DATE-GCN: Dynamic Untangled Adjacency Learning with Adaptive Temporal Encoding for Skeleton-Based Action Recognition

Anonymous authors
Paper under double-blind review

Abstract

Graph convolutional networks (GCNs) are commonly used for skeleton-based action recognition and have achieved remarkable performance. However, existing GCN-based approaches still face two major limitations: (1) Most models use predefined or fixed adjacency matrices to represent inter-joint relationships, which restricts their ability to adapt to the dynamic and context-dependent nature of human actions, and (2) their temporal modeling modules often fail to explicitly disentangle motion dynamics from feature extraction, resulting in limited capacity to capture complex and non-linear temporal dependencies. In this work, we present two main contributions: (1) a dynamic untangled multi-scale adjacency mechanism (DUMA) that adaptively learns action-dependent spatial relationships between body joints, and (2) an adaptive temporal encoding module (ATEM) that improves temporal modeling by decoupling temporal and feature channel interactions. Our full model, DATE-GCN, achieves state-of-the-art performance in skeleton-based action recognition across major benchmark datasets: NTU RGB+D 60, NTU RGB+D 120, and Northwestern-UCLA. The source code of this work is available at: anonymous.4open.science/r/DATE-GCN.

1 Introduction

Human action recognition (HAR) aims to classify human activities from video sequences for applications, such as human computer interaction, intelligent surveillance, and virtual reality. Recent advances have focused on two primary input modalities: RGB-based (Abbas et al., 2025; Sun & Sun, 2025) and skeleton-based representations (Wu et al., 2025; Ji et al., 2026; Su et al., 2026; Zhu & Liu, 2026). While RGB-based methods have shown promising results, their performance often degrades in real-world conditions due to sensitivity to environmental factors such as background clutter, lighting variations, and clothing diversity. In contrast, skeleton-based methods, which represent human actions as sequences of 2D or 3D joint coordinates, have gained increasing attention for their robustness and compact representation of human motion. By filtering out visual noise and capturing only structural movements, skeleton-based methods provide an alternative for action recognition by minimizing sensitivity to visual disturbances, making them effective in challenging environments.

Graph convolutional networks (GCNs) have been used for skeletal action recognition that inputs skeletons as a graph structure(Gao et al., 2025; Li et al., 2024a). A GCN performs convolutions on a graph structure to extract features. By utilizing a GCN, features that capture the relationships between joints can be extracted, enabling the recognition of complex actions. A typical GCN-based method is the spatial-temporal GCN (ST-GCN) (Yan et al., 2018). ST-GCN achieves high performance by representing skeletons through two graph structures: spatial and temporal graphs. In ST-GCN, the graph connections are predetermined based on human skeleton patterns. In practice, the importance and correlation between joints can change depending on the action being performed. For instance, when a person throws an object using the right arm while stabilizing with the left leg, the interaction between these two distant joints becomes functionally significant, even though they are not directly connected in the predefined skeleton graph (see Fig. 1). Similarly, during actions such as running, both arms and legs move in coordinated but alternating patterns, meaning that

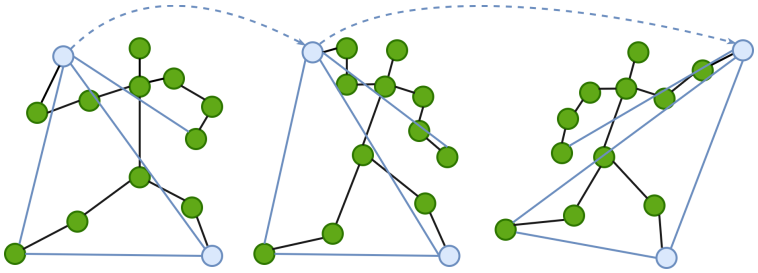

Figure 1: Illustration of joint relevance during a throwing action. Notice that the most significant motion occurs in the right arm, which initiates the throwing movement, while the left leg provides balance and support. The dynamic connections (in blue) highlight how non-adjacent joints, such as the right arm and left leg, become functionally correlated during this action, emphasizing that inter-joint relationships can vary depending on the performed motion.

the most informative joint dependencies are context-dependent rather than anatomically fixed. Thus, using fixed connection patterns in advance limits the ability to capture action-specific joint relationships that are critical for accurate action recognition.

To solve this problem, some graph convolutional network approaches automatically obtain graph connection patterns from feature maps (Zhou et al., 2024; 2023; gun Chi et al., 2022; Dong et al., 2025). They acquire the connection pattern from the feature map representing the optimal relationship between joints for each action. However, these methods fail to fully exploit the inherent skeletal structure due to two limitations.

The first limitation lies in the use of a fixed adjacency matrix $\mathbf{A}$ derived from the anatomical structure of the human body. Although this prior provides a good starting point for modeling spatial dependencies between joints, it remains fixed across all action classes and temporal contexts. As a result, it cannot adapt to the dynamic nature of human motion, where the relevance between joints changes according to the action and its phase. To capture broader structural information, some approaches (Jang et al., 2024; Altassan et al., 2024) extend this matrix by computing its powers, where each entry reflects the number of paths of a certain length between joints. This technique allows for multi-scale aggregation by enabling feature propagation across larger neighborhoods. However, higher-order graph convolutions often introduce a strong bias toward short-range connections, as self-loops and short cycles tend to dominate the graph spectrum in undirected skeletons. Consequently, nearby joints disproportionately influence the learned representations, regardless of their actual relevance to the action being performed. This uniform treatment of all joints fails to account for action-specific dynamics, where the importance of each joint may vary significantly depending on the motion context.

The second limitation is the rigidity of temporal modeling in existing approaches. Most models apply fixed-size 1D convolutions uniformly across joints and time, ignoring the fact that different actions unfold at varying temporal rhythms and involve different joint dynamics. This results in overly generic temporal features that lack adaptability. Although multi-branch temporal modules improve expressiveness by capturing patterns at different temporal scales, they do so at the cost of increased parameters and computational complexity, limiting their efficiency in real-time applications.

In this work, we address the above limitations with two main contributions:

- First, we propose a dynamic untangled multi-scale adjacency learning mechanism (DUMA) that enables the model to learn action-dependent spatial relationships between body joints (Sec. 3.2). Unlike fixed topologies, this mechanism adaptively reconfigures the graph structure based on the input motion, allowing the model to emphasize contextually relevant joints while suppressing irrelevant ones (Illustrated in Fig. 2).
- Second, we introduce an adaptive temporal encoding module (ATEM) that improves temporal modeling by decoupling temporal and feature channel interactions (Sec. 3.3). Inspired by transformer-style feedforward networks, this module enhances non-linear dynamics across time and feature chan-

nels while maintaining computational efficiency, enabling the model to better capture diverse motion patterns across different actions.

We conducted extensive experiments on three benchmark action recognition datasets: NTU-RGB+D 60 (Shahroudy et al., 2016), NTU-RGB+D 120 (Liu et al., 2019), and Northwestern-UCLA (Wang et al., 2014). Our method, DATE-GCN, is establishing a new state-of-the-art performance on these benchmarks without reliance on extra supervision. It demonstrates consistent improvements, averaging over 1.0% in accuracy against previous best-performing approaches.

2 Related Work

2.1 Skeleton-based Action Recognition

Early work in skeleton-based action recognition predominantly utilized Recurrent Neural Networks (RNNs) (Du et al., 2015; Song et al., 2016; Zhang et al., 2017), leveraging their strength in capturing temporal relationships over time. Convolutional Neural Networks (CNNs) were also explored, but they lacked effectiveness in explicitly modeling spatial relationships between body joints (Wang et al., 2024; Cheng et al., 2023). 3D convolutional neural networks (3D-CNNs) extended traditional 2D-CNNs by adding a temporal dimension, allowing them to capture both spatial and temporal features from video data. Since their early use in action recognition tasks (Ji et al., 2013; Tran et al., 2014), 3D-CNNs have become a dominant framework, particularly following the success of I3D (Carreira & Zisserman, 2017), which inflated 2D ImageNet pretrained filters into 3D. Subsequent architectures, such as R(2+1)D (Tran et al., 2017), SlowFast (Feichtenhofer et al., 2018), and X3D (Feichtenhofer, 2020), have further improved performance and efficiency through more refined temporal modeling and architectural optimizations. However, despite their effectiveness, 3D-CNNs typically require large-scale video datasets and high computational resources due to the increased number of parameters. Moreover, their reliance on appearance-based features makes them susceptible to variations in lighting, background, and viewpoint, which limits their robustness in real-world applications.

More recently, skeleton-based action recognition has shifted toward more adaptive and semantically rich paradigms. LaDy (Ji et al., 2026) introduces a dynamic formulation for modeling skeleton motion, aiming to better capture complex temporal structure. PCLE (Su et al., 2026) explores prompt-based contrastive learning to improve representation learning by leveraging semantic guidance and cross-modal alignment. DSTAFormer (Zhu & Liu, 2026) builds on dynamic spatial-temporal attention ideas to more effectively model joint dependencies and motion variations.

