2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Human Action Recognition with Multi-Level Granularity and Pair-wise Hyper GCN

Tamam Alsarhan[1,2], Syed Sadaf Ali[1], Ayoub Alsarhan[3], Iyyakutti Iyappan Ganapathi[1], and Naoufel Werghi[1]
[1]C2PS and Department of Computer Science Khalifa University, Abu Dhabi, UAE
[2]King Abdullah || school of Information Technology, The University of Jordan, Jordan
[3]The Hashemite University, Zarqa, Jordan
Email: {syed.ali, iyyakutti.ganapathi, naoufel.werghi}@ku.ac.ae, t_alsarhan@ju.edu.jo, ayoubm@hu.edu.jo

***Abstract*— Lately, there has been a surge in interest in utilizing Graph Convolutional Networks (GCNs) for the purpose of action recognition using skeletal data. In order to achieve optimal results, it is crucial to generate high-quality representations of the skeletal graph. Graph Convolutional Networks (GCNs) often employ the Message-Passing Mechanism (MPM) to acquire knowledge about various components of the skeleton by iteratively computing new features at each step. However, the interconnections between joints in the skeletal structure are intricate and extend beyond mere proximity. In order to address this issue, we propose the implementation of our Disassembled Hyper-Graph (DH-Graph), which draws inspiration from hyper-graph edges. The process of constructing the DH-network entails a few steps: partitioning the skeleton network into clusters of hyper-edges according to their semantic significance and relevance to action recognition, arranging these clusters in a hierarchical structure to enhance granularity, and establishing connections between joints within these clusters to discover hidden relationships. The DH-Graph employs a spatial domain GCN technique to construct the Pair-wise Hyper Hierarchical GCN (PH-GCN). In addition, we incorporate the HyperAttention module, which employs Multi-scale Representative Spatial Average Pooling and Edge Convolution techniques to emphasize significant sets of hyper-hierarchical information. Extensive experiments demonstrate that PH-GCN achieves remarkable performance on challenging NTU RGB+D and Northwestern UCLA datasets.**

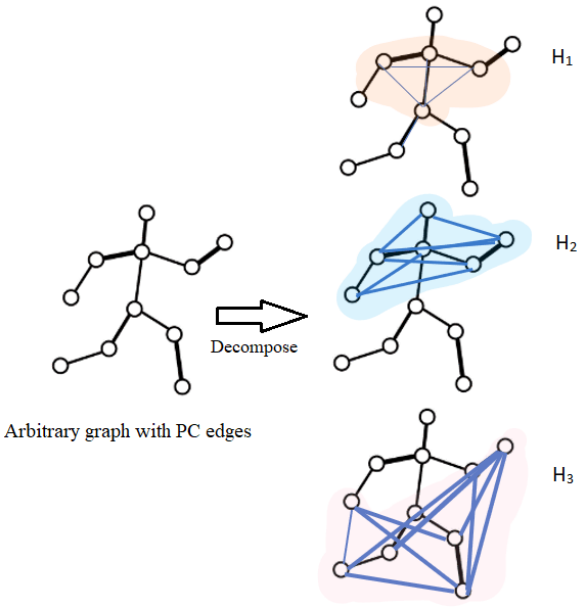


Fig. 1. Our PH-GCN illustration uses color-coded areas to represent connections in distinct hierarchical-hyperedge groups, with line thickness emphasizing edge importance. The *running action* highlights the set containing knee, hands, and feet joints. (PC) indicates physical connection.

I. Introduction

Automated categorization of human movements using visual data analysis is a crucial task in computer vision, known as skeleton-based action recognition [1]–[3], [16]. This field has gained significant attention due to its applications in human-computer interaction [8], robotics [13], and intelligent video surveillance [12] over the past decade. Graph Convolutional Networks (GCNs) have become a robust approach for modeling human skeletons as graphs and capturing spatiotemporal patterns [14], [21], [22]. GCNs aim to learn effective representations of non-Euclidean data, such as graphs, which are vital for skeleton-based action recognition [27]. These networks employ a message-passing mechanism (MPM) with three main steps: node initialization using initial features, node updating via neighborhood feature aggregation, and readout to obtain node representations for the final task. However, GCNs mainly focus on direct relationships between adjacent nodes, overlooking crucial distant interactions required for understanding complex actions. Actions involving multiple body parts and movements, like throwing a ball or performing jumping jacks, necessitate capturing relationships beyond immediate neighbors. For example, to throw a ball effectively, the coordination of distant joints (elbow, shoulder, wrist) is vital. Therefore, graph modeling's effectiveness is closely tied to how well it captures these intricate node connections.

Graph modeling in skeleton-based action recognition, which defines structural dependencies in GCNs through adjacency matrices, remains an open challenge. Current approaches like [50], and [51] rely on Yan's predefined graph [49], using a fixed adjacency matrix, limiting their focus to Physically Connected (PC) joints in the graph. However, this approach restricts the receptive field to adjacent joints and

overlooks the importance of distant connections for many actions. Recent efforts have introduced learnable graphs to address these limitations, yet the influence of the fixed graph [49] persists, and physically unconnected joints are underemphasized [50], [51]. In [84], authors learn compositional relationships among body parts of different semantic levels and then exploit multilevel structural reasoning to reduce the depth uncertainty.

Hypergraph, a specialized graph model that groups vertices into hyperedges, has gained attention in various domains, such as image matching [39], multi-label classification [41], video object segmentation [45], mesh segmentation [42], [43],and image retrieval [46]. Notably, techniques like clique and star expansion [47] and probabilistic models [46] have enhanced hypergraph structures. Learning optimal hyperedge weights, crucial for data correlation modeling, has also been studied, with methods like L2 regularization proposed by Gao et al. [54].

This work introduces a novel Pair-wise Hyper Graph Convolutional Network (PH-GCN) that leverages a carefully crafted Disassembled Hypergraph (DH-Graph) to extend traditional GCNs. Unlike the conventional approach, which uses Yan's predefined graph focusing solely on physically connected joints, DH-Graph addresses the challenge of distant node connectivity. The DH-Graph construction unfolds through three key steps: hyper-edges formulation, grouping joints based on function and importance; hierarchy formulation, organizing hyper-edge groups hierarchically in a semantic space representing body parts; and edge construction, connecting nodes within adjacent hierarchical sets to highlight significant distant joints within the same semantic space. The overall view of our model is illustrated in Fig.1.

