Partial Skeleton Visibility for Action Recognition: A Constrained Field-of-View Approach

Yingjie Dai[1], Tianyang Xu[1⋆], Yanglin Deng[1], Xiao-Jun Wu[1], and Josef Kittler[2]

[1] Jiangnan University, Wuxi, China
[2] University of Surrey, Guildford, UK
daiyingjie2024@gmail.com; {tianyang.xu,wu_xiaojun}@jiangnan.edu.cn; yanglin_deng@163.com; j.kittler@surrey.ac.uk;

Abstract. Skeleton-based action recognition has achieved remarkable success by exploiting joint coordinates and their topological connections, yet prevailing methods overwhelmingly assume complete and clean skeleton inputs. In real-world deployments, such as egocentric vision, crowded surveillance, wearable devices, or edge robotics, limited field-of-view (FoV) frequently causes substantial joint visibility dropout, leading to severe performance degradation that existing models are largely unprepared to handle. To bridge this critical yet underexplored gap, we introduce PartialVisGraph, a novel hypergraph framework tailored for robust skeleton action recognition under constrained FoV. We first construct highly expressive hypergraphs by introducing learnable virtual hyperedges that form a soft incidence matrix, capturing flexible high-order dependencies beyond conventional pairwise graphs. We then propose the Single-Head Sample-Adaptive Transformer, which adaptively aggregates joint features onto hyperedges while explicitly incorporating a visibility prior. This prior selectively gates information flow, preventing occluded or out-of-view joints from corrupting reliable feature propagation. We further establish rigorous evaluation protocols with realistic FoV simulation benchmarks on NTU RGB+D 60 and 120. Extensive experiments demonstrate that PartialVisGraph consistently achieves state-of-the-art accuracy under partial visibility, with gains of up to 68.8% on subsets with severe FoV restrictions compared to recent strong baselines, while remaining superior on full-visibility settings. Our approach offers a principled and practical pathway toward deployable skeleton-based action understanding in unconstrained environments. Resources related to the constrained FoV setting used in this work are available at: https://github.com/yaa1haa1/PartialVisGraph.

Keywords: Skeleton-based Action Recognition · Constrained Field of View · Hypergraph

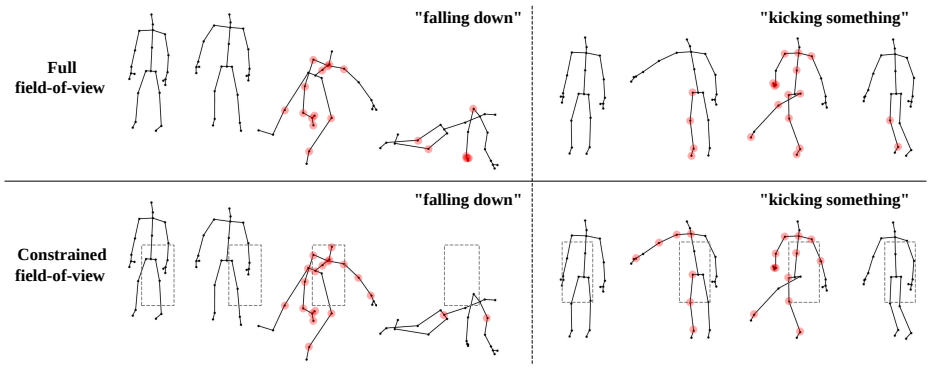


Fig. 1: Skeleton-based action recognition under constrained field-of-view. The dashed box indicates the restricted visible region, and joints whose features in the final layer contribute most to the model's prediction are highlighted in red.

1 Introduction

Recently, skeleton-based human action recognition [5, 29, 31, 38, 47, 52, 53] has demonstrated significant potential across a wide range of applications, including medical rehabilitation, virtual reality, intelligent surveillance, and autonomous driving. By representing human motion through joints and their topological structure, skeleton sequences provide a structured and geometry-aware description of body dynamics. This representation directly captures the kinematic relationships among joints and the temporal evolution of movements. Owing to its abstraction from appearance, skeleton data is inherently more robust to complex backgrounds, illumination variations, and viewpoint changes. Meanwhile, its compact and low-dimensional structure substantially reduces computational overhead. With the continuous advancement of depth sensors and 3D pose estimation algorithms [2, 11, 17, 26, 28], the accuracy and stability of skeleton acquisition have steadily improved, further accelerating the development of skeleton-based action recognition in both theoretical research and real-world deployment.

Looking ahead, within the emerging paradigm of embodied intelligence [37], reliable human action recognition is a prerequisite for natural human-machine interaction [27], for example, in service robotics and assistive devices. In embodied settings, sensors are often mounted on moving agents or constrained platforms, so camera observations are commonly partial and occluded. Methods [12,13,20,21,25,36,43] that assume complete skeletons are therefore ill-suited for real-world embodied deployment, motivating our work to explicitly address the constrained FoV setting.

From a methodological perspective, the human skeleton naturally forms a standard graph [42], where joints and their physical connections correspond to vertices and edges, respectively. Most existing approaches [7,9,10,14,16,19,41,45]

* Corresponding author.

are therefore designed upon this normal graph structure. They typically employ graph convolutions to exchange information among spatially connected joints within a single frame, and temporal convolutions to capture motion dynamics across frames. However, in a normal graph, each edge connects only two vertices, and information propagates primarily along physical connections [34]. As a result, modelling relationships between physically distant joints (*e.g.*, the two hands during clapping) can be suboptimal. In reality, human actions are jointly defined by multiple body parts [39]. Beyond pairwise interactions, many actions involve higher-order relationships among several joints (*e.g.*, standing up requires coordinated movement of the legs, spine, and other joints). Under constrained FoV conditions, it becomes even more critical to reason about information flow from a multi-joint perspective. Joint modelling over multiple joints may enable the aggregation of weak and spatially distributed cues into discriminative composite features.

To address these challenges, we adopt hypergraphs for skeleton-based action recognition under constrained FoV. Only a few prior works [49,50] have explored hypergraphs for this task, and these existing methods often have limitations in hypergraph construction. To this end, we propose a novel hypergraph construction strategy and introduce a Single-Head Sample-Adaptive Transformer (SHSAT) to aggregate node features onto hyperedges. In addition, we explicitly constrain information propagation to better suit constrained FoV scenarios.

