Pattern Recognition 175 (2026) 113146

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Frequency-aware spatio-temporal topology learning for skeleton-based human activity recognition

Yi Xia [a], Sira Yongchareon [a,*], Raymond Lutui [a], Quan Z. Sheng [b]

[a] School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand
[b] School of Computing, Faculty of Science and Engineering, Macquarie University, Sydney, Australia

ARTICLE INFO

Keywords:
Skeleton-based activity recognition
Graph convolutional networks
Spatio-temporal modeling
Attention mechanisms
Dynamic topology
Frequency analysis

ABSTRACT

Skeleton-based human activity recognition (HAR) has made significant progress through graph convolutional networks (GCNs) and Transformer architectures for spatiotemporal modeling. However, existing methods either employ predefined static graph topologies that cannot adapt to heterogeneous skeleton data or learn dynamic topologies based solely on local spatiotemporal features, thereby overlooking the global temporal frequency features of joint movements that are important for discovering semantically meaningful spatial relationships. We propose Frequency-Aware Topology Learning Graph Convolutional Network (FATL-GCN), a novel architecture that integrates frequency-aware temporal context to guide adaptive learning of spatial topology. Our approach leverages Time-to-Vector linear frequency encoding to capture both periodic and non-periodic motion patterns, employs frequency-guided topology learning to generate action-specific graphs through temporal-context-driven attention, and incorporates hierarchical multi-scale fusion for robust feature extraction across scales. Extensive experiments achieved top-1 accuracies of 93.8% (cross-subject) and 97.5% (cross-view) on NTU-60, 91.9% (cross-subject) and 93.1% (cross-setup) on NTU-120, and 51.7% on Kinetics-Skeleton. Ablation studies confirm the critical role of our components, with removing the dynamic graph topology causing a 3.5% accuracy drop and removing frequency-aware encoding causing a 2.1% drop.

1. Introduction

Human activity recognition (HAR) based on skeleton data has become a widely used technology in real-world application domains, including healthcare monitoring, human-computer interaction, surveillance systems, and fitness applications. The emergence of inexpensive non-invasive sensors and robust action recognition algorithms has made skeleton-based methods increasingly popular due to their computational efficiency, privacy protection, and robustness to appearance changes [1]. Despite significant progress in recent years, robust and generalizable action recognition from skeleton sequences remains a challenge due to the spatiotemporal complexity of human actions and the limitations of current modeling methods [2]. The computational demands of existing approaches further complicate their deployment in resource-constrained environments [3].

Graph Convolutional Networks (GCNs) have transformed skeleton-based HAR by naturally modeling the human body as a graph structure, where joints serve as nodes and bones as edges [4]. Early approaches such as ST-GCN [5], AS-GCN [6], and 2s-AGCN [7] demonstrate the effectiveness of applying graph convolutions to capture spatial relationships between joints, achieving substantial improvements over traditional methods. Subsequent work focuses on enhancing graph topology modeling through various strategies. MS-G3D [8] introduces multi-scale aggregation, CTR-GCN [9] proposes channel-wise topology refinement, achieving 92.4% accuracy on NTU RGB + D 60, and InfoGCN [10] leverages information-theoretic objectives to learn more discriminative representations. Recent advances, including HD-GCN [11], BlockGCN [12], and DeGCN [13], have achieved significant performance improvements on challenging benchmarks by improving graph convolutional structures and feature extraction mechanisms.

Subsequently, Transformer architectures have become widely used in skeleton-based HAR [14]. Pure attention-based methods were initially used [15,16], while hybrid methods combining GCNs with Transformer models have proven to be more effective. Recent work, such as SkateFormer [17], introduces skeletal-temporal relation partitioning to model different types of joint relationships systematically. MSAST [18] employs multi-scale attention mechanisms for hierarchical feature extraction. These hybrid architectures have achieved state-of-the-art

* Corresponding author.
E-mail addresses: yi.xia@autuni.ac.nz (Y. Xia), sira.yongchareon@aut.ac.nz (S. Yongchareon), raymond.lutui@aut.ac.nz (R. Lutui), michael.sheng@mq.edu.au (Q.Z. Sheng).

https://doi.org/10.1016/j.patcog.2026.113146
Received 28 September 2025; Received in revised form 23 December 2025; Accepted 20 January 2026
Available online 22 January 2026

performance on established benchmarks, demonstrating the complementary nature of local graph structures and global attention mechanisms.

However, our review of prior work still shows a key limitation in spatiotemporal feature encoding: existing mainstream approaches predominantly model spatial and temporal dependencies in a decoupled or loosely coupled manner, hindering their ability to capture the intricate interplay between these dimensions. Traditional methods usually adopt spatial GCN for joint relationship modeling, followed by temporal convolution or RNN for sequence modeling, treating these dimensions as independent processing streams [7]. Even recent advances in adaptive graph learning [19] and attention mechanisms [20] largely maintain this separation, learning spatial topologies primarily based on instantaneous joint features or simple temporal fusions. This decoupling fails to capture the inherent interdependence between spatial joint relationships and temporal dynamics, where the same spatial configuration may correspond to different actions depending on the temporal evolution patterns.

The limitations of decoupled spatiotemporal modeling are particularly evident in graph topology learning strategies [21]. Static GCN methods use fixed, predefined topologies and cannot adapt to the joint relationships of specific actions. Although dynamic GCN methods learn adaptive topologies, the adaptations are typically based on local spatiotemporal features, and spatial and temporal information are often processed in separate paths. This separation prevents the model from leveraging rich temporal context when determining the most relevant spatial relations for a given action. In addition, existing temporal modeling methods in skeleton-based HAR primarily rely on simple temporal convolutions or basic RNN variants, which struggle to capture the complex temporal patterns inherent in human motion, including both periodic and non-periodic components that characterize different actions [22].

We propose the Frequency-Aware Topology Learning Graph Convolutional Network (FATL-GCN), a novel architecture that rethinks the relationship between spatial and temporal modeling in skeleton-based HAR. Unlike previous methods that treat the dimensions separately, FATL-GCN introduces a coupled learning approach where temporal dynamics directly guide the construction of spatial graph topology. Our key insight is that the frequency features of joint movements captured through an enhanced Time2Vec [23] representation provide rich contextual information about the action being performed, which should inform how joints are connected in the spatial graph. Our approach addresses several challenges in current skeleton-based HAR. First, we address the limitation of fixed or locally adaptive graph topologies by introducing a global, frequency-aware context that captures the overall temporal signature of an action. Second, we enhance temporal representation beyond simple convolutions through Time2Vec [23], which provides a theoretically grounded method for encoding both periodic patterns (e.g., repetitive motions in “clapping”) and non-periodic progressions (e.g., the continuous motion of “standing up”). Third, we create an integrated spatio-temporal learning framework where spatial graph construction is explicitly conditioned on temporal understanding, enabling more semantically meaningful joint relationships that adapt based on the action’s temporal features. The overall design of our FATL-GCN is illustrated in Fig. 1, and the primary contributions of this work are summarized as follows:

- We propose FATL-GCN for skeleton-based human activity recognition, which leverages frequency-aware temporal context to guide adaptive spatial topology learning, enabling action-specific joint dependencies to emerge from temporal dynamics.
- We employ Time2Vec encoding to capture both periodic and non-periodic motion patterns, generating frequency-aware temporal context that guides spatial topology learning.
- We design context-guided dynamic graph convolution that translates this temporal context into adaptive spatial topology through multi-head attention, enabling joint relationships to adapt based on temporal dynamics rather than anatomical constraints alone.
- State-of-the-art results are achieved on NTU RGB+D 60, NTU RGB+D 120, and Kinetics-Skeleton, with ablation studies confirming the essential role of frequency-aware temporal encoding.

The remainder of this paper is organized as follows. Section 2 reviews related work in graph-based skeleton action recognition, attention mechanisms, and temporal modeling. Section 3 presents the detailed architecture of FATL-GCN, including the theoretical foundations of frequency-aware topology learning. Section 4 describes our experimental setup, presents evaluations on multiple benchmarks, and provides detailed ablation studies analyzing the contribution of key components. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related work

2.1. Early approaches in skeleton-based HAR

Early applications of deep learning to skeleton-based HAR relied on established models from sequence and image processing, but these methods were not inherently well-suited to the geometric properties of skeleton data. RNN-based methods [24] treat skeleton sequences as temporal sequences, processing joint coordinates in temporal order. However, the dimensionality reduction operation of flattening 3D coordinates into 1D vectors fundamentally destroys the spatial topology of the skeleton. For example, anatomically adjacent joints (such as the “wrist” and “elbow”) may appear far apart in the vector representation, whereas non-adjacent joints may appear close together. On the other hand, due to the widespread application of CNNs in computer vision, CNN-based methods [25] convert skeleton sequences into 2D pseudoimages with both temporal and joint dimensions. This strategy artificially imposes a regular Euclidean grid structure on non-Euclidean skeletal data, causing convolutions to aggregate features from anatomically meaningless neighborhoods. Furthermore, due to the importance of temporal features, such methods require larger convolution kernels to handle long-range joint dependencies. Both paradigms impose priors that are inconsistent with the human skeleton: RNNs assume a one-dimensional sequence, while CNNs assume a two-dimensional grid. This mismatch has led to a shift towards more reasonable graph-based approaches that naturally preserve the graph structure of the human skeleton through GCNs [5].