The contribution of each hierarchy edge set in a skeletal graph varies depending on the action. For instance, recognizing *walking* relies on the hip and leg hierarchy edge set, emphasizing the relevant hierarchy edge set enhances action recognition. A HyperAttention module integrates correlations from different hyper-hierarchical edge sets, emphasizing the most vital hierarchy edge set of DH-Graph. The attention map, obtained using multi-scale representative spatial average pooling (MS-RSAP) and hierarchical edge Convolution, improves action recognition. PH-GCN disassembles the graph into smaller sub-graphs, capturing fine-grained details and high-level context in actions. It accommodates actions of varying complexity, efficiently modeling complex relationships between body parts. This approach enhances action recognition by providing both fine-grained details and high-level context, enabling nuanced understanding and accurate recognition of actions.

This paper delves into graph modeling for skeleton-based action recognition by introducing hyper-graphs representing human skeletons and assessing their influence on action recognition. Experimental findings establish PH-GCN's superior performance over conventional GCNs, particularly for intricate actions involving multiple body parts. Our hyper-hierarchical graph employs multiple hierarchical levels, enabling the representation of even more intricate node relationships by capturing hierarchical connections between levels. The key contributions include:

1) We introduce a meticulously-designed Disassembled Hyper-Graph, structured to highlight the most significant distant edges within each hyper-edge group. This innovative graph model enhances the representation of skeletal relationships, particularly emphasizing crucial distant connections.
2) Our proposed HyperAttention strategy establishes a fusion mechanism for hyper-edge groups. This careful fusion incorporates correlations from various hyper-edge groups and accentuates the most vital hyper-hierarchical set within the DH-Graph. The Group Correlation aspect enhances the model's ability to capture significant relationships among joints.
3) Rigorous experimentation on benchmark datasets, including NTU-RGB+D 60 and Northwestern-UCLA, validates the effectiveness of our framework. The empirical results showcase the superior performance of our Pair-wise Hyper Hierarchical GCN (PH-GCN) in the domain of action recognition. Notably, our framework achieves remarkable accuracy on these challenging datasets, as evidenced by the outcomes of extensive experiments.

II. Related work

Graph Neural Networks (GNNs): Graph neural networks (GNNs) represent a specialized category of feed-forward neural networks recognized for their proficiency in graph-based learning [34], [35]. These networks efficiently condense complex patterns of connectedness within neighborhoods into low-dimensional embeddings, playing a pivotal role in diverse downstream tasks. Notable variations in this domain include Graph Convolutional Networks, which leverage mean pooling, and GraphSAGE, incorporating enhanced aggregation techniques by combining node features through mean, max, and LSTM pooled neighborhood information [44]. Additionally, Graph Attention Networks stand out by utilizing trainable attention weights to intelligently aggregate neighborhood information [10]. This variety of GNN architectures underscores the dynamic landscape of techniques employed to harness and process graph-based data for improved learning outcomes.

Skeleton-Based Action Recognition: Skeleton-based action recognition detects specific human motion from skeletal data [5], [9], [79]. Most techniques represent the human skeleton as a graph with joints as nodes and bones as edges [73]–[76], frequently employing [73]'s graph. Despite efforts to introduce learnable graphs emphasizing distant joint relationships, handcrafted graphs like [73]'s remain dominant due to their superior relationship highlighting. Attention mechanisms [83] enhance GCN models by emphasizing crucial information along specific dimensions. For instance, Li et al. [85] utilized the non-local network [86] for an attention-based graph. Chen et al. [88] used channel-wise

attention to create channel-wise topologies, enriching the shared graph.

Hyper-Graph Representation: Hyper-graph representation groups multiple nodes/vertices with hyper-edges, as shown in Fig. 2. While akin to graph learning, hyper-graphs generalize graphs, allowing for structured data analysis in node classification [58], link prediction [60], and community detection [61]. In computer vision tasks like person re-identification [62], video segmentation [68], image retrieval [64], 3D object classification [69], hyperspectral image analysis [60], landmark retrieval [71], and visual tracking [72], hyper-graph structures model high-order relationships. Zhou et al. [55] pioneered hyper-graph learning, while [56] proposed a Chebyshev expansion-based hyper-graph neural network. Sophisticated hyper-graph structures, incorporating global, local visual features, and tag information, have been developed for tasks like image retrieval [57] to determine image relevance.

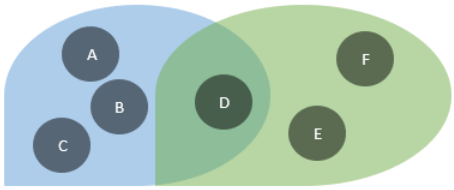


Fig. 2. An example of a hypergraph structure comprises 6 nodes (A, B, C, D, E, F) and 2 hyperedges ({A, B, C, D} and {D, E, F}).

Acknowledging the importance of how data is represented in Graph Convolutional Network (GCN) methods, we drew inspiration from the idea of hypergraphs to create our skeleton graph. This approach was influenced by exploring hypergraph concepts, aiming to make our graph representation more nuanced and effective within the GCN framework.

III. Proposed Methods

A. Revisiting action recognition in GCN

A graph is denoted as:

$$G = (V, E) \in a_1, a_2, ... \tag{1}$$

where G denotes the graph, V represents the vertices (joints), E denotes the edges, and a_i, (where $i \in 1, 2, 3..$) is one of the labeled human actions. Given a new skeletal graph $\hat{G}$, our goal is to predict or classify which action is being performed based on the structural information encoded in the skeletal graph. Typically, most of the mainstream approaches consider graph convolutional networks as a way to extract features from the given graph, as stated in Eq. 2 :

