applied sciences

MDPI

Article

Exploring High-Order Skeleton Correlations with Physical and Non-Physical Connection for Action Recognition

Cheng Wang [1], Nan Ma [2,3,*] and Zhixuan Wu [4]

[1] Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China; 15056271781@163.com
[2] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[3] Engineering Research Center of Intelligence Perception and Autonomous Control, Ministry of Education, Beijing University of Technology, Beijing 100124, China
[4] School of computer science, Beijing University of Posts and Telecommunications, Beijing 100876, China; zhixuanwuyslya@163.com
* Correspondence: manan123@bjut.edu.cn; Tel.: +86-139-11155086

Abstract: Hypergraphs have received widespread attention in modeling complex data correlations due to their superior performance. In recent years, some researchers have used hypergraph structures to characterize complex non-pairwise joints in the human skeleton and model higher-order correlations of the human skeleton. However, traditional methods of constructing hypergraphs based on physical connections ignore the dependencies among non-physically connected joints or bones, and it is difficult to model the correlation among joints or bones that are highly correlated in human action but are physically connected at long distances. To address these issues, we propose a skeleton-based action recognition method for hypergraph learning based on skeleton correlation, which explores the effects of physically and non-physically connected skeleton information on accurate action recognition. Specifically, in this paper, spatio-temporal correlation modeling is performed on the natural connections inherent in humans (physical connections) and the joints or bones that are more dependent but not directly connected (non-physical connection) during human actions. In order to better learn the hypergraph structure, we construct a spatio-temporal hypergraph neural network to extract the higher-order correlations of the human skeleton. In addition, we use an attentional mechanism to compute the attentional weights among different hypergraph features, and adaptively fuse the rich feature information in different hypergraphs. Extensive experiments are conducted on two datasets, NTU-RGB+D 60 and Kinetics-Skeleton, and the results show that compared with the state-of-the-art skeleton-based methods, our proposed method can achieve an optimal level of performance with significant advantages, providing a more accurate environmental perception and action analysis for the development of embodied intelligence.

Keywords: action recognition based on skeleton; multi-channel features; spatio-temporal hypergraph neural network; cross-channel attention mechanism; high-order semantic correlation; adaptive fusion

Citation: Wang, C.; Ma, N.; Wu, Z. Exploring High-Order Skeleton Correlations with Physical and Non-Physical Connection for Action Recognition. *Appl. Sci.* **2024**, *14*, 3832. https://doi.org/10.3390/app14093832

Academic Editor: João M. F. Rodrigues

Received: 12 December 2023
Revised: 3 February 2024
Accepted: 3 February 2024
Published: 30 April 2024

1. Introduction

Action recognition [1], a pivotal area within the realm of computer vision, has been extensively applied across various domains, including human–computer interaction [2], virtual reality [3], video surveillance [4], and the field of autonomous driving [5]. Action recognition refers to analyzing and classifying human motion in specific environments. One important method is action recognition based on skeleton, where the skeleton sequence represents compact information on human motion as high-level information, providing an effective and robust representation form for describing human action. With the continuous development of 3D depth sensor technology, researchers can easily obtain the three-dimensional coordinates of human key joints, and thus obtain human

skeleton data. Therefore, in recent years, more attention has been paid to some methods based on skeleton in the field of action recognition [6,7].

In real-life scenarios, apart from pairwise relationships, there also exist multiple types of data that contain numerous non-pairwise relationships. However, these intricate relationships cannot be effectively represented using graph structures. Additionally, due to the need for coordination among multiple joints in human actions, these joints may have significant distances in the graph structure of the human skeleton. The graph neural network methods only focus on the local physical connections between joints, while overlooking the non-physical dependency relationships among them. Furthermore, each action has varying degrees of impact on these joints, and the degree of influence of each joint should be captured during the network training process. Therefore, one of the main challenges in skeleton-based action recognition is how to utilize the graph structure of the skeleton data while fully considering the impact of different joints on human action. As a result, researchers have introduced the concept of hypergraphs [8] to represent complex non-pairwise relationships. Hypergraphs can effectively represent higher-order relationships in the data. Feature learning is performed using hypergraph neural networks [9] to obtain the feature representations of complex data. Compared to traditional graph neural networks, hypergraph neural networks are a more versatile representation learning framework that can effectively handle complex high-order correlations through the use of hypergraph structures, thus enabling the efficient processing of diverse types of complex data. Consequently, the utilization of hypergraph neural networks for achieving action recognition in complex environments has become a widely studied problem in recent years. An increasing number of researchers are attempting to model human skeletons using hypergraphs and employ hypergraph neural networks to obtain the feature representations of human actions, thereby facilitating more efficient human action recognition.

To address the above issues, this paper proposes a framework to explore high-order skeleton correlations with physical and non-physical connections for action recognition. Specifically, the contributions of this paper are as follows:

- Firstly, this paper proposes a spatio-temporal hypergraph modeling method of human skeleton correlation. This method focuses on the inherent physical connections and non-physically connected relationships among human skeletons. For different actions, different weights are assigned to different vertex in the hypergraph, which highlights the regions that have a significant impact on human action.
- Secondly, this paper proposes an adaptive multi-channel spatio-temporal hypergraph neural network (AMC-STHGNN). The network captures the high-order correlation among human skeletons, and the features of different data channels are adaptively fused, making full use of the complementarity and diversity among three types of features.
- Finally, our proposed method is tested on two public datasets and shows a superior performance, effectively enhancing the ability of environmental perception and action analysis.

The rest of this paper is organized as follows. We first introduce related work in Section 2, including action recognition based on skeleton and hypergraph neural networks. Then, in Section 3, we review the basic theory of hypergraphs and hypergraph neural networks. Section 4 provides detailed implementation details of the proposed method. The experimental results and discussions of the proposed method are presented in Section 5. Finally, a summary of the entire paper is given in Section 6.

2. Related Work

2.1. Action Recognition Based on Skeleton

Due to the rich spatial and temporal information present in deep skeleton sequences, many researchers have utilized skeleton data as input for neural networks in human action recognition tasks. Xu et al. [10] proposed an architecture called topology-aware CNN for

skeleton-based action recognition. This architecture enhances performance through a novel cross-channel feature enrichment module and SkeletonMix strategy. Ref. [11] introduced a design method based on temporal convolutional neural networks (TCNs) to analyze and recognize human activities. The emergence of graph convolutional neural networks (GCNs) has significantly improved the ability to handle non-Euclidean data. Chen et al. [12] proposed the CTR-GCN method, which dynamically learns joint information in the spatial domain to achieve channel topology modeling. Song et al. [13] presented an effective GCN baseline method with high accuracy and fewer trainable parameters, employing a spatial-temporal separation learning strategy for model configuration. Although significant progress has been made in skeleton-based action recognition methods, traditional methods based on skeleton information cannot meet practical requirements for extracting action features in complex environments.

2.2. Hypergraph Neural Networks

As a pioneering work in hypergraph neural networks, Feng et al. [14] introduced a hypergraph neural network framework (HGNN) that leverages the hypergraph structures for feature learning and effectively extracts high-order correlations in the data. HGNN extends spectral-domain convolutional operations from graph learning to hypergraph learning, performing convolution in the spectral domain using the hypergraph Laplacian operator. Jiang et al. [15] proposed a dynamic hypergraph neural network framework (DHGNN) to address the limitation of fixed hypergraph structures in hypergraph neural networks, which hinders the representation capacity for varying data features. Bai et al. [16] introduced two end-to-end operators, which can be inserted into most graph neural networks when non-pairwise relationships exist in the data for model training. Gao et al. [17] proposed a hypergraph neural network framework called HGNN+, which can be extended to directed hypergraphs. Due to the powerful representation capability of complex relationships, hypergraphs have been widely applied in the field of computer vision [18,19]. It is worth noting that research on action recognition based on hypergraph neural networks [20–22] has achieved certain progress. However, further optimization is needed to improve the ability to extract advanced features in the process of human motion, in order to enhance the accuracy of action recognition.

3. Hypergraph Theory

A hypergraph is characterized by the triple $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}}, \mathbf{W})$, where $\boldsymbol{\mathcal{V}}$ represents the set of vertices, the vertex denoted by $v \in \boldsymbol{\mathcal{V}}$; $\boldsymbol{\mathcal{E}}$ is the set of hyperedges, the hyperedge denoted by $e \in \boldsymbol{\mathcal{E}}$; and $\mathbf{W}$ is the matrix of hyperedge weights, indicating the weight associated with each hyperedge, represented as $\omega(e)$. The incidence between hyperedges and vertices is captured through an incidence matrix $\mathbf{H}$, which has dimensions $|\boldsymbol{\mathcal{V}}| \times |\boldsymbol{\mathcal{E}}|$. In this matrix, if a vertex v is part of a hyperedge e, the corresponding element $h(v, e)$ is set to 1; if not, $h(v, e)$ is set to 0. The elements of the incidence matrix $\mathbf{H}$ are defined such that:

$$h(v, e) = \begin{cases} 1, & v \in e \\ 0, & v \notin e \end{cases}. \tag{1}$$

In addition, $h(v, e)$ can also represent the possibility of assigning vertex v to hyperedge e or the importance of vertex v with respect to hyperedge e, using a range of [0,1] for representation. The relationship between the hypergraph and the incidence matrix is illustrated in Figure 1.

The hyperedge degree represents the number of vertices associated with a hyperedge e, which can be defined as

$$\delta(e) = \sum_{v \in \boldsymbol{\mathcal{V}}} h(v, e) \tag{2}$$

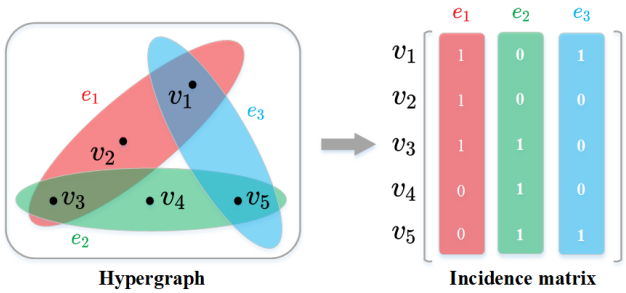


Figure 1. Hypergraph and incidence matrix.