2.2. Graph convolutional networks in skeleton-based HAR

The human skeleton’s natural graph structure, with joints as vertices and bones as edges, aligns seamlessly with GCNs. By performing convolutions directly on skeletal graphs, GCNs preserve anatomical connectivity while overcoming the limitations of RNN and CNN approaches.

Predefined Graphs. Early GCN methods relied on fixed topologies from physical skeletal connections. ST-GCN [5] pioneered this approach with spatial graphs based on anatomical connections and temporal edges that link joints in frames. AS-GCN [6] extends this with action links that capture implicit dependencies along with structure links for high-order relationships. 2s-AGCN [7] introduces two-stream processing for joint and bone information simultaneously. Although these methods achieved 81–85% accuracy on NTU RGB+D 60, significantly improving over previous methods, their fixed topologies cannot adapt to different datasets.

Adaptive and Dynamic Graphs. Recent research has shifted to learnable graph structures, addressing limitations of fixed topology. MS-G3D [8] introduces multi-scale graph convolutions with learnable adjacency matrices capturing dependencies at different spatial scales. CTR-GCN [9] revolutionizes topology learning through channel-wise refinement, learning different structures per feature channel, and achieving 92.4% accuracy on NTU RGB+D 60, an 11% improvement over

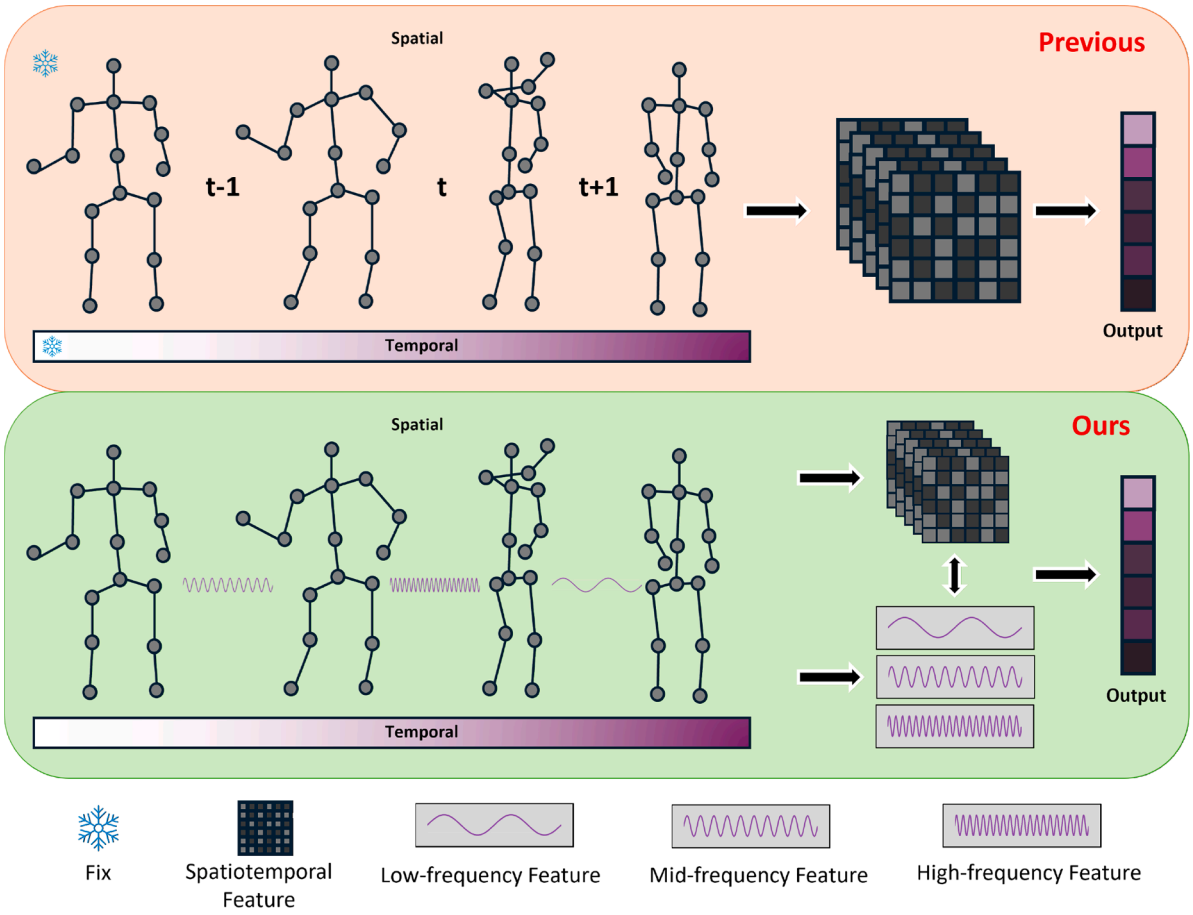

Fig. 1. Comparison between previous approaches and our proposed FATL-GCN. **Top:** Traditional approaches process spatial and temporal features separately with fixed or locally-adaptive topologies. **Bottom:** Our method extracts multi-frequency temporal patterns (low, mid, high frequencies) that directly guide adaptive spatial topology learning, enabling dynamic joint relationships based on temporal dynamics.

ST-GCN. InfoGCN [10] leverages mutual information maximization for semantically meaningful structures. Recent advances include HD-GCN [11] with hierarchical graph decomposition, BlockGCN [12] featuring block-wise adaptive learning, and DeGCN [13] employing deformable graph convolutions. Despite achieving over 93% accuracy on NTU RGB+D 60, current adaptive methods primarily learn spatial topology from instantaneous or locally aggregated features, lacking rich temporal context for more semantically meaningful graph structures. Recent methods such as GSTLN [26], SPIANet [27], and LG-SGNet [28] have explored synergistic topology learning and hierarchical feature extraction. However, these approaches still treat spatial topology learning and temporal feature extraction as largely independent, motivating our frequency-aware approach.

2.3. Attention and transformers in skeleton-based HAR

Transformer-based methods utilize self-attention mechanisms to capture arbitrary joint dependencies across spatial and temporal dimensions, thereby flexibly modeling complex spatiotemporal features, whereas graph convolutions struggle to extract long-range dependencies.

Pure Attention Models. ST-TR [15] proposes pure attention mechanisms for skeleton HAR, abandoning graph structures for self-attention. Its spatial and temporal self-attention modules model intra-frame and inter-frame relationships without skeletal connection constraints. DSTA-Net [16] advances this through spatiotemporal attention decoupling, separated position encoding, and spatial global regularization, achieving SOTA performance on NTU RGB+D 60/120. Recent variants explore hierarchical attention, multi-scale temporal modeling, and efficient attention approximations. While excelling at long-range dependencies and action-specific relationships without topology restrictions, these models generally require more computational resources than GCNs.

Hybrid GCN-Transformer Models. Recent hybrid architectures combine GCN's efficient local extraction with Transformer's global modeling. Notable examples include SkateFormer [17] with skeleton-temporal relation partitioning, MSAST [18] with multi-scale attention, and HAM-HGNet [29] with hypergraph convolution. These hybrids consistently outperform pure GCN or Transformer methods.

2.4. Temporal modeling in skeleton-based HAR

In skeleton-based HAR, the quality of temporal modeling is as important as spatial modeling. While spatial features capture the instantaneous configuration of joints, temporal features encode the dynamics of motion, fundamentally distinguishing different actions. However, current approaches face significant limitations in capturing complex temporal patterns.

Temporal Convolutions. TCNs dominate temporal modeling, with ST-GCN [5] establishing alternating spatiotemporal processing. Recent advances include multi-scale convolutions [8,30] and adaptive kernels [9]. However, fixed kernels struggle to capture periodicity in repetitive actions or nonlinear progression in complex motions [31].