$$F_{out} = \sum_{S \in s} A_s X W_s \tag{2}$$

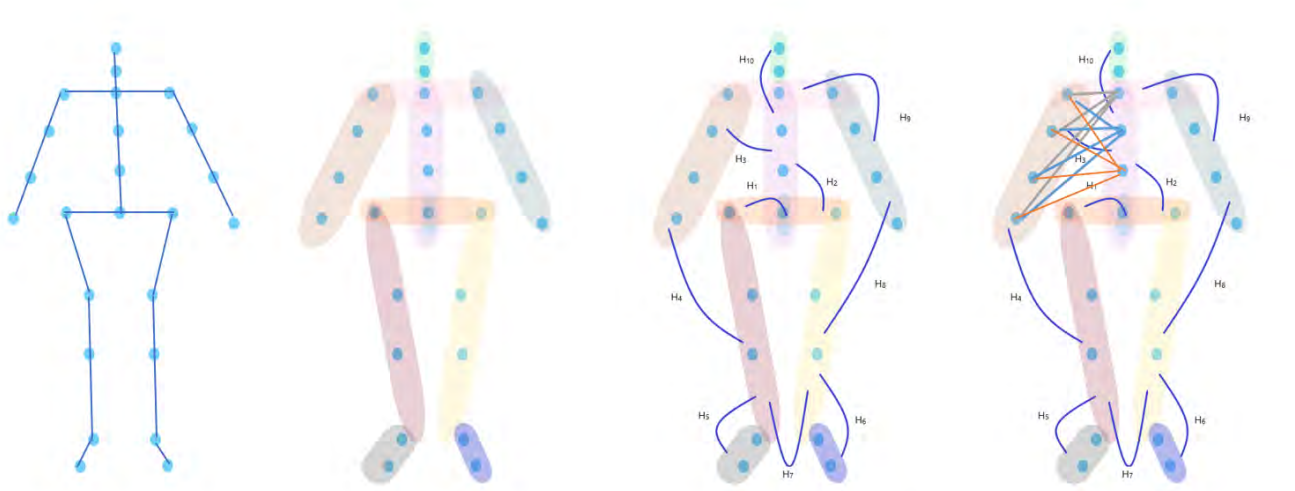
where the adjacency matrix, which is a $V \times V$ matrix that reflects the physical relationship between nodes in the skeletal graph, is represented by A. If a physical connection between any two nodes (i, j) is found, then $A_{i,j}$ equals 1, otherwise, $A_{i,j}$ is 0. Following [85], we normalized A and initialized it as $\Lambda^{-1/2} A \Lambda^{-1/2}$, where Λ is the diagonal matrix. X is the 3D time-series skeletal data, where $X \in \mathbb{R}^{C \times T \times V}$. Here, C is the number of channels, T is the number of frames, and V is the number of joints.

Consider a scenario where we are meticulously recording the movements of an individual engaged in a specific action, let's say *walking*. In this context, each joint of the human body, ranging from the head to the knees, is meticulously represented as a point in the expansive realm of three-dimensional space. This motion data is not captured in isolation but unfolds over a duration, resulting in a sequence of frames. Each frame serves as a temporal snapshot, encapsulating the spatial arrangement of joints at precise moments during the action. To comprehend and process this wealth of information, we introduce the concepts of channels, frames, and joints.

B. Disassembled hyper-graph

In the realm of skeleton-based action recognition, graphs have proven effective in capturing intricate node-to-node interactions. However, the limitations of conventional graphs, primarily utilizing Physically Connected edges (PC) [49], become apparent in their sparsity and inadequate representation of distant relationships. Consider the action of "clapping," where precise coordination between two distant joints, such as the hands, is crucial. Conventional graphs struggle to accurately depict such collaboration due to the substantial spatial gap between these joints. To overcome these challenges, we draw inspiration from the realm of graph theory, specifically the concept of hyper-edge groups. Leveraging these ideas, we introduce our novel hyper-graph, aiming to provide a more expressive representation that can effectively capture and articulate distant relationships within the skeletal structure.

In this paper, the primary objective is to enhance the conventional graph structure, which relies on physically connected (PC) edges, by introducing a new graph architecture featuring more meaningful edges. The goal is to facilitate the capture of richer and more distant relationships between joints in a skeletal structure. To achieve this objective, we develop a new graph structure referred to as the DH-Graph unfolded in three key stages. Initially, the focus was on the formulation of hyper-edges, termed the *Hyper-edges formulation*, aimed to understand and define the creation of hyper-edges, which are sets of edges that go beyond the traditional physically connected links. In the next stage, *Hyper-edges organization* is developed to organize hyper-edges in a hierarchical manner. This step is another crucial for establishing a structured and layered representation of relationships between joints. By arranging hyper-edges hierarchically, the aim is to capture not only the direct connections between joints but also more intricate relationships that extend across different levels of granularity. Lastly, in the *Edge construction* phase, we built connections between joints in neighboring hierarchy edge



a) A graph with PC edges b) A graph with hyper-edge groups c) A graph with hierarchal sets d) Our disassembled hyper-graph (DH-Graph) with distant edges

Fig. 3. Illustration of graph structures. (a) represents the Yan's graph structure [49], which only considers PC edges. (b) represents a disassembled graph with hyper edges. (c) represents the hierarchical hyper edges structure. (d) represents our DH-Graph structure which contains distant edges.

sets. Here are the specifics:

Hyper-edges formulation: During this phase, we undertook the segmentation of Yan's graph [49], a widely employed structure in prevalent models for skeleton-based human action recognition, into multiple sub-graphs. This segmentation involved categorizing joints into hyper-edges, taking into account their functional associations and relevance to the action recognition task, as depicted in Fig. 3-b. For instance, we distinguished between the upper and lower limbs, as well as the head and neck from the spine. The introduction of hyper-edges proves crucial in refining feature depiction within human action recognition. Their capacity to cluster joints based on function contributes to creating richer and more pertinent feature representations for the action recognition task. By separating the movements of the upper limbs from those of the lower limbs, we can more effectively identify distinctive motion patterns within each group. This strategic use of hyper-edges becomes instrumental in enhancing feature representation for human action recognition. Notably, the grouping of joints based on similar functionalities through hyper-edges yields more insightful feature representations, as exemplified by the improved capture of differences in motion patterns between the upper and lower limbs.

Hierarchy formulation: In this phase, our objective is to systematically categorize hyper-edge groups that share a common semantic space into cohesive hierarchical sets. For clarity, consider Fig. 3-c as a visual reference. The first hierarchy set, denoted as H_1, is crafted to include joints that occupy a central semantic space, such as the Center of Mass joint (CoM) and the hip joints. Building on this organizational principle, the second hierarchy set H_2 consolidates the hyper-edge sets related to hips and spine. This hierarchical structuring is consistently extended, resulting in the formulation of N hierarchy sets, where $N = H - 1$. Each hierarchy set encapsulates hyper-edge groups with shared semantic characteristics, fostering a more nuanced and organized representation of the skeletal structure. This method ensures that each hierarchy set aligns with the distinct semantic spaces delineated by the hyper-edge groups, enhancing the interpretability and effectiveness of our proposed hierarchical framework.