As illustrated in Fig. 1, our method continues to extract useful discriminative information even under a constrained FoV. For example, in the "kicking something" sample, the model perceives the right arm lifting and moving out of view, a motion used to maintain balance during a kick, and it is able to extract additional information from the otherwise weak right-arm joint signals. In summary, our contributions are as follows:

- We are the first to introduce the task of skeleton-based action recognition under constrained FoV, which relaxes the deployment constraints of action recognition systems and provides a practical solution for potential field-of-view limitations in future embodied intelligence scenarios.
- We propose a novel hypergraph construction scheme and design a corresponding feature aggregation method tailored to this formulation.
- Our model achieves superior performance compared with state-of-the-art methods under constrained FoV settings. In addition, when trained and evaluated under full FoV conditions, it also surpasses existing approaches.

2 Related Works

2.1 GCN-Based Approaches

Since the physical topology of human joints and bones can be naturally represented as a graph, Graph Convolutional Networks (GCNs) [18] have been widely adopted for skeleton-based action recognition. [42] was the first to introduce GCNs to jointly model spatial and temporal features, demonstrating their

strong performance for action recognition. Subsequent studies observed that a fixed topology defined by natural anatomical connections is inherently limited. As a result, most later works [4, 6, 19, 22, 44, 51] focused on learning adaptive graph structures. [34] incorporated adaptive graph convolution to dynamically learn the underlying topology, achieving superior performance compared to earlier methods.

However, these approaches typically model relationships between joints as strictly pairwise interactions. In reality, human actions arise from coordinated movements of multiple joints, which entail not only binary dependencies between joint pairs but also complex higher-order interactions among groups of joints.

2.2 Hypergraph-Based Approaches

Recognising that pairwise connections are insufficient to capture collaborative interactions among multiple joints, recent studies have introduced hypergraphs [1, 15] into skeleton-based action recognition. [50] constructs a hypergraph and proposes a hypergraph-based self-attention mechanism (HyperSA) to explicitly model high-order dependencies among joints, thereby overcoming the limitation of normal graph convolution, which is restricted to pairwise modelling. Similarly, [49] adaptively optimises the hypergraph structure during training, enabling joint-level feature aggregation and relational modelling at a higher semantic level, and dynamically uncovering action-driven high-order correlations.

Despite their promising performance, these methods primarily target action recognition with complete skeleton data, overlooking skeleton inputs under constrained FoV conditions that commonly arise in real-world scenarios and future embodied intelligence applications. Moreover, their hypergraph construction comes with structural constraints: [50] restricts each node to belong to only one hyperedge, while [49] requires the number of hyperedges to equal the number of joints.

2.3 Occluded Human Action Recognition

Occluded skeleton-based human action recognition has attracted increasing attention in recent years. [35] designs a multi-branch architecture tailored to different occlusion scenarios and integrates multi-scale motion features, effectively improving recognition performance under occluded skeleton conditions. [8] addresses performance degradation caused by occlusion or noise by introducing a dual inhibition training strategy. Through simulating key parts and random occlusions, combined with global-local part modelling, the method significantly enhances robustness to incomplete skeleton inputs.

Although these approaches improve robustness to occlusion to a certain extent, they still exhibit limitations in terms of architectural complexity and generalisation capability. More importantly, their focus is on occlusion-specific scenarios rather than skeleton data under constrained FoV settings, which fundamentally distinguishes them from our work.

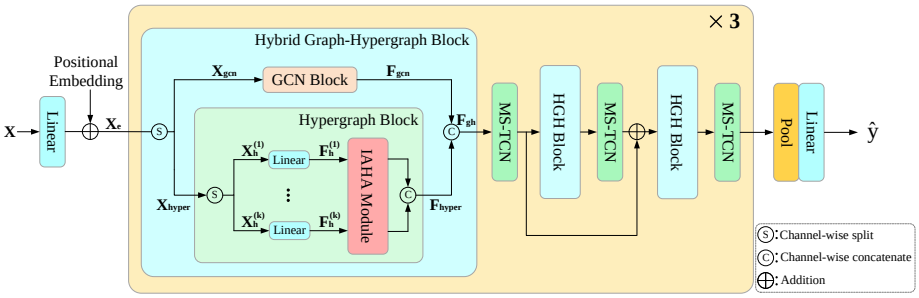


Fig. 2: Overview of our PartialVisGraph approach

3 Method

3.1 Preliminaries

Graph Convolutional Network. In skeleton-based action recognition, the human body is naturally represented as a graph, where joints are treated as nodes and bones as edges. Let $\mathbf{A} \in \mathbb{R}^{N\times N}$ denote the adjacency matrix of the skeleton graph with N joints. Following the standard formulation, we first add self-connections to obtain $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The corresponding degree matrix is defined as $\tilde{\mathbf{D}} \in \mathbb{R}^{N\times N}$ with diagonal entries $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. Given the input feature matrix $\mathbf{X} \in \mathbb{R}^{N\times C_{\text{in}}}$, a graph convolutional layer can be written as

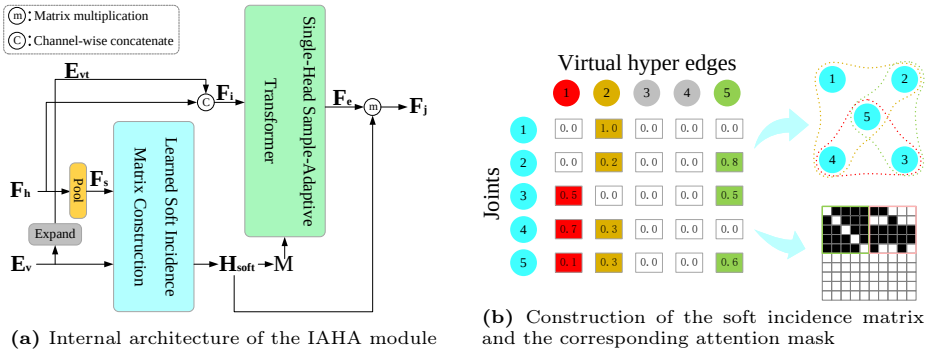
$$\mathbf{F} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}\right), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{C_{\text{in}}\times C_{\text{out}}}$ is a learnable weight matrix for feature transformation, and $\sigma(\cdot)$ denotes a nonlinear activation function.