Time2Vec. To overcome these limitations, we propose leveraging Time2Vec [23], which provides model-agnostic vector representations for time indices, capturing both periodic and non-periodic patterns through learnable sinusoidal and linear functions [32]. Its temporal en-

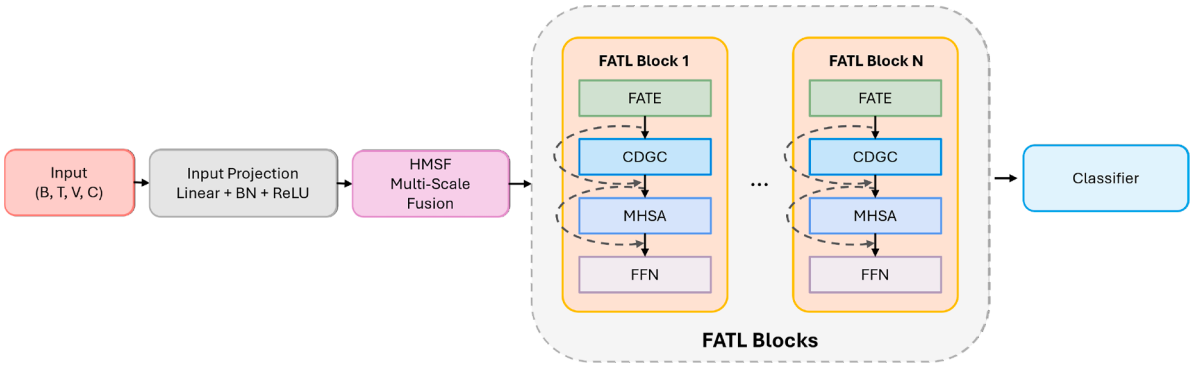

Fig. 2. The detailed architecture of FATL-GCN. An input skeleton sequence is first processed by the HMSF module. The output is then fed into a series of N FATL blocks. Each FATL block follows a pre-norm residual architecture. The FATE module generates a frequency-aware temporal context vector $\mathbf{H}_{\text{context}}$, which is used by the CDGC module to dynamically generate a spatial graph topology. Standard MHSA and FFN layers follow to capture global dependencies.

coding invariance and seamless neural architecture integration enable FATL-GCN to capture richer temporal dynamics, ranging from repetitive action frequencies to linear progressions, and subsequently use these dynamics to construct more semantically meaningful spatial graphs. To our knowledge, this represents the first application of Time2Vec to skeleton-based HAR, addressing the critical gap where existing methods use sophisticated spatial GCNs but relatively simple temporal modeling.

3. Methodology

This section provides a detailed technical description of our proposed FATL-GCN. FATL-GCN is designed as a multi-stage architecture for processing skeleton sequences to learn deeply coupled spatio-temporal representations. In Section 3.1, we first introduce the overall framework, which outlines the complete data flow from input to classifier. Subsequently, in Section 3.2, we provide a detailed explanation of the Hierarchical Multi-Scale Fusion (HMSF) module, an initial module for multi-scale feature enhancement. Finally, the core innovation of this work is detailed in Section 3.3, which describes the Frequency-Aware Topology Learning (FATL) module and its constituent components, responsible for implementing our novel frequency-guided learning mechanism.

3.1. Overall framework

The input to our model is a skeleton sequence represented as a tensor $\mathbf{X} \in \mathbb{R}^{B\times T\times V\times C_{\text{in}}}$, where B is the batch size, T is the number of temporal frames, V is the number of joints, and C_{in} is the dimension of the input coordinates (e.g., 3 for x, y, z). The overall architecture of FATL-GCN, depicted in Fig. 2, consists of four main stages:

1. Input Embedding: The input tensor $\mathbf{X}$ is first projected to a higher-dimensional feature space C_{model} and normalized. A physical adjacency matrix based on the skeletal structure is also initialized to provide structural priors.
2. Hierarchical Multi-Scale Fusion (HMSF): An initial feature enrichment module that captures spatio-temporal features at multiple receptive fields.
3. FATL Blocks: A stack of N_{blocks} core Frequency-Aware Topology Learning blocks that perform the main spatio-temporal feature extraction.
4. Classification Head: A global average pooling layer followed by a linear classifier to produce the final action classifications.

3.2. Hierarchical multi-scale fusion (HMSF)

Robust skeleton-based HAR requires the ability to extract multi-scale spatio-temporal features. Human actions exhibit inherent multi-scale features: fine-grained actions like "typing" require precise local joint movements, while dynamic actions like "jumping" depend on large-scale global body coordination. To address this challenge and provide rich, scale-invariant feature representations from the input stage, we propose the Hierarchical Multi-Scale Fusion (HMSF) module as the initial component of our architecture.

As depicted in Fig. 3, the HMSF module adopts a parallel-branch structure. The embedded features from the input projection are permuted to $\mathbf{X}' \in \mathbb{R}^{B\times C_{\text{model}}\times T\times V}$ to facilitate efficient 2D convolution operations along the temporal and spatial dimensions. The module processes this feature map through four concurrent transformation pathways designed to capture features at different scales:

$$\mathbf{Y}_k = f_k(\mathbf{X}'), k \in \{1, 3, 5, \text{Pool}\} \tag{1}$$

where f_k represents the transformation function of each branch. Specifically, the module comprises: (1) a 1×1 convolution for channel interaction; (2, 3) factorized depthwise convolutions (kernel sizes 3×3 and 5×5) to capture local and mid-range dependencies; and (4) a global average pooling branch to incorporate scene-level context. Notably, the factorized design significantly lowers the computational complexity from $\mathcal{O}(k^2C^2)$ to $\mathcal{O}(2kC)$ without sacrificing representational power.

Each branch projects features to an intermediate channel dimension $C_{\text{mid}} = C_{\text{model}}/4$. The features from all four pathways are then concatenated along the channel dimension:

$$\mathbf{Y}_{\text{cat}} = [\mathbf{Y}_1; \mathbf{Y}_3; \mathbf{Y}_5; \mathbf{Y}_{\text{Pool}}] \in \mathbb{R}^{B\times 4C_{\text{mid}}\times T\times V} \tag{2}$$

where the concatenated dimension $4C_{\text{mid}} = C_{\text{model}}$ preserves the original feature dimensionality, and subsequently fused via a final 1×1 convolution followed by batch normalization and ReLU activation:

$$\mathbf{Y}_{\text{HMSF}} = \text{ReLU}(\text{BN}(\text{Conv}_{1\times 1}(\mathbf{Y}_{\text{cat}}))) \tag{3}$$

The HMSF module is an efficient feature processing module that provides semantically rich multi-scale feature representations for the core FATL module. By capturing local joint relations and global skeleton context in parallel, the HMSF module enables the subsequent frequency-guided topology learning component to focus on adaptive graph structure discovery rather than low-level feature extraction.

3.3. The frequency-aware topology learning (FATL) block

The FATL block is the core innovation of our architecture, designed to overcome the fundamental limitation of existing methods that decouple spatial and temporal modeling. Each FATL block consists of three synergistic components that progressively refine spatio-temporal representations: (1) a Frequency-Aware Temporal Encoding (FATE) module that generates a global temporal context capturing action-specific motion patterns through frequency-domain analysis; (2) a Context-Guided Dynamic Graph Convolution (CDGC) module that leverages this temporal context to adaptively modulate spatial graph topology; and (3) global

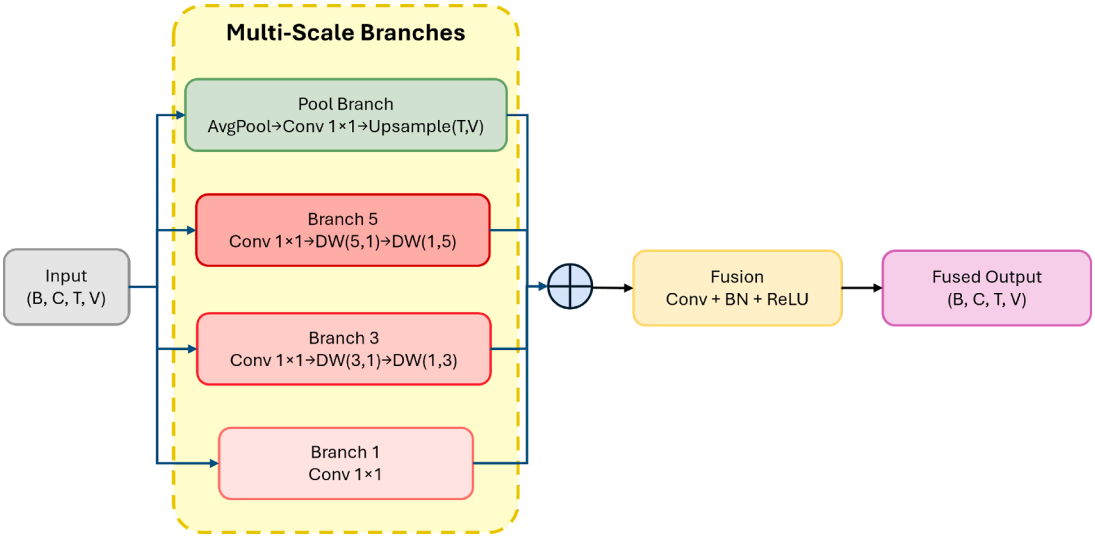

Fig. 3. Architecture of the HMSF module. Input (B, C, T, V) is processed through four parallel branches with different receptive fields, each producing $(B, C_{\text{mid}}, T, V)$ where $C_{\text{mid}} = C/4$: direct 1×1 convolution, factorized 3×3 and 5×5 depthwise convolutions, and global pooling with upsampling. Features are concatenated to $(B, 4C_{\text{mid}}, T, V)$ and fused via 1×1 convolution, batch normalization, and ReLU to output (B, C, T, V).

dependency refinement layers comprising Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN) that capture complementary patterns beyond the graph structure.