Edge construction: Ultimately, our approach involves establishing pair-wise connections (edges) among all nodes within neighboring hierarchical sets. This strategic connection scheme accentuates the relationships among distant joint nodes, particularly those sharing the same semantic context, resulting in a fully connected structure, as illustrated in Fig. 3-d. For clarity, we've specifically highlighted edges within H_3 for simplicity, although this concept extends across all hierarchical sets. By deconstructing the conventional graph with physically connected (PC) joints and introducing our disassembled graph, we inherently divide the standard adjacency matrix A into multiple sub-adjacency matrices. Each of these matrices is constructed by establishing edges within the hierarchical sets. Mathematically, the conventional $A \in \mathbb{R}^{V \times V}$ is reformulated as $A \in \mathbb{R}^{N \times V \times V}$, with the formulation expressed as:

$$A = \{e(H_1, H_2)|| \dots ||e(H_H, H_{H-1})\}$$

The notation used is as follows: H1, H2, etc.,representst the hierarchy sets, and e signifies the edges formed between nodes in each pair of consecutive hierarchy sets. Finally, the adjacency matrix $\mathbf{A}$ is formed by concatenating all subsets.

In Fig. 3-d, we depict the connections in a singular

direction. However, to align with established methodologies [50] [51], we extend beyond forward connections between any two joints. Instead, we incorporate identity connections and backward connections. Each edge set encompasses a concatenation of identity (id), centripetal (cp), and centrifugal (cf) edge subsets, as illustrated in Fig. 4 and represented by the equation:

$$A_i = [A_{id}||A_{cf}||A_{cp}], \quad i = 1 \text{ to } N$$

The resulting graph configuration, termed DH-Graph, seamlessly integrates both physically connected (PC) and substantial distant relationships. This synthesis enhances the overall density of the graph and extends its receptive range, contributing to a more comprehensive and effective representation of skeletal interactions.

C. Graph convolution

As illustrated in Fig. 4, for each hierarchical set N, we employ Graph Convolution on our DH-Graph for the three components of A, as outlined in Eq. 4. Subsequently, we concatenate the results according to the following equation:

$$F^{(N)} = \sum_{S \in s} A_s^{(N)} X W_s^{(N)}, \quad s = 3 \tag{3}$$

Specifically, the output feature map F is generated by passing the input feature map X through a Graph Convolution layer, utilizing the adjacency matrix A and weight matrix W. The final output is derived by concatenating the output feature maps obtained by utilizing the three edge subsets (identity (id), centripetal (cp), and centrifugal (cf)) for each hierarchy set, along with EdgeConv [94]. EdgeConv is adopted to capture key sample-wise relationships between all nodes in the feature space, reflecting affinity relationships not captured by our GCN layer. It is essential to note that we apply a linear transformation to X before inputting it into the Graph Convolution layer to enhance computational efficiency.

D. HyperAttention Module

After applying Graph Convolution to each of the N hierarchical sets, we introduce our HyperAttention Module, specifically designed to highlight the most informative hierarchical sets. Given that specific body parts collaborate to execute distinct human actions, it becomes crucial to identify the most critical set(s) for the given task. For instance, when distinguishing running, the upper and lower limbs are the most relevant subsets. Similarly, in actions like clapping, emphasizing the coordination of both hands outweighs the importance of the neck and head.

Our HyperAttention method comprises three key steps: edge convolution, multi-scale representative spatial average pooling (MS-RSAP), and temporal frame selection. In the *Temporal frame selection* step, temporal pooling is employed to identify the frame with the highest feature map score F. This selected frame then serves as a reference frame for the subsequent selection of representative nodes in each hierarchy layer, as illustrated in Fig. 5.

$$f^N = \text{MaxPooling}_t(F^N) \tag{4}$$

where N represents the N-th hierarchy set, and f is the tensor obtained after temporal pooling. Subsequently, we employ MS-RSAP to extract representative nodes from the input tensor across various hierarchical levels. We ensure consistency across all samples in the dataset by extracting representative nodes from each hierarchy layer. This step is essential for preventing scaling bias, a potential issue when each node has a variable number of connecting edges. Applying spatial average pooling directly to the feature map F without this extraction phase could result in varying edge counts per node, potentially compromising accuracy in subsequent computations.

$$\zeta(f_N) = \frac{1}{v_k + v_{k+1}} \sum_{v \in h_{N,N+1}} f(v) \tag{5}$$

where v_k is the number of nodes in each set. The attention weights are computed by applying an edge convolution to the sampled feature map, followed by an aggregation operation and re-weighting the result. This generates a tensor with dimensions $N \times L \times 1 \times 1$. Finally, the output feature map is obtained by element-wise multiplication of the input feature map F with the sigmoid of the attention weights, followed by a summation operation over the L dimension, resulting in a tensor with dimensions $N \times C \times 1 \times 1$.

E. Network Architecture

Our final model comprises nine stacked GCN blocks, each incorporating a DH-Graph Convolution layer, HyperAttention, and temporal convolution. Blocks 1-3 generate 64 output channels, blocks 4-6 produce 128, and blocks 5-9 yield 256. A residual connection is applied, and the temporal convolution from [22] is adopted in our model.

IV. Experiments

In this section, we conduct a comprehensive evaluation of our PH-GCN against state-of-the-art models for skeleton-based human action recognition. Our assessments are performed on two widely recognized datasets to showcase the efficacy of our proposed model.

A. Experimental Settings

1) Datasets: To evaluate the performance of our PH-GCN, we assess its capabilities on two prominent datasets: NTU-RGB 60 [91] and Northwestern-UCLA [101].

