NeuroPath: Brain-Inspired Dual-Pathway Graph Convolutional Networks for Skeleton-Based Action Recognition

Kanglei Zhou[a], Ruizhi Cai[b], Hubert P. H. Shum[c], Frederick W. B. Li[c], Xiaohui Liang[b,d,*]

[a]*Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing 100084, China*
[b]*State Key Laboratory of VR Technology and Systems, Beihang University, Beijing 100191, China*
[c]*Department of Computer Science, Durham University, DH1 3LE, United Kingdom*
[d]*Zhongguancun Laboratory, Beijing 100190, China*

Abstract

Skeleton-based action recognition aims to recognize human actions from sequences of human joint coordinates. Most existing Spatial-Temporal Graph Convolutional Networks (STGCNs) have achieved promising results by modeling skeletal structures with implicit spatial–temporal representations. However, our empirical study reveals a clear performance imbalance across different skeletal modalities, indicating that implicitly coupling spatial and temporal information limits the full exploitation of complementary structural and motion cues. Inspired by the ventral and dorsal pathways in human perception, we propose Dual-Pathway Graph Convolutional Networks (NeuroPath), which adopt a dual-pathway architecture for separate yet collaborative modeling of spatial and temporal information. Specifically, transformation units first convert the input into pathway-specific skeletal representations, allowing each pathway to focus on complementary aspects of human motion. To further capture coordinated joint behaviors and their interrelationships, we introduce a group graph convolution block that dynamically identifies key body parts and models their spatial-temporal dependencies. In addition, inter-pathway dynamic fusion modules integrate complementary inter-modal information across pathways, facilitating higher-level semantic interpretation of actions. Extensive experiments on Kinetics Skeleton 400, NTU RGB+D 60, and NTU RGB+D 120 demonstrate consistent performance improvements, validating the effectiveness of dual-pathway spatial-temporal modeling for skeleton-based action recognition.

*Corresponding author
Email address: liang_xiaohui@buaa.edu.cn (Xiaohui Liang)

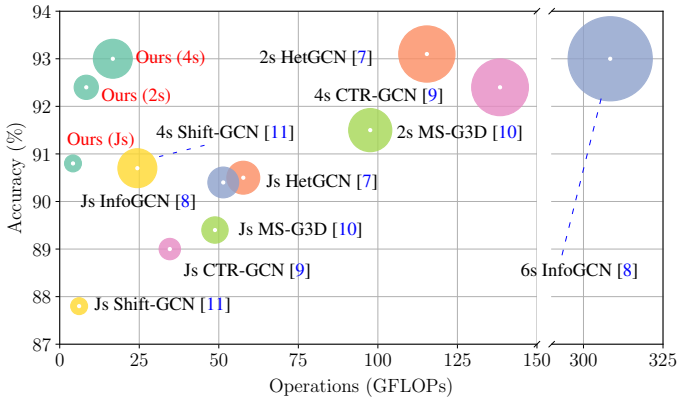


Figure 1: **The bubble plot of model computational performance in the X-Sub setting on the NTU RGB+D 60 dataset.** The horizontal axis represents computational operations (GFLOPs), the vertical axis represents accuracy (%), and the size of the bubbles represents the number of parameters (M). Different bubble colors represent different kinds of models.

Keywords: Spatial-Temporal Graph Convolutional Networks, Skeleton-Based Action Recognition

1. Introduction

Human action recognition aims to recognize human behaviors [1, 2], and is fundamental to applications such as security surveillance, autonomous driving, and human–computer interaction [3, 4]. Skeletal data provide a compact and structured representation of human motion, and are more robust to background clutter and illumination variations than raw video data [5]. These advantages have led to the widespread adoption of skeleton-based action recognition [6]. However, accurately modeling skeletal data remains challenging, as human actions involve complex spatial dependencies among joints and long-term temporal dynamics that must be jointly captured.

Early deep learning methods [12] using RNNs or CNNs have shown clear performance advantages over handcrafted approaches. However, these methods typically represent skeleton sequences as vectors or pseudo-graphs, without explicitly modeling the structural constraints among body joints. To better capture such structural information, Graph Convolutional Networks (GCNs) have attracted increasing attention due to their ability to naturally encode joint dependencies [13]. The vanilla GCN [14], how-

ever, focuses primarily on spatial relationships and neglects the temporal dynamics that are critical for modeling sequential human actions. To address this limitation, Spatial-Temporal GCNs (STGCNs) [15, 16] have been proposed to jointly model spatial and temporal information, and have become a dominant paradigm for skeleton-based action recognition. Despite these advances, effectively capturing the interplay between spatial structure and temporal dynamics in skeletal data remains a non-trivial challenge.

