2022b] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. Group contextualization for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 928–938, 2022.

[Lee *et al.*, 2017] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1012–1020, 2017.

[Li *et al.*, 2019] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019.

[Liu *et al.*, 2016] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016.

[Liu *et al.*, 2017] Zhenguang Liu, Zepeng Wang, Luming Zhang, Rajiv Ratn Shah, Yingjie Xia, Yi Yang, and Xuelong Li. Fastshrinkage: Perceptually-aware retargeting toward mobile platforms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 501–509, 2017.

[Liu *et al.*, 2019a] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

[Liu *et al.*, 2019b] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10004–10012, 2019.

[Liu *et al.*, 2020] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.

[Liu *et al.*, 2021] Zhenguang Liu, Kedi Lyu, Shuang Wu, Haipeng Chen, Yanbin Hao, and Shouling Ji. Aggregated multi-gans for controlled 3d human motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2225–2232, 2021.

[Liu *et al.*, 2022] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang

Wang. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11016, 2022.

[Lu *et al.*, 2018] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1466–1474. IEEE, 2018.

[Shahroudy *et al.*, 2016] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[Shi *et al.*, 2019a] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019.

[Shi *et al.*, 2019b] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

[Si *et al.*, 2018] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.

[Si *et al.*, 2019] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. 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*, pages 1227–1236, 2019.

[Singh *et al.*, 2018] Amarjot Singh, Devendra Patil, and SN Omkar. Eye in the sky: Real-time drone surveillance system (dss) for violent individuals identification using scatternet hybrid deep learning network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1629–1637, 2018.

[Song *et al.*, 2022] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Tan *et al.*, 2021] Yi Tan, Yanbin Hao, Xiangnan He, Yinwei Wei, and Xun Yang. Selective dependency aggregation for action classification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 592–601, 2021.

[Tian *et al.*, 2021] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell

to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1522–1531, 2021.

[Veeriah *et al.*, 2015] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 4041–4049, 2015.

[Wang *et al.*, 2014] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2649–2656, 2014.

[Wang *et al.*, 2016] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[Wang *et al.*, 2019] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1227–1236, 2019.

[Xu *et al.*, 2022] Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2866–2874, 2022.

[Yan *et al.*, 2018] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[Ye *et al.*, 2020] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 55–63, 2020.

[Yin *et al.*, 2019] Yifang Yin, Zhenguang Liu, Ying Zhang, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. Gps2vec: Towards generating worldwide gps embeddings. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 416–419, 2019.

[Zhang *et al.*, 2020] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1112–1121, 2020.