2.2 Graph Topology for Action Recognition

A human skeleton can be naturally modeled as a graph, where joints are connected by predefined edges to capture structural relationships. Both GCNs (Gao et al., 2025; Song et al., 2021; Li et al., 2024b) and Transformers (Do & Kim, 2024)(Chen et al., 2025) (Xin et al., 2023) have been explored for semantic aggregation.

Graph convolutional networks (GCNs) model the human body as a spatiotemporal graph, where joints are represented as nodes and bones or temporal links as edges. ST-GCN (Yan et al., 2018) is one of the earliest and most influential models for GCN-based approaches, combining spatial graph convolutions with temporal convolutions to capture motion dynamics over time. Building upon this baseline, several works have proposed extensions using adjacency matrix powers to enable multi-scale spatial modeling (Fang et al., 2022; Zhang et al., 2023a), while others have integrated self-attention mechanisms to enhance the model representational capacity (Wu et al., 2024; Xu et al., 2024). Pioneering work on capturing groups of nodes across time uses tensors to represent the 3D human body joints to exploit the kinematic relations among the adjacent and non-adjacent joints (Koniusz et al., 2020). In (Liu et al., 2020b), the human body is modeled as a hypergraph using a semi-dynamic hypergraph neural network, which enables the capture of more detailed information compared to GCNs. A hypergraph GNN, as discussed in (Hao et al., 2021), is capable of capturing both spatio-temporal features and higher-order dependencies for skeleton-based action recognition. In GNNs, attention can be defined over edges (Velickovic et al., 2017) (Zhang et al., 2018), or over nodes (Lee et al.,

2018). Despite these advancements, conventional GCN-based methods often rely on fixed graph structures and uniform temporal processing, which limits their adaptability to diverse motion patterns and reduces robustness under challenging conditions. Furthermore, extending these models to incorporate multimodal information, such as RGB or depth data often requires complex fusion strategies, making them less scalable and harder to generalize.

In Transformer-based methods, IIP-Transformer (Wang et al., 2021) utilizes a self-attention mechanism to capture intra-part and inter-part joint dependencies. IGFormer (Pang et al., 2022) extends this by learning inter-person topologies at both semantic and spatial distance levels. Similarly, STTFormer (Qiu et al., 2022) introduces a structure designed to model correlations between joints over consecutive frames.

The two paradigms mentioned above build the skeleton topology from joint pairs, considering only binary relations between two vertices. In practice, the topology is represented through an adjacency matrix or attention map. However, such formulations dismiss higher-order joint interactions, thereby neglecting the cooperative information shared among multiple joints.

2.3 Multi-Scale Graph Learning

To enhance feature discrimination, SEFN (Kong et al., 2021) introduces the Multi-perspective Attention Fusion Module (MPAFM), which was the first to integrate information from spatial, channel, and temporal branches through an attention mechanism. Nonetheless, relying on single-scale features limits its ability to capture action patterns, as human movements often occur across multiple temporal ranges concurrently. ScaleGNN (Li et al., 2025) and HiGCN (Wen, 2025) explore multi-scale feature extraction by employing higher-order polynomial expansions of the adjacency matrix to model non-local neighbors. Such formulations are prone to biased weighting, largely due to the presence of self-connections. MS-G3D, proposed by Liu et al. (Liu et al., 2020a), aims to reduce redundant dependencies and is the first to employ differentiable dilated convolutions for capturing multi-scale temporal dynamics. Chen et al. (Chen et al., 2021b) explored the MS-GC and MT-GC modules for multi-scale spatial–temporal modeling, though in MS-GC the temporal scale is primarily controlled by a scale factor. CTR-GCN (Chen et al., 2021a) employed a multi-branch structure with varying dilations, similar to MS-G3D, and achieved improved performance. Shu et al. (2023) used a multi-scale dilated attention graph convolutional network, and Zhang et al. (2023b) used a multi-stream with a multi-scale dilated spatial-temporal graph convolution network for skeleton-based action recognition. Despite their effectiveness, the above approaches obtain multi-scale features using either predefined topological structures or fixed dilations. Such rigid designs restrict scale adaptability, making it difficult to model discriminative features for subtle actions. To overcome this drawback, we propose a dynamic untangled multi-scale adjacency learning mechanism and an adaptive temporal encoding module capable of learning action-dependent spatial relationships between body joints.

3 Method

In this section, we first present relevant mathematical definitions (Sec. 3.1). We then present our Dynamic Untangled Multi-Scale Aggregation (DUMA) module that disentangles and adaptively learns spatial relationships across multiple joint neighborhoods (Sec. 3.2). Then, we provide a detailed explanation of our Adaptive Temporal Encoding Module (ATEM) that captures hierarchical temporal patterns through a streamlined convolutional design (Sec. 3.3). The above contributions are the core building block of our DATE-GCN (Sec. 3.4), as shown in Fig. 3.

3.1 Preliminaries

A skeleton sequence is modeled as a sequence of spatial graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \ldots, v_N\}$ is the set of N vertex representing body joints and $\mathcal{E}$ denotes the set of edges capturing structural relationships between joints. These spatial relationships are encoded in an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where $\mathbf{A}_{i,j} = 1$ if an edge exists between nodes v_i and v_j, and $\mathbf{A}_{i,j} = 0$ otherwise. We assume $\mathbf{A}$ to be symmetric since the human skeleton graph is undirected. The temporal aspect is introduced by observing a sequence of such skeletons over T frames. Note that each skeleton is encoded in a separate graph. Let the input node features

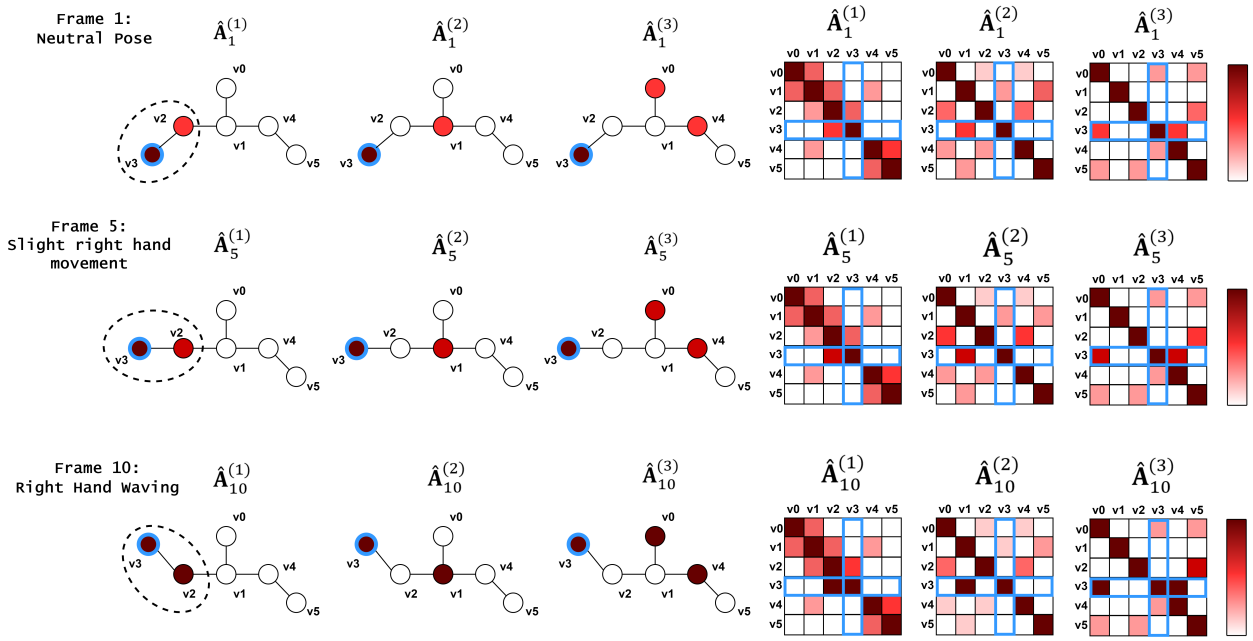

Figure 2: Proposed scheme for Dynamic Untangled Multi-Scale Adjacency learning (DUMA) module. Node colors reflect joint-level attention, and darker cells indicate stronger message passing. As the action "Right Hand Waving" unfolds, attention shifts toward the moving hand, demonstrating the model ability to adaptively capture motion-relevant joints across time and scale. $\hat{\mathbf{A}}_t^{(s)}$ is the attention-modulated adjacency matrix, with t being the frame number and s being the scale number.

be denoted as $\mathcal{X} = \{x_{t,n} \in \mathbb{R}^C \mid t \in \mathbb{Z}, 1 \leq t \leq T, 1 \leq n \leq N\}$, where $x_{t,n}$ is the C-dimensional feature vector of joint v_n at time step t. The full sequence is represented as a tensor $\mathbf{X} \in \mathbb{R}^{T \times N \times C}$, where C is the number of channels, N is the number of joints in one frame and T is the number of frames. The input action is therefore described both structurally by $\mathbf{A}$ and dynamically by $\mathbf{X}$. The goal of the model is to learn a function that maps the input sequence to an action class using a spatio-temporal neural network.