NTU RGB+D: Comprising 56,880 skeletal action sequences, the NTU RGB+D dataset [91] is a widely utilized benchmark for human action recognition. Video frames are captured using three Microsoft Kinect-V2 depth sensors placed at different horizontal viewpoints. The dataset offers two primary benchmarks for evaluation: Cross-Subject (X-Sub) and Cross-View (X-View). The X-View subset comprises 37,920 samples from camera views 2 and 3, with the test set containing 18,960 sequences captured by camera

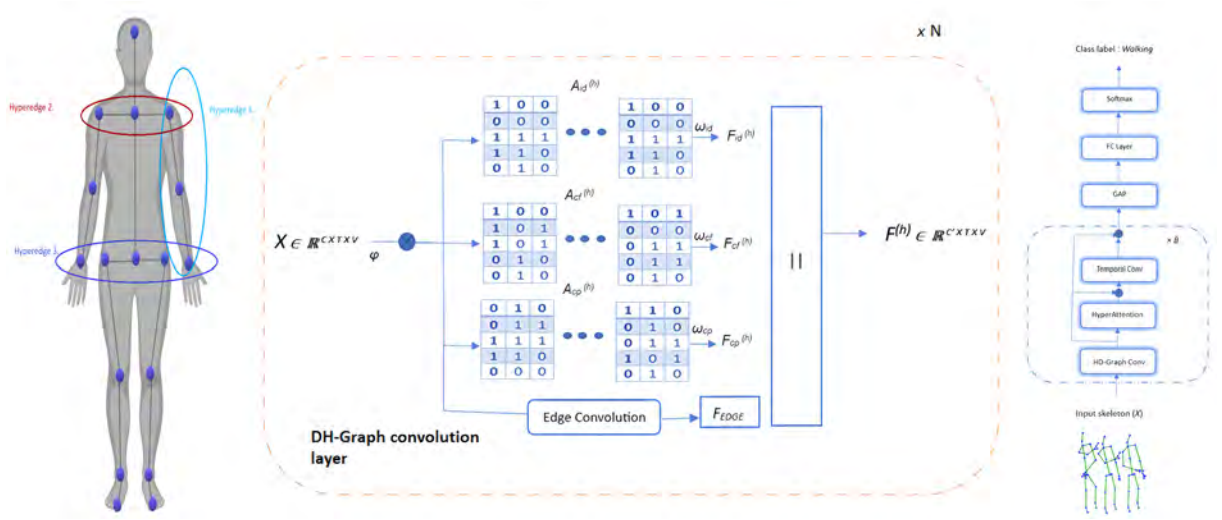


Fig. 4. Overview of our PH-GCN. The left side displays the human skeleton with hyper-edge groups, showcasing the operation of a single hyper-hierarchical edge set (highlighted by the orange dotted rectangle). On the right side, the overall architecture of our PH-GCN is presented. It takes in the skeleton sequence and processes it through our Spatial-temporal block, which comprises HD-GC, the HyperAttention module, and a temporal convolution layer (B blocks in total). The class label is obtained through a Global Average Pooling (GAP) layer, fully connected (FC) layer, and Softmax layer.

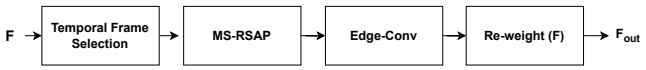


Fig. 5. Illustration of our HyperAttention module. F is the input feature map after the HD-Graph Convolution.

view 1. The Cross-Subject (X-Sub) training set involves 20 subjects, with the testing set captured from an additional 20 subjects.

Northwestern-UCLA: This dataset features 1,494 video clips across ten different classes, with each action performed by a distinct subject. Following the methodology established in [59], we utilize training data from the first two cameras and validation samples from the third.

2) Implementation details: The backbone of our model is inspired by [7]. For optimization, we employed Stochastic Gradient Descent (SGD) with a weight decay of 0.0004 and Nesterov momentum of 0.9. The model was trained over 100 epochs, implementing a warm-up strategy for the initial five epochs. Cosine annealing [89] was employed to progressively reduce the maximum learning rate, set at 0.1. The chosen loss function for our experiments is cross-entropy loss, with a batch size of 64. Each sample was standardized to 64 frames. Preprocessing has been conducted using [66].

TABLE I

PERFORMANCE EVALUATION OF SKELETON-BASED ACTION RECOGNITION IN TOP-1 ACCURACY (%) ON NTU RGB+D DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, WHILE THE SECOND BEST IS IN BLUE.

Category	Model	NTU RGB+D X-Sub(%)	NTU RGB+D X-View(%)
RNN	VA-LSTM [70]	79.4	87.6
	AGC-LSTM [63]	89.2	95.0
	H-RNN [67]	59.1	64.0
	PA-LSTM [91]	62.9	70.3
CNN	Ta-CNN+ [40]	90.7	95.1
	VA-CNN [82]	88.7	94.3
Transformers	ST-TR [15]	89.9	96.1
	Hyperformer [17]	90.7	95.1
	DSTA [59]	91.5	96.4
GCN	Shift-GCN [87]	90.7	96.5
	AM-GCN [48]	90.3	95.2
	InfoGCN [78]	89.8	95.2
	MS-G3D [81]	91.5	96.2
	Dynamic GCN [80]	91.5	96.0
	ST-GCN [49]	81.5	88.3
	AS-GCN [93]	86.8	94.2
	2s-AGCN [7]	88.5	95.1
	RA-GCN [65]	87.3	93.6
	Our PH-GCN (2-streams)	92.3	96.5
	Our PH-GCN (4-streams)	92.9	97.2
	Our PH-GCN (6-streams)	**93.3**	**97.5**

B. Comparison with state-of-the-art approaches

To showcase the competitive performance of our proposed model, we conducted a thorough comparison with state-of-the-art models on the NTU RGB+D dataset and the Northwestern-UCLA dataset. The results presented in Table 1 and Table 2 demonstrate that our model outperforms existing models, achieving state-of-the-art results across all datasets. This highlights the significance of graph modeling in optimizing Graph Convolutional Networks (GCNs) [88] [76]. Our Disassembled Hyper-Graph is specifically designed to capture both adjacent and distant relationships among human joints in the skeletal graph. The observed improvement in accuracy and efficiency in our results emphasizes the impact of well-designed graph models on GCN performance. Following established state-of-the-art methodologies [88], [87] [85] [81], we adopted a multi-stream fusion framework. Specifically, we utilized three joints (CoM nodes: chest, hip, belly) to create three distinct graph structures, in addition to the joint and bone streams.