Hypergraph. A hypergraph is a generalised form of a graph that allows each edge to connect an arbitrary number of vertices. Formally, a hypergraph is defined as $\mathcal{G}_h = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a finite set of vertices and $\mathcal{E}$ is a set of non-empty subsets of $\mathcal{V}$, referred to as hyperedges. Unlike normal graphs, where each edge connects exactly two vertices, a hyperedge can connect any number of vertices, enabling the modelling of higher-order relationships.

Analogous to the adjacency matrix in a standard graph, a hypergraph is commonly represented by an incidence matrix $\mathbf{H} \in \mathbb{R}^{N\times K}$, where $N = |\mathcal{V}|$ denotes the number of vertices and $K = |\mathcal{E}|$ denotes the number of hyperedges. The entries of the incidence matrix are defined as

$$\mathbf{H}_{v,e} = \begin{cases} 1, & \text{if } v \in e, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$



(a) Internal architecture of the IAHA module

(b) Construction of the soft incidence matrix and the corresponding attention mask

Fig. 3: Incidence-aware hypergraph attention module

3.2 Overview

The overall architecture of our method is illustrated in Fig. 2. Given a skeleton sequence $\mathbf{X} \in \mathbb{R}^{T,V,C_r}$ as input, where T denotes the number of frames, V the number of joints per frame, and C_r the number of input channels, we first project the low-dimensional raw skeleton data into a higher-dimensional feature space through a linear layer. The positional embedding is then added to the projected features to preserve structural and temporal information, resulting in the embedded representation $\mathbf{X}_e \in \mathbb{R}^{T,V,C_f}$, where C_f denotes the feature dimension after projection. The embedded features are subsequently fed into a Hybrid Graph-Hypergraph block (HGH block), which models spatial relationships among joints within each frame. To capture temporal dependencies across multiple frames, we employ a Multi-Scale Temporal Convolutional Network (MS-TCN) [24]. After stacking multiple HGH blocks and MS-TCN layers, the final feature representation is subjected to temporal pooling followed by a linear projection to produce the prediction output $\hat{\mathbf{y}} \in \mathbb{R}^{N_{\text{class}}}$, where N_{class} denotes the number of action categories.

3.3 Hybrid Graph-Hypergraph Block

To exploit the inherent topology of the human skeleton and thus provide a useful inductive bias, the HGH block splits the input $\mathbf{X}_e$ along the channel dimension into two parts, namely $\mathbf{X}_{\text{gcn}}$ with $C_f/4$ channels and $\mathbf{X}_{\text{hyper}}$ with $3C_f/4$ channels, and processes them with a GCN branch and a hypergraph branch respectively.

The GCN branch provides the explicit structural inductive bias. It treats the skeleton as a graph with adjacency $\mathbf{A} \in \mathbb{R}^{V \times V}$ and performs graph convolution, producing spatially-aware joint features $\mathbf{F}_{\text{gcn}} = \text{GCN}(\mathbf{X}_{\text{gcn}}, \mathbf{A})$.

The hypergraph branch first evenly splits $\mathbf{X}_{\text{hyper}}$ into k channel groups $\{\mathbf{X}_{\text{h}}^{(i)}\}_{i=1}^{k}$, each of shape $\mathbb{R}^{T \times V \times C_v}$ with $C_v = 3C_f/(4k)$. Each group is linearly projected and fed to the Incidence-Aware Hypergraph Attention module (IAHA module), which is responsible for constructing the hypergraph in a data-driven

manner, aggregating joint information onto hyperedges, and performing feature transformation via attention-style operations. The outputs of the IAHA module are concatenated to form $\mathbf{F}_{\text{hyper}}$.

Finally, we concatenate $\mathbf{F}_{\text{gcn}}$ and $\mathbf{F}_{\text{hyper}}$ along the channel dimension to obtain the block output $\mathbf{F}_{\text{gh}} \in \mathbb{R}^{T\times V\times C_{\text{gh}}}$, which is passed to subsequent modules.

3.4 Incidence-Aware Hypergraph Attention Module

As shown in Fig. 3a, the IAHA module introduces a set of learnable virtual hyperedge tokens $\mathbf{E}_v \in \mathbb{R}^{K_v\times C_v}$, where K_v denotes the number of virtual hyperedges (we set $K_v = 10$). Given an input feature $\mathbf{F}_h \in \mathbb{R}^{T\times V\times C_v}$, we first aggregate information along the temporal axis using average pooling to obtain a joint-level summary $\mathbf{F}_s = \text{Pool}_t(\mathbf{F}_h) \in \mathbb{R}^{V\times C_v}$, and then construct a soft incidence matrix $\mathbf{H}_{\text{soft}} \in [0,1]^{V\times K_v}$ jointly from the joint summary $\mathbf{F}_s$ and the virtual hyperedge bank $\mathbf{E}_v$. Unlike a conventional incidence matrix whose entries are binary, the elements of $\mathbf{H}_{\text{soft}}$ are continuous membership weights in the interval $[0,1]$. To limit computational cost, the same $\mathbf{H}_{\text{soft}}$ is shared across all frames of a given sample.

Next, $\mathbf{H}_{\text{soft}}$ is converted into an attention mask $\mathbf{M} \in \{0,1\}^{L\times L}$ with $L = K_v + V$, which encodes allowable attention links between the K_v virtual hyperedge tokens and the V joint tokens. This mask is applied during attention computation so that each joint token is allowed to transmit information only to the virtual hyperedge tokens it belongs to, ensuring that feature aggregation follows the learned hyperedge assignments.

To perform joint-hyperedge attention per frame, we replicate the virtual-token bank across the temporal dimension to form $\mathbf{E}_{vt} \in \mathbb{R}^{T\times K_v\times C_v}$ (*i.e.*, $\mathbf{E}_{vt}$ contains T identical copies of $\mathbf{E}_v$), and then concatenate the token sets to produce the attention input $\mathbf{F}_i = \text{concat}\left(\mathbf{E}_{vt},\ \mathbf{F}_h\right) \in \mathbb{R}^{T\times(K_v+V)\times C_v}$. A Single-Head Sample-Adaptive Transformer (SHSAT) is applied to $\mathbf{F}_i$ using the mask $\mathbf{M}$. From the attention output, we retain only the first K_v tokens (the updated virtual hyperedge tokens) to obtain $\mathbf{F}_e \in \mathbb{R}^{T\times K_v\times C_v}$.