The FATL block employs a pre-layer normalization (Pre-LN) residual architecture for enhanced training stability. Given an input tensor $\mathbf{X}_{\text{block}} \in \mathbb{R}^{B\times T\times V\times C_{\text{model}}}$ from either the HMSF module or a previous FATL block, the processing pipeline follows:

3.3.1. *Frequency-aware temporal encoding (FATE)*

The FATE module generates a frequency-aware global temporal context that captures both periodic and non-periodic motion patterns inherent in human actions. This context serves as the guidance signal for subsequent spatial topology learning, establishing the critical link between temporal dynamics and spatial relationships.

The FATE module, shown in the left panel of Fig. 4, builds upon the Time2Vec [23] to create temporal embeddings that explicitly encode frequency information. For each temporal position $t \in \{0, 1, \ldots, T-1\}$, we compute a normalized time index $\tau_t = t/(T-1)$ and generate its frequency-aware embedding:

$$\mathbf{t2v}(\tau_t) = [\mathbf{v}_{\text{linear}}(\tau_t); \mathbf{v}_{\text{periodic}}(\tau_t)] \in \mathbb{R}^{d_{\text{t2v}}} \tag{4}$$

where the total embedding dimension is $d_{\text{t2v}} = d_{\text{linear}} + K$. The linear component $\mathbf{v}_{\text{linear}}(\tau_t) = \mathbf{w}_0\tau_t + \boldsymbol{\varphi}_0 \in \mathbb{R}^{d_{\text{linear}}}$ captures monotonic temporal progression (essential for non-periodic actions like sitting or standing), and the periodic component:

$$\mathbf{v}_{\text{periodic}}(\tau_t) = [\sin(\omega_1\tau_t + \varphi_1), \ldots, \sin(\omega_K\tau_t + \varphi_K)]^T \in \mathbb{R}^K \tag{5}$$

models rhythmic patterns through K learnable sinusoidal bases. Here, ω_k and φ_k are learnable frequency and phase parameters. We set $K = 5$ and $d_{\text{linear}} = 3$, yielding $d_{\text{t2v}} = 8$, based on empirical analysis showing human actions typically exhibit dominant frequencies in the 0.5–5 Hz range [33].

The temporal embeddings are integrated with the input features through concatenation and projection. First, we normalize the input:

$$\mathbf{X}_{\text{norm}} = \text{LayerNorm}(\mathbf{X}_{\text{block}}) \in \mathbb{R}^{B\times T\times V\times C_{\text{model}}} \tag{6}$$

Then, for each time step, we concatenate the normalized features with the temporal embedding and project to maintain dimensional consistency:

$$\mathbf{X}^{(t)}_{\text{temp}} = [\mathbf{X}^{(t)}_{\text{norm}}; \mathbf{T}^{(t)}]\mathbf{W}^T_{\text{proj}} + \mathbf{b}_{\text{proj}} \tag{7}$$

where $\mathbf{X}^{(t)}_{\text{norm}} \in \mathbb{R}^{B\times V\times C_{\text{model}}}$ denotes the features at time t, $\mathbf{T}^{(t)} \in \mathbb{R}^{B\times V\times d_{\text{t2v}}}$ is the temporal embedding $\mathbf{t2v}(\tau_t)$ broadcast across all V joints, $[\cdot;\cdot]$ denotes concatenation along the feature dimension, and $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{C_{\text{model}}\times(C_{\text{model}}+d_{\text{t2v}})}$ is the learnable projection matrix.

The global temporal context is obtained through temporal pooling:

$$\mathbf{H}_{\text{context}} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{X}^{(t)}_{\text{temp}} \in \mathbb{R}^{B\times V\times C_{\text{model}}} \tag{8}$$

This pooling operation produces a temporally-invariant representation where each joint's feature vector encodes the overall temporal dynamics of its movement throughout the sequence. Joints exhibiting similar temporal patterns will have similar representations in $\mathbf{H}_{\text{context}}$, providing a principled basis for topology learning.

The effectiveness of using frequency-domain temporal context to guide spatial topology learning rests on a fundamental observation about human motion: joints that participate in coordinated movements exhibit correlated temporal patterns, and this correlation manifests most clearly in their frequency characteristics. We formalize this intuition through the following analysis.

Consider two joints i and j with temporal feature sequences $\mathbf{f}_i(t)$ and $\mathbf{f}_j(t)$ over T time steps. Traditional approaches compute spatial affinity based on instantaneous features or simple temporal aggregations. In contrast, our frequency-aware approach leverages the temporal frequency signature encoded in $\mathbf{H}_{\text{context}}$. The Time2Vec encoding effectively projects each joint's temporal trajectory into a frequency-enriched space where the representation $\mathbf{h}_i \in \mathbb{R}^{C_{\text{model}}}$ captures both periodic oscillations (through sinusoidal components) and directional trends (through linear components).

The key insight is that functional correlation between joints, defined as their tendency to move together during specific actions, manifests as similarity in their frequency signatures. For repetitive actions like "clapping" or "waving", coordinated joints exhibit similar dominant frequencies in their periodic components. For progressive actions like "standing up" or "sitting down", functionally related joints share similar temporal pacing captured by the linear components. This relationship can be expressed formally: joints with high temporal correlation $\rho_{ij} = \text{corr}(\mathbf{f}_i, \mathbf{f}_j)$ will have similar frequency-domain representations, leading to high attention scores S_{ij} in our CDGC module.

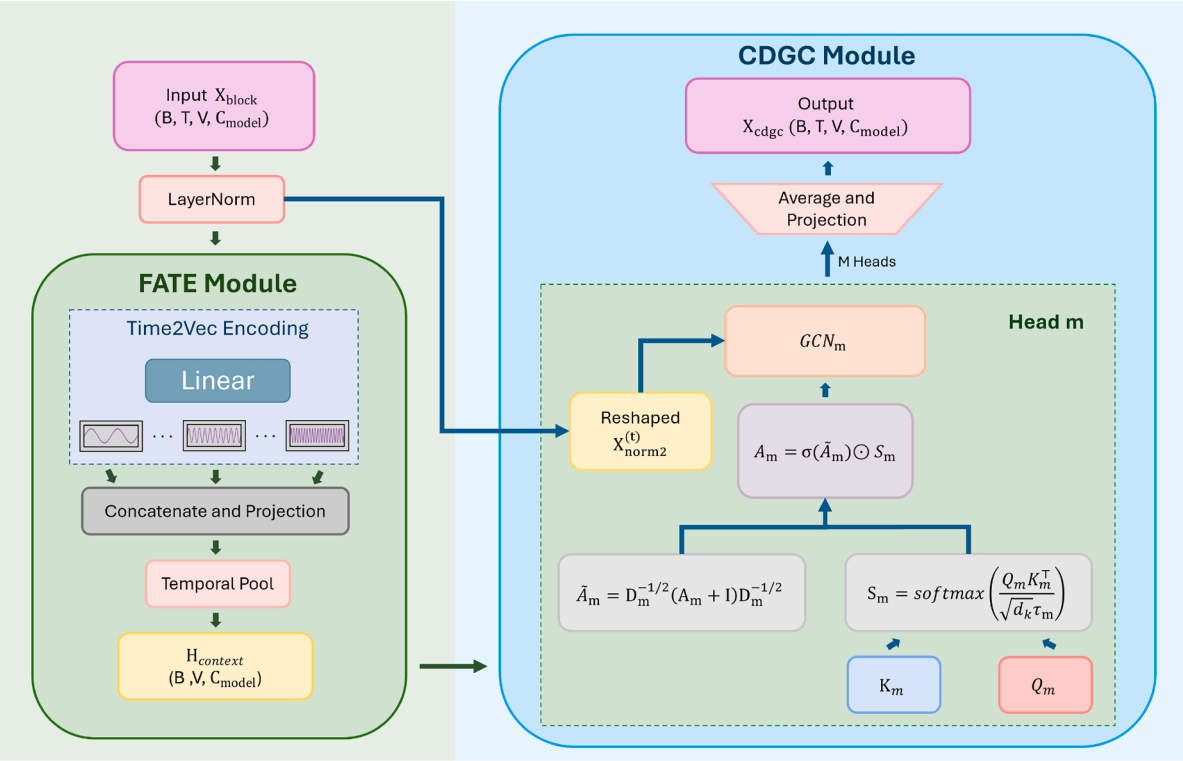

Fig. 4. Core components of the FATL block: FATE module (Left) generates frequency-aware context $\mathbf{H}_{\text{context}} \in \mathbb{R}^{B\times V\times C}$ from input $\mathbf{X}_{\text{block}} \in \mathbb{R}^{B\times T\times V\times C}$ through Time2Vec encoding and temporal pooling. CDGC module (Right) uses $\mathbf{H}_{\text{context}}$ to compute M attention heads with query/key $\mathbf{Q}_m, \mathbf{K}_m \in \mathbb{R}^{B\times V\times d_k}$ producing dynamic adjacency $\mathbf{A}_m \in \mathbb{R}^{B\times V\times V}$ for graph convolution, outputting $\mathbf{X}_{\text{cdgc}} \in \mathbb{R}^{B\times T\times V\times C}$. Here B = batch size, T = 64 frames, V = 25 joints, $C = C_{\text{model}}$ = 128 or 256, $d_k = C/M$.