Given a set of joint features $\mathbf{X}_t \in \mathbb{R}^{N \times C}$ at time step t, and a graph structure defined by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, the spatial graph convolution operation at layer l can be expressed as

$$\mathbf{X}_t^{(l+1)} = \sigma\left(\tilde{\mathbf{M}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{M}}^{-\frac{1}{2}} \mathbf{X}_t^{(l)} \Theta^{(l)}\right), \tag{1}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ denotes the adjacency matrix augmented with self-loops to preserve node identity, and $\tilde{\mathbf{M}}$ is its degree matrix with entries $\tilde{M}_{ii} = \sum_j \tilde{A}_{ij}$. The term $\Theta^{(l)} \in \mathbb{R}^{C_l \times C_{l+1}}$ is a learnable weight matrix, and $\sigma(\cdot)$ is an activation function.

3.2 Dynamic Untangled Adjacency Learning

The Dynamic Untangled Multi-Scale Aggregation (DUMA) module unifies untangled multi-hop graph aggregation with dynamic graph adaptation, allowing the model to learn multi-scale and action-dependent spatial dependencies. It processes and encodes the spatial joint relationships of each graph in the input sequence. Instead of using adjacency powers to reach distant neighbors, which inherently overlaps neighborhoods and amplifies close-range influence, we untangle each scale by explicitly constructing a set of s-hop adjacency matrices $\{\tilde{\mathbf{A}}^{(s)}\}_{s=0}^{S}$, where each $\tilde{\mathbf{A}}^{(s)} \in \mathbb{R}^{N \times N}$ encodes only those connections between joints that are exactly

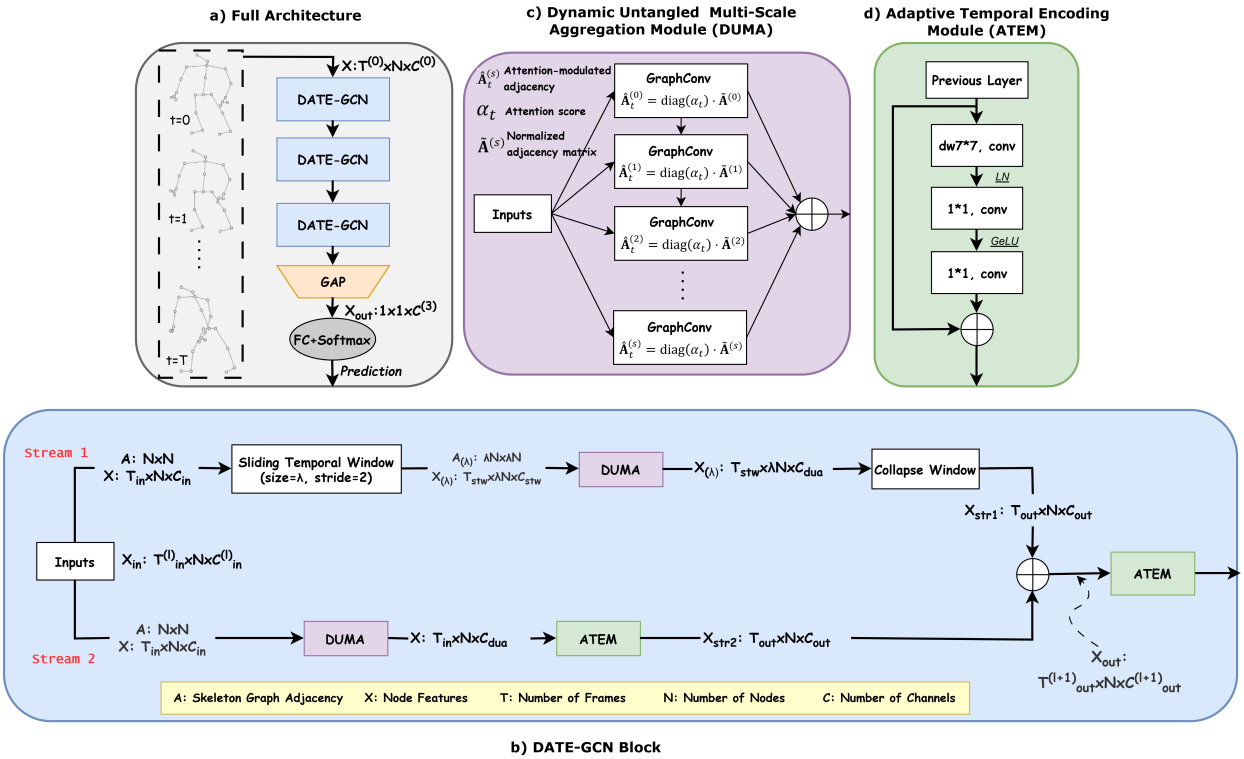

Figure 3: Our proposed approach: (a) is the overall model architecture, (b) is the DATE-GCN block, (c) is the dynamic untangled multi-scale aggregation module (DUMA) in the DATE-GCN, and (d) is the adaptive temporal encoding (ATEM) module in the DATE-GCN.

s hops apart. Formally, the entries of $\tilde{\mathbf{A}}^{(s)}$ are defined as:

$$\tilde{\mathbf{A}}^{(s)}_{i,j} = \begin{cases} 1, & \text{if } d(v_i, v_j) = s \text{ or } i = j \text{ (self-loop)}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

To ensure numerical stability and consistent scaling across different neighborhoods, we normalize each matrix using its corresponding degree matrix $\mathbf{M}^{(s)}$:

$$\bar{\mathbf{A}}^{(s)} = \mathbf{M}^{(s)^{-\frac{1}{2}}} \tilde{\mathbf{A}}^{(s)} \mathbf{M}^{(s)^{-\frac{1}{2}}}, \tag{3}$$

Each normalized matrix $\bar{\mathbf{A}}^{(s)}$ serves as a scale-specific aggregation operator, focusing only on a fixed topological neighborhood without the recursive overlap introduced by adjacency matrix powers. This formulation ensures that each scale contributes equally and independently, avoiding the over-smoothing problem and allowing meaningful multi-scale learning. To enhance the flexibility of this multi-scale aggregation, we extend the formulation by incorporating input-adaptive dynamics into the graph structure. We introduce a set of frame-wise, joint-specific attention coefficients that modulate the importance of each node in the graph, enabling the network to adaptively adjust spatial dependencies according to the motion context. For a given time step t, we compute attention scores $\boldsymbol{\alpha}_t \in \mathbb{R}^N$, where each coefficient $\alpha_{t,i}$ reflects the saliency of joint v_i at that frame t. Let $\mathbf{X}_t \in \mathbb{R}^{N \times C}$ denote the joint features at frame t, where C is the number of feature channels. The attention scores are generated via a learnable attention function:

$$\boldsymbol{\alpha}_t = \text{Softmax}\left(\mathbf{w}_2 \cdot \sigma\left(\mathbf{w}_1 \cdot \mathbf{X}_t\right)\right), \tag{4}$$

where $\mathbf{w}_1 \in \mathbb{R}^{C' \times C}$ and $\mathbf{w}_2 \in \mathbb{R}^{1 \times C'}$ $(C' = C/2)$ are learnable linear projections, $\sigma(\cdot)$ is a non-linear ReLU activation function, and the Softmax function ensures that attention values sum to 1 across joints:

$\sum_{i=1}^{V} \alpha_{t,i} = 1$. The attention scores α_t are then used to modulate the untangled adjacency matrices, resulting in frame-specific, attention-weighted graphs. Specifically, we scale each row of $\bar{\mathbf{A}}^{(s)}$, which determines how each joint aggregates information from its neighbors, by the corresponding attention coefficient. This results in an attention-modulated adjacency matrix $\hat{\mathbf{A}}_t^{(s)} \in \mathbb{R}^{N \times N}$ computed as:

$$\hat{\mathbf{A}}_t^{(s)} = \operatorname{diag}(\boldsymbol{\alpha}_t) \cdot \bar{\mathbf{A}}^{(s)}, \tag{5}$$

This modulation controls message passing toward the most informative joints at frame t. It allows the model to strengthen or suppress connections contextually, depending on the spatial configuration and dynamics of the input sequence.