The vertex degree represents the sum of the hyperedge weights associated with a vertex v, which can be defined as

$$d(v) = \sum_{e \in \mathcal{E}} w(e) \times h(v, e) \tag{3}$$

In addition, $\mathbf{D}_e$ and $\mathbf{D}_v$ are defined to denote the diagonal matrices of hyperedge degree and vertex degree, respectively. In [18], the features of the 3D object are used to construct the hypergraph by the k-NN method, where each vertex is connected to its k nearest neighbors. In [23], the l-1 sparse representation of the vertex is learned to construct the hyperedges.

After constructing the hypergraph, hypergraph learning is performed using a deep learning method [24], which is known as a hypergraph neural network. Specifically, according to [25], the Laplacian matrix of the hypergraph is defined as

$$\Delta = \mathbf{I} - \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{D}_v^{-1/2} \tag{4}$$

where $\mathbf{I}$ is the unit matrix. The hypergraph convolution on the spectral domain is parametrized as

$$\mathbf{X}^{t+1} = \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{D}_v^{-1/2}\mathbf{X}^t\mathbf{\Theta} \tag{5}$$

where $\mathbf{\Theta} \in \mathbb{R}^{d_{in} \times d_{out}}$ is the trainable parameter in the hypergraph convolution layer. In [17], inspired by hyperpaths in hypergraphs, the spatial hypergraph-based convolution HGNNConv+ can be written in matrix format as

$$\mathbf{X}^{t+1} = \mathbf{D}_v^{-1}\mathbf{H}\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{X}^t\mathbf{\Theta} \tag{6}$$

In the process of hypergraph convolution, vertex features are transformed into hyperedge features, which are then converted into new vertex features as shown in Figure 2. Hypergraph neural networks include spectral domain-based methods [14], where convolution uses a pre-defined hypergraph Laplacian matrix for vertex-to-vertex feature smoothing, and space-based methods [17], where a two-stage message passing method (vertex-edge-vertex) is used, which is more flexible and can be easily extended to more types of higher-order structures, such as directed hypergraphs.

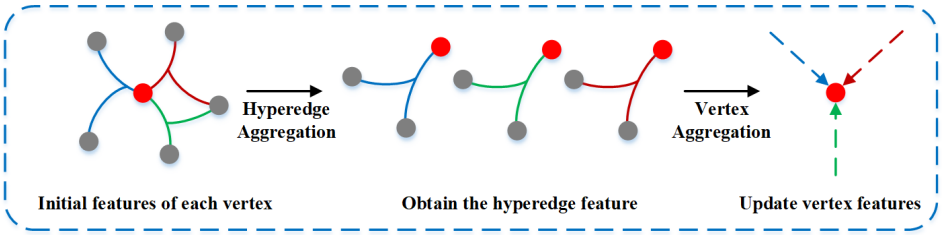


Figure 2. Hypergraph and incidence matrix.

4. The Proposed Method

4.1. Pipeline Overview

In order to explore high-order skeleton correlations with physical and non-physical connections for action recognition, this paper proposes a framework, which is shown in Figure 3. Firstly, this paper uses both human joint and bone to construct a physically connected joint (PCJ) spatio-temporal hypergraph, a non-physically connected joint (N-PCJ) spatio-temporal hypergraph, and a non-physically connected bone (N-PCB) spatio-temporal hypergraph as model inputs to generate three different data channels. Notably, we not only use the inherent physical connectivity associations between joints as a priori knowledge to construct hyperedges, but also use a clustering algorithm to obtain the non-physical connectivity associations between joints to construct hyperedges, where the non-physical connectivity hyperedges are constructed to differentiate the importance of different joints or bones in action recognition and assign different weights to different joints or bones. Then, this paper proposes a network called AMC-STHGNN. In this network, the spatio-temporal information-based hypergraph convolution module is used to more accurately capture the higher-order spatio-temporal feature information during human action; the multi-feature fusion module based on cross-channels attention mechanism makes full use of the complementarity and diversity among different data channels, and obtains the final prediction results by adaptively fusing the rich feature information in the multi-channel data.

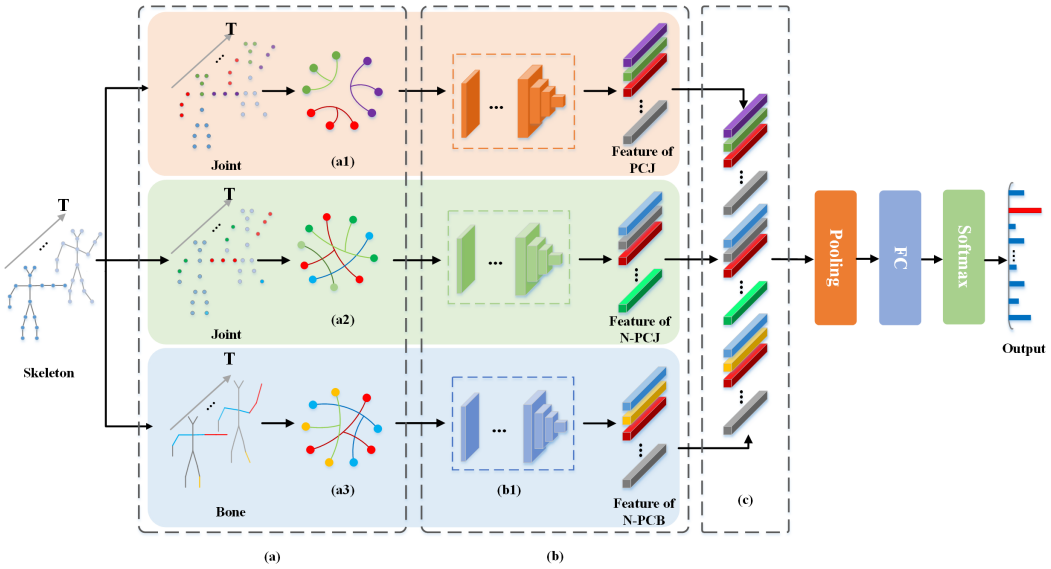


Figure 3. The pipeline of the proposed method: (**a**) spatio-temporal hypergraph modeling of human skeleton correlation; (**b**) adaptive multi-channel spatio-temporal hypergraph neural network; (**a1**) physically connected joint spatio-temporal hypergraph; (**a2**) non-physically connected joint spatio-temporal hypergraph; (**a3**) non-physically connected bone spatio-temporal hypergraph; (**b1**) hypergraph convolution module based on spatio-temporal information; and (**c**) multi-feature fusion module based on cross-channels attention mechanism.

4.2. Spatio-Temporal Hypergraph Modeling of Human Skeleton Correlation

In this paper, a video sequence is used as the model input, which is cropped into T frames containing information about the human skeleton, each frame is noted as $\{I_t\}_{t=1,2,\cdots,T}$. Feature extraction is performed for each frame according to the method in [26], with the initial feature vector of the i-th joint $v_t^{(i)}$ in the t-th frame image noted as $\boldsymbol{v}_i^t \in \mathbb{R}^{C_1}$ and the initial feature vector of the j-th bone $b_t^{(j)}$ in the t-th frame image noted as $\boldsymbol{b}_j^t \in \mathbb{R}^{C_2}$, where C_1 denotes the dimension of the feature vector $\boldsymbol{v}_i^t$ and C_2 denotes the dimension

of the feature vector b_j^t, and the hypergraphs are constructed separately using the joints information and the bones information according to different strategies.

4.2.1. Physically Connected Joint Spatio-Temporal Hypergraph

For the physically connected joint (PCJ) spatio-temporal hypergraph, the inherent topology of the human joints is used as a basis to divide them into different regions, including the head, trunk, left hand (*Lhand*), right hand (*Rhand*), left leg (*Lleg*), and right leg (*Rleg*) in six regions $\mathcal{V}_{area}$, $area = \{head, trunk, Lhand, Rhand, Lleg, Rleg\}$. $\mathcal{V}_{area}$ represents the set of all joints belonging to a certain area, and the joints of the human in consecutive frames of the same area are connected by a hyperedge to construct a spatio-temporal hypergraph as a fixed information representation of the human body structure, as shown in Figure 4a.

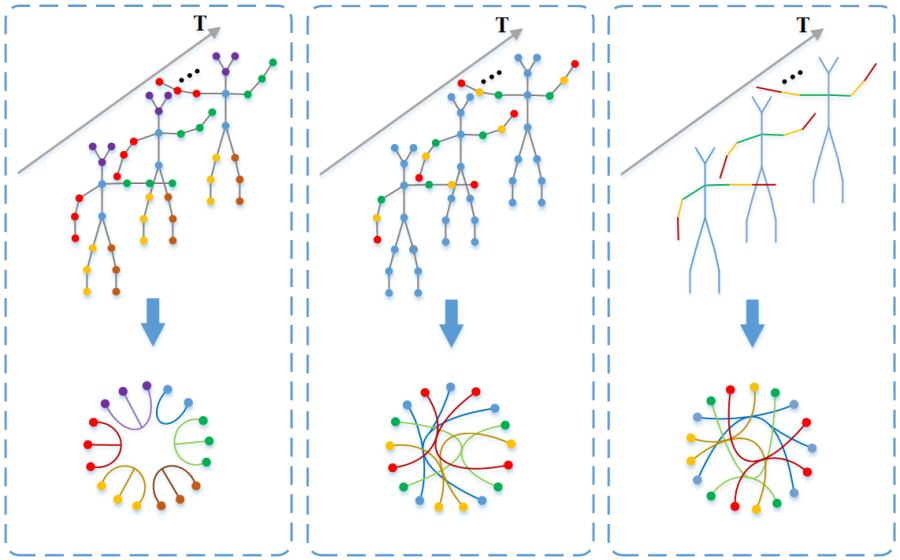


Figure 4. Visualization of different hypergraph constructions: (**a**) the construction of a physically connected joint spatio-temporal hypergraph; (**b**) the construction of a non-physically connected joint spatio-temporal hypergraph; (**c**) the construction of a non-physically connected bone spatio-temporal hypergraph.