Mathematically, the attention mechanism in CDGC computes similarity as:

$$S_{ij} \propto \exp\left(\frac{\langle \mathbf{Q}\mathbf{h}_i, \mathbf{K}\mathbf{h}_j \rangle}{\sqrt{d_k}\cdot\tau}\right) \tag{9}$$

where $\mathbf{h}_i \in \mathbb{R}^{C_{\text{model}}}$ denotes the ith row of $\mathbf{H}_{\text{context}}$ (the context vector of joint i), and $\mathbf{Q}$, $\mathbf{K}$ are learned projection matrices corresponding to $\mathbf{W}_Q^{(m)}$, $\mathbf{W}_K^{(m)}$ in the formal definition below. When $\mathbf{h}_i$ and $\mathbf{h}_j$ encode similar frequency patterns, their projected representations align in the query-key space, yielding high S_{ij}. This effectively discovers edges in the spatial graph that correspond to functional rather than purely anatomical relationships.

This frequency-based approach provides several advantages over topology learning from instantaneous features. First, it is inherently robust to temporal phase shifts: two joints performing the same cyclic motion with different phase offsets will still exhibit similar frequency components, ensuring stable edge detection. Second, the temporal pooling operation provides noise suppression, as transient perturbations are attenuated while consistent frequency patterns are preserved. Third, the learnable sinusoidal bases in Time2Vec can capture diverse frequency ranges across the action spectrum, enabling the model to emphasize relevant temporal scales for different action categories.

3.3.2. Context-guided dynamic graph convolution (CDGC)

The CDGC module represents our key innovation: using frequency-aware temporal context to dynamically modulate spatial graph topology. Unlike standard graph convolutions with fixed adjacency matrices or self-attention mechanisms where queries and keys derive from the same source, our approach generates attention weights from the global temporal context, ensuring that spatial relationships are determined by temporal motion patterns.

As depicted in the right panel of Fig. 4, the module employs M parallel heads to capture diverse relational patterns. For each head $m \in \{1, \ldots, M\}$, we generate context-guided query and key matrices:

$$\mathbf{Q}_m = \text{LayerNorm}(\mathbf{H}_{\text{context}}\mathbf{W}_Q^{(m)}), \quad \mathbf{K}_m = \text{LayerNorm}(\mathbf{H}_{\text{context}}\mathbf{W}_K^{(m)}) \tag{10}$$

where $\mathbf{W}_Q^{(m)}, \mathbf{W}_K^{(m)} \in \mathbb{R}^{C_{\text{model}}\times d_k}$ are learnable projection matrices with $d_k = C_{\text{model}}/M$, and the resulting $\mathbf{Q}_m, \mathbf{K}_m \in \mathbb{R}^{B\times V\times d_k}$.

The context-guided attention scores are computed as:

$$\mathbf{S}_m = \text{softmax}\left(\frac{\mathbf{Q}_m\mathbf{K}_m^T}{\sqrt{d_k}\cdot\tau_m}\right) \in \mathbb{R}^{B\times V\times V} \tag{11}$$

where $\tau_m \in [0.1, 10.0]$ is a learnable temperature parameter that controls the sharpness of attention distribution. These attention scores represent the temporal similarity between joints and form the basis for dynamic topology.

To preserve anatomical priors while enabling adaptive learning, we combine the attention scores with a learnable base topology:

$$\mathbf{A}_m = \sigma(\mathbf{A}_{\text{base}}^{(m)}) \odot \mathbf{S}_m \in \mathbb{R}^{B\times V\times V} \tag{12}$$

where $\mathbf{A}_{\text{base}}^{(m)} \in \mathbb{R}^{V\times V}$ is a learnable parameter initialized from the physical skeleton structure, $\sigma(\cdot)$ denotes the sigmoid function that constrains values to $[0, 1]$, and $\odot$ represents element-wise multiplication. This multiplicative fusion enables the model to retain anatomical constraints while adapting connections based on action-specific temporal patterns.

The adjacency matrix undergoes symmetric normalization to ensure numerical stability:

$$\bar{\mathbf{A}}_m = \mathbf{A}_m + \mathbf{I} \tag{13}$$

$$\hat{\mathbf{A}}_m = \mathbf{D}_m^{-1/2}\bar{\mathbf{A}}_m\mathbf{D}_m^{-1/2} \tag{14}$$

where $\bar{\mathbf{A}}_m$ incorporates self-connections through the identity matrix $\mathbf{I} \in \mathbb{R}^{V\times V}$, and $\mathbf{D}_m \in \mathbb{R}^{V\times V}$ is the diagonal degree matrix with $[\mathbf{D}_m]_{ii} = \sum_j [\bar{\mathbf{A}}_m]_{ij}$. This normalization bounds the magnitude of aggregated features regardless of node degree.

Graph convolution is then applied to the normalized input features at each time step:

$$\mathbf{Z}_m^{(t)} = \hat{\mathbf{A}}_m \mathbf{X}_{\text{norm}}^{(t)} \mathbf{W}_g^{(m)} \tag{15}$$

where $\mathbf{X}_{\text{norm}}$ reused here to maintain computational efficiency, and $\mathbf{W}_g^{(m)} \in \mathbb{R}^{C_{\text{model}} \times C_{\text{model}}}$ are learnable graph convolution weights. Note that while the FATE module uses $\mathbf{X}_{\text{norm}}$ to generate the frequency-aware context $\mathbf{H}_{\text{context}}$, the graph convolution operates on the same normalized features without the Time2Vec enrichment, allowing the topology learning signal to remain decoupled from the feature transformation path.

The multi-head outputs are aggregated through averaging and projection:

$$\mathbf{X}_{\text{cdgc}} = \mathbf{W}_{\text{out}} \left(\frac{1}{M} \sum_{m=1}^{M} \mathbf{Z}_m \right) + \mathbf{b}_{\text{out}} \in \mathbb{R}^{B \times T \times V \times C_{\text{model}}} \tag{16}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{C_{\text{model}} \times C_{\text{model}}}$ performs feature mixing across heads. This multi-head design enables specialization, with different heads potentially focusing on different body parts or motion patterns.

3.3.3. Global dependency refinement

While CDGC effectively captures frequency-guided spatial relationships, certain long-range dependencies may not conform to any graph structure. We therefore incorporate standard MHSA and FFN Transformer components to model arbitrary global patterns that complement the structured graph convolutions.

The complete FATL block employs a three-stage residual architecture with learnable scaling parameters. Denoting the combined FATE and CDGC operations as FATL-GC(·), the processing flow is:

$$\mathbf{X}_{\text{res1}} = \mathbf{X}_{\text{block}} + \alpha_1 \cdot \text{Dropout}_{p_1}(\text{FATL-GC}(\text{LayerNorm}(\mathbf{X}_{\text{block}}))) \tag{17}$$

$$\mathbf{X}_{\text{res2}} = \mathbf{X}_{\text{res1}} + \alpha_2 \cdot \text{Dropout}_{p_2}(\text{MHSA}(\text{LayerNorm}(\mathbf{X}_{\text{res1}}))) \tag{18}$$

$$\mathbf{X}_{\text{out}} = \mathbf{X}_{\text{res2}} + \alpha_3 \cdot \text{Dropout}_{p_3}(\text{FFN}(\text{LayerNorm}(\mathbf{X}_{\text{res2}}))) \tag{19}$$

where FATL-GC(·) encompasses both the frequency-aware context generation and context-guided graph convolution as detailed above. The parameters $\alpha_1, \alpha_2, \alpha_3$ are learnable scalars initialized to 1.0, enabling the model to adaptively weight the contribution of each component during training. We employ progressive dropout rates $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$ to account for increasing feature abstraction depth.

The MHSA module operates on the flattened spatiotemporal sequence of shape $(B, T \times V, C_{\text{model}})$, computing:

$$\text{MHSA}(\mathbf{X}) = \text{MultiHeadAttention}(\mathbf{X}_{\text{flat}}, \mathbf{X}_{\text{flat}}, \mathbf{X}_{\text{flat}}) \tag{20}$$

where the output is reshaped back to $(B, T, V, C_{\text{model}})$. The FFN consists of two linear transformations with ReLU activation and an expansion ratio of 4:

$$\text{FFN}(\mathbf{X}) = \text{ReLU}(\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{21}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C_{\text{model}} \times 4C_{\text{model}}}$ and $\mathbf{W}_2 \in \mathbb{R}^{4C_{\text{model}} \times C_{\text{model}}}$ are the expansion and projection weights, respectively.

This three-stage design creates a hierarchical feature refinement process: CDGC captures structured, frequency-guided spatial relationships; MHSA models arbitrary global dependencies; and FFN performs channel-wise feature transformation. The learnable scaling parameters allow the model to dynamically balance these complementary mechanisms based on the specific requirements of different actions, resulting in a flexible yet principled approach to spatio-temporal modeling in skeleton-based action recognition.