Extending Eq. 1 to multiscale, the spatial graph convolution operation at layer l for frame t and scale s can be expressed as:

$$\mathbf{X}_t^{(l+1)} = \sigma\left(\sum_{s=0}^{S} \tilde{\mathbf{M}}_{(s)}^{-\frac{1}{2}} \hat{\mathbf{A}}_t^{(s)} \tilde{\mathbf{M}}_{(s)}^{-\frac{1}{2}} \mathbf{X}_t^{(l)} \Theta_{(s)}^{(l)}\right), \tag{6}$$

where $\tilde{\mathbf{M}}_{(s)}^{-\frac{1}{2}} \hat{\mathbf{A}}_t^{(s)} \tilde{\mathbf{M}}_{(s)}^{-\frac{1}{2}}$ is the normalized s-adjacency at frame t and $\Theta_{(s)}^{(l)} \in \mathbb{R}^{C \times C'}$ is a learnable transformation matrix for layer l at scale s. Unlike the previous methods, where possible length s walks are predominantly conditioned on length $s-1$ walks, the proposed dynamic untangled formulation in Eq. 5 addresses the static biased weighting problem by eliminating redundant dependencies between distant and nearby neighborhoods. Instead, it assigns context-dependent weights to each joint based on the input features at every time step t. This results in s-adjacency matrices that are more sparse (see Fig. 2), leading to more efficient and expressive representations.

3.3 Adaptive Temporal Encoding Module

Most existing works (Zhou et al., 2024; Liu et al., 2020a; Dong et al., 2025) inherently involve fixed-scale convolutional modeling or multi-branch temporal designs, which require carefully coordinated branches to capture different temporal scales. To overcome this, our Adaptive Temporal Encoding Module (ATEM) forgoes traditional multi-branch structures in favor of a streamlined, hierarchical design that combines large-kernel depthwise temporal convolutions, channel-wise normalization, and a nonlinear feedforward transformation as illustrated in Fig. 3 (d). The input to the ATEM block is a sequence represented as a tensor $\mathbf{X} \in \mathbb{R}^{T \times N \times C}$. We first apply a depthwise convolution along the temporal axis, using a large kernel size $k = 7$, to model long-range dependencies across time. This operation is performed independently for each channel and joint, allowing our final model to learn temporal dynamics specific to each spatial location.

To better accommodate temporal variations across different actions, the output is normalized using Layer Normalization over the channel dimension. Unlike Batch Normalization employed in previous works, which is sensitive to batch statistics, Layer Normalization operates independently at each temporal and spatial location, making it well-suited for sequence modeling tasks. The normalized features are subsequently processed by a lightweight MLP feedforward network, inspired by the transformer architecture (Vaswani et al., 2017). It consists of two pointwise linear layers separated by a GeLU activation, where the first layer expands the feature dimensionality to $C' = 4C$ and the second projects it back to C. Unlike MS-TCN (Farha & Gall, 2019), which relies on implicit channel mixing through convolutional layers, our design introduces explicit non-linear modeling of channel-wise relationships at each joint and time step, enabling more expressive temporal representations. To preserve the representational capacity of the input and promote stable gradient flow, a residual connection is employed. The output of the feedforward network is scaled by a learnable parameter $\gamma \in \mathbb{R}$ and added to the original input, yielding the final output of the block:

$$\mathbf{Y} = \mathbf{X} + \gamma \cdot (\mathbf{W}_2 \cdot \phi(\mathbf{W}_1 \cdot \operatorname{LayerNorm}(\operatorname{DWConv}(\mathbf{X})))), \tag{7}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C \times 4C}$ and $\mathbf{W}_2 \in \mathbb{R}^{4C \times C}$ are learnable weights of the two-layer feedforward MLP, and $\phi(\cdot)$ is the GeLU activation function. In contrast to conventional multi-branch temporal modules, our ATEM

block offers a streamlined and scalable alternative for temporal feature extraction in skeleton-based action recognition.

3.4 Full Model Architecture

As depicted in Fig. 3, our final model architecture is composed of a series of three stacked DATE-GCN blocks designed to extract discriminative features from skeleton sequences. The feature representations generated by these blocks are subsequently aggregated using a global average pooling layer, and a softmax classifier is applied to produce the final action prediction.

Each DATE-GCN block incorporates two complementary processing streams to enhance the representation of skeleton-based action dynamics: (1) The first stream, referred to as the short-term stream, receives an input tensor $X_{\text{in}} \in \mathbb{R}^{T_{\text{in}} \times N \times C_{\text{in}}}$ where T_{in}, N, and C_{in} denote the number of frames, joints, and feature channels, respectively. A sliding temporal window of size $\lambda = 3$ and stride $\mathcal{S} = 2$ is applied along the temporal axis, generating overlapping local segments $X_{(\lambda)} \in \mathbb{R}^{T_{\text{stw}} \times \lambda N \times C_{\text{stw}}}$. Each segment is processed by the DUMA module, which performs Dynamic Untangled Multi-Scale Aggregation to adaptively refine spatial dependencies within each local time window, yielding refined features $X_{\text{DUMA}} \in \mathbb{R}^{T_{\text{stw}} \times \lambda N \times C_{\text{DUMA}}}$. The outputs of overlapping windows are then aggregated and temporally aligned back to the original frame positions through a collapse and fusion step, producing short-term representations $X_{\text{str1}} \in \mathbb{R}^{T_{\text{out}} \times N \times C_{\text{out}}}$ with $T_{out} = T_{in}$. This design preserves temporal resolution while enriching the representation of each frame with fine-grained motion context from its neighboring frames. (2) The second stream, referred to as the long-term stream, complements the first by modeling extended temporal dependencies across the entire sequence. It takes as input the same sequence $X_{\text{in}} \in \mathbb{R}^{T_{\text{in}} \times N \times C_{\text{in}}}$ and first applies the DUMA module to refine spatial correlations among joints in each frame, while keeping the temporal resolution unchanged. This operation produces feature maps $X_{\text{DUMA}} \in \mathbb{R}^{T_{\text{in}} \times N \times C_{\text{DUMA}}}$, where DUMA dynamically adjusts the inter-joint relationships at multiple spatial scales for every time step. The output of DUMA is then passed to the Adaptive Temporal Encoding Module (ATEM), which operates exclusively along the temporal dimension, resulting in $X_{\text{ATEM}} \in \mathbb{R}^{T_{\text{in}} \times N \times C_{\text{out}}}$. This module adaptively learns the temporal importance of each frame, emphasizing salient motion intervals. The output features X_{str2} thus encode coherent long-term motion patterns and global temporal dependencies. Finally, these long-range temporal representations are fused with the short-term features of the first stream along the channel dimension, yielding a combined tensor $X_{\text{out}} \in \mathbb{R}^{T_{\text{out}} \times N \times C_{\text{out}}}$ where $C_{\text{out}} = C_{\text{str1}} + C_{\text{str2}}$. This fusion preserves the temporal and spatial dimensions while integrating complementary short and long-range motion features into a unified representation. This final ATEM stage ensures that the fused multi-scale representations are temporally smoothed and globally coherent before being propagated through the subsequent DATE-GCN blocks.

4 Experiments

4.1 Datasets

The NTU RGB+D 60 dataset (Shahroudy et al., 2016) is one of the largest benchmarks for human action recognition. It includes 56,880 skeleton-based action sequences performed by 40 different individuals, covering 60 distinct action categories. Each action was recorded using three Microsoft Kinect v2 cameras positioned at different viewpoints. To evaluate recognition methods, two standard protocols are suggested: 1) Cross-subject (X-sub): data from 20 subjects are used for training, while the remaining 20 subjects are reserved for testing. This setup measures how well a model generalizes to unseen people. 2) Cross-view (X-view): sequences from camera views 2 and 3 are used for training, and those from camera view 1 are used for testing. This benchmark focuses on assessing robustness across different viewpoints.

The NTU RGB+D 120 dataset (Liu et al., 2019) extends the original NTU RGB+D collection and is currently the largest resource for human action recognition with 3D joint annotations. It adds 57,367 new skeleton sequences spanning 60 additional action categories, resulting in a total of 113,945 samples carried out by 106 participants. The recordings were made from three different camera viewpoints across 32 distinct setups, each corresponding to a unique environment and background. To assess recognition models, two evaluation protocols are proposed. 1) Cross-subject (X-sub): actions from 53 subjects are used for training, while the

remaining 53 subjects form the test set, assessing generalization to unseen individuals. 2) Cross-setup (X-set): samples recorded in setups with even IDs are used for training, while those from odd IDs are reserved for testing, evaluating robustness to different environmental conditions.

The Northwestern-UCLA dataset (Wang et al., 2014) was collected using three Kinect cameras that simultaneously recorded actions from multiple viewpoints. It comprises 1,494 video sequences spanning 10 action classes, with each action carried out by 10 subjects. For evaluation, the dataset adopts the same protocol as a prior study: data from the first two cameras are used for training, while data from the third camera are reserved for testing.