The physically connected joint spatio-temporal hypergraph is denoted as $\mathcal{G}_{PCJ} = (\mathcal{V}_{PCJ}, \mathcal{E}_{PCJ}, \mathbf{W}_{PCJ})$, where $\mathcal{V}_{PCJ}$ denotes the set of vertices of this hypergraph, $\mathcal{E}_{PCJ}$ denotes the set of hyperedges of this hypergraph, and $\mathbf{W}_{PCJ}$ denotes the diagonal matrix of hyperedge weights in this hypergraph. Specifically, $\mathcal{V}_{PCJ}$ is the set of all joints $v_t^{(i)}$ (the i-th joint in the image of frame t) in all frames; the elements in $\mathcal{E}_{PCJ}$ are noted as $e_{area} = \{\forall v_t^{(i)} | v_t^{(i)} \in \mathcal{V}_{area}\}$, $\mathcal{E}_{PCJ} = \{e_{area} | area = head, trunk, Lhand, Rhand, Lleg, Rleg\}$. The incidence matrix $\mathbf{H}_{PCJ}$ of the spatio-temporal hypergraph of physically connected joints is constructed based on the association between $\mathcal{V}_{PCJ}$ and $\mathcal{E}_{PCJ}$, and the hyperedge degree $\delta(e_{area})$ and vertex degree $d(v_t^{(i)})$ are calculated using Equations (2) and (3), following which the diagonal matrices $\mathbf{D}_{e_{area}}$ and $\mathbf{D}_{v_t^{(i)}}$ of the hyperedge degree and vertex degree are obtained. In order to optimize the network using higher-order information, the incidence matrix $\mathbf{H}_{PCJ}$ is transformed by Laplace to generate the Laplace matrix $\mathbf{G}_{PCJ}$, which is calculated as

$$\mathbf{G}_{PCJ} = \mathbf{D}_{v_t^{(i)}}^{-1/2} \mathbf{H}_{PCJ} \mathbf{W}_{PCJ} \mathbf{D}_{e_{area}}^{-1} (\mathbf{H}_{PCJ}^{\top}) \mathbf{D}_{v_t^{(i)}}^{-1/2} \tag{7}$$

4.2.2. Non-Physically Connected Joint Spatio-Temporal Hypergraph

For example, in a traffic police gesture, the arm of the command 'go straight' changes more, while the torso and legs change less, as shown in Figure 5. Therefore, we propose to add an attention mechanism to obtain the key areas of the action, so as to highlight the influence of the key areas on human action recognition. This paper proposes an attention-based representation of non-physically connected skeletal information, aiming to capture the key regions that have a greater impact on action recognition, and thus increase the efficiency of action recognition.

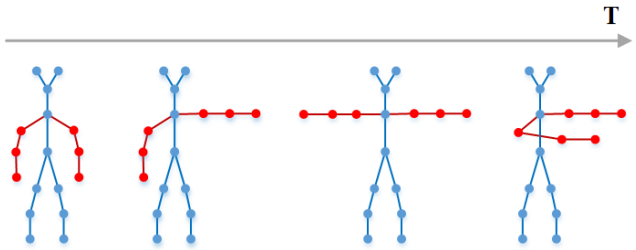


Figure 5. Illustration of the skeletal changes in traffic police gesture of 'go straight'. Red changes more, which indicates the key area, and blue changes less.

For non-physical connections, this paper introduces the concept of magnitude of change values to calculate the magnitude of change that occurs in the temporal sequence for each joint. Each joint has a magnitude of change value, which indicates the degree of influence that a particular joint has on human action recognition. Specifically, a joint is taken and its magnitude of change between the first and second frames is calculated using two-parametric numbers, where the joint is represented using xy-coordinates, such as (x_0, y_0), representing the two-dimensional coordinates of the joint. The magnitude of change of the joint between two consecutive frames is calculated as

$$d(\boldsymbol{v}_i^t, \boldsymbol{v}_i^{t+1}) = \sqrt{(\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t)^\top (\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t)} \tag{8}$$

where $\boldsymbol{v}_i^t$ denotes the coordinates of the i-th joint at time t and $\boldsymbol{v}_i^{t+1}$ denotes the coordinates of the i-th joint at time $t+1$. The value of the magnitude of change of the joint v_i over the whole time series is

$$d_{sum}^{v_i} = \sum_{t=1}^{T-1} d(\boldsymbol{v}_i^t, \boldsymbol{v}_i^{t+1}) \tag{9}$$

where T denotes the number of frames of the video data segmentation. After obtaining the change amplitude value of each joint over the whole time series, it is normalized and the normalization process of the change amplitude value of the joint over the whole time series is as follows:

$$d(v_i) = \frac{d_{sum}^{v_i}}{\sum_{i=1}^{I} d_{sum}^{v_i}} \tag{10}$$

where I indicates the number of human joints in a frame.

The association is modeled for joints that are more dependent but not directly physically connected in the human action, using a clustering algorithm that clusters joints according to the value of the magnitude of change of each joint. Joints of the same class in consecutive frames are connected by a hyperedge, and the joints in the same hyperedge have a similar degree of influence on the action. In general, for a given action, there is a dependency among the joints connected by the hyperedge, and this is used to construct a spatio-temporal hypergraph representing the association information among the non-physically connected joints in the different actions of the human body. Assuming that the clustering algorithm divides the joints into a total of Q classes, denoted by the set $\{\boldsymbol{\mathcal{V}}_q\}_{q=1,2,\cdots,Q}$, the spatio-temporal hypergraph of the non-physically connected joint (N-PCJ) is denoted by $\boldsymbol{\mathcal{G}}_{N-PCJ} = (\boldsymbol{\mathcal{V}}_{N-PCJ}, \boldsymbol{\mathcal{E}}_{N-PCJ}, \mathbf{W}_{N-PCJ})$, where $\boldsymbol{\mathcal{V}}_{N-PCJ}$ denotes the set of vertices of this

hypergraph, $\boldsymbol{\mathcal{E}}_{N-PCJ}$ denotes the set of hyperedges of this hypergraph, and $\mathbf{W}_{N-PCJ}$ denotes the diagonal matrix of hyperedge weights in this hypergraph. Specifically, $\boldsymbol{\mathcal{V}}_{N-PCJ}$ is the set of all joints $v_t^{(i)}$ in all frames; the elements in $\boldsymbol{\mathcal{E}}_{N-PCJ}$ are denoted by $e_q = \{v_t^{(i)} | v_t^{(i)} \in \boldsymbol{\mathcal{V}}_q\}$ and $\boldsymbol{\mathcal{E}}_{N-PCJ} = \{e_q | q = 1, 2, \cdots, Q\}$. According to the steps in Section 4.2.1, the incidence matrix $\mathbf{H}_{N-PCJ}$ of the non-physically connected joint spatio-temporal hypergraph is obtained. After calculating the hyperedges and vertex degrees of this hypergraph, the Laplace matrix $\mathbf{G}_{N-PCJ}$ is generated by the Laplace transform. The process of constructing the spatio-temporal hypergraph of non-physically connected joint is shown in Figure 4b.

It is worth noting that, in the constructed non-physically connected joint hypergraph, the incidence matrix of the hypergraph is not a simple 0, 1 matrix; the value of each element in the matrix corresponds to the value of the magnitude of change (normalized value) of each joint or bone, and the elements in the incidence matrix $\mathbf{H}_{N-PCJ}$ of the non-physically connected joint spatio-temporal hypergraph are represented as follows:

$$h(v_t^{(i)}, e_q) = \begin{cases} d(v_i), & v_t^{(i)} \in e_q \\ 0, & v_t^{(i)} \notin e_q \end{cases}. \tag{11}$$

The normalized magnitude of change is used as the weight of a hypergraph vertex to represent the degree of influence of an articulation point or bone on human movement; the sum of the weights of all the vertices in the hypergraph is used as the weight of the hypergraph edge to differentiate the influence of different parts on action recognition.

4.2.3. Non-Physically Connected Bone Spatio-Temporal Hypergraph

For the non-physically connected bone (N-PCB) spatio-temporal hypergraph, the construction process is similar to that for the non-physically connected joint spatio-temporal hypergraph, as shown in Figure 4c. In our study, the bones of the human body are used as the vertices of the hypergraph, and bones with similar values of change in successive frames are connected by a single hyperedge, which characterizes both the spatial and temporal information of the human bones.

In our method, bones are represented using a vector representation, pointing from the source joint to the target joint, such as $(x_1, y_1) \rightarrow (x_2, y_2)$, and the magnitude of change of the bones between two consecutive frames is calculated as

$$d(\boldsymbol{b}_j^t, \boldsymbol{b}_j^{t+1}) = \sqrt{(\boldsymbol{b}_j^{t+1} - \boldsymbol{b}_j^t)^\top (\boldsymbol{b}_j^{t+1} - \boldsymbol{b}_j^t)} \tag{12}$$

where $\boldsymbol{b}_j^t$ denotes the coordinate representation of the j-th bone at moment t and $\boldsymbol{b}_j^{t+1}$ denotes the coordinate representation of the j-th bone at moment $t+1$. The value of the magnitude of change of the bone b_j over the whole time series is

$$d_{sum}^{b_j} = \sum_{t=1}^{T-1} d(\boldsymbol{b}_j^t, \boldsymbol{b}_j^{t+1}) \tag{13}$$

The process of normalizing the magnitude of change values of bones over the entire time series is as follows:

$$d(b_j) = \frac{d_{sum}^{b_j}}{\sum_{j=1}^{J} d_{sum}^{b_j}} \tag{14}$$

where J denotes the number of human bones in a frame.