4. Experiments

We conduct extensive experiments to validate the performance of FATL-GCN on three large-scale benchmark datasets. We compare our model with state-of-the-art methods and perform detailed ablation studies to analyze the contribution of each component.

4.1. Datasets and protocols

NTU RGB+D 60 [34] is a widely used benchmark containing 56,880 skeleton sequences across 60 action classes. We follow the two standard evaluation protocols: Cross-Subject (X-Sub), where 40 subjects are split into training and testing sets, and Cross-View (X-View), where samples from camera 1 are used for testing and samples from cameras 2 and 3 are for training.

NTU RGB+D 120 [35] is a large-scale extension of NTU-60, containing 114,480 videos across 120 action classes performed by 106 subjects. This dataset introduces greater diversity and complexity. We report results on the two official protocols: Cross-Subject (X-Sub), with a 53/53 subject split for training/testing, and Cross-Setup (X-Set), where training and testing data are split by camera setup IDs.

Kinetics-Skeleton [32] is a large-scale, "in-the-wild" dataset derived from the Kinetics-400 video dataset. It contains approximately 240,000 training clips and 20,000 validation clips across 400 action classes. The skeletons are estimated using pose estimation tools, which introduce noise and make it a challenging benchmark to evaluate. We report Top-1 and Top-5 accuracy on the validation set, following standard protocol.

4.2. Implementation details

To evaluate the performance-efficiency trade-off of our architecture, we define two model configurations:

- FATL-GCN-B (Base): An efficient configuration with 8 FATL blocks, a model dimension (C_{model}) of 128, 4 topology heads, and 4 MHSA heads.
- FATL-GCN-L (Large): A larger configuration designed for maximum accuracy, with 12 FATL blocks, a model dimension (C_{model}) of 256, 8 topology heads, and 8 MHSA heads.

Our model is implemented using PyTorch on NVIDIA A100 GPUs. We adopt the SGD optimizer with momentum of 0.9 and weight decay of 0.0004. The initial learning rate is set to 0.1 and decays by a factor of 0.1 at epochs 110 and 120. A warmup strategy is applied in the first 5 epochs. All input skeleton sequences are resized to 64 frames. We train separate models for four modalities (joint, bone, joint motion, bone motion) and fuse them by averaging softmax scores at inference. For NTU RGB+D datasets, the batch size is 64 and training runs for 140 epochs; for Kinetics-Skeleton, batch size is 64 with 120 epochs.

4.3. Comparison with state-of-the-art methods

We evaluate FATL-GCN against state-of-the-art skeleton-based HAR methods on three benchmark datasets. Tables 1, 2, and 3 present the comparative results, demonstrating that our approach achieves new state-of-the-art performance across all benchmarks.

Results on NTU RGB+D 60. As shown in Table 1, FATL-GCN-L achieves 93.8% and 97.5% accuracy on the X-Sub and X-View protocols, respectively, outperforming the previous best method DeGCN [13] by 0.2% and 0.1%. Even our base model, FATL-GCN-B, surpasses most existing approaches while maintaining computational efficiency. Moreover, our method outperforms recent advanced architectures, including BlockGCN [12] and LG-SGNet [28], which also explore enhanced topology modeling. The superior performance validates the effectiveness of our frequency-aware temporal guidance for learning dynamic joint relationships.

Results on NTU RGB+D 120. The larger and more challenging NTU-120 dataset provides a more rigorous evaluation of model generalization. As reported in Table 2, FATL-GCN-L achieves 91.9% on X-Sub and 93.1% on X-Set, establishing new benchmarks with improvements of 0.9% and 1.0% over the previous best method DeGCN [13]. The substantial performance gains on this dataset, which contains twice as many

Table 1
Performance comparison (Top-1 Acc. %) on NTU RGB+D 60[a].

Method	X-Sub	X-View
Shift-GCN [4]	90.7	96.5
MS-G3D [8]	91.5	96.2
DSTA-Net [16]	91.5	96.4
CTR-GCN [9]	92.4	96.8
InfoGCN [10]	93.0	97.1
ML-STGNet [36]	91.9	96.2
SAN-GCN [37]	92.1	96.2
GSTLN [26]	91.9	96.6
SPIANet [27]	92.8	96.8
FR-Head [38]	92.8	96.8
BlockGCN [12]	93.1	97.0
DeGCN [13]	93.6	97.4
LG-SGNet [28]	93.1	96.7
FATL-GCN-B (Ours)	93.3	97.2
FATL-GCN-L (Ours)	**93.8**	**97.5**

[a]All methods use four-stream fusion (4s).

Table 2
Performance comparison (Top-1 Acc. %) on NTU RGB+D 120[a].

Method	X-Sub	X-Set
Shift-GCN [4]	85.9	87.6
MS-G3D [8]	86.9	88.4
DSTA-Net [16]	86.6	89.0
CTR-GCN [9]	88.9	90.6
InfoGCN [10]	89.4	90.7
ML-STGNet [36]	88.6	90.0
SAN-GCN [37]	88.7	90.1
GSTLN [26]	88.1	89.3
SPIANet [27]	89.2	90.4
FR-Head [38]	89.5	90.9
BlockGCN [12]	90.3	91.5
DeGCN [13]	91.0	92.1
LG-SGNet [28]	89.4	91.0
MSAST [18]	88.7	91.6
HAM-HGNet [29]	90.1	90.8
FATL-GCN-B (Ours)	91.1	92.5
FATL-GCN-L (Ours)	**91.9**	**93.1**

[a] All methods use four-stream fusion (4s).

Table 3
Performance comparison on Kinetics-Skeleton.

Method	Top-1 Acc. (%)	Top-5 Acc. (%)
ST-GCN [5]	30.7	52.8
AS-GCN [6]	34.8	56.5
MS-G3D [8]	38.0	60.9
MST-GCN [30]	38.1	60.8
PYSKL [39]	49.1	-
HD-GCN [11]	40.9	63.5
DS-GCN [19]	50.6	-
FATL-GCN-B (Ours)	50.4	67.5
FATL-GCN-L (Ours)	**51.7**	**68.1**

action classes and greater intra-class variation, demonstrate that our frequency-aware topology learning effectively captures discriminative spatio-temporal patterns even in complex scenarios. Compared to recent transformer-based approaches, such as MSAST [18], our method shows better performance while maintaining a more efficient architecture.

Results on Kinetics-Skeleton. Table 3 presents the results on the challenging Kinetics-Skeleton dataset, where skeleton data are obtained through pose estimation and contain significant noise. FATL-GCN-L achieves 51.7% Top-1 and 68.1% Top-5 accuracy, outperforming the recent DS-GCN [19] by 1.1% on Top-1 accuracy. The performance of our model on this noisy, in-the-wild dataset shows that our frequency-aware approach is effective in coping with imperfect skeleton data, as the temporal-frequency patterns help eliminate the impact of noise on spatial-temporal feature extraction.

Table 4
Model complexity and efficiency analysis. FLOPs are reported for a single clip on NTU-120. Acc. is Top-1 on NTU-120 X-Sub.

Method	Params (M)	FLOPs (G)	Acc. (%)
Shift-GCN [4]	2.8	10.0	85.9
CTR-GCN [9]	1.46	7.88	88.9
MS-G3D [8]	6.4	48.5	86.9
DSTA-Net [16]	14.0	64.7	86.6
InfoGCN [10]	2.1	16.5	89.4
HD-GCN [11]	10.08	9.6	90.1
MSAST [18]	4.14	11.60	88.7
HAM-HGNet [29]	2.73	9.85	90.1
FATL-GCN-B (Ours)	5.8	14.2	91.1
FATL-GCN-L (Ours)	11.26	17.15	91.9

Table 4 analyzes computational efficiency. FATL-GCN-B (5.8M parameters, 14.2 GFLOPs) achieves 91.1% accuracy, outperforming MS-G3D [8] (6.4M, 48.5G, 86.9%) with significantly lower cost. Compared to Transformer-based methods, FATL-GCN-B improves upon MSAST [18] by 2.4% while adding only 2.6 GFLOPs. FATL-GCN-L (11.26M, 17.15G) reaches 91.9%, achieving 5.3% higher accuracy than DSTA-Net [16] (14.0M, 64.7G) at less than one-third the computational cost. This efficiency stems from our design where frequency-guided graph convolutions handle structured spatial modeling while attention layers capture only residual dependencies, avoiding full spatiotemporal attention overhead.

These efficiency characteristics make FATL-GCN particularly suitable for deployment in resource-constrained environments, such as edge devices or real-time systems, without sacrificing recognition accuracy. The Base configuration provides an optimal balance for applications prioritizing inference speed, while the Large configuration serves scenarios where accuracy is paramount and additional computational resources are available.