Finally, the hyperedge features are redistributed to the joint domain using the soft incidence matrix. For each frame t, we compute $\mathbf{F}_j^{(t)} = \mathbf{H}_{\text{soft}}\,\mathbf{F}_e^{(t)} \in \mathbb{R}^{V\times C_v}$, and aggregating over all frames yields the module output $\mathbf{F}_j \in \mathbb{R}^{T\times V\times C_v}$.

3.5 Learned Soft Incidence Matrix Construction

To represent and construct the hypergraph more flexibly than prior work, we propose to build a learned soft incidence matrix by pairing a bank of virtual hyperedge tokens $\mathbf{E}_v \in \mathbb{R}^{K_v\times C_v}$ with the temporally pooled joint summaries $\mathbf{F}_s \in \mathbb{R}^{V\times C_v}$. An illustrative example for $V = 5$ and $K_v = 5$ is shown in Fig. 3b.

Concretely, we compute the cosine similarity between each joint token and each virtual hyperedge token to obtain

$$\mathbf{S} \in \mathbb{R}^{V\times K_v}, \qquad S_{j,k} = \frac{\langle \mathbf{F}_s^{(j)},\ \mathbf{E}_v^{(k)} \rangle}{\|\mathbf{F}_s^{(j)}\|\,\|\mathbf{E}_v^{(k)}\|}, \tag{3}$$

where $\mathbf{F}_s^{(j)}$ is the pooled feature of joint j and $\mathbf{E}_v^{(k)}$ is the k-th virtual hyperedge token. We then apply a row-wise sparse normalisation to induce sparsity and produce the soft incidence matrix, $\mathbf{H}_{\text{soft}} = \text{sparsemax}_{\text{row}}(\mathbf{S}) \in [0,1]^{V \times K_v}$, so that each row sums to one and entries are in $[0,1]$. The sparsity encourages compact joint-hyperedge memberships and reduces spurious associations.

In this formulation, a nonzero entry $\mathbf{H}_{\text{soft}}(j,k) > 0$ indicates that joint j is assigned to hyperedge $\mathbf{E}_v^{(k)}$. Conversely, a column of $\mathbf{H}_{\text{soft}}$ that is identically zero denotes an absent hyperedge, which is then ignored during subsequent hyperedge-to-joint feature redistribution.

3.6 Single-Head Sample-Adaptive Transformer

The SHSAT is a single-head transformer encoder [40] whose attention mask is generated per-sample from the soft incidence matrix $\mathbf{H}_{\text{soft}}$. The mask enforces that joint features are aggregated onto their corresponding hyperedge tokens. Therefore, when a joint is not assigned to a given virtual hyperedge, it is masked out during attention. Fig. 3b shows a simple example of the mask construction. Since SHSAT uses a single layer and we retain only the first K_v output tokens (the updated virtual hyperedge tokens) for downstream use, only the first K_v rows of the mask are essential, and the remaining rows are inconsequential.

We also impose a visibility prior bias to suppress information flow from invisible joints. Let $\mathbf{vis} \in [0,1]^{T \times V}$ denote the joint visibility over an entire sequence, where T is the number of frames and V is the number of joints, with entries equal to 1 for visible joints and 0 for invisible ones. To limit contributions from invisible joints, we compute a per-entry bias by taking the natural logarithm after adding a small stability constant $\varepsilon > 0$, *i.e.*, $\mathbf{bias} = \log(\mathbf{vis} + \varepsilon)$. This bias is added to the attention logits before the softmax so that joints with low visibility receive negligible attention. It is worth noting that the entries of **vis** may take non-integer values. As the network depth increases, the MS-TCN progressively downsamples the temporal dimension, and the visibility map is downsampled accordingly to maintain alignment with the feature representation. This temporal downsampling may result in fractional visibility values.

In practice, the masked, visibility-biased single-head attention is computed as

$$\text{Attention}(Q,K,V) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d}} + \mathbf{M} + \mathbf{B}_{\text{vis}}\Big)V, \tag{4}$$

where Q, K, and V are the query, key, and value matrices derived from linear projections of the input tokens, and $\mathbf{B}_{\text{vis}}$ contains the log-visibility terms broadcast to the corresponding joint-to-hyperedge logits.

3.7 Training Strategy

Curriculum learning. We adopt curriculum learning [3] to gradually expose the model to harder visibility conditions. Samples with most joints inside the FoV are treated as easy, while samples with most joints outside the FoV are

treated as hard. Training starts with easy samples and progressively reduces the observable FoV to synthesise harder samples as training proceeds.

Temporal CutMix. During training, we apply temporal CutMix [46] with probability 0.3. Given two samples a and b, we randomly select a contiguous segment of length t from a and a complementary segment of length $T - t$ from b, and concatenate them to form a mixed sample. The cross-entropy losses computed from the model logits on the mixed sample with respect to a's label and b's label are denoted as $\mathcal{L}_a$ and $\mathcal{L}_b$, respectively. The final loss for the mixed sample is the length-weighted combination:

$$\mathcal{L}_{\text{mix}} \;=\; \frac{t}{T}\mathcal{L}_a \;+\; \frac{T-t}{T}\mathcal{L}_b. \tag{5}$$

3.8 Training Loss

We train the model using the standard cross-entropy loss:

$$\mathcal{L}_{\text{ce}} = -\sum_{c=1}^{N_{\text{class}}} y_c \log \hat{y}_c, \tag{6}$$

where $\hat{y_c}$ denotes the predicted logits after softmax, y_c is the one-hot ground-truth label, and N_{class} is the number of action classes.

In addition to $\mathcal{L}_{\text{ce}}$, we design dedicated loss terms for the virtual hyperedge tokens $\mathbf{E}_v$ and the soft incidence matrix $\mathbf{H}_{\text{soft}}$ to stabilise hypergraph construction and encourage meaningful hyperedge-to-joint assignments.