The non-physically connected bone spatio-temporal hypergraph is denoted by $\boldsymbol{\mathcal{G}}_{N-PCB} = (\boldsymbol{\mathcal{V}}_{N-PCB}, \boldsymbol{\mathcal{E}}_{N-PCB}, \mathbf{W}_{N-PCB})$, where $\boldsymbol{\mathcal{V}}_{N-PCB}$ denotes the set of vertices of this hypergraph, $\boldsymbol{\mathcal{E}}_{N-PCB}$ denotes the set of hyperedges of this hypergraph, and $\mathbf{W}_{N-PCB}$ denotes the diagonal matrix of hyperedge weights in this hypergraph. Similarly to the process of constructing a spatio-temporal hypergraph of non-physically connected joint,

the incidence matrix $\mathbf{H}_{N-PCB}$ of this hypergraph is constructed and the Laplacian matrix $\mathbf{G}_{N-PCB}$ of this hypergraph is calculated.

The above process of constructing a hypergraph correlates the temporal and spatial relationships of the target objects. For spatial information, the hyperedge connects joints or bones with similar values of change in the same frame, characterizing the dependency among the joints or bones in the action; for temporal information, the hyperedge connects the same joints or bones in different frames, characterizing the temporal information of the joints or bones in the change of the action. In addition, the joints and bones of the human action are also considered, with the joints as the basic information of the human action and the bones as the complementary information, the modeling process fully reflects the completeness of the data. For the action recognition in complex scenes, the use of both joint and bone can obtain more effective information from the human skeleton than a single joint, which is beneficial to improve the accuracy of human action recognition. In addition, the modeling process characterizes the association among physical connections and the dependency among non-physical connections in the human skeleton, taking into account both the fixed human topology and the dependency among different joint and bone in the changing human action, and more accurately solving the problem of the inaccurate recognition of human action in complex scenes.

4.3. Adaptive Multi-Channel Spatio-Temporal Hypergraph Neural Network

After constructing the hypergraph, this paper proposes an adaptive multi-channel spatio-temporal hypergraph neural network for hypergraph learning, which consists of a hypergraph convolution module based on spatio-temporal information and a multi-feature fusion module based on cross-channel attention mechanism. Specifically, the spatio-temporal information-based hypergraph convolution module is used to iteratively update the features of each vertex to fully learn the spatio-temporal associations and the higher-order semantics of human joints and bones during motion; the multi-feature fusion module based on cross-channel attention mechanism is used to adaptively fuse features from different channels to maximize the use of rich information in each channel and to weaken conflicts and interferences among different channels. The module can also be used to improve the accuracy and robustness of action recognition.

4.3.1. Hypergraph Convolution Module Based on Spatio-Temporal Information

The hypergraph convolution module based on spatio-temporal information consists of a spatial convolution block, a temporal convolution block, and a global pooling layer, with a spatial convolution block and a temporal convolution block representing one layer in series, for a total of L layers, as shown in Figure 6. Among them, the spatial convolutional block is used to aggregate the semantic features of the hypergraph in the spatial dimension, and the temporal convolutional block is used to learn the temporal information of the hypergraph in the temporal dimension. Finally, a global pooling layer is connected to output the updated feature vector.

In order to better learn the temporal information of the human skeleton in the process of movement, the incidence matrix of the constructed hypergraph is split according to the temporal dimension in this paper, and the incidence matrix of the spatio-temporal hypergraph is dimensionally transformed, and the incidence matrix of the spatio-temporal hypergraph of each channel after the transformation is expressed as

$$\tilde{\mathbf{H}}_{*} = reshape(\mathbf{H}_{*}) = \{\mathbf{H}_{*}^{(t)} | t = 1, 2, \cdots, T\} \tag{15}$$

where $\tilde{\mathbf{H}}_{*}$ can be viewed as the subincidence matrix of $\mathbf{H}_{*}$, and $*$ denotes the different strategies for constructing the hypergraph (including PCJ, N-PCJ, and N-PCB). The Laplacian matrix $\mathbf{G}_{*}^{(t)}$ of each subincidence matrix is calculated according to the transformation pro-

cess of Laplacian matrices in Section 4.2, and stitched according to the temporal dimension to obtain the 3D matrix $\tilde{\mathbf{G}}_*$, denoted by

$$\tilde{\mathbf{G}}_* = \{\mathbf{G}_*^{(t)} | t = 1, 2, \cdots, T\} \tag{16}$$

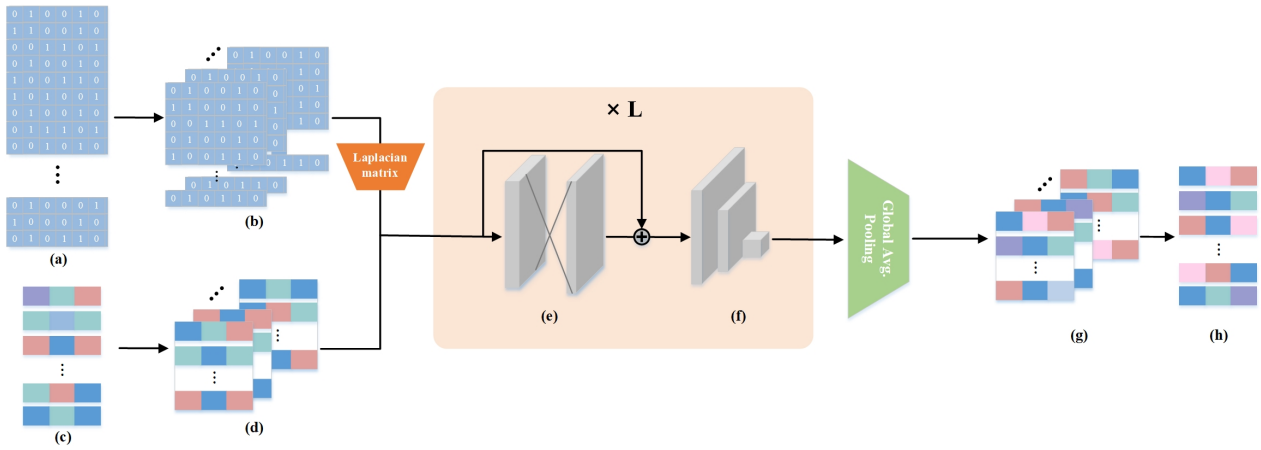


Figure 6. Hypergraph convolution module based on spatio-temporal information: (**a**) incidence matrix; (**b**) subincidence matrix; (**c**) initial features of hypergraph vertices; (**d**) initial features of hypergraph vertices after splitting by time dimension; (**e**) spatial convolution block; (**f**) temporal convolution block; (**g**) hypergraph vertex features after splitting by temporal dimension; and (**h**) trained hypergraph vertex features.

In addition, the initial joint features and bone features are also represented according to the temporal dimension. Specifically, the initial feature matrix of the joints in frame t is denoted by $\mathbf{V}_t = \{v_1^t, v_2^t, \cdots, v_I^t\}$, and the initial feature matrix of the joints in all frames is denoted by $\mathbf{V} = \{\mathbf{V}_t | t = 1, 2, \cdots, T\}$; the initial feature matrix of the bones in frame t is denoted by $\mathbf{B}_t = \{\boldsymbol{b}_1^t, \boldsymbol{b}_2^t, \cdots, \boldsymbol{b}_J^t\}$, and the initial feature matrix of the bones in all frames is denoted by $\mathbf{B} = \{\mathbf{B}_t | t = 1, 2, \cdots, T\}$.

Taking the physically connected joint spatio-temporal hypergraph as an example, $\tilde{\mathbf{G}}_{PCJ}$ and $\mathbf{V}$ are first input to the spatial convolution block for spatial feature learning, and the vertex features after Laplacian matrix learning are weighted and fused with the original vertex features to form the input feature matrix for the l-th temporal convolution block:

$$\tilde{\mathbf{V}}^{(l)} = \sigma(\mathbf{V}^{(l)} + \tilde{\mathbf{G}}_{PCJ} \odot \mathbf{V}^{(l)} \boldsymbol{\Theta}^{(l)}) \tag{17}$$

where σ is the activation function, $\odot$ denotes the element multiplication, $\mathbf{V}^{(l)}$ denotes the input features of the l-th spatial convolution block, and $\boldsymbol{\Theta}^{(l)}$ denotes the learnable parameters. $\tilde{\mathbf{V}}^{(l)}$ is then fed into the temporal convolution block for training, inspired by [27], to obtain the new vertex feature matrix as follows:

$$\mathbf{V}^{(l+1)} = tanh(\Gamma_1^{(l)} \ast \tilde{\mathbf{V}}^{(l)}) \odot \Phi(\Gamma_2^{(l)} \ast \tilde{\mathbf{V}}^{(l)}) \tag{18}$$

where $\ast$ is the convolution operation, $\Gamma_1^{(l)}$ and $\Gamma_2^{(l)}$ are two different learnable parameters in the l-th temporal convolution block, and $\Phi(\cdot)$ is the gating unit to control the utilization of historical information. After L iterations of updating, the vertex feature matrix $\mathbf{V}^{(L)}$ is obtained, and the final output feature matrix $\tilde{\mathbf{X}}_{PCJ}$ for this channel is obtained after a global pooling layer. Similarly, the output feature matrices $\tilde{\mathbf{X}}_{N-PCJ}$ and $\tilde{\mathbf{X}}_{N-PCB}$ for the other two channels are obtained.

4.3.2. Multi-Feature Fusion Module Based on Cross-Channel Attention Mechanism

After obtaining the spatio-temporal features of each channel, the features $\tilde{\mathbf{X}}_{PCJ}$, $\tilde{\mathbf{X}}_{N-PCJ}$ and $\tilde{\mathbf{X}}_{N-PCB}$ of the three channels are adaptively fused by a multi-feature fusion module

based on the cross-channel attention mechanism, which fully considers the complementarity among the data features of different channels. The structure of the module is shown in Figure 7.