Fig. 5 presents visualization results on NTU RGB+D datasets. We apply t-SNE to compare the feature distributions of ST-GCN (baseline) and our FATL-GCN on NTU RGB+D 120 X-Sub with 10 randomly selected classes under identical settings. As shown in Fig. 5(a), the baseline model produces overlapping clusters for several action categories, indicating confusion in the learned feature space. In contrast, Fig. 5(b) demonstrates that FATL-GCN achieves more compact intra-class clustering and larger inter-class margins, suggesting that frequency-aware temporal context helps the model learn more discriminative representations. Moreover, Fig. 5(c) reports per-class accuracy on NTU RGB+D 60. Our method achieves consistently high performance across action categories, with particularly strong results on actions involving complex temporal dynamics (e.g., “reading”, “writing”). The improvement from Base to Large configuration is relatively uniform across classes, indicating that the additional capacity benefits all action types rather than overfitting to specific categories.

4.4. Ablation studies

To understand the contribution of each component in FATL-GCN and validate our design choices, we conducted extensive ablation studies on the NTU-120 X-Sub benchmark using the FATL-GCN-B configuration.

4.4.1. Component analysis

Table 5 presents component ablation results. Replacing CDGC with a fixed graph causes the largest drop (3.5%), validating that dynamic topology learning is essential. Removing FATE results in 2.1% degradation, confirming that temporal context fundamentally shapes spatial relationship learning. HMSF contributes 0.7% through multi-resolution

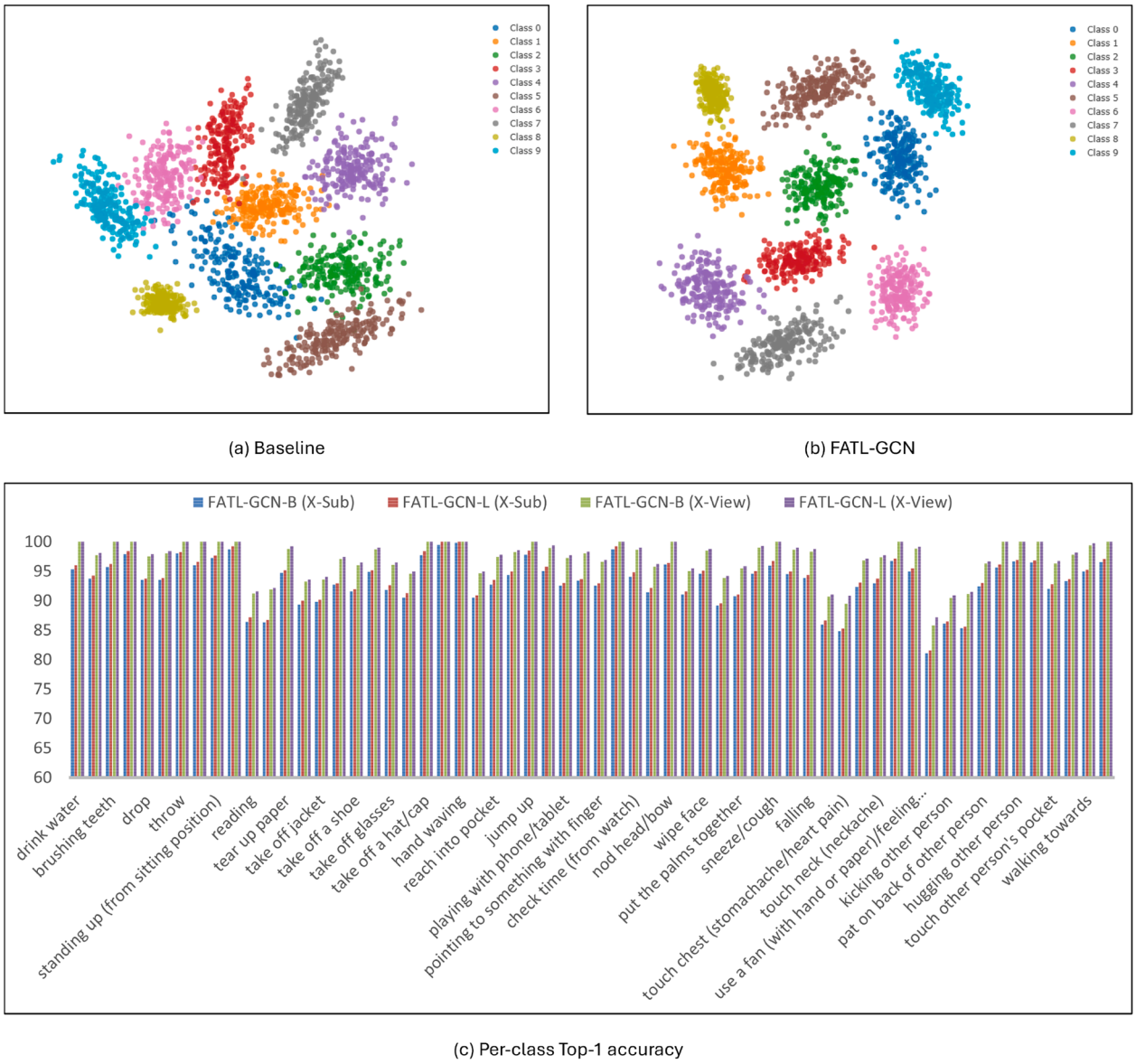

Fig. 5. Visualization results on NTU RGB + D 60 and NTU RGB + D 120. (a) t-SNE visualization of features from the baseline model (ST-GCN) for 10 randomly selected classes on NTU RGB + D 120 X-Sub. (b) t-SNE visualization of features from FATL-GCN for the same classes, showing improved cluster separation. (c) Per-class Top-1 accuracy (%) of FATL-GCN-B and FATL-GCN-L on NTU RGB + D 60 X-Sub and X-View benchmarks.

Table 5
Ablation study on core components of FATL-GCN-B on NTU-120 X-Sub.

Model Configuration	Top-1 Acc. (%)
Full FATL-GCN-B	**91.1**
w/o HMSF (use linear projection)	90.4 (−0.7) ↓
w/o FATE (no Time2Vec guidance)	89.0 (−2.1) ↓
w/o CDGC (use fixed physical graph)	87.6 (−3.5) ↓
w/o MHSA and FFN	89.9 (−1.2) ↓

Table 6
Analysis of HMSF branches on NTU-120 X-Sub.

HMSF Branch Configuration	Top-1 Acc. (%)
1×1 conv only	89.7 (−1.4) ↓
3×3 factorized only	90.0 (−1.1) ↓
5×5 factorized only	89.8 (−1.3) ↓
Global pooling only	88.5 (−2.6) ↓
$1 \times 1 + 3 \times 3 + 5 \times 5$	90.6 (−0.5) ↓
All branches (Ours)	**91.1**

features, while MHSA and FFN add 1.2% by capturing residual global dependencies.

4.4.2. Multi-Scale feature analysis

Table 6 shows that each HMSF branch captures complementary features. Global pooling alone performs worst (−2.6%), but removing it from the full configuration still causes 0.5% drop, indicating its value for scene-level context. The full multi-scale fusion enables adaptive feature weighting across scales.

4.4.3. Temporal encoding and topology learning

Table 7 shows two critical design decisions: the choice of temporal encoding method and the topology learning strategy. These results

Table 7
Ablation on temporal encoding, topology learning and pooling strategies on NTU-120 X-Sub.

Temporal Encoding Method	Top-1 Acc. (%)
Time2Vec (Ours)	**91.1**
Fixed Sinusoidal Positional Encoding	89.8 (−1.3) ↓
Learned Positional Embedding	89.5 (−1.6) ↓
No Explicit Temporal Encoding	88.2 (−2.9) ↓
Topology Learning Strategy	**Top-1 Acc. (%)**
Multiplicative ($\tilde{A} \times S$) (Ours)	**91.1**
Fixed Physical Graph Only	87.6 (−3.5) ↓
Learnable Static Only	89.2 (−1.9) ↓
Attention Dynamic Only	88.9 (−2.2) ↓
Pooling Strategy	**Top-1 Acc. (%)**
Average Pooling (Ours)	**91.1**
Max Pooling	90.4 (−0.7) ↓
Attention-based Pooling	90.8 (−0.3) ↓
Learnable Weighted Pooling	90.9 (−0.2) ↓

Table 8
Ablation on Time2Vec parameters on NTU-120 X-Sub.

K (periodic)	d_{linear}	Top-1 Acc. (%)
3	3	90.3 (−0.8) ↓
5	1	90.5 (−0.6) ↓
5	**3 (Ours)**	**91.1**
5	5	91.0 (−0.1) ↓
7	3	90.9 (−0.2) ↓
9	3	90.6 (−0.5) ↓

provide insights into how temporal and spatial modeling interact within our framework.

Time2Vec outperforms alternatives by 1.3–2.9%, combining learnable linear and periodic components to capture both progressive and cyclical motion patterns. For topology learning, our multiplicative fusion ($\tilde{A} \times S$) balances anatomical priors with dynamic adaptation, outperforming both fixed graphs (−3.5%) and pure attention (−2.2%).