4.2 Implementation Details

The model was optimized via Stochastic Gradient Descent (SGD) with a learning rate of 0.05, Nesterov momentum of 0.9 and a weight decay of 0.0005 for NTU RGB+D and NTU RGB+D 120, and 0.0002 for Northwestern-UCLA. The learning rate is set to 0.05 initially, and then scheduled using a Cosine Annealing strategy without epoch-based scheduling. Our experiments were conducted on an NVIDIA A100 GPU, with training carried out for a maximum of 50 epochs. Evaluations were performed at the end of every epoch, utilizing top-1 accuracy as the primary metric to assess model performance. For both NTU RGB+D and NTU RGB+D 120, we adopt the data splits and evaluation settings described in (Zhang et al., 2019) to ensure a fair comparison with prior work. The batch size is set to 16 during training. For NW-UCLA, we adopt the data splits settings in (Cheng et al., 2020b), and a batch size of 8.

4.3 Comparison With the State-of-the-Art Methods

We evaluate the performance of our proposed approach against a broad range of state-of-the-art skeleton-based action recognition methods on all three datasets in Table 1. To establish a fair comparison, we adopt a multi-stream fusion strategy and utilize four different modalities for building the graph: joint, bone, joint motion, and bone motion. Specifically, the joint modality corresponds to the raw skeletal coordinates, the bone modality is derived from the spatial differences between connected joints, and the joint motion and bone motion modalities capture the differential along the temporal dimension of the joint and bone modalities, respectively. The outputs from these modalities are then aggregated to generate the final prediction using score fusion.

As shown in Table 1, the performance of our DATE-GCN is consistently strong across all tested modalities. Whether using all four modalities (J+B+JM+BM), two modalities (J+B), or only the joint modality (J), DATE-GCN achieves the best results on every benchmark except for the NTU-120 X-Sub with J+B+JM+BM modality. Our DATE-GCN shows an improvement of 0.9% accuracy on average over the previous state-of-the-art methods. This consistency highlights the generalization capacity of our model architecture and its resilience to variations in input representations. These observations validate the design choices made in DATE-GCN and demonstrate its robustness under different levels of skeletal information. A more detailed breakdown of each module impact of our DATE-GCN will be provided in the following section through an ablation study, further showing the effectiveness of our architectural design.

4.4 Performance on Semantically Overlapping and Low-Saliency Actions

Despite significant progress, existing methods still struggle to accurately recognize semantically overlapping and low-saliency actions. Semantically overlapping actions are particularly challenging because they exhibit similar spatial configurations or joint interactions across different classes. These can be grouped into three subtypes: (1) actions involving interactions with the same body parts, such as drinking water versus eating a meal; (2) actions involving interactions with anatomically similar body parts, such as neck pain versus headache or sneeze/cough versus nausea/vomiting; and (3) actions exhibiting opposite temporal dynamics, such as putting on a shoe versus taking off a shoe. In contrast, low-saliency actions correspond to motion patterns with inherently weak cues, including short-duration actions such as sneezing or nausea, where the joint displacements are subtle and difficult to capture reliably.

Methods	Publication	Modalities	Params	NTU-60 X-Sub	NTU-60 X-View	NTU-120 X-Sub	NTU-120 X-Set	NW-UCLA
DC-GCN+ADG (Cheng et al., 2020a)	ECCV20	J+B+JM+BM	4.9M	90.8	96.6	86.5	88.1	95.3
MS-G3D (Liu et al., 2020a)	CVPR20	J+B+JM+BM	2.8M	91.5	96.2	86.9	88.4	–
MST-GCN (Chen et al., 2021b)	AAAI21	J+B+JM+BM	12.0M	91.5	96.6	87.5	88.3	–
STF (Ke et al., 2022)	AAAI22	J+B+JM+BM	–	92.5	96.9	88.9	89.9	–
CTR-GCN (Chen et al., 2021a)	ICCV21	J+B+JM+BM	1.5M	92.4	96.9	88.9	90.6	96.5
EfficientGCN-B4 (Song et al., 2021)	TPAMI22	J+B+JM+BM	2.0M	91.7	96.2	88.3	89.9	96.6
InfoGCN (gun Chi et al., 2022)	CVPR22	J+B+JM+BM	1.6M	92.7	97.0	89.0	91.2	96.6
FR Head (Zhou et al., 2023)	CVPR23	J+B+JM+BM	2.0M	92.8	96.6	90.0	91.2	96.8
SkeletonGCL (Huang et al., 2023)	ICLR23	J+B+JM+BM	–	92.8	97.1	89.8	91.2	–
GAP (Xiang et al., 2022)	ICCV23	J+B+JM+BM	–	92.9	97.0	89.9	91.1	–
MSS-GCN (Jang et al., 2024)	TCSVT24	J+B+JM+BM	–	92.7	96.9	88.9	90.6	96.6
SaPR-GCN (Li et al., 2024a)	TCSVT24	J+B+JM+BM	8.3M	92.4	96.4	88.7	90.3	–
DS-GCN (Xie et al., 2024)	AAAI24	J+B+JM+BM	–	93.1	97.5	89.2	91.1	–
BlockGCN (Zhou et al., 2024)	CVPR24	J+B+JM+BM	1.3M	93.1	97.0	90.3	91.2	$96.9^{\uparrow \mathbf{0.1}}$
DSTSA-GCN (Cui et al., 2025)	NEUROC25	J+B+JM+BM	–	92.8	97.0	89.1	90.9	–
LG-SGNet (Wu et al., 2025)	PR25	J+B+JM+BM	1.7M	93.1	96.7	89.4	91.0	96.8
LaDy (Ji et al., 2026)	CVPR26	J+B+JM+BM	1.8M	92.9	–	–	–	–
PCLE (Su et al., 2026)	TCSVT26	J+B+JM+BM	–	93.1	97.2	90.2	91.5	–
DSTAFormer (Zhu & Liu, 2026)	TCSVT26	J+B+JM+BM	2.7M	93.5	97.2	90.3	91.5	–
DATE-GCN (ours)	–	J+B+JM+BM	2.5M	$94.0^{\uparrow \mathbf{0.5}}$	$98.5^{\uparrow \mathbf{1.0}}$	$89.3^{\downarrow \mathbf{1.0}}$	$91.5^{\uparrow \mathbf{0.3}}$	$96.9^{\uparrow \mathbf{0.1}}$
2s-AGCN (Shi et al., 2018)	CVPR19	J+B	6.9M	88.5	95.1	82.9	84.9	–
ShiftGCN (Cheng et al., 2020b)	CVPR20	J+B	–	90.7	96.5	85.9	87.6	–
SANGCN (Tian et al., 2023)	TMM23	J+B	–	92.1	96.2	88.7	90.1	–
DATE-GCN (ours)	–	J+B	2.5M	$93.8^{\uparrow \mathbf{1.7}}$	$98.4^{\uparrow \mathbf{1.9}}$	$88.9^{\uparrow \mathbf{0.2}}$	$90.3^{\uparrow \mathbf{0.2}}$	–
ST-GCN (Yan et al., 2018)	AAAI18	J	–	81.5	88.0	71.7	72.2	–
AS-GCN (Li et al., 2019)	CVPR19	J	9.5M	86.8	94.2	78.2	77.7	–
InfoGCN (gun Chi et al., 2022)	CVPR22	J	1.6M	89.8	95.2	85.1	86.3	–
STA-GCN (Shiraki et al., 2020)	ACCV20	J	–	90.1	95.6	83.9	86.5	–
HDGCN (Lee et al., 2022)	ICCV23	J	1.7M	–	–	85.7	87.3	–
FR Head (Zhou et al., 2023)	CVPR23	J	2.0M	90.3	95.5	85.6	87.3	–
BlockGCN (Zhou et al., 2024)	CVPR24	J	1.3M	90.9	95.4	86.9	88.2	95.5
DSTSA-GCN (Cui et al., 2025)	NEUROC25	J	–	90.1	95.5	85.4	87.4	–
LaDy (Ji et al., 2026)	CVPR26	J	1.8M	90.3	–	–	–	–
DSTAFormer (Zhu & Liu, 2026)	TCSVT26	J	2.7M	90.7	95.2	87.2	87.1	–
DATE-GCN (ours)	–	J	2.5M	$92.7^{\uparrow \mathbf{1.8}}$	$97.2^{\uparrow \mathbf{1.7}}$	$87.7^{\uparrow \mathbf{0.5}}$	$88.6^{\uparrow \mathbf{0.4}}$	$96.3^{\uparrow \mathbf{0.8}}$

Table 1: Comparison of DATE-GCN with recent GCN and hybrid models. Modalities: J = Joint, B = Bone, JM = Joint Motion, BM = Bone Motion. "–" indicates that results are missing from the original references. The green and yellow cells indicate the best and second-best performances respectively. ↑ or ↓ indicates the variation w.r.t. previous SOTA.