The module takes the spatio-temporal features of the respective channels, the incidence matrix of the spatio-temporal hypergraph, and the corresponding Laplacian matrix as input, and learns the spatio-temporal features of each channel once on the entire spatio-temporal hypergraph separately, using the Laplacian matrix while supplementing the original information in the incidence matrix to make full use of the incidence matrix and Laplacian matrix of the hypergraph to update the spatio-temporal features of each channel. The updated data features of each channel are represented as follows:

$$\mathbf{F}_{PCJ} = (\mathbf{H}_{PCJ}\mathbf{H}_{PCJ}^{\top} + \beta_1 \mathbf{G}_{PCJ})\tilde{\mathbf{X}}_{PCJ} \tag{19}$$

$$\mathbf{F}_{N-PCJ} = (\mathbf{H}_{N-PCJ}\mathbf{H}_{N-PCJ}^{\top} + \beta_2 \mathbf{G}_{N-PCJ})\tilde{\mathbf{X}}_{N-PCJ} \tag{20}$$

$$\mathbf{F}_{N-PCB} = (\mathbf{H}_{N-PCB}\mathbf{H}_{N-PCB}^{\top} + \beta_3 \mathbf{G}_{N-PCB})\tilde{\mathbf{X}}_{N-PCB} \tag{21}$$

where $\mathbf{H}_{*}^{\top}$ denotes the transpose of the correlation matrix $\mathbf{H}_{*}$ and β_1, β_2, β_3 are learnable scalars.

Considering that different channel features have different degrees of influence on action recognition, in order to make full use of the correlation and complementarity among different channel features, this paper uses a cross-channel attention mechanism to calculate the attention score of each channel feature in human action, uses this score as the weight of each channel feature, and performs a weighted fusion of the features of each channel to adaptively obtain the final fused features. Specifically, first, the features $\mathbf{F}_{*}$ of each channel are compressed into a one-dimensional vector f_{*} for attention weighting using a one-dimensional convolutional layer, and the process is represented as

$$f_{*} = Conv1d(\mathbf{F}_{*}) \tag{22}$$

where $Conv1d$ denotes a one-dimensional convolutional layer. Next, the attention matrix between each of the two channel features is calculated separately, as follows:

$$\mathbf{A}_{PCJ,N-PCJ} = Softmax\left(\frac{(\mathbf{W}_{q_1} f_{PCJ})(\mathbf{W}_{k_1} f_{N-PCJ})^{\top}}{\sqrt{d}}\right) \tag{23}$$

$$\mathbf{A}_{N-PCJ,N-PCB} = Softmax\left(\frac{(\mathbf{W}_{q_2} f_{N-PCJ})(\mathbf{W}_{k_2} f_{N-PCB})^{\top}}{\sqrt{d}}\right) \tag{24}$$

$$\mathbf{A}_{PCJ,N-PCB} = Softmax\left(\frac{(\mathbf{W}_{q_3} f_{PCJ})(\mathbf{W}_{k_3} f_{N-PCB})^{\top}}{\sqrt{d}}\right) \tag{25}$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are trainable parameter matrices, respectively, and d denotes the dimensionality of f_{*}. Then, the attention score of each channel feature is calculated separately according to the attention matrix, and the calculation process is represented as follows:

$$\alpha_{PCJ} = \sigma(\mathbf{A}_{PCJ,N-PCJ}\mathbf{W}_{v_1} f_{N-PCJ} + \mathbf{A}_{PCJ,N-PCB}\mathbf{W}_{v_1} f_{N-PCB}) \tag{26}$$

$$\alpha_{N-PCJ} = \sigma(\mathbf{A}_{N-PCJ,N-PCB}\mathbf{W}_{v_2} f_{N-PCB} + \mathbf{A}_{PCJ,N-PCJ}\mathbf{W}_{v_2} f_{PCJ}) \tag{27}$$

$$\alpha_{N-PCB} = \sigma(\mathbf{A}_{N-PCJ,N-PCB}\mathbf{W}_{v_3} f_{N-PCJ} + \mathbf{A}_{PCJ,N-PCB}\mathbf{W}_{v_3} f_{PCJ}) \tag{28}$$

where $\mathbf{W}_v$ is the trainable parameter matrix and σ is the activation function. It is worth noting that the attention score is used to adjust the contribution of different channel features to action recognition, thereby highlighting the features that have a greater impact on human

action recognition. The attention score is then used to weight and sum the features of the three channels, and the adaptive fusion of the multi-channel features is represented as

$$\mathbf{F}_{fusion} = \alpha_{PCJ} \odot \mathbf{F}_{PCJ} + \alpha_{N-PCJ} \odot \mathbf{F}_{N-PCJ} + \alpha_{N-PCB} \odot \mathbf{F}_{N-PCB} \tag{29}$$

Finally, the fused features are globally pooled and passed through the fully connected layer before $Softmax$ is used to calculate the action prediction probability values, with the predicted category being the action category z with the highest probability value:

$$z = Softmax\Big(FCL(GAP(\mathbf{F}_{fusion}))\Big) \tag{30}$$

where GAP denotes the global average pooling operation and FCL denotes full connection layer.

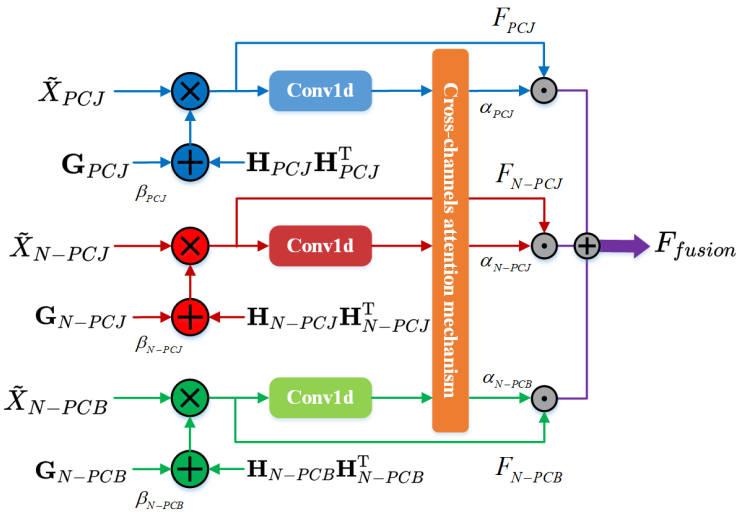


Figure 7. Multi-feature fusion module based on cross-channel attention mechanism.

The multi-feature fusion module based on cross-channel attention mechanism achieves the effective fusion of information features of each channel by learning the features of the three channels, reasonably using the diversity and complementarity of physically connected joints, non-physically connected joint, and non-physically connected bone in human action, weakening the interference of noise and effectively improving the accuracy of action recognition.

5. Experimental Results and Analysis

In this paper, we propose a framework for action recognition, which uses an adaptive multi-channel spatio-temporal hypergraph neural network (AM-STGNN) to learn hypergraphs based on human skeleton information to obtain higher-order semantic features of human actions for accurate action recognition. Then, we conduct experiments using two publicly available datasets, the NTU RGB+D 60 dataset [28] and the Kinetics-Skeleton dataset [29], respectively, and extensive experimental results show that our method achieves significant performance improvements in action recognition on both publicly available datasets. We will then present the datasets and relevant experimental details, compare the action recognition results of our proposed method with recent state-of-the-art methods, and perform ablation experiments and visualization analysis.

5.1. Dataset

NTU RGB+D 60: The dataset has 60 categories of actions, such as drinking, sitting down and standing up, with a total of 56,880 samples. These samples were performed by 40 volunteers and captured simultaneously by 3 Microsoft Kinect v2 cameras from different viewpoints. Each sample contains one action and is guaranteed to have a maximum of two objects. The dataset uses two benchmarks: (1) cross-subject (CS): training data from

20 volunteers and test data from 20 other volunteers; and (2) cross-view (CV): training data from camera views 2 and 3, test data from camera view 1.

Kinetics-skeleton: The dataset is a large-scale human action dataset containing 300,000 video clips in 400 categories. The video clips are derived from YouTube videos and are highly diverse. The dataset only provides raw video clips and does not include skeleton data. [30] used the publicly available OpenPose toolbox [31] to estimate the positions of 18 articulation points on each clip frame. Two individuals were selected for multiplayer clip cropping based on average joint point confidence. We used their published data to evaluate our model. The dataset was divided into a training set (240,000 fragments) and a validation set (20,000 fragments). The model was trained on the training set and top-1 and top-5 accuracies were reported on the validation set, following the evaluation method described in [30].

5.2. Experimental Settings

All experiments were performed on two 2080Ti GPUs with a batch size of 64, trained with the SGD optimization algorithm (momentum of 0.9), with weight decay set to 0.0001, using a cross-entropy loss function to back-propagate the gradient, and the hyperedge weight of the physically connected joint hypergraph set to 1 by default. In addition, considering that the number of bones is one less than the number of articulation points, an empty bone is added to make the number of bones consistent with the number of articulation points.

For the NTU RGB+D 60 dataset, the maximum number of frames in each sample is 300. For samples with less than 300 frames, we repeated these samples until 300 frames were reached. The learning rate was set to 0.1 and divided by 10 at rounds 30 and 40, with the training process ending at round 50. For the Kinetics-Skeleton dataset, the size of the Kinetics input tensor was the same as in [30], and 150 frames were randomly selected from the input skeleton sequence and processed by slightly perturbing the joint coordinates as well as rotation and translation operations. The learning rate was also set to 0.1 and divided by 10 at rounds 45 and 55. the training process ended at round 65.