A closer examination of existing STGCNs reveals that this challenge is typically addressed by learning different skeletal modalities using separate streams, followed by simple score-level ensemble fusion [17, 10]. Here, a **skeletal modality** denotes a specific representation derived from skeletal data, including joint positions, bone vectors, joint motion, and bone motion, which predominantly emphasize either spatial structure or temporal dynamics. While such multi-stream ensemble strategies can improve overall performance, they implicitly assume that different modalities contribute independently. However, our empirical analysis exposes a **clear performance imbalance**, where joint- and bone-based streams consistently outperform motion-based streams, indicating that motion-only representations lack sufficient structural support. At the same time, the results also reveal strong **modality complementarity**: although motion-based streams perform poorly in isolation, integrating structural and motion modalities leads to substantial performance gains, whereas fusing motion-only modalities yields the lowest accuracy. These findings motivate us to explicitly model the imbalanced yet complementary spatial and temporal information within a unified representation, rather than relying on loosely coupled ensemble strategies.

Motivated by the above observations and analysis, we revisit spatial–temporal modeling in skeleton-based action recognition from a modality-aware perspective. We propose NeuroPath, a brain-inspired dual-pathway network that explicitly disentangles and collaboratively models spatial structure and temporal dynamics. By separating yet coordinating spatial-aware and temporal-aware representations within a unified framework, NeuroPath aims to overcome the limitations of loosely coupled multi-branch designs and enable more effective exploitation of complementary skeletal modalities.

To mitigate inadequate inter-modal fusion, we design a modality transformation unit for each pathway, which decomposes the input skeletal data into complemen-

tary representations. Specifically, one pathway is guided to focus on spatial-aware information, while the other emphasizes temporal-aware information, resembling the functional segregation of visual stimuli in the human brain. Based on these pathway-specific representations, we further introduce a dynamic fusion module to explicitly model inter-modal dependencies at multiple stages of the network, enabling effective information exchange across pathways [18].

To improve spatial–temporal cohesion within each pathway, we propose a spatial–temporal group graph convolution block. First, the group aggregation module incorporates temporal cues into the spatial domain, facilitating the identification of key body parts. Then, the adaptive structural aggregation module effectively infers the inter-part dependencies while avoiding redundancy. Finally, a spatial-temporal attention module prevents discriminative information loss.

Extensive experiments on Kinetics Skeleton 400, NTU RGB+D 60, and NTU RGB+D 120 demonstrate consistent performance improvements, while achieving notable computational efficiency (see Fig. 1). The code is publicly available at https://github.com/ZhouKanglei/NeuroPath. The main contributions of this work are:

- We empirically uncover and analyze an overlooked yet critical phenomenon in skeleton-based action recognition, showing that different skeletal modalities exhibit intrinsic performance imbalance while remaining strongly complementary.
- Motivated by these observations, we introduce a brain-inspired dual-pathway framework that explicitly disentangles spatial-aware and temporal-aware representations while enabling their coordinated interaction within a single network.
- We provide an efficient and practical implementation of this framework, achieving consistent performance improvements with favorable computational efficiency on multiple skeleton-based action recognition benchmarks.

2. Related Work

STGCNs [15, 19, 20] are widely adopted for skeleton-based human action analysis tasks, including action recognition and time-series forecasting. Typically, STGCNs

model spatial features via graph convolution, while temporal dependencies are modeled in a separate stage. For spatial modeling, some methods [17, 21] treat joints as graph nodes, while others [17, 22] additionally construct bone-based graphs. Since human actions often involve the coordinated motion of different limbs, part-level graph representations are effective for extracting semantic features [23]. For temporal modeling, temporal convolution and RNNs are commonly adopted. The key challenge is integrating spatial and temporal information effectively. Early approaches [15, 16, 17] focused mainly on spatial modeling and incorporated temporal information separately. Some recent methods [10, 7] construct 3D spatial-temporal graphs by combining consecutive spatial graphs, but this introduces substantial computational overhead. To address these issues, this work proposes a novel spatial-temporal module that injects temporal cues into graph convolution, enabling efficient spatial-temporal cohesion.

Skeleton-Based Action Recognition aims to extract spatial–temporal patterns from skeletal data for action classification [12]. Early approaches mainly relied on handcrafted features, while recent deep learning methods have achieved superior performance, including both non-graph-based approaches [6, 2, 24] and graph-based methods [25, 23]. Early non-graph-based methods [24] commonly represent skeleton sequences as ordered vectors or pseudo-images, and use RNNs, CNNs, or their variants to capture temporal and spatial patterns. Meanwhile, hybrid designs have also been explored, where graph-based modules are used to model spatial joint relationships, while RNNs, temporal convolutions, or attention mechanisms are adopted to capture temporal dynamics. STGCN-based methods [26, 27] further integrate graph convolution with temporal modeling modules, enabling explicit topology-aware spatial modeling together with temporal dependency learning. Recent STGCN-based approaches further enhance spatial–temporal modeling through improved attention mechanisms [28] or topology-aware designs [29, 30]. Existing STGCN-based methods mainly improve performance by either deepening network architectures [15, 16] or widening them via multi-stream or disentangled designs [21, 10]. However, deeper models often suffer from over-smoothing and increased training complexity, while wider architectures tend to emphasize coarse postural information rather than fine-grained motion patterns. In this context, we propose an efficient dual-pathway architecture with dynamic fusion,

which enables complementary modeling of spatial-temporal information.