To evaluate the recognition of these challenging actions, we compared the performance of our DATE-GCN and other state of the art methods on eight selected action categories from the NTU RGB+D 60 dataset. These categories were grouped into four action pairs including drink water / eat meal, put on a shoe / take off a shoe, headache / neck pain, and sneeze / nausea as illustrated in Figure 4. Each pair represents a type of challenging case: similar limb interactions, reversed temporal dynamics, or subtle motion cues. On these eight actions of NTU RGB+D 60 dataset, DATE-GCN achieves the highest accuracy on seven over eight classes and is behind on one (sneeze/cough). Aggregated by action pairs, DATE-GCN leads drink water / eat meal and put on a shoe / take off a shoe with pair means of 87.5% and 87.0%, providing margins of +5.5 % over the next-best method in both cases. It also tops headache / neck pain with a pair mean of 90.0% (+2.0 %). The only pair where it is not the best is sneeze-cough / nausea-vomiting, where SaPR-GCN is higher (pair mean 86.5%) and DATE-GCN reaches 84.0% (-2.5 %). Against strong baselines, DATE-GCN improves over MS-G3D by +7.3 % per class on average (pair-wise gains of +9.0, +10.5, +4.5, +5.0 % across the four pairs), and over MSS-GCN by +5.0 % per class on average (pair-wise gains of +7.0, +8.5, +3.0, +1.5 % across the four pairs). Qualitatively, the largest margins occur in pairs that require disambiguating reversed temporal dynamics (put on vs. take off a shoe) and similar limb interactions (drink vs. eat), consistent with DATE-GCN design, dynamic untangled adjacency learning reduces inter-joint interference among neighboring joints, while adaptive temporal encoding helps align motion phases.

4.5 Ablation Study

4.5.1 Search for the best number of scales S.

In table 2, we validate the effectiveness of our proposed dynamic untangled multi-scale aggregation scheme through experiments with varying numbers of scales S. Initially, we note that the untangled formulation provides up to a 0.8% improvement compared to simple adjacency powering when $S = 5$. This demonstrates that incorporating multiple scales enhances the ability of our model to capture diverse features. However,

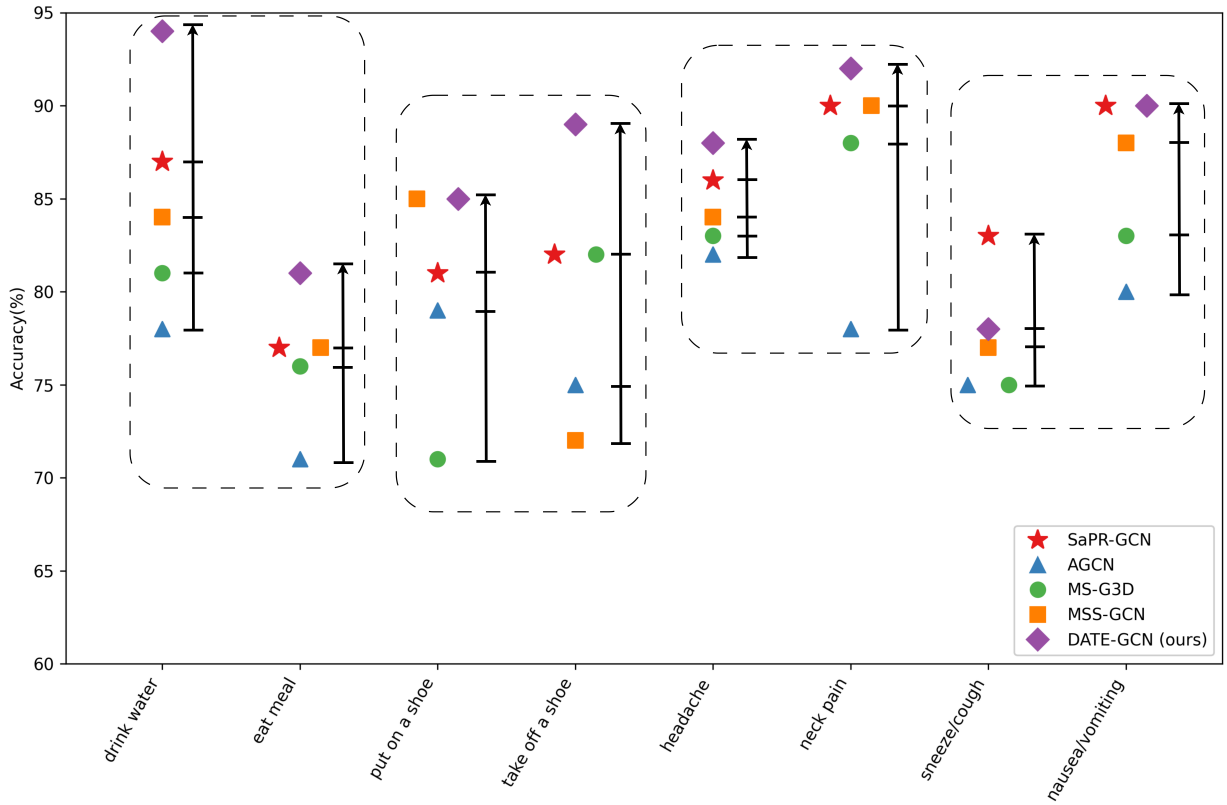

Figure 4: The effectiveness of our DATE-GCN in recognizing semantically overlapping and low-saliency Actions from NTU RGB+D 60 dataset. We select 8 actions as test samples and divide them into four pairs marked in dotted rectangles. The figure mirrors the results of other state of the art methods such as SaPR-GCN (Li et al., 2024a), AGCN (Shi et al., 2018), MS-G3D (Liu et al., 2020a), and MSS-GCN (Jang et al., 2024).

as we increase the number of scales beyond $S = 5$, the improvements become less pronounced. Specifically, for $S = 8$ and $S = 12$, the top-1 accuracy reaches 92.4% and 92.3%, respectively, showing only a marginal decrease from the peak accuracy of 92.7% at $S = 5$. This suggests that while additional scales provide slight benefits, the model reaches a point of diminishing returns beyond a certain scale threshold. Therefore, our results indicate that a balance must be struck between the number of scales and computational efficiency, with $S = 5$ achieving the optimal performance.

In Figure 5, we further illustrate the effectiveness of DATE-GCN by conducting a comparative analysis involving its performance evaluation with MS-G3D (Liu et al., 2020a) and MSS-GCN (Jang et al., 2024) across various scales. For the baseline model, we adopt 'GCN ($S = 1$) + TCN'. This baseline reaches 90.8%, establishing a lower bound for models that aggregate only immediate neighbors. MS-G3D (Liu et al., 2020a) exhibits the expected monotonic but saturating gain: 91.1% to 91.7% as S increases from 3 to 12. This shows that enlarging the receptive field improves coverage of medium range joint relations, but benefits attenuate as neighborhoods become denser. MSS-GCN (Jang et al., 2024) is comparatively insensitive to S, it peaks at 92.0% around S = 6-7 and slightly dips to 91.8–91.9% for larger S. The small variance suggests that multi-scale aggregation already supplies some longer-range context, so further expanding the hop size yields limited additional gain. Our DATE-GCN consistently outperforms the other models at moderate scales, achieving 92.7% at $S = 5$ and 92.4-92.3% at S = 8-12. The shallow decline beyond $S = 5$ indicates that DATE-GCN benefits most from mid-range interactions while avoiding the over-smoothing that can arise when too many distant joints are linked.

	Number of Scales (S)			
	$S = 1$	$S = 5$	$S = 8$	$S = 12$
Top-1 Accuracy	91.9%	92.7%	92.4%	92.3%

Table 2: Model accuracy with various number of scales S on NTU RGB+D 60 X-sub with Joint modality.