TABLE II

VALIDATION ACCURACY COMPARISON WITH STATE-OF-THE-ART METHODS ON THE NORTHWESTERN-UCLA DATASET.

Model	Northwestern-UCLA Top-1 (%)
Lie Group [95]	72.2
Action Ensemble [97]	76.0
H-RNN [98]	78.5
Ensemble TS-LSTM [99]	89.2
AGC-LSTM [96]	93.3
CTR-GCN [100]	96.5
Our PH-GCN (2-streams)	96.8
Our PH-GCN (4-streams)	97.0
Our PH-GCN (6-streams)	**97.3**

C. Effect of different graph architectures

As previously mentioned, we constructed distinct graph architectures by selecting the Center of Mass (CoM) to be either the hip, chest, or belly. Subsequently, we evaluated the two streams, specifically the joint and bone streams, using these three architectures. The results presented in Table 3 indicate variations in accuracy among the three architectures, emphasizing that the set of edges constructed between each hyperedge set differs. Notably, the highest performance was achieved when the hip was chosen as the Center of Mass. This outcome underscores that the three graph architectures exhibit different learning patterns, highlighting the significant impact of the chosen graph architecture on model performance.

TABLE III

VALIDATION ACCURACY COMPARISON ON NTU RGB+D DATASET WITH DIFFERENT DH-GRAPH IMPLEMENTATIONS.

CoM Joint	X_sub Bone	X_view
Hip	**90.6**	**95.7**
Chest	90.4	95.3
Belly	90.5	95.6

D. HyperAttention module

Table 4 presents the outcomes obtained through the manipulation or elimination of specific components within our attention module, illustrating the efficacy of our HyperAttention module. The deliberate choice of employing Multi-scale Representative Spatial Average Pooling (MS-RSAP) in our attention module is highlighted. This decision was motivated by the limitation of traditional spatial average pooling layers, which solely calculate the mean along a single axis without the ability to select representative nodes. This deficiency often leads to a scaling bias problem that could potentially compromise accuracy.

E. Performance on Ambiguous Action

The level of ambiguity varies among different action classes. Upon scrutinizing misclassified actions, a pattern emerged, predominantly occurring in cases of actions with inherent ambiguity. For instance, actions like *writing* and *typing on a keyboard* are susceptible to misclassification due to their pronounced similarities. However, our well-designed graph plays a pivotal role in emphasizing connections between the two hands in the case of *typing on a keyboard*, which are comparatively weaker in the case of *writing*. To provide a more granular understanding, we conducted a detailed analysis of the class-wise accuracy difference (%) between our PH-GCN and the baseline 2s-AGCN for the x-sub setting on NTU-60. The results, depicted in Fig. 6, highlight the superior performance of our PH-GCN across most classes, particularly for those characterized by ambiguous actions with distant correlations between joints. Notable accuracy gains were observed for actions such as touching the neck (+7.1%), taking off a shoe (+6%), and rubbing two hands together (+5.2%). This performance improvement can be attributed to the explicit modeling of long-range dependencies achieved by our PH-GCN model.

V. CONCLUSIONS AND FUTURE WORK

This study introduces the Pair-wise Hyper Graph Convolutional Network (PH-GCN) built upon the innovative Disassembled Hypergraph (DH-Graph), which enriches conventional Graph Convolutional Networks (GCNs) by incorporating hyper-edge modeling. The addition of the Hyper-Attention module further enhances the DH-Graph by intelligently fusing correlations from hyper-edge groups. Rigorous experiments conducted on challenging datasets showcase the PH-GCN's exceptional performance in human action recognition. However, it's important to note a key limitation, namely the model's reliance on the hyper-graph architecture. Future research directions will focus on exploring learning-based graph topology designs, refining model efficiency, and investigating adaptive mechanisms for action-specific hierarchies.

VI. ACKOWLEDGEMENT

This research is funded by research grants from the Advanced Technology Research Center Program (ASPIRE) under Ref: AARE20-279 and Khalifa University Internal Fund CIRA-2021-052.

REFERENCES

[1] Liu, Jian, Naveed Akhtar, and Ajmal Mian. "Adversarial attack on skeleton-based human action recognition." IEEE Transactions on Neural Networks and Learning Systems 33.4 (2020): 1609-1622.

[2] Y. Xiao, P. Siebert N. Werghi, "Topological segmentation of discrete human body shapes in various postures based on geodesic distance", Proceedings of the 17th International Conference on Pattern Recognition, 2004, 131-135.

[3] Li, Ce, et al. "Memory attention networks for skeleton-based action recognition." IEEE Transactions on Neural Networks and Learning Systems 33.9 (2021): 4800-4814.

[4] Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3288–3297 (2017)

[5] A. Zhu, Q. Ke, M. Gong and J. Bailey, "Adaptive Local-Component-aware Graph Convolutional Network for One-shot Skeleton-based Action Recognition," 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2023, pp. 6027-6036, doi: 10.1109/WACV56688.2023.00598.

TABLE IV

COMPARISON OF VARIOUS ATTENTION STRUCTURES ON THE NTU RGB+D DATASET

Model	EdgeConv	X_sub (%)	X_view (%)
Baseline	-	91.5	95.5
DH-GCN w/o Hyper-Attention		92.9	97.0
DH-GCN w/ MS-SAP		93.1	97.1
DH-GCN w/ MS-SAP and Hyper-Attention	✓	**93.3**	**97.5**