Dual-Pathway Networks are inspired by the structure and function of the human visual system, which comprises the dorsal stream, responsible for spatial processing, and the ventral stream, involved in object recognition and perception. In dual-pathway networks, input data is processed independently by two parallel pathways before the extracted features are merged for downstream tasks. This architecture has been successfully applied in various domains, including computer vision [31, 32**?** , 33] and medical care [34, 35, 36]. In the realm of skeleton-based action recognition, dual-pathway networks [17] or multi-pathway networks [9, 10] offer a promising approach for effectively integrating spatial and temporal cues, thereby enhancing performance and robustness. Recent evidence suggests that the dorsal and ventral pathways interact significantly, playing crucial roles in various aspects of cognition [18]. However, existing networks do not fully model these pathway interactions, limiting overall ensemble performance. To address this gap, we propose a novel dual-pathway network with dynamic fusion to fully leverage the complementary nature of skeletal data.

3. Problem Formulation and Empirical Analysis

Notations and Task Definition. Skeleton-based action recognition aims to predict an action label from a sequence of human skeletons. We denote a skeleton sequence as $\mathbf{X} \in \mathbb{R}^{T\times N\times C}$, where T is the number of frames, N is the number of joints, and C is the feature dimension. The skeletal structure is modeled as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes represent joints and each edge (bone) connects two joints. The input features $\mathbf{X}$ are defined on the graph $\mathcal{G}$ and evolve over time.

Definition of the Skeletal Modality. We define the *skeletal modality* as a family of representations derived from skeletal data, which is distinct from other input modalities such as RGB or depth and is referred to as the *modality* for brevity. As illustrated in Fig. 2(a), skeletal modalities can be broadly categorized into *spatial-aware modalities*, which encode human limb configurations using joint positions or bone vectors, and *temporal-aware modalities*, which characterize motion dynamics through temporal variations of joint or bone representations. In practice, we consider four commonly

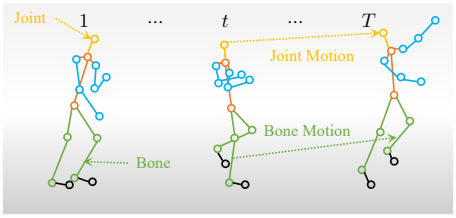
used skeletal modalities: *joint position*, *bone position*, *joint motion*, and *bone motion*. Given the joint position sequence $\mathbf{X}_{\text{jp}}$, the remaining modalities are deterministically derived via simple spatial and temporal differencing operations: joint motion is defined as the temporal difference of joint positions, bone position is obtained from the two joints connected by each bone, and bone motion is computed as the temporal difference of bone positions, which can be represented as:

$$\begin{aligned} \mathbf{X}_{\text{jm}}(t) &= \mathbf{X}_{\text{jp}}(t) - \mathbf{X}_{\text{jp}}(t-1), \quad t = 2, \ldots, T, \\ \mathbf{X}_{\text{bp}}(t, e) &= \mathbf{X}_{\text{jp}}(t, v_1(e)) - \mathbf{X}_{\text{jp}}(t, v_2(e)), \\ \mathbf{X}_{\text{bm}}(t) &= \mathbf{X}_{\text{bp}}(t) - \mathbf{X}_{\text{bp}}(t-1), \quad t = 2, \ldots, T, \end{aligned} \tag{1}$$

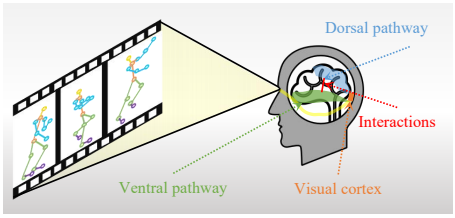
where $v_1(e)$ and $v_2(e)$ denote the two joints incident to bone e.