Table 8 analyzes the impact of Time2Vec hyperparameters on recognition accuracy. The number of periodic components K determines the capacity to model different frequency patterns in human motion. Our choice of $K = 5$ achieves optimal performance, aligning with biomechanical studies showing that human movements exhibit dominant frequencies in the 0.5–5 Hz range [33]. Fewer periodic components ($K = 3$) result in 0.8% accuracy drop, suggesting insufficient capacity to capture the diverse frequency characteristics of different actions. Conversely, excessive periodic components ($K = 9$) lead to 0.5% degradation, likely due to overfitting to spurious frequency patterns or increased optimization difficulty.

The linear dimension d_{linear} controls the capacity for modeling non-periodic temporal progression. We find $d_{\text{linear}} = 3$ provides sufficient expressiveness while maintaining parameter efficiency. Reducing to $d_{\text{linear}} = 1$ causes 0.6% accuracy loss, indicating that a single linear component cannot adequately capture the varied pacing of different actions. Increasing to $d_{\text{linear}} = 5$ yields only marginal changes (−0.1%), suggesting diminishing returns beyond our selected configuration.

Moreover, Table 7 compares different strategies for aggregating temporal features into the global context $\mathbf{H}_{\text{context}}$. Average pooling achieves the best performance, effectively capturing the overall temporal characteristics without emphasizing extreme values that may represent outliers or noise. Max pooling performs worst (−0.7%), as it focuses solely on peak activations and discards information about the temporal distribution of features. Attention-based pooling (−0.3%) and learnable weighted pooling (−0.2%) underperform despite their added complexity, suggesting that the uniform weighting of average pooling provides appropriate inductive bias for global temporal context extraction. This finding indicates that all temporal positions contribute meaningfully to characterizing action-specific motion patterns, and selective emphasis through learned weights does not improve representation quality.

Ablation studies not only validate our architectural decisions but also provide practical guidance for future work in skeleton-based action recognition, demonstrating that effective models must jointly consider temporal dynamics, spatial relationships, and their frequency-domain characteristics.

5. Conclusion

In this work, we propose FATL-GCN for skeleton-based action recognition with three components. The FATE module uses Time2Vec encoding to capture temporal frequency patterns from skeleton sequences, generating context that reflects both periodic and non-periodic motion characteristics. The CDGC module leverages this frequency context to dynamically construct spatial graph topologies via multi-head attention, allowing joint connections to adapt based on temporal motion patterns rather than fixed anatomical structure. As shown in the experiments, our FATL-GCN achieves the current state-of-the-art results on NTU RGB+D 60, NTU RGB+D 120, and Kinetics-Skeleton, and ablation studies also demonstrate the effectiveness of the proposed method. However, our method is currently limited by its need for sufficient temporal frames for frequency analysis, which reduces effectiveness on very short action sequences. The multi-head topology learning also increases memory consumption compared to simpler GCN variants, which may constrain deployment on resource-limited edge devices. Additionally, performance may degrade on skeleton data with severe occlusions or significant noise from pose estimation errors. Future research directions include developing efficient variants through pruning or knowledge distillation for mobile deployment, and investigating fusion with RGB or depth modalities to improve robustness in challenging scenarios.

CRediT authorship contribution statement

Yi Xia: Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization; **Sira Yongchareon:** Writing – review & editing, Supervision, Conceptualization; **Raymond Lutui:** Writing – review & editing, Conceptualization; **Quan Z. Sheng:** Writing – review & editing.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: a review, IEEE Trans. Pattern Anal. Mach. Intell. 45 (3) (2022) 3200–3225.
[2] Y. Zhao, Q. Gao, Z. Ju, J. Zhou, Y. Guo, Sharing-Net: lightweight feedforward network for skeleton-based action recognition based on information sharing mechanism, Pattern Recognit. 146 (2024) 110050.
[3] H. Qiu, B. Hou, Multi-grained clip focus for skeleton-based action recognition, Pattern Recognit. 148 (2024) 110188.
[4] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 183–192.
[5] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 32, 2018, pp. 7444–7452.
[6] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.

[7] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12026–12035.
[8] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 143–152.
[9] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13359–13368.
[10] H.-g. Chi, M.H. Ha, S. Chi, S.W. Lee, Q. Huang, K. Ramani, InfoGCN: representation learning for human skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20186–20196.
[11] J. Lee, M. Lee, D. Lee, S. Lee, Hierarchically decomposed graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10444–10453.
[12] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, X.-S. Hua, BlockGCN: redefine topology awareness for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 2049–2058.
[13] W. Myung, N. Su, J.-H. Xue, G. Wang, DeGCN: deformable graph convolutional networks for skeleton-based action recognition, IEEE Trans. Image Process. 33 (2024) 2477–2490.
[14] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2969–2978.
[15] C. Plizzari, M. Cannici, M. Matteucci, Skeleton-based action recognition via spatial and temporal transformer networks, Comput. Vision Image Understanding 208 (2021) 103219.
[16] L. Shi, Y. Zhang, J. Cheng, H. Lu, Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition, in: Asian Conference on Computer Vision (ACCV), 2020, pp. 627–643.
[17] J. Do, M. Kim, SkateFormer: skeletal-temporal transformer for human action recognition, in: European Conference on Computer Vision, Springer, 2024, pp. 401–420.
[18] X. Wang, K. Chen, Z. Zhao, G. Shi, X. Xie, X. Jiang, Y. Yang, Multi-scale adaptive skeleton transformer for action recognition, Comput. Vision Image Understanding 250 (2025) 104229.
[19] J. Xie, Y. Meng, Y. Zhao, A. Nguyen, X. Yang, Y. Zheng, Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 38, 2024, pp. 6225–6233.
[20] F. Xu, P. Shi, X. Zhang, Skeleton-based human action recognition with spatial and temporal attention-enhanced graph convolution networks, J. Adv. Computat. Intell. Intell. Inf. 28 (6) (2024) 1367–1379.
[21] B. Xu, X. Shu, J. Zhang, G. Dai, Y. Song, Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition, IEEE Trans. Neural Netw. Learn. Syst. 35 (8) (2023) 11035–11048.
[22] H. Le, C.-K. Lu, C.-C. Hsu, S.-K. Huang, Skeleton-based human action recognition using LSTM and depthwise separable convolutional neural network, Appl. Intell. 55 (5) (2025) 298.
[23] S.M. Kazemi, R. Goel, S. Eghbali, J. Ramanan, J. Sahota, S. Thakur, S. Wu, C. Smyth, P. Falt, P. Poupart, Time2Vec: learning a vector representation of time, arXiv preprint arXiv:1907.05321 (2019).
[24] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive recurrent neural networks for high performance human action recognition from skeleton data, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2117–2126.
[25] P. Wang, W. Li, C. Li, Y. Hou, Action recognition based on joint trajectory maps with convolutional neural networks, Knowl. Based Syst. 158 (2018) 43–53.
[26] M. Dai, Z. Sun, T. Wang, J. Feng, K. Jia, Global spatio-temporal synergistic topology learning for skeleton-based action recognition, Pattern Recognit. 140 (2023) 109540.
[27] X. Yin, J. Zhong, D. Lian, W. Cao, Spatiotemporal progressive inward-outward aggregation network for skeleton-based action recognition, Pattern Recognit. 150 (2024) 110262.
[28] Z. Wu, Y. Ding, L. Wan, T. Li, F. Nian, Local and global self-attention enhanced graph convolutional network for skeleton-based action recognition, Pattern Recognit. 159 (2025) 111106.
[29] H. Yang, S. Wang, L. Jiang, Y. Su, Y. Zhang, Hierarchical adaptive multi-scale hypergraph attention convolution network for skeleton-based action recognition, Appl. Soft Comput. 172 (2025) 112855.
[30] Z. Chen, S. Li, B. Yang, Q. Li, H. Liu, Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 35, 2021, pp. 1113–1122.
[31] L. Hedegaard, N. Heidari, A. Iosifidis, Continual spatio-temporal graph convolutional networks, Pattern Recognit. 140 (2023) 109528.
[32] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv preprint arXiv:1705.06950 (2017).
[33] J. Nilsson, A. Thorstensson, Adaptability in frequency and amplitude of leg movements during human locomotion at different speeds, Acta Physiol. Scand. 129 (1) (1987) 107–114.
[34] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, NTU RGB+ D: a large scale dataset for 3D human activity analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1010–1019.
[35] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, NTU RGB+ D 120: a large-scale benchmark for 3D human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2019) 2684–2701.
[36] Y. Zhu, H. Shuai, G. Liu, Q. Liu, Multilevel spatial-temporal excited graph network for skeleton-based action recognition, IEEE Trans. Image Process. 32 (2023) 496–508. https://doi.org/10.1109/TIP.2022.3230249
[37] H. Tian, X. Ma, X. Li, Y. Li, Skeleton-based action recognition with select-assemble-normalize graph convolutional networks, IEEE Trans. Multimed. 25 (2023) 8527–8538. https://doi.org/10.1109/TMM.2023.3318325
[38] H. Zhou, Q. Liu, Y. Wang, Learning discriminative representations for skeleton based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10608–10617.
[39] H. Duan, J. Wang, K. Chen, D. Lin, PYSKL: towards good practices for skeleton action recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7351–7354.