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

In order to verify the effectiveness of the proposed method, this paper compares it with other state-of-the-art methods. For the NTU RGB+D 60 dataset, experiments are conducted in both CS (cross-subject) protocols and CV (cross-view) protocols, and the experimental results are shown in Table 1. The methods used for comparison include handcraft-feature-based methods [32], RNN-based methods [33–35], CNN-based methods [10,36–38], GCN-based methods [30,39–41], and the HGNN-based method [42–45]. The experimental results show that the accuracy of our proposed method achieves 91.2% and 96.7% for action recognition in CS (cross-subject) protocols and CV (cross-view) protocols, respectively, and the accuracy of human action recognition is significantly improved, even compared with other hypergraph neural network-based methods. For the Kinetics-Skeleton dataset, the experimental results are shown in Table 1. The methods used for comparison include handcraft-feature-based methods [46], RNN-based methods [28], CNN-based methods [36], GCN-based methods [30,39,40,47] and HGNN-based methods [42–45]. The experimental results show that the accuracy of our proposed method is 39.1% and 61.3% in Top-1 and Top-5, respectively. Compared with other methods based on skeleton, our model achieves the state-of-the-art. From the above experimental results, it is easy to see that our proposed method can achieve a better performance by constructing hypergraphs from skeleton information in human actions. Thus, capturing higher-order correlations among skeletons can extract richer information, which can further improve the performance of action recognition.

Table 1. Comparison of classification accuracy with state-of-the-art methods on the NTU RGB+D 60 dataset and the Kinetics-Skeleton dataset.

Type	Method	NTU RGB+D 60		Kinetics-Skeleton	
		CS(%)	CV(%)	TOP-1(%)	TOP-5(%)
Handcraft feature based	Lie Group [32]	50.1	82.5	-	-
	Feature Enc [46]	-	-	14.9	25.8
RNN based	ST-LSTM [33]	69.2	77.7	-	-
	VA-LSTM [34]	79.4	87.6	-	-
	AGC-LSTM [35]	89.2	95.0	-	-
	Deep LSTM [28]	-	-	16.4	35.3
CNN based	TCN [36]	74.3	83.1	20.3	40.0
	HCN [37]	86.5	91.1	-	-
	VA-CNN [38]	88.7	94.3	-	-
	Ta-CNN+ [10]	90.7	95.1	-	-
GCN based	ST-GCN [39]	81.5	88.3	30.7	52.8
	2S-AGCN [30]	88.5	95.1	36.1	58.7
	MS-AAGCN [40]	90.0	96.2	37.8	61.0
	Shift-GCN [41]	90.7	96.5	-	-
	Sym-GNN [47]	-	-	37.2	58.1
HGNN based	Hyper-GCN(3S) [42]	89.5	95.7	37.1	60.0
	DHGCN [43]	90.7	96.0	37.7	60.6
	Selective-HCN [44]	90.8	96.6	38.0	61.1
	SD-HGCN [45]	90.9	96.7	37.4	60.5
	Ours	91.2	96.7	39.1	61.3

5.4. Ablation Experiments

In order to verify the validity of the different modules in the proposed network model in this paper, without loss of generality, we conducted a series of ablation experiments on the CS (cross-subject) protocols and CV (cross-view) protocols of the NTU RGB+D 60 dataset. In this subsection, we will analyze whether the modules in the proposed model affect the performance of action recognition in order to verify the plausibility of the network model.

5.4.1. Different Channel Data Channels

The influence of data about joints and bones on final experimental results. On the NTU RGB+D 60 dataset, we compared using only the non-physically connected joint hypergraph as model input, using only the non-physically connected bone hypergraph as model input using both the non-physically connected joint hypergraph and the non-physically connected bone hypergraph as model input, and the experimental results are shown in Table 2. The results show that the action recognition accuracy of using both the non-physically connected joint hypergraph and the non-physically connected bone hypergraph as model input was 90.3% and 95.2% for the CS (cross-subject) protocols and CV (cross-view) protocols, respectively, which were better than that of only using the joint or bone. This shows that the use of multi-channel data containing both joints and bones can provide richer feature information than single-channel data only containing joints or bones, thus enhancing the perceptual capability of the model and achieving a complete representation of the data, thus improving the accuracy of action recognition.

The influence of different hypergraph construction methods on experimental results. In the NTU RGB+D 60 dataset, where we compared only using physically connected joint hypergraphs as model input, only using non-physically connected joint hypergraphs as model input, and using both non-physically connected joint hypergraphs and physically connected joint hypergraphs as model input. The experimental results for this are shown in

Table 2. The experimental results show that the action recognition accuracy of using both non-physically connected joint hypergraphs and physically connected joint hypergraphs as model inputs was 90.7% and 95.4% for CS (cross-subject) protocols and CV (cross-view) protocols, respectively, both of which were better than that of using only non-physically connected or physically connected joints. This shows that both the physical connection inherent in the human skeleton and the non-physical connection that are more relevant during movement can provide effective information for action recognition, and their simultaneous use can improve the learning ability of skeletal features in different movements.

Table 2. Comparison of action recognition accuracy of different hypergraph construction methods and different input data information.

Hypergraph Construction Method			Input Data		NTU RGB+D 60	
PCJ	N-PCJ	N-PCB	Joint	Bone	CS(%)	CV(%)
√			√		88.6	94.0
	√		√		88.7	93.8
		√		√	89.2	94.2
√	√		√		90.7	95.4
	√	√	√	√	90.3	95.2
√	√	√	√	√	91.7	96.7

5.4.2. Different Neural Network Structures

To verify the learning effect of the spatio-temporal hypergraph neural network, we conducted experiments on different hypergraph neural networks on CS (cross-subject) protocols and CV (cross-view) protocols of the NTU RGB+D 60 dataset, respectively. As shown in Table 3, for CS (cross-subject) protocols, the accuracy of action recognition was 81.2% when only using spatial convolutional blocks in the hypergraph neural network; 82.1% when using only temporal hypergraph convolutional blocks in the hypergraph neural network; and 91.2% when using both temporal and spatial hypergraph convolutional blocks. Compared to the former, the accuracy of action recognition was improved by about 10% and 9.1%, respectively. For CV (cross-view) protocols, the accuracy of action recognition was 86.3% when only using spatial convolutional blocks in the hypergraph neural network, 87.7% when only using temporal hypergraph convolutional blocks in the hypergraph neural network, and 96.7% when using both temporal and spatial hypergraph convolutional blocks, with an improvement of about 10.4% and 9%, respectively. This shows that the spatio-temporal hypergraph neural network can further improve the mining ability of temporal and spatial information in human actions, and achieve the purpose of improving the accuracy and model generalization ability.

Table 3. Comparison of the action recognition accuracy of different structural neural networks.

Method	NTU RGB+D 60	
	CS(%)	CV(%)
Only temporal convolution blocks	81.2	86.3
Only spatial convolution blocks	82.1	87.7
Temporal convolution blocks + spatial convolution blocks	91.2	96.7

5.4.3. Different Multi-Channel Fusion Modules

To verify the effect of multi-feature fusion module based on the cross-channel attention mechanism, we compared the module with the traditional natural splicing approach on the NTU RGB+D 60 dataset, and the experimental results are shown in Table 4. The experimental results show that, when a multi-feature fusion module based on a cross-channel attention mechanism is not used and the simple coequal fusion method is used to fuse multi-

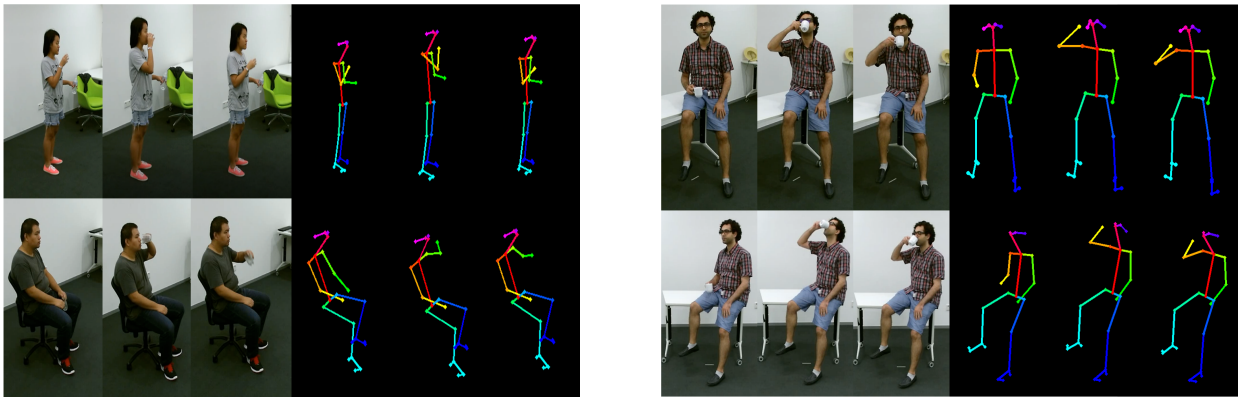
channel features, the accuracy of action recognition on CS (cross-subject) protocols and CV (cross-view) protocols is 88.4% and 93.2%, respectively; when the multi-feature fusion module based on cross-channel attention mechanism is used, the accuracy of action recognition on CS (cross-subject) protocols and CV (cross-view) protocols is 91.2% and 96.7%, respectively, an improvement of 2.8% and 3.5% compared to the former. This shows that although the introduction of higher-order correlations can improve the action recognition performance, the traditional coequal fusion method does not yield the best combination of multi-channel feature performance, suggesting that different higher-order correlations have different effects on representation learning. Therefore, this paper correlates each channel feature with an attention score, thus adaptively identifying the importance of different channel features to the overall action recognition and making better use of the complementary representations of multi-channel features, which helps improve the accuracy of action recognition.

Table 4. Comparison of action recognition accuracy of different ways of fusing multi-channel features.

Method	NTU RGB+D 60	
	CS(%)	CV(%)
Coequal fusion method	88.4	93.2
Multi-feature fusion module based on cross-channel attention mechanism	91.2	96.7

5.5. Visualization

In order to better illustrate the need for modeling non-physical connections, this section provides visualizations for some samples on the NTU RGB+D 60 dataset. From Figure 8, it can be seen that the action of 'drinking water' is mainly an upper limb movement, and the upper limb should be given more attention in the hypergraph. Meanwhile, it can also be observed that when drinking water, the arm and the head have a strong correlation, but they are not directly physically connected. In conclusion, it is shown that constructing a non-physically connected hypergraph can capture more complex and higher-order dependence among the joints and further improve the recognition accuracy of the model. In addition, the proposed method performs well in both CS (cross-subject) protocols and CV (cross-view) protocols, indicating that the proposed method has excellent robustness.