Architectural Challenges. From an architectural perspective, existing STGCNs can be broadly categorized into (2+1)D and 3D paradigms, depending on how spatial and temporal modeling are coupled within a single branch [15, 17, 10]. (2+1)D STGCNs [15, 17] sequentially apply spatial graph convolution and temporal convolution, which is computationally efficient but enforces a rigid separation between spatial structure and temporal dynamics, limiting their interaction within a unified feature space. In contrast, 3D STGCNs [37, 10] jointly model spatial and temporal dependencies by constructing spatiotemporal graphs over multiple frames, but the rapidly growing spatiotemporal neighborhood incurs high computational cost, thereby constraining the temporal receptive field and hindering long-range temporal modeling. As a result, despite their distinct coupling strategies, both paradigms face inherent difficulties in achieving efficient spatial-temporal modeling for complex human actions.

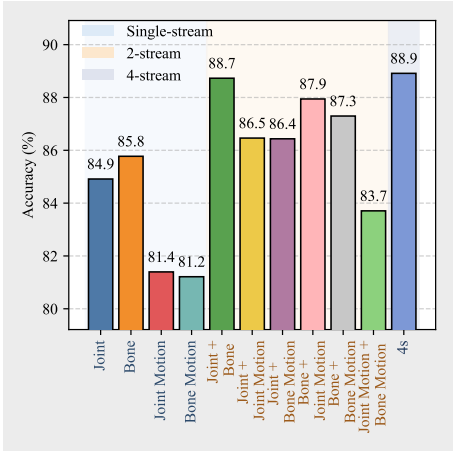
Empirical Study and Key Observations. To examine how existing STGCNs exploit different skeletal modalities, we conduct an empirical analysis using CTR-GCN [9] on the NTU RGB+D 120 X-Sub benchmark. Following common practice [32], different skeletal modalities are processed by independent streams, and ensemble performance is obtained by averaging prediction scores across multiple streams. As shown in Fig. 2(c), single-stream results exhibit a **clear performance imbalance**: joint and bone modalities achieve higher accuracy (84.91% and 85.77%) than motion-



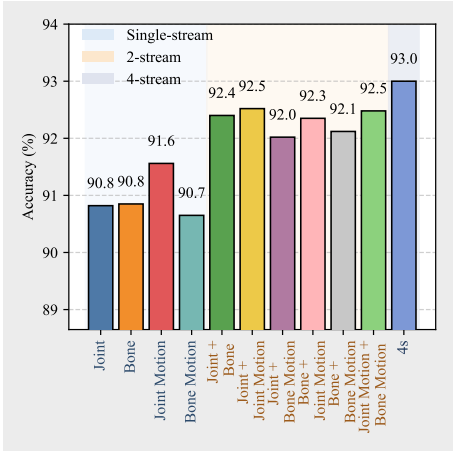
(a) Skeletal modality representations



(b) Dual-pathway inspiration



(c) Modality imbalance and complementarity



(d) Improved modality balance

Figure 2: **Motivation and empirical observations.** (a) Common skeletal modality representations derived from joint sequences, including joint positions, bone vectors, joint motion, and bone motion. (b) Inspiration from the two-stream hypothesis of the human visual system, where spatial structure and motion are processed by two parallel yet interacting pathways. (c) Empirical results of CTR-GCN [9] on the X-Sub protocol of NTU RGB+D 120 under different skeletal modalities. We report Top-1 Accuracy (denoted as Accuracy). Single-stream results exhibit clear performance imbalance across modalities, while combining structural and motion information consistently improves performance, highlighting their complementarity. (d) Results of our method, which significantly alleviates modality imbalance and achieves more balanced performance across different skeletal modalities.

based modalities (81.40% and 81.21%). This indicates that motion-only representations, which primarily encode temporal differences, lack sufficient spatial structure to reliably distinguish fine-grained actions when learned in isolation. Two-stream fusion further reveals **modality complementarity**. While combining structural modalities (joint + bone) yields a substantial improvement (88.73%), fusing motion-only modalities (joint motion + bone motion) results in the lowest performance (83.71%), suggesting that temporal cues alone are insufficient even when aggregated. In contrast, integrating structural and motion modalities consistently improves accuracy, implying

that motion information is most effective when grounded in explicit spatial structure. These observations highlight that the imbalance across modalities is intrinsic rather than incidental, and that simply averaging predictions cannot fully resolve it. Instead, they suggest that explicitly modeling the complementary spatial and temporal information within a unified representation is key to improving single-stream efficiency and reducing reliance on heavy multi-stream ensembles.

Motivation of the Dual-Pathway Network. Motivated by the above empirical findings, we seek a unified modeling framework that can explicitly exploit the complementary nature of spatial structure and temporal dynamics within a single network. Our design is inspired by the two-stream hypothesis [38], which posits that the human visual cortex processes spatial information through the ventral pathway and motion information through the dorsal pathway. As illustrated in Fig. 2(b), these two pathways operate in parallel and interact to support coherent action perception. Analogously, we introduce a dual-pathway architecture that separates spatial-aware and temporal-aware processing while enabling their collaboration for holistic