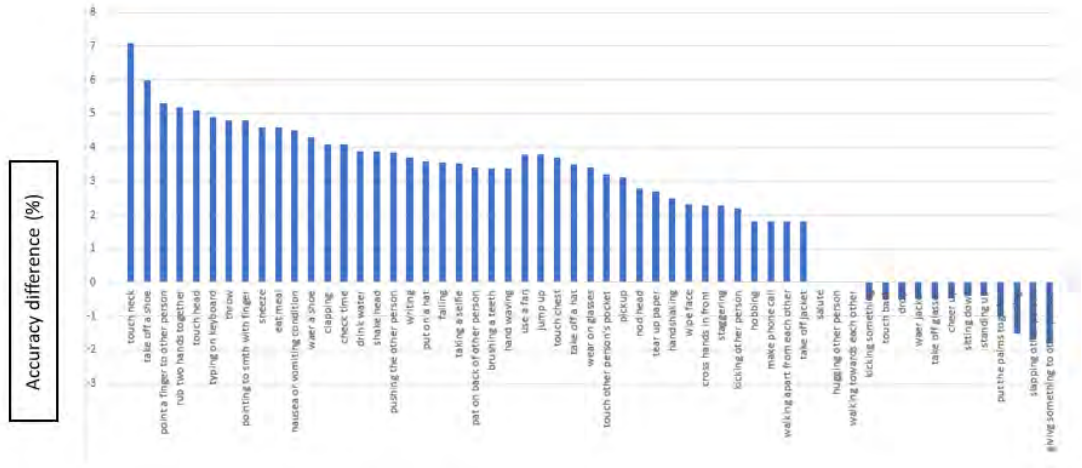


Fig. 6. Class-wise accuracy difference (%) between our PH-GCN and the baseline 2s-AGCN [85] for the x-sub setting on NTU-60. Please zoom in for better visibility.

[6] 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 (CVPR) (June 2016)

[7] 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. pp. 12026–12035 (2019)

[8] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proc. CVPR. pp. 6299–6308 (2017)

[9] J. Cai, N. Jiang, X. Han, K. Jia and J. Lu, "JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 2734-2743, doi: 10.1109/WACV48630.2021.00278.

[10] Velickovic, Petar, et al. "Graph attention networks." stat 1050.20 (2017): 10-48550.

[11] Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In European semantic web conference, 593–607. Springer

[12] Poppe, R.: A survey on vision-based human action recognition. Image and vision computing 28(6), 976–990 (2010)

[13] Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S.: Rgb-d-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding 171, 118–139 (2018)

[14] Lee, Jungho, et al. "Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition." arXiv preprint arXiv:2208.10741 (2022).

[15] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In International Conference on Pattern Recognition Workshops and Challenges, pages 694–701. Springer, 2021.

[16] Zhou, Y., Li, C., Cheng, Z. Q., Geng, Y., Xie, X., Keuper, M. (2022). Hypergraph Transformer for Skeleton-based Action Recognition. arXiv preprint arXiv:2211.09590.

[17] Xin, Wentian, et al. "Transformer for Skeleton-based Action Recognition: A Review of Recent Advances." Neurocomputing (2023)

[18] @inproceedingsyu2018generative, title=Generative image inpainting with contextual attention, author=Yu, Jiahui and Lin, Zhe and Yang, Jimei and Shen, Xiaohui and Lu, Xin and Huang, Thomas S, booktitle=Proceedings of the IEEE conference on computer vision and pattern recognition, pages=5505–5514, year=2018

[19] Alsarhan T, Harfoushi O, Shdefat AY, Mostafa N, Alshinwan M, Ali A. Improved Graph Convolutional Network with Enriched Graph Topology Representation for Skeleton-Based Action Recognition. Electronics. 2023; 12(4):879. https://doi.org/10.3390/electronics12040879.

[20] Alsarhan, T., Lu, H. (2021). Collaborative Positional-Motion Excitation Module for Efficient Action Recognition. In: Pham, D.N., Theeramunkong, T., Governatori, G., Liu, F. (eds) PRICAI 2021: Trends in Artificial Intelligence. PRICAI 2021. Lecture Notes in Computer Science(), vol 13033. Springer, Cham.

[21] Alsarhan, T., Ali, U., Lu, H. (2022). Enhanced discriminative graph convolutional network with adaptive temporal modelling for skeleton-based action recognition. Computer Vision and Image Understanding, 216, 103348.

[22] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13359-13368).

[23] Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst. "Convolutional neural networks on graphs with fast localized spectral filtering." Advances in neural information processing systems 29 (2016).

[24] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5115–5124, 2017.

[25] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," IEEE Data Engineering Bulletin, vol. 40, no. 3, pp. 52–74, 2017.

[26] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[27] Lee, Jungho, et al. "Hierarchically Decomposed Graph Convolutional