Hyperedge Diversity Loss. Let the virtual-hyperedge bank be $\mathbf{E}_v \in \mathbb{R}^{K_v \times C_v}$ (denoted as a pool P of K_v vectors of dimension C_v). To encourage diversity among the K_v vectors in the pool and avoid redundant hyperedges, we apply an orthogonality loss $\mathcal{L}_{\text{pool}}$. For a single pool $P \in \mathbb{R}^{K_v \times C_v}$, first normalise each row to unit ℓ_2 length (with a small constant ε for stability) and form the Gram matrix

$$\tilde{P}_i = \frac{P_i}{\|P_i\|_2 + \varepsilon}, \qquad \mathbf{G} = \tilde{P}\,\tilde{P}^\top \in \mathbb{R}^{K_v \times K_v}. \tag{7}$$

The orthogonality loss for this pool is the mean squared off-diagonal entry of $\mathbf{G}$,

$$\mathcal{L}_{\text{pool}}(P) \;=\; \frac{1}{K_v(K_v-1)} \sum_{i \neq j} G_{ij}^2 \;=\; \frac{\|\mathbf{G} - \mathbf{I}\|_F^2}{K_v(K_v-1)}. \tag{8}$$

Assignment Regularisation Loss. Let $\mathbf{H}_{\text{soft}} \in \mathbb{R}^{V \times K_v}$ denote the soft incidence matrix for one sample, where V is the number of joints and K_v is the number of virtual hyperedges. We introduce a composite assignment loss $\mathcal{L}_{\text{assign}}$ that prevents degenerate and overly imbalanced assignments.

First, compute column sums over joints, $s_k = \sum_{j=1}^{V} (\mathbf{H}_{\text{soft}})_{j,k}$, $\mathbf{s} = (s_1, \ldots, s_{K_v})$, and normalise to obtain a distribution over edges

$$\mathbf{p} = \frac{\mathbf{s}}{\sum_{k=1}^{K_v} s_k + \varepsilon}. \tag{9}$$

The loss comprises two terms:

1. **Batch-level balance.** Average $\mathbf{p}$ over the batch to get $\bar{\mathbf{p}}$, and penalise deviation from the uniform distribution $\mathbf{u} = (1/K_v, \ldots, 1/K_v)$:

$$\mathcal{L}_{\text{balance}} \;=\; \frac{1}{K_v} \sum_{k=1}^{K_v} \big(\bar{p}_k - u_k\big)^2. \tag{10}$$

2. **Per-sample max-hinge.** For each sample, take the maximum assignment probability $\max_k p_k$ and apply a hinge loss with threshold τ :

$$\mathcal{L}_{\text{hinge}} \;=\; \mathbb{E}\big[\max\big(0,\; \max_k p_k - \tau\big)\big], \tag{11}$$

where the expectation is over samples.

The combined assignment loss is the weighted sum

$$\mathcal{L}_{\text{assign}} \;=\; w_{\text{balance}} \mathcal{L}_{\text{balance}} \;+\; w_{\text{hinge}} \mathcal{L}_{\text{hinge}}, \tag{12}$$

where w_{balance} and w_{hinge} denote the corresponding weighting coefficients. This regulariser encourages balanced and non-collapsed hyperedge assignments while allowing inactive (all-zero) columns to occur.

Class-Centre Clustering Loss. We project the flattened soft incidence matrix of each sample into a compact embedding and denote the resulting per-sample features by $\mathbf{z}_n \in \mathbb{R}^C$. The loss enforces that $\mathbf{z}_n$ clusters around learnable class centres while keeping distinct class centres separated. Centres are maintained as parameters $\mathbf{c}_k \in \mathbb{R}^C$ for each class k. The loss has two terms:

$$\mathcal{L}_{\text{pull}} \;=\; \frac{1}{N} \sum_{n=1}^{N} \big\|\mathbf{z}_n - \mathbf{t}_n\big\|_2^2, \tag{13}$$

where N denotes the number of samples in the mini-batch. The target $\mathbf{t}_n$ is the (possibly mixed) centre corresponding to the sample label: $\mathbf{t}_n = \lambda\, \mathbf{c}_{k_a} + (1-\lambda)\, \mathbf{c}_{k_b}$ to handle CutMix, where λ denotes the proportion of sample a in the mixed input (and $\lambda = 1$ when CutMix is not applied).

The repel term pushes different class centres apart. We compute pairwise centre distances and apply a hinge repulsion:

$$\mathcal{L}_{\text{repel}} \;=\; \frac{1}{\binom{N_{\text{class}}}{2}} \sum_{i<j} \text{ReLU}\big(\Delta - \|\mathbf{c}_i - \mathbf{c}_j\|\big)^2. \tag{14}$$

Here Δ is the repel margin.

The combined loss is

$$\mathcal{L}_{\text{cluster}} \;=\; \mathcal{L}_{\text{pull}} \;+\; \gamma\, \mathcal{L}_{\text{repel}}, \tag{15}$$

with repel weight γ.

4 Experiments

4.1 Datasets

NTU RGB+D. The NTU RGB+D dataset [32] contains 60 action categories, including daily activities such as drinking water, eating, brushing teeth, and throwing objects, with a total of 56,880 video samples. The data were collected from 40 subjects under 155 camera viewpoints using the Kinect v2 sensor, which simultaneously captured RGB, infrared, depth, and 3D skeleton sequences. Two standard evaluation protocols are provided, namely Cross-Subject (X-Sub 60) and Cross-View (X-View 60). Among the 60 action classes, only 11 correspond to two-person interaction actions, while the majority are single-person activities.

NTU RGB+D 120. NTU RGB+D 120 [23] is an extended version of NTU RGB+D, expanding the number of action categories to 120 and increasing the dataset size to 114,480 video samples. The data cover 106 subjects and 155 camera viewpoints. The introduced categories include more complex daily behaviours and sports-related activities, such as wearing headphones, shooting a basketball, and moving heavy objects. This dataset adopts Cross-Subject (X-Sub 120) and Cross-Setup (X-Set 120) evaluation protocols, placing greater emphasis on generalisation across different acquisition environments and device configurations. Among the 120 action classes, approximately 26 involve two-person interactions, while the remaining classes are predominantly single-person actions.

4.2 Experimental Settings

Due to the difficulty of collecting diverse human activity data [30], we simulate constrained FoV scenarios by applying a FoV window to complete skeleton sequences. The generation of constrained FoV scenarios involves two key factors: the selection of the FoV centre and the size of the FoV window.

To define the FoV centre, we select a joint from the first frame as the centre point. Since the choice of centre largely determines which body parts remain visible, we construct three types of test data by selecting representative joints, namely the head joint, the spine joint, and the base-of-spine joint, as the FoV centre. In addition, we create a fourth test setting where the centre joint is randomly sampled per sample, with detailed results provided in the supplementary material. The size of the FoV window directly determines how many joints remain visible. To systematically evaluate the model under varying visibility constraints, we design three levels of difficulty: easy, medium, and hard. A smaller window corresponds to a more challenging scenario. In the easy setting, approximately 75% of the joints in a skeleton sequence are visible. In the medium and hard settings, only about 50% and 25% of the joints are visible, respectively.