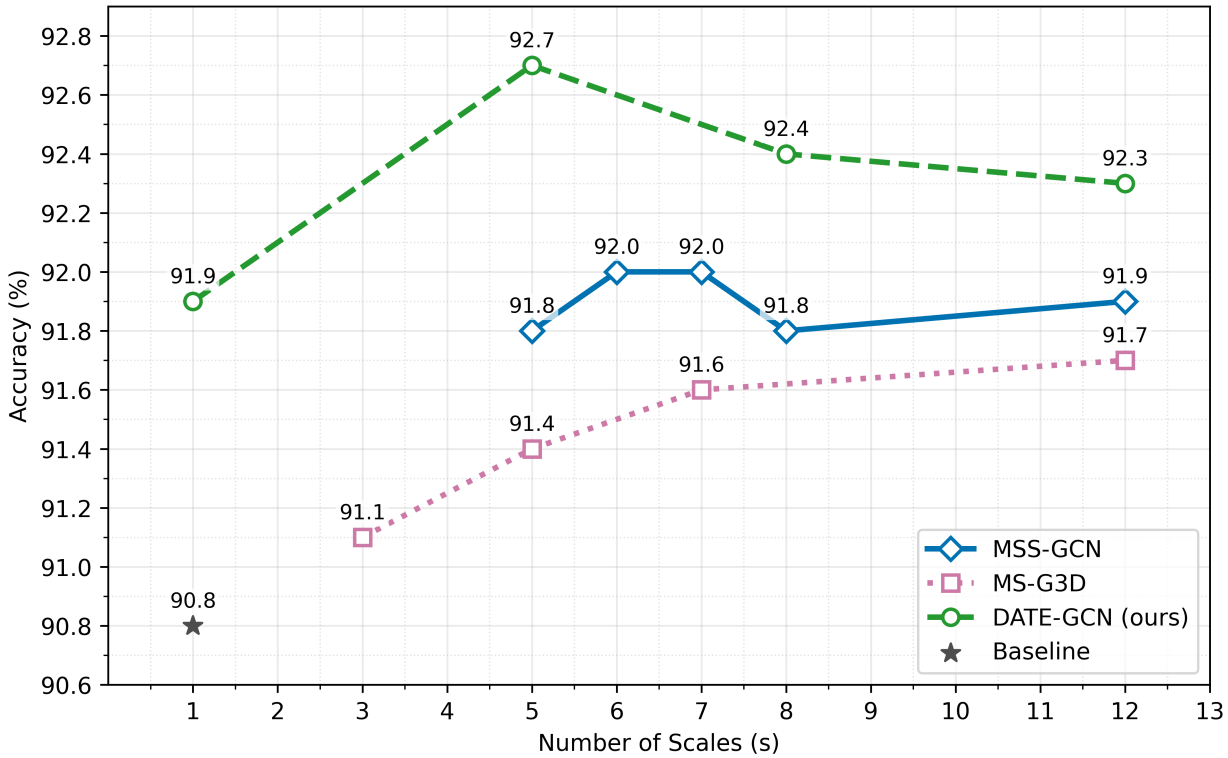

Figure 5: Comparison of Top-1 accuracy between our DATE-GCN and other multi-scale methods: MS-G3D and MSS-GCN across various number of scales on the NTU RGB+D 60 X-sub with Joint modality.

4.5.2 Discussion on model variants.

To evaluate the effectiveness of our DATE-GCN and the contribution of its modules, we tested different configurations where we remove in turn a module, and we analyze the size and accuracy of the various models. The comparison includes four configurations: (A) our complete model with DUMA and ATEM modules, (B) we remove from our complete model the ATEM module, (C) we remove from our complete model the DUMA module, and (D) we remove from our complete model both the DUMA and ATEM modules. When we remove the ATEM module, we use instead Multi-Stage Temporal Convolutional Encoding (MS-TCN) (Farha & Gall, 2019) as in (Liu et al., 2020a), and when we remove the DUMA module, we perform instead a static untangled adjacency learning. All ablation experiments in Table 3 are conducted on the X-Sub benchmark of NTU RGB+D 60. Only the joint modality is used as input for these experiments.

We initiate the study by examining the impact of our Adaptive Temporal Encoding Module (ATEM) by comparing models A and B in Table 3. We did that by replacing ATEM with the MS-TCN module employed in most of the recent methods, which yields a performance drop of 0.53% and an increase of model parameters (3.12M vs. 2.52M). This result demonstrates that incorporating the ATEM module in our DATE-GCN enhances the accuracy and reduce the parameter count of the model. Next, we assess the contribution of our DUMA module by comparing models A and C. Keeping the ATEM module but disabling the DUMA module causes a drop of 0.9% in the accuracy and a slight increase of the parameter count by 0.2M. This confirms that DUMA significantly enhances spatial modeling by learning action-specific topologies. Finally, for model

Model Variant	Temporal Encoding Module	Adjacency Learning Mechanism	Params	Accuracy (%)
A (DATE-GCN)	*ATEM*	*DUMA*	2.52M	92.69
B	MS-TCN	*DUMA*	3.12M	$92.16^{\downarrow 0.53}$
C	*ATEM*	Static Untangled	2.74M	$91.79^{\downarrow 0.90}$
D	MS-TCN	Static Untangled	2.92M	$91.69^{\downarrow 1.00}$

Table 3: Top-1 accuracy comparison with different components on NTU RGB+D 60 X-sub with Joint modality.

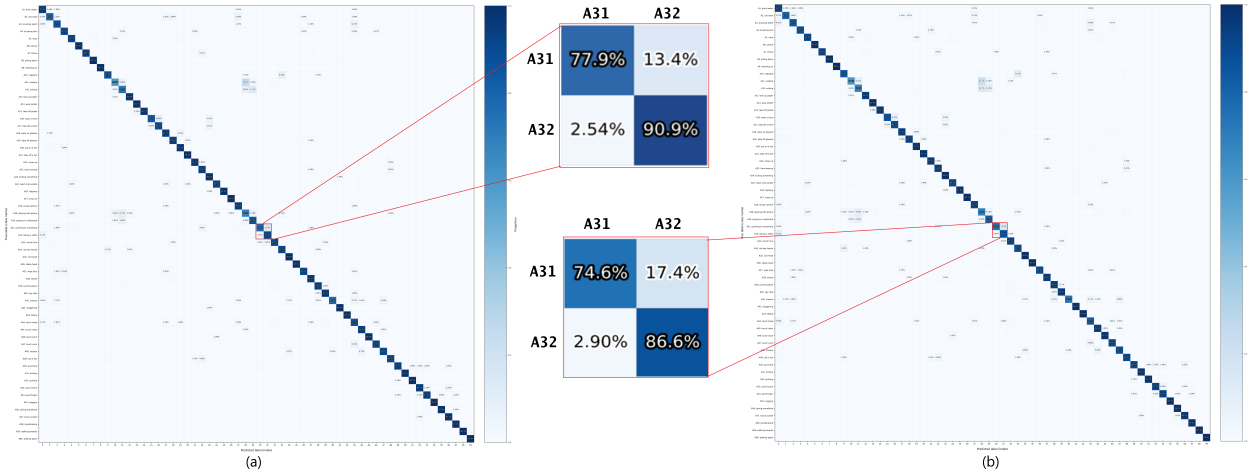

Figure 6: Comparison of the confusion matrix on NTU RGB+D 60 X-sub with Joint modality. The action index is as follows: Pointing to something (A31) and Taking a selfie (A32). (a) the confusion matrix for our DATE-GCN, (b) the confusion matrix for model variant D.

D, both DUMA and ATEM modules are disabled. This reduces the accuracy by 1.0%. Overall, using both our proposed modules versus using standard modules allows accuracy improvement, and furthermore, this improvement is achieved with the smallest parameter count among all configurations (2.52M). This highlights the compactness and synergy of the proposed modules. These results validate the effectiveness of our approach. By leveraging ATEM for temporal encoding and DUMA for graph aggregation, our approach significantly improves skeleton-based action recognition performance.

To analyze the classification performance in more details, the confusion matrices of our DATE-GCN and Variant D on NTU RGB+D 60 X-Sub were generated as shown in Figure 6. Consider, for instance, the actions "pointing to something" (index A31 in the confusion matrix) and "taking a selfie" (index A32 in the confusion matrix). Both involve movements of the arm and hand, but with different semantic intent: pointing to something is characterized by a straightened arm extending outward, whereas taking a selfie involves raising the arm toward the face while holding an imagined device. DATE-GCN yields clear gains compared with Variant D, recall on A31 rises from 74.6% to 77.9% (+3.3 pp), and on A32 from 86.6% to 90.9% (+4.3 pp), giving a balanced accuracy across the pair of 84.4% vs. 80.6% (+3.8 pp). Cross-class confusions are reduced in both directions: A31 →A32 drops from 17.4% to 13.4% (-4.0 pp, -23% relative), and A32 →A31 drops 2.90% to 2.54% (-0.36 pp, -12% relative). The error asymmetry is also mitigated: the ratio of A31 →A32 and A32 →A31 mistakes falls from ∼6.0× to ∼5.3×. These reductions indicate that DATE-GCN better captures the discriminative cues between a straight outward arm extension and face-directed arm-raising with a handheld device, leading to fewer swaps and higher correct detections for both classes.

5 Conclusion

In this work, we presented DATE-GCN, a novel graph architecture for skeleton-based action recognition that jointly addresses limitations in both spatial and temporal modeling. To overcome the rigidity of fixed graph structures, we introduced a dynamic adjacency learning module that adapts spatial connections based on action context. As well, we proposed an adaptive temporal encoding module that decouples temporal and channel interactions, enabling more expressive and efficient modeling of motion dynamics. Our approach demonstrates state-of-the-art performance across multiple datasets, confirming the effectiveness of our design.

References

Baqar Abbas, Abderrazak Chahi, and Yassine Ruichek. Context-aware 3d cnn for action recognition based on semantic segmentation (cars). *Comput. Vis. Image Underst.*, 263:104570, 2025.

Alaa Altassan, Hilal Ahmad Ganie, and Yilun Shang. On the extended adjacency eigenvalues of a graph. *Information*, 2024.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017.