Networks for Skeleton-Based Action Recognition." arXiv preprint arXiv:2208.10741 (2022).
[28] 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.
[29] 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.
[30] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in International Conference on Machine Learning, 2017, pp. 1263–1272
[31] Conor A Bradley. 2020. A statistical framework for rare disease diagnosis. NAT REV GENET 21, 1 (2020), 2–3.
[32] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673, 2019.
[33] You, J., Ying, R., Ren, X., Hamilton, W., and Leskovec, J. Graphrnn: Generating realistic graphs with deep autoregressive models. In International Conference on Machine Learning, 2018b.
[34] Zonghan Wu et al. "A comprehensive survey on graph neural networks". In: IEEE Transactions on Neural Networks and Learning Systems (2020).
[35] Palash Goyal and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey". In: Knowledge-Based Systems 151 (2018)
[36] ameer Agarwal, Kristin Branson, and Serge Belongie. 2006. Higher order learning with graphs. In Proceedings of the 23rd international conference on Machine learning (ICML '06). Association for Computing Machinery, New York, NY, USA, 17–24.
[37] Zhang, C., Hu, S., Tang, Z.G. amp; Chan, T.H.. (2017). Re-revisiting Learning on Hypergraphs: Confidence Interval and Subgradient Method.
[38] D. C. G. Pedronette, L. P. Valem, J. Almeida and R. da S. Torres, "Multimedia Retrieval Through Unsupervised Hypergraph-Based Manifold Ranking," in IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 5824-5838, Dec. 2019, doi: 10.1109/TIP.2019.2920526.
[39] R. Zass and A. Shashua, "Probabilistic Graph and Hypergraph Matching," in Proc. IEEE Conf. Computer Vision Pattern Recognition, 2008
[40] Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition
[41] L. Sun, S. Ji, and J. Ye, "Hypergraph Spectral Learning for Multi-Label Classification," in Proc. ACM SIG KDD, 2008.
[42] N. Werghi and M. Rahayem and J. Kjellander, "An ordered topological representation of 3D triangular mesh facial surface: concept and applications", EURASIP Journal on Advances on Signal Processing, 2012.
[43] N. Werghi, C. Tortorici, S. Berretti, A.D Bimbo "Local binary patterns on triangular meshes: Concept and applications", Computer Vision and Image Understanding, 2015, 139, 161–177
[44] Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems, pp. 1024–1034, 2017a
[45] Y. Huang, Q. Liu, and D. Metaxas, "Video Object Segmentation by Hypergraph Cut," in Proc. IEEE Conf. Computer Vision Pattern Recognition, 2009.
[46] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas, "Image Retrieval via Probabilistic Hypergraph Ranking," in Proc. IEEE Conf. Computer Vision Pattern Recognition, 2010.
[47] L. Sun, S. Ji, and J. Ye, "Hypergraph Spectral Learning for Multi-Label Classification," in Proc. ACM SIG KDD, 2008.
[48] Kang, Min-Seok, Dongoh Kang, and HanSaem Kim. "Efficient Skeleton-Based Action Recognition via Joint-Mapping Strategies." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[49] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In AAAI, pages 7444–7452, 2018
[50] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition
[51] 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.
[52] Osadchiy, Timur, et al. "Recommender system based on pairwise association rules." Expert Systems with Applications 115 (2019): 535-542.
[53] Humski, Luka, Damir Pintar, and Mihaela Vranić. "Exploratory analysis of pairwise interactions in online social networks." Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije 58.4 (2017): 422-428.
[54]] Gao, Y.; Wang, M.; Zha, Z.-J.; Shen, J.; Li, X.; and Wu, X. 2013. Visual-Textual Joint Relevance Learning for Tag-based Social Image Search. IEEE Transactions on Image Processing 22(1):363–376.
[55] Zhou, D.; Huang, J.; ¨ and Scholkopf, B. 2007. Learning with Hypergraphs: Clus- ¨ tering, Classification, and Embedding. In NIPS
[56] Feng, Yifan, et al. "Hypergraph neural networks." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.
[57] Y. Wang, L. Zhu, X. Qian and J. Han, "Joint Hypergraph Learning for Tag-Based Image Retrieval," in IEEE Transactions on Image Processing, vol. 27, no. 9, pp. 4437-4451, Sept. 2018, doi: 10.1109/TIP.2018.2837219.
[58] Q. Fang, J. Sang, C. Xu, and Y. Rui, "Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning," IEEE Trans. Multimedia, vol. 16, no. 3, pp. 796–812, Apr. 2014.
[59]] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action recognition
[60] D. Li, Z. Xu, S. Li, and X. Sun, "Link prediction in social networks based on hypergraph," in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 41–42.
[61] L.-E. Martinet et al., "Robust dynamic community detection with applications to human brain functional networks," Nat. Commun., vol. 11, no. 1, pp. 1–13, 2020.
[62] W. Zhao et al., "Learning to Map Social Network Users by Unified Manifold Alignment on Hypergraph," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 12, pp. 5834-5846, Dec. 2018, doi: 10.1109/TNNLS.2018.2812888.
[63] 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
[64] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2010, pp. 3376–3383
[65] Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Richly activated graph convolutional network for 507 robust skeleton-based action recognition. IEEE Transactions on Circuits and Systems for Video 508 Technology 2020, 31, 1915–1925.
[66] 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 [15] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In International Conference on Pattern Recognition Workshops and Challenges, pages 694–701. Springer, 2021
[67] Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action 359 recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern 360 Recognition (CVPR), 2015, pp. 1110–111
[68] Y. Huang, Q. Liu, and D. Metaxas, "Video object segmentation by hypergraph cut," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1738–1745.
[69] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," IEEE Trans. Image Process., vol. 21, no. 9, pp. 4290–4303, Sep. 2012.
[70] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In Proceedings of the IEEE International Conference on Computer Vision, pages 2117–2126, 2017.
[71] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, "Feature learning using spatial-spectral hypergraph discriminant analysis for hyperspectral image," IEEE Trans. Cybern., vol. 49, no. 7, pp. 2406–2419, Jul. 2019.

[72] Du, Dawei, Honggang Qi, Longyin Wen, Qi Tian, Qingming Huang, and Siwei Lyu. "Geometric hypergraph learning for visual tracking." IEEE transactions on cybernetics 47, no. 12 (2016): 4182-4195.

[73] 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.

[74] 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

[75] 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.

[76] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[77] 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.

[78] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20186–20196, 2022

[79] B. Xu, X. Shu and Y. Song, "X-Invariant Contrastive Augmentation and Representation Learning for Semi-Supervised Skeleton-Based Action Recognition," in IEEE Transactions on Image Processing, vol. 31, pp. 3852-3867, 2022, doi: 10.1109/TIP.2022.3175605.

[80] 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

[81] 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

[82] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. IEEE transactions on pattern analysis and machine intelligence, 41(8):1963–1978, 2019

[83] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018.

[84] Zou, Zhiming, et al. "Compositional Graph Convolutional Networks for 3D Human Pose Estimation." 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE, 2021.

[85] 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

[86] Wang, Xiaolong, et al. "Non-local neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[87]] 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

[88] 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

[89] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016

[90] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying WEI, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 12559–12571. Curran Associates, Inc., 2020.

[91] Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity 483 analysis. In Proceedings of the Proceedings of the IEEE conference on computer vision and 484 pattern recognition, 2016, pp. 1010–1019.

[92] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A largescale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10):2684–2701, 2019

[93] Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional 495 networks for skeleton-based action recognition. In Proceedings of the Proceedings of the 496 IEEE/CVF conference on computer vision and pattern recognition, 2019,

[94] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. Acm Transactions On Graphics (tog), 38(5):1–12, 2019

[95] Liu, Ziyu, et al. "Disentangling and unifying graph convolutions for skeleton-based action recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[96] Si, Chenyang, et al. "An attention enhanced graph convolutional lstm network for skeleton-based action recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[97] Wang, Jiang, et al. "Learning actionlet ensemble for 3D human action recognition." IEEE transactions on pattern analysis and machine intelligence 36.5 (2013): 914-927.

[98] Du, Yong, Wei Wang, and Liang Wang. "Hierarchical recurrent neural network for skeleton based action recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[99] Lee, Inwoong, et al. "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks." Proceedings of the IEEE international conference on computer vision. 2017.

[100] Chen, Yuxin, et al. "Channel-wise topology refinement graph convolution for skeleton-based action recognition." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[101] Wang, Jiang, et al. "Cross-view action modeling, learning and recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.