4.3 Implementation Details

During training, we feed complete skeleton sequences to the model with a probability of 0.1, while randomly generated samples under constrained FoV are used

Table 1: Performance comparison on NTU RGB+D under constrained FoV settings with different FoV centres. Results are reported on easy, medium, and hard splits using X-Sub and X-View protocols. Bold font denotes the best result, and underline denotes the second-best result.

FoV centre	Method	Easy (%)		Medium (%)		Hard (%)	
		X-Sub	X-View	X-Sub	X-View	X-Sub	X-View
Head (3)	FR-Head [48]	83.2	88.2	57.3	62.7	26.5	25.7
	HD-GCN [19]	81.9	85.9	53.4	58.7	23.5	24.7
	Hyper-GCN [49]	<u>85.4</u>	<u>91.3</u>	<u>65.9</u>	<u>72.4</u>	<u>31.6</u>	<u>34.4</u>
	Hyperformer [50]	84.5	89.5	64.3	69.2	29.4	32.9
	PartialVisGraph	**90.9**	**95.8**	**88.5**	**93.2**	**79.2**	**84.5**
Spine (20)	FR-Head [48]	80.6	85.3	51.3	54.7	21.7	20.5
	HD-GCN [19]	78.5	82.7	46.4	51.2	18.8	17.7
	Hyper-GCN [49]	<u>82.1</u>	<u>89.0</u>	<u>58.5</u>	<u>64.3</u>	<u>24.1</u>	<u>27.8</u>
	Hyperformer [50]	<u>82.1</u>	87.4	56.7	61.9	23.7	27.2
	PartialVisGraph	**90.6**	**95.4**	**87.1**	**92.1**	**77.7**	**83.2**
Base of the spine (0)	FR-Head [48]	61.9	65.9	33.8	34.6	9.0	7.2
	HD-GCN [19]	57.3	59.1	30.8	29.9	7.2	6.6
	Hyper-GCN [49]	65.1	<u>74.3</u>	34.1	38.3	7.3	7.7
	Hyperformer [50]	<u>68.8</u>	73.5	<u>40.0</u>	<u>44.1</u>	<u>10.8</u>	<u>9.5</u>
	PartialVisGraph	**89.2**	**94.5**	**83.8**	**89.1**	**71.4**	**76.5**

with a probability of 0.9. As curriculum learning is adopted, the visibility ratio of these samples is gradually adjusted. In the first epoch, the visibility ratio is set between 95% and 100%. After 70 epochs, it progressively decreases to a range of 20% to 80%, and remains unchanged thereafter. The model is trained for 200 epochs in total, with the first 5 epochs serving as a warm-up. The initial learning rate is set to 0.05 and is multiplied by 0.1 at the 110th and 120th epochs. We optimise the model using stochastic gradient descent (SGD) with Nesterov momentum of 0.9 and weight decay of 0.0004. The batch size is set to 64.

4.4 Experimental Results

Experiments under Constrained Field-of-View. Multi-stream ensemble methods are widely adopted to boost performance in prior works [10, 33, 51], and in our experiments, the ensemble remains beneficial under constrained FoV, Tab. 1 and Tab. 2 report results using the commonly used 4-stream ensemble. Results show that, regardless of which joint is selected as the FoV centre and regardless of the severity of the constrained FoV, PartialVisGraph consistently performs well. In particular, under the hard difficulty level, almost all competing methods collapse, whereas PartialVisGraph maintains stable performance.

Experiments under Full Field-of-View. We train and evaluate PartialVisGraph on full FoV skeleton sequences, and for fair comparison, we adopt the widely used 4-stream ensemble strategy. Tab. 3 presents a comprehensive comparison on NTU RGB+D and NTU RGB+D 120 against a large set of benchmark methods. PartialVisGraph outperforms all compared approaches, demon-

Table 2: Performance comparison on NTU RGB+D 120 under constrained FoV settings with different FoV centres. Results are reported on easy, medium, and hard splits using X-Sub and X-Set protocols. Bold font denotes the best result, and underline denotes the second-best result.

FoV centre	Method	Easy (%)		Medium (%)		Hard (%)	
		X-Sub	X-Set	X-Sub	X-Set	X-Sub	X-Set
Head (3)	FR-Head [48]	74.2	75.5	45.9	46.0	12.7	14.5
	HD-GCN [19]	74.3	75.0	44.6	45.5	12.2	14.4
	Hyper-GCN [49]	78.6	81.8	52.4	54.2	12.7	16.8
	Hyperformer [50]	79.0	80.3	57.1	55.1	21.0	19.8
	PartialVisGraph	**87.8**	**89.2**	**84.1**	**85.9**	**72.9**	**74.4**
Spine (20)	FR-Head [48]	69.2	69.8	40.3	39.9	10.0	10.9
	HD-GCN [19]	68.5	68.9	38.7	38.3	9.2	10.9
	Hyper-GCN [49]	74.4	77.4	43.8	44.0	9.1	12.1
	Hyperformer [50]	75.5	76.1	50.5	47.7	16.6	15.9
	PartialVisGraph	**87.2**	**88.6**	**82.1**	**83.9**	**71.5**	**73.1**
Base of the spine (0)	FR-Head [48]	53.1	47.6	29.0	23.9	4.3	3.9
	HD-GCN [19]	51.3	47.4	24.8	22.3	2.8	2.7
	Hyper-GCN [49]	58.6	58.7	24.6	22.6	1.6	2.0
	Hyperformer [50]	63.8	60.5	32.6	30.0	4.7	3.4
	PartialVisGraph	**85.3**	**86.7**	**79.2**	**80.9**	**66.7**	**67.4**

strating its effectiveness and superiority even with complete skeleton observations, and showing that our learnable hypergraph construction and feature aggregation remain advantageous on full skeleton inputs.