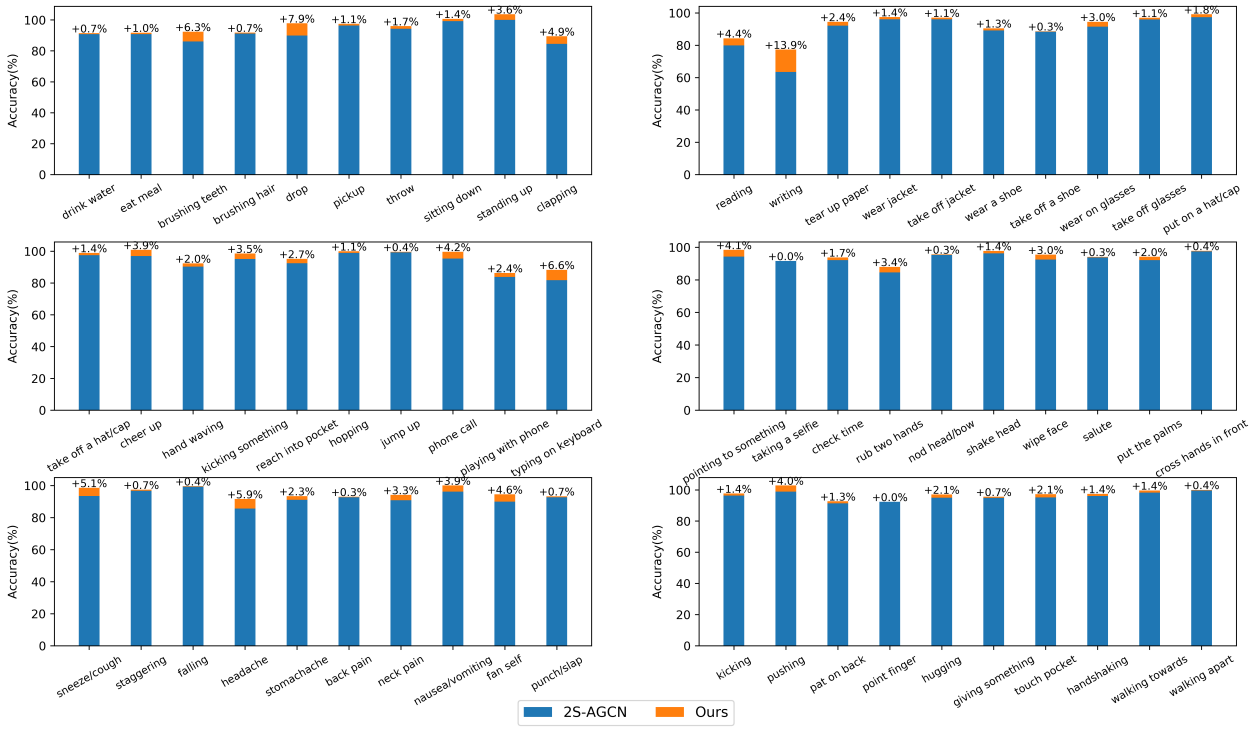
(**a**) (**b**)

Figure 8. Visual examples on the NTU RGB + D dataset: (**a**) drink water in CS (cross-subject) protocols; and (**b**) drink water in CV (cross-view) protocols.

The NTU RGB+D 60 dataset contains common human actions such as drinking water, eating a meal, etc., and we analyze the accuracy of the action recognition for each category on this dataset. In order to visualize the improvement in the accuracy on different actions

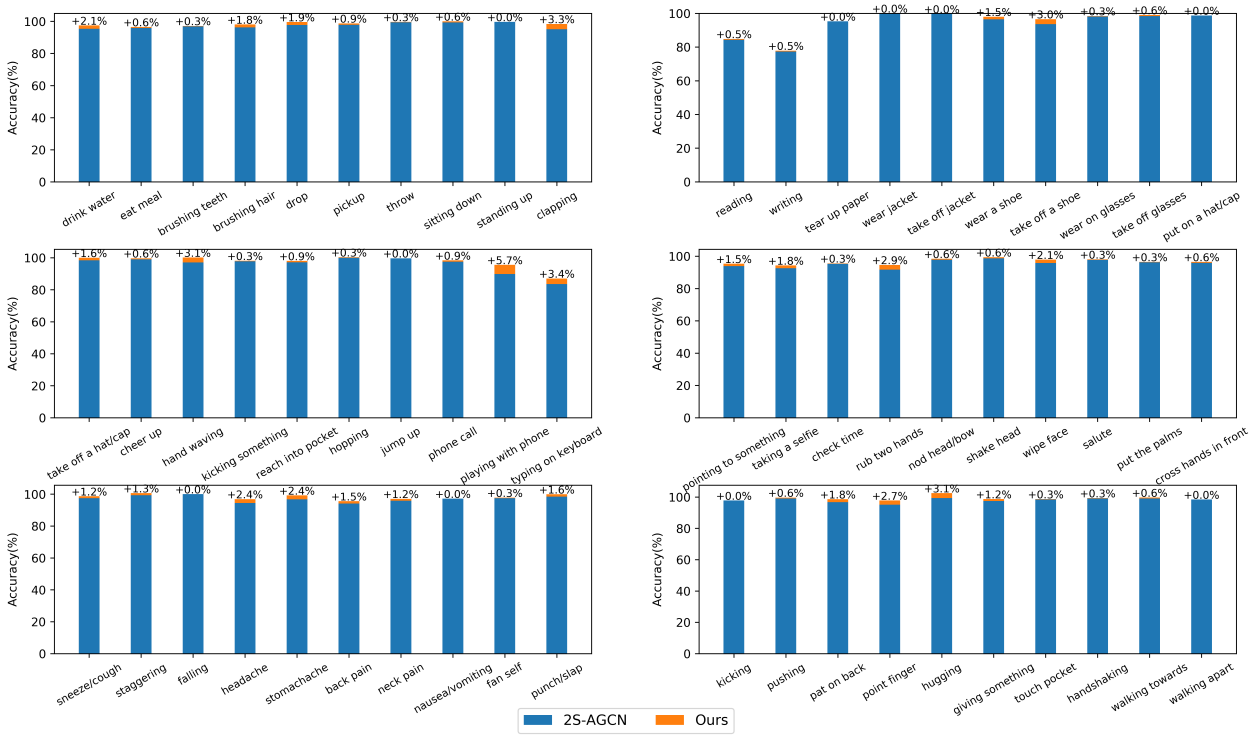
and to analyze the causes of misclassified actions, we plot bar charts based on the accuracy of 2S-AGCN and our proposed method on the NTU RGB+D 60 dataset for 60 action categories. As shown in Figure 9, the blue bars indicate the performance of 2S-AGCN, and the orange bars and the numbers on them indicate the performance improvement of our proposed method compared to using 2S-AGCN. Observation reveals that our proposed method has some performance improvement over 2S-AGCN for the vast majority of action categories. Figure 9a shows the CS (cross-subject) protocols of the NTU RGB+D 60 dataset. We can observe that the recognition accuracy of 'write' and 'read' is lower compared to other actions when using 2S-AGCN. For these two actions, the recognition accuracy is improved by about 13.9% and 4.4%, respectively, when using our proposed method. Figure 9b shows the CV (cross-view) protocols of the NTU RGB+D 60 dataset. We found that the recognition accuracy of some action, such as 'writing', 'reading', 'playing with the phone', and 'typing on keyboard' is lower when using 2S-AGCN. The accuracy of action recognition is improved in all cases after using our proposed method. From the above experimental results, it can be seen that, compared to 2S-AGC, there is a large performance improvement for actions such as 'writing', 'reading', etc., which only use part of the human body region. This is due to the fact that the proposed method can focus on regions with large variation and focus on the effective information of the action.

In addition, in order to comprehensively evaluate the model, it is necessary to consider not only the accuracy but also the convergence speed. To better demonstrate the comparative performance with other methods, we conducted accuracy tests on the NTU RGB+D 60 dataset using the CS (cross-subject) and CV (cross-view) protocols, and generated corresponding charts. As shown in Figure 10a,b, it can be observed that, as the number of epochs increases, all curves tend to reach their maximum values and stabilize. Our proposed method exhibits the fastest convergence speed, while Hyper-GNN, 2S-AGCN, AS-GCN, and ST-GCN show slower convergence rates. This indicates that our proposed network can achieve optimal results at the fastest speed compared to other methods.



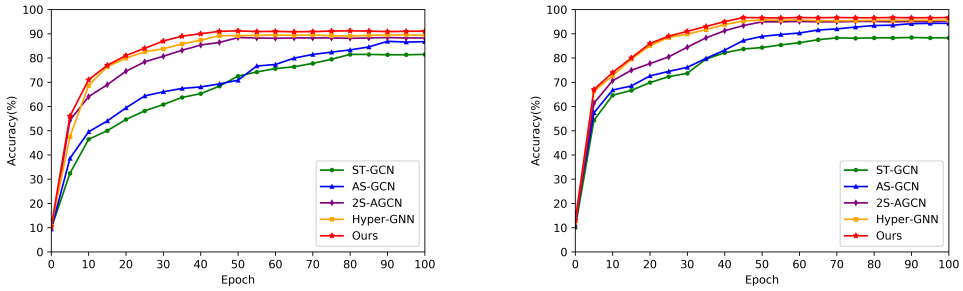
(**a**)

Figure 9. *Cont.*



(**b**)

Figure 9. Accuracy in each action category of our method and 2S-AGCN on the NTU RGB+D 60 dataset. The horizontal coordinates indicate different action categories and the vertical coordinates indicate the accuracies of different actions: (**a**) CS (cross-subject) protocols; (**b**) CV (cross-view) protocols.