Dong Chen, Mingdong Chen, Peisong Wu, Mengtao Wu, Tao Zhang, and Chuanqi Li. Two-stream spatio-temporal gcn-transformer networks for skeleton-based action recognition. *Scientific Reports*, 15, 2025.

Yuxin Chen, Ziqi Zhang, Chunfen Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13339–13348, 2021a.

Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*, 2021b.

Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *European Conference on Computer Vision*, 2020a.

Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 180–189, 2020b.

Qin Cheng, Jun Cheng, Ziliang Ren, Qieshi Zhang, and Jianming Liu. Multi-scale spatial–temporal convolutional neural network for skeleton-based action recognition. *Pattern Analysis and Applications*, 26:1303 – 1315, 2023.

Hu Cui, Renjing Huang, Ruoyu Zhang, and Tessai Hayama. Dstsa-gcn: Advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling. *Neurocomputing*, 637:130066, 2025.

Jeonghyeok Do and Munchurl Kim. Skateformer: Skeletal-temporal transformer for human action recognition. *ArXiv*, abs/2403.09508, 2024.

Pengcheng Dong, Wenbo Wan, Huaxiang Zhang, Shuai Li, Sujuan Hou, and Jiande Sun. Hfgcn:hypergraph fusion graph convolutional networks for skeleton-based action recognition. *ArXiv*, abs/2501.11007, 2025.

Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, 2015.

Zheng Fang, Xiongwei Zhang, Tieyong Cao, and Yunfei Zheng. The adjacency matrix with self-loop and adaptive weights for skeleton-based action recognition. *SSRN Electronic Journal*, 2022.

Yazan Abu Farha and Juergen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3570–3579, 2019.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 200–210, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6201–6210, 2018.

Yuxin Gao, Xiaodong Duan, and Qiguo Dai. Skeleton-based action recognition using graph convolutional network with pose correction and channel topology refinement. *Computers, Materials & Continua*, 2025.

Hyung gun Chi, Myoung Hoon Ha, Seung geun Chi, Sang Wan Lee, Qi-Xing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20154–20164, 2022.

Xiaoke Hao, Jie Li, Yingchun Guo, Tao Jiang, and Ming Yu. Hypergraph neural network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 30:2263–2275, 2021.

Xiaohui Huang, Hao Zhou, Bin Feng, Xinggang Wang, Wenyu Liu, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Graph contrastive learning for skeleton-based action recognition. In *International Conference on Learning Representations*, 2023.

Sungjun Jang, Heansung Lee, Woojin Kim, Jungho Lee, Sungmin Woo, and Sangyoun Lee. Multi-scale structural graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34:7244–7258, 2024.

Haoyu Ji, Xueting Liu, Yu Gao, Wenze Huang, Zhihao Yang, Weihong Ren, Zhiyong Wang, and Honghai Liu. Lady: Lagrangian-dynamic informed network for skeleton-based action segmentation via spatial-temporal modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34449–34459, June 2026.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. doi: 10.1109/TPAMI.2012.59.

Lipeng Ke, Kuan-Chuan Peng, and Siwei Lyu. Towards to-a-t spatio-temporal focus for skeleton-based action recognition. *ArXiv*, abs/2202.02314, 2022.

Jun Kong, Haoyang Deng, and Min Jiang. Symmetrical enhanced fusion network for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31:4394–4408, 2021.

Piotr Koniusz, Lei Wang, and Anoop Cherian. Tensor representations for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:648–665, 2020.

John Boaz Lee, Ryan A. Rossi, and Xiangnan Kong. Graph classification using structural attention. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

Junghoon Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoon Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10410–10419, 2022.

Chang Li, Yingchi Mao, Qian Huang, Xiaowei Zhu, and Jie Wu. Scale-aware graph convolutional network with part-level refinement for skeleton-based human action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34:4311–4324, 2024a.

Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3590–3598, 2019.

Min Li, Ke Chen, Yunqing Bai, and Jihong Pei. Skeleton action recognition via graph convolutional network with self-attention module. *Electronic Research Archive*, 2024b.

Xiang Li, Jianpeng Qi, Haobing Liu, Yuan Cao, Guoqing Chao, Zhongying Zhao, Junyu Dong, Xinwang Liu, and Yanwei Yu. Scalegnn: Towards scalable graph neural networks via adaptive high-order neighboring feature fusion. *Proceedings of the ACM Web Conference 2026*, 2025.

Jun Liu, Amir Shahroudy, Mauricio Perez, G. Wang, Ling yu Duan, and Alex Chichung Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2684–2701, 2019.

Ken Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 140–149, 2020a.

Shengyuan Liu, Pei Lv, Yuzhen Zhang, Jie Fu, Junjin Cheng, Wanqing Li, Bing Zhou, and Mingliang Xu. Semi-dynamic hypergraph neural network for 3d pose estimation. In *International Joint Conference on Artificial Intelligence*, 2020b.

Yunsheng Pang, Qiuhong Ke, H. Rahmani, James Bailey, and J. Liu. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *European Conference on Computer Vision*, 2022.

Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-temporal tuples transformer for skeleton-based action recognition. *ArXiv*, abs/2201.02849, 2022.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010–1019, 2016.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12018–12027, 2018.

Katsutoshi Shiraki, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Spatial temporal attention graph convolutional networks with mechanics-stream for skeleton-based action recognition. In *Asian Conference on Computer Vision*, 2020.

Yang Shu, Wanggen Li, Doudou Li, Kun Gao, and Biao Jie. Multi-scale dilated attention graph convolutional network for skeleton-based action recognition. In *Chinese Conference on Pattern Recognition and Computer Vision*, 2023.

Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *ArXiv*, abs/1611.06067, 2016.

Yisheng Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 1474–1488, 2021.

Taiyi Su, Zheng Liu, and Hanli Wang. Prompted contrastive learning for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Maojin Sun and Luyi Sun. A histogram transformer approach using attention-based 3d residual network for human action recognition. *PLOS One*, 20, 2025.

Haoyu Tian, Xin Ma, Xiang Li, and Yibin Li. Skeleton-based action recognition with select-assemble-normalize graph convolutional networks. *IEEE Transactions on Multimedia*, 25:8527–8538, 2023.

Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2014.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2017.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. Graph attention networks. *ArXiv*, abs/1710.10903, 2017.

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2649–2656, 2014.

Jingyao Wang, Emmanuel Bergeret, and Issam Falih. Skeleton-based action recognition with spatial-structural graph convolution. *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2024.

Qingtian Wang, Jianlin Peng, Shuze Shi, Tingxi Liu, Jiabin He, and Renliang Weng. Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition. *2023 IEEE International Conference on Big Data (BigData)*, pp. 936–945, 2021.

Qi Wen. Modeling higher-order interactions in graphs through combinatorial arc-transitive structure using graph convolutional network. *IEEE Access*, 13:55899–55914, 2025.

Zhize Wu, Peng Sun, Xin Chen, Keke Tang, Tong Xu, Le Zou, Xiaofeng Wang, Ming Tan, Fan Cheng, and Thomas Weise. Selfgcn: Graph convolution network with self-attention for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 33:4391–4403, 2024.

Zhize Wu, Yue Ding, Long Wan, Teng Li, and Fudong Nian. Local and global self-attention enhanced graph convolutional network for skeleton-based action recognition. *Pattern Recognit.*, 159:111106, 2025.

Wangmeng Xiang, C. Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10242–10251, 2022.

Jianyang Xie, Yanda Meng, Yitian Zhao, Anh Nguyen, Xiaoyun Yang, and Yalin Zheng. Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition. In *AAAI Conference on Artificial Intelligence*, 2024.

Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing*, 537:164–186, 2023.

Fen Xu, Pengfei Shi, and Xiaoping Zhang. Skeleton-based human action recognition with spatial and temporal attention-enhanced graph convolution networks. *J. Adv. Comput. Intell. Intell. Informatics*, 28: 1367–1379, 2024.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*, 2018.

Daqing Zhang, Hongmin Deng, and Yong Zhi. Enhanced adjacency matrix-based lightweight graph convolution network for action recognition. *Sensors (Basel, Switzerland)*, 23, 2023a.

Haiping Zhang, Xu Liu, Dongjin Yu, Lim Sie Guan, Dongjing Wang, Conghao Ma, and Zepeng Hu. Skeleton-based action recognition with multi-stream, multi-scale dilated spatial-temporal graph convolution network. *Applied Intelligence*, pp. 1–15, 2023b.

Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and D. Y. Yeung. Gaan: Gated attention networks for learning on large and spatiotemporal graphs. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

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. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2136–2145, 2017.

Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1109–1118, 2019.

Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10608–10617, 2023.

Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xiansheng Hua. Blockgcn: Redefine topology awareness for skeleton-based action recognition. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2049–2058, 2024.

Yisheng Zhu and Guangcan Liu. Dynamic prompting spatial temporal actor transformer for fine-grained skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 36: 5923–5935, 2026.