4.5 Ablation Study

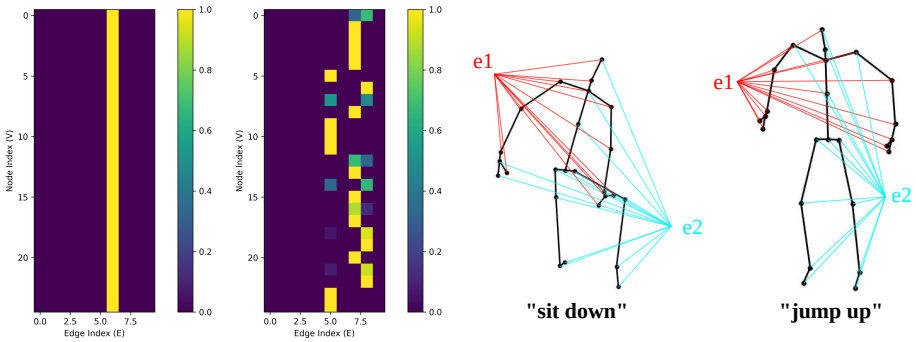
All ablation experiments are evaluated on the joint modality. Since temporal CutMix, $\mathcal{L}_{\text{pool}}$, $\mathcal{L}_{\text{assign}}$ and $\mathcal{L}_{\text{cluster}}$ apply to both constrained and full FoV data, Tab. 4 reports their effects under full FoV. The constrained FoV results shown are averaged across the three difficulty levels.

Table 3: Performance comparison on NTU RGB+D and NTU RGB+D 120 under full FoV settings. For fair comparison, all methods adopt a 4-stream ensemble. Bold font denotes the best result, and underline denotes the second-best result.

Method	Publication	Params (M)	GFLOPs	NTU RGB+D (%)		NTU RGB+D 120 (%)	
				X-Sub	X-View	X-Sub	X-Set
CTR-GCN [6]	ICCV 2021	1.5	1.97	92.4	96.4	88.9	90.6
Hyperformer [50]	arXiv 2022	2.6	–	92.9	96.5	89.9	91.3
FR-Head [48]	CVPR 2023	2.0	–	92.8	96.8	89.5	90.9
HD-GCN [19]	ICCV 2023	1.7	1.77	93.0	97.0	89.8	91.2
SkateFormer [12]	ECCV 2024	2.0	3.62	93.5	**97.8**	89.8	91.4
BlockGCN [51]	CVPR 2024	1.3	1.63	93.1	97.0	90.3	91.5
ProtoGCN [22]	CVPR 2025	–	–	93.5	97.5	90.4	91.9
Hyper-GCN [49]	ICCV 2025	2.3	2.88	93.7	**97.8**	90.9	92.0
PartialVisGraph	–	4.0	2.90	**93.8**	97.6	**91.0**	**92.3**

Table 4: Ablation results under full FoV and constrained FoV settings

Full FoV (%)		Constrained FoV (%)	
PartialVisGraph	92.0	PartialVisGraph	78.8
Wo $\mathcal{L}_{\text{pool}}$	91.8	Wo $\mathbf{B}_{\text{vis}}$	78.4
Wo $\mathcal{L}_{\text{assign}}$ and $\mathcal{L}_{\text{cluster}}$	91.2	Wo Curriculum learning	78.4
Wo CutMix	91.0	–	–



(a) Soft incidence matrix without $\mathcal{L}_{\text{assign}}$

(b) Soft incidence matrix with $\mathcal{L}_{\text{assign}}$

(c) Constructed hypergraph of our model

Fig. 4: Visualisation of the hypergraph

Removing $\mathcal{L}$pool leads to a 0.2% drop in accuracy, indicating that promoting diversity among hyperedges contributes to performance gains. When $\mathcal{L}$assign and $\mathcal{L}_{\text{cluster}}$ are removed jointly, accuracy decreases by 0.8%. As shown in Fig. 4a, the model then collapses to a single hyperedge covering all joints, confirming that these regularisation terms prevent assignment collapse. Further insight is provided in Fig. 4c, which visualises hypergraphs constructed in intermediate layers. Here, $e1$ and $e2$ denote two hyperedges connected to their member joints. The hypergraph construction realises a structural reorganisation of the skeleton. As illustrated by the two samples, they group joints into semantically coherent parts (one group comprising the legs and spine and another comprising both arms). This reorganisation provides a structured basis for subsequent feature aggregation across related joints.

From Tab. 4, we further observe that omitting temporal CutMix reduces accuracy by 1%, suggesting that temporal CutMix contributes positively to the model's performance. Under constrained FoV, disabling the visibility prior lowers accuracy by 0.4%, highlighting the importance of the visibility prior. Finally, curriculum learning yields a 0.4% accuracy gain, indicating that progressively introducing more challenging visibility conditions benefits the model's performance on the test set.

5 Conclusion

In this paper, we investigated skeleton-based human action recognition under constrained FoV, and proposed a novel hypergraph-based framework that learns soft incidence matrices via a bank of virtual hyperedge tokens, aggregates joint features onto hyperedges with a Single-Head Sample-Adaptive Transformer (SHSAT), and applies a visibility prior to restrict information flow from invisible joints. To ensure a thorough evaluation, we design multiple experimental protocols that mimic varying degrees of view restriction and benchmark our method against recent state-of-the-art approaches. Empirical results show that PartialVisGraph achieves leading performance across these protocols. Moreover, when trained and evaluated under the full FoV setting, it also outperforms all compared methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62576152, 62332008), the Basic Research Program of Jiangsu (BK20250104), and the Fundamental Research Funds for the Central Universities (JUSRP202504007).

References

1. Bai, S., Zhang, F., Torr, P.H.: Hypergraph convolution and hypergraph attention. Pattern Recognition **110**, 107637 (2021)
2. Banik, S., Gschoßmann, P., García, A.M., Knoll, A.: Occlusion robust 3d human pose estimation with stridedposegraphformer and data augmentation. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2023)
3. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
4. Chan, W., Tian, Z., Wu, Y.: Gas-gcn: Gated action-specific graph convolutional networks for skeleton-based action recognition. Sensors **20**(12), 3499 (2020)
5. Chen, Y., Guo, J., Guo, S., Tao, D.: Neuron: Learning context-aware evolving representations for zero-shot skeleton action recognition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8721–8730 (2025)
6. 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. pp. 13359–13368 (2021)
7. Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1113–1122 (2021)
8. Chen, Z., Wang, H., Gui, J.: Occluded skeleton-based human action recognition with dual inhibition training. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2625–2634 (2023)

9. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., Lu, H.: Decoupling gcn with dropgraph module for skeleton-based action recognition. In: European conference on computer vision. pp. 536–553. Springer (2020)
10. Chi, H.g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20186–20196 (2022)
11. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: European Conference on Computer Vision. pp. 769–787. Springer (2020)
12. Do, J., Kim, M.: Skateformer: skeletal-temporal transformer for human action recognition. In: European Conference on Computer Vision. pp. 401–420. Springer (2024)
13. Do, J., Kim, M.: Bridging the skeleton-text modality gap: Diffusion-powered modality alignment for zero-shot skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12757–12768 (2025)
14. Duan, H., Wang, J., Chen, K., Lin, D.: Pyskl: Towards good practices for skeleton action recognition. In: Proceedings of the 30th ACM international conference on multimedia. pp. 7351–7354 (2022)
15. Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3558–3565 (2019)
16. Gao, Y., Duan, X., Dai, Q.: Skeleton-based action recognition using graph convolutional network with pose correction and channel topology refinement. Computers, Materials & Continua **83**(1) (2025)
17. Hao, X., Li, H.: Perspose: 3d human pose estimation with perspective encoding and perspective rotation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8110–8119 (2025)
18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
19. Lee, J., Lee, M., Lee, D., Lee, S.: Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10444–10453 (2023)
20. Li, S.W., Wei, Z.X., Chen, W.J., Yu, Y.H., Yang, C.Y., Hsu, J.Y.j.: Sa-dvae: Improving zero-shot skeleton-based action recognition by disentangled variational autoencoders. In: European conference on computer vision. pp. 447–462. Springer (2024)
21. Li, Z., Chang, X., Li, Y., Su, J.: Skeleton-based group activity recognition via spatial-temporal panoramic graph. In: European Conference on Computer Vision. pp. 252–269. Springer (2024)
22. Liu, H., Liu, Y., Ren, M., Wang, H., Wang, Y., Sun, Z.: Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29248–29257 (2025)
23. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
24. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 143–152 (2020)
25. Ma, N., Zhang, H., Li, X., Zhou, S., Zhang, Z., Wen, J., Li, H., Gu, J., Bu, J.: Learning spatial-preserved skeleton representations for few-shot action recognition. In: European Conference on Computer Vision. pp. 174–191. Springer (2022)
26. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2640–2649 (2017)
27. Mehak, S., Kelleher, J.D., Guilfoyle, M., Leva, M.C.: Action recognition for human–robot teaming: Exploring mutual performance monitoring possibilities. Machines **12**(1), 45 (2024)
28. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7753–7762 (2019)
29. Presti, L.L., La Cascia, M.: 3d skeleton-based human action classification: A survey. Pattern Recognition **53**, 130–147 (2016)
30. Rajendran, M., Tan, C.T., Atmosukarto, I., Ng, A.B., See, S.: Review on synergizing the metaverse and ai-driven synthetic data: enhancing virtual realms and activity recognition in computer vision. Visual Intelligence **2**(1), 27 (2024)
31. Ren, B., Liu, M., Ding, R., Liu, H.: A survey on 3d skeleton-based action recognition using learning method. Cyborg and Bionic Systems **5**, 0100 (2024)
32. 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. pp. 1010–1019 (2016)
33. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7912–7921 (2019)
34. 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)
35. Shi, W., Li, D., Wen, Y., Yang, W.: Occlusion-aware graph neural networks for skeleton action recognition. IEEE Transactions on Industrial Informatics **19**(10), 10288–10298 (2023)
36. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Richly activated graph convolutional network for robust skeleton-based action recognition. IEEE Transactions on Circuits and Systems for Video Technology **31**(5), 1915–1925 (2020)
37. Sun, F., Chen, R., Ji, T., Luo, Y., Zhou, H., Liu, H.: A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives. CAAI Artificial Intelligence Research **3**(9150042), 1 (2024)
38. Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J.: Human action recognition from various data modalities: A review. IEEE transactions on pattern analysis and machine intelligence **45**(3), 3200–3225 (2022)
39. Thakkar, K., Narayanan, P.: Part-based graph convolutional network for action recognition. arXiv preprint arXiv:1809.04983 (2018)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
41. Wang, Q., Zhang, K., Asghar, M.A.: Skeleton-based st-gcn for human action recognition with extended skeleton graph and partitioning strategy. IEEE Access **10**, 41403–41410 (2022)

42. 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. vol. 32 (2018)
43. Yan, T., Zeng, W., Xiao, Y., Tong, X., Tan, B., Fang, Z., Cao, Z., Zhou, J.T.: Crossglg: Llm guides one-shot skeleton-based 3d action recognition in a cross-level manner. In: European Conference on Computer Vision. pp. 113–131. Springer (2024)
44. Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., Tang, H.: Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In: Proceedings of the 28th ACM international conference on multimedia. pp. 55–63 (2020)
45. Yu, Q., Dai, Y., Hirota, K., Shao, S., Dai, W.: Shuffle graph convolutional network for skeleton-based action recognition. Journal of Advanced Computational Intelligence and Intelligent Informatics **27**(5), 790–800 (2023)
46. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
47. Zheng, Q., Yu, Y., Yang, S., Liu, J., Lam, K.Y., Kot, A.: Towards physical world backdoor attacks against skeleton action recognition. In: European Conference on Computer Vision. pp. 215–233. Springer (2024)
48. Zhou, H., Liu, Q., Wang, Y.: Learning discriminative representations for skeleton based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10608–10617 (2023)
49. Zhou, Y., Xu, T., Wu, C., Wu, X., Kittler, J.: Adaptive hyper-graph convolution network for skeleton-based human action recognition with virtual connections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12648–12658 (2025)
50. Zhou, Y., Cheng, Z.Q., Li, C., Fang, Y., Geng, Y., Xie, X., Keuper, M.: Hypergraph transformer for skeleton-based action recognition. arXiv preprint arXiv:2211.09590 (2022)
51. Zhou, Y., Yan, X., Cheng, Z.Q., Yan, Y., Dai, Q., Hua, X.S.: Blockgcn: Redefine topology awareness for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2049–2058 (2024)
52. Zhu, A., Ke, Q., Gong, M., Bailey, J.: Part-aware unified representation of language and skeleton for zero-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18761–18770 (2024)
53. Zhu, A., Zhu, J., Bailey, J., Gong, M., Ke, Q.: Semantic-guided cross-modal prompt learning for skeleton-based zero-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13876–13885 (2025)