(**c**) (**d**)

Figure 10. Classification accuracy of different methods on the NTU RGB+D 60 dataset. The x axis represents Epoch in (**a**,**b**), and the y axis represents accuracy.

6. Conclusions

In this paper, we explore the higher-order correlations of skeleton-based action recognition in physical and non-physical connections. For human actions, the higher-order correlations of skeleton sequences is characterized using hypergraphs to model the physical and non-physical connections of the skeleton before forming different data. Then, AMC-STHGNN effectively fuses the features of each channel, reasonably using the diversity and complementarity of physically connected joints, non-physically connected joints, and non-physically connected bone in human action, weakening the interference of noise and effectively improving the accuracy of action recognition. Experimental results using the NTU RGB+D 60 dataset and the Kinetics-Skeleton dataset demonstrate the superior accuracy and robustness of the proposed model compared to mainstream human action recognition methods.

Future research directions include exploring lightweight models to improve the network computation and real-time performance. The optimization of the hypergraph structure and reduction in noise impact on hypergraph modeling are also important areas of focus. Moreover, the application of hypergraph neural network-based action recognition methods in unmanned fields, such as self-driving vehicles and interactive wheeled robots, shows promising potential [48,49].

Author Contributions: Conceptualization, C.W.; methodology, C.W. and Z.W.; validation, C.W. and Z.W.; investigation, C.W.; data curation, C.W.; writing—original draft preparation, C.W., N.M. and Z.W.; writing—review and editing, C.W. and N.M.; visualization, Z.W.; supervision, N.M.; project administration, N.M.; funding acquisition, N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (No.2023YFF0615800), the Beijing Natural Science Foundation (No. 4222025), the National Natural Science Foundation of China (No. 61931012), and QIYUAN LAB Innovation Foundation (Innovation Research) Project (No. S20210201107).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: NTU RGB+D 60 dataset was analyzed in this study. This data can be found here: [https://rose1.ntu.edu.sg/dataset/actionRecognition/]. Kinetics-skeleton was analyzed in this study. This data can be found here: https://github.com/activitynet/ActivityNet/tree/master/Crawler/Kinetics.

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

References

1. Ma, N.; Wu, Z.; Cheung, Y.m.; Guo, Y.; Gao, Y.; Li, J.; Jiang, B. A survey of human action recognition and posture prediction. *Tsinghua Sci. Technol.* **2022**, *27*, 973–1001. [CrossRef]
2. Zhang, R.; Jiang, C.; Wu, S.; Zhou, Q.; Jing, X.; Mu, J. Wi-Fi sensing for joint gesture recognition and human identification from few samples in human-computer interaction. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 2193–2205. [CrossRef]
3. Dallel, M.; Havard, V.; Dupuis, Y.; Baudry, D. Digital twin of an industrial workstation: A novel method of an auto-labeled data generator using virtual reality for human action recognition in the context of human–robot collaboration. *Eng. Appl. Artif. Intell.* **2023**, *118*, 105655. [CrossRef]
4. Mabrouk, A.B.; Zagrouba, E. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Syst. Appl.* **2018**, *91*, 480–491. [CrossRef]
5. Wu, Z.; Ma, N.; Gao, Y.; Li, J.; Xu, X.; Yao, Y.; Chen, L. Attention Mechanism Based on Improved Spatial-Temporal Convolutional Neural Networks for Traffic Police Gesture Recognition. *Int. J. Pattern Recognit. Artif.* **2022**, *36*, 2256001. [CrossRef]
6. Xu, B.; Shu, X.; Zhang, J.; Dai, G.; Song, Y. Spatiotemporal Decouple-and-Squeeze Contrastive Learning for Semisupervised Skeleton-Based Action Recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. [CrossRef]
7. Dai, M.; Sun, Z.; Wang, T.; Feng, J.; Jia, K. Global spatio-temporal synergistic topology learning for skeleton-based action recognition. *Pattern Recognit.* **2023**, *140*, 109540. [CrossRef]
8. Gao, Y.; Zhang, Z.; Lin, H.; Zhao, X.; Du, S.; Zou, C. Hypergraph Learning: Methods and Practices. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2548–2566. [CrossRef] [PubMed]
9. Wang, C.; Ma, N.; Wu, Z.; Zhang, J.; Yao, Y. Survey of Hypergraph Neural Networks and Its Application to Action Recognition. In Proceedings of the CAAI International Conference on Artificial Intelligence, Beijing, China, 27–28 August 2022; pp. 387–398.
10. 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, Vancouver, BC, Canada, 28 February–1 March 2022; Volume 36, pp. 2866–2874.
11. Andrade-Ambriz, Y.A.; Ledesma, S.; Ibarra-Manzano, M.A.; Oros-Flores, M.I.; Almanza-Ojeda, D.L. Human activity recognition using temporal convolutional neural network architecture. *Expert Syst. Appl.* **2022**, *191*, 116287. [CrossRef]
12. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 13359–13368.
13. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1474–1488. [CrossRef] [PubMed]
14. Feng, Y.; You, H.; Zhang, Z.; Ji, R.; Gao, Y. Hypergraph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HA, USA, 29–31 January 2019; Volume 33, pp. 3558–3565.

15. Jiang, J.; Wei, Y.; Feng, Y.; Cao, J.; Gao, Y. Dynamic Hypergraph Neural Networks. In Proceedings of the IJCAI, Macao, China, 10–16 August 2019; pp. 2635–2641.
16. Bai, S.; Zhang, F.; Torr, P.H. Hypergraph convolution and hypergraph attention. *Pattern Recognit.* **2021**, *110*, 107637. [CrossRef]
17. Gao, Y.; Feng, Y.; Ji, S.; Ji, R. HGNN+: General Hypergraph Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3181–3199. [CrossRef] [PubMed]
18. Zhang, Z.; Lin, H.; Zhao, X.; Ji, R.; Gao, Y. Inductive Multi-Hypergraph Learning and Its Application on View-Based 3D Object Classification. *IEEE Trans. Image Process.* **2018**, *27*, 5957–5968. [CrossRef] [PubMed]
19. Shi, H.; Zhang, Y.; Zhang, Z.; Ma, N.; Zhao, X.; Gao, Y.; Sun, J. Hypergraph-induced convolutional networks for visual classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 2963–2972. [CrossRef] [PubMed]
20. Chen, Y.; Li, Y.; Zhang, C.; Zhou, H.; Luo, Y.; Hu, C. Informed Patch Enhanced HyperGCN for skeleton-based action recognition. *Inf. Process. Manag.* **2022**, *59*, 102950. [CrossRef]
21. Wang, S.; Zhang, Y.; Qi, H.; Zhao, M.; Jiang, Y. Dynamic Spatial-temporal Hypergraph Convolutional Network for Skeleton-based Action Recognition. *arXiv* **2023**, arXiv:2302.08689.
22. Zhou, Y.; Li, C.; Cheng, Z.Q.; Geng, Y.; Xie, X.; Keuper, M. Hypergraph transformer for skeleton-based action recognition. *arXiv* **2022**, arXiv:2211.09590.
23. Wang, M.; Liu, X.; Wu, X. Visual classification by ℓ_1-hypergraph modeling. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 2564–2574. [CrossRef]
24. Zhang, Z.; Feng, Y.; Ying, S.; Gao, Y. Deep Hypergraph Structure Learning. *arXiv* **2022**, arXiv:2208.12547.
25. Zhou, D.; Huang, J.; Schölkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 1601–1608.
26. Lu, J.; Wan, H.; Li, P.; Zhao, X.; Ma, N.; Gao, Y. Exploring High-Order Spatio–Temporal Correlations From Skeleton for Person Re-Identification. *IEEE Trans. Image Process.* **2023**, *32*, 949–963. [CrossRef]
27. He, B.; Guan, Y.; Dai, R. Convolutional gated recurrent units for medical relation classification. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain , 3–6 December 2018; pp. 646–650. [CrossRef]
28. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
29. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
30. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12026–12035.
31. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HA, USA, 21–26 July 2017; pp. 7291–7299.
32. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3d skeletons as points in a lie group. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 588–595.
33. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 816–833.
34. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2117–2126.
35. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1227–1236.
36. Soo Kim, T.; Reiter, A. Interpretable 3d human action analysis with temporal convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 20–28.
37. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv* **2018**, arXiv:1804.06055.
38. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1963–1978. [CrossRef]
39. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
40. 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]
41. 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/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 183–192.
42. Hao, X.; Li, J.; Guo, Y.; Jiang, T.; Yu, M. Hypergraph neural network for skeleton-based action recognition. *IEEE Trans. Image Process.* **2021**, *30*, 2263–2275. [CrossRef] [PubMed]

43. Wei, J.; Wang, Y.; Guo, M.; Lv, P.; Yang, X.; Xu, M. Dynamic hypergraph convolutional networks for skeleton-based action recognition. *arXiv* **2021**, arXiv:2112.10570.
44. Zhu, Y.; Huang, G.; Xu, X.; Ji, Y.; Shen, F. Selective hypergraph convolutional networks for skeleton-based action recognition. In Proceedings of the 2022 International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022; pp. 518–526.
45. He, C.; Xiao, C.; Liu, S.; Qin, X.; Zhao, Y.; Zhang, X. Single-skeleton and dual-skeleton hypergraph convolution neural networks for skeleton-based action recognition. In Proceedings of the Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, 8–12 December 2021; pp. 15–27.
46. Fernando, B.; Gavves, E.; Oramas, J.M.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5378–5387.
47. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3316–3333. [CrossRef] [PubMed]
48. Ma, N.; Li, D.; He, W.; Deng, Y.; Li, J.; Gao, Y.; Bao, H.; Zhang, H.; Xu, X.; Liu, Y.; et al. Future vehicles: Interactive wheeled robots. *Sci. China Inf. Sci.* **2021**, *64*, 156101. [CrossRef]
49. Li, D.; Ma, N.; Gao, Y. Future vehicles: Learnable wheeled robots. *Sci. China Inf. Sci.* **2020**, *63*, 193201. [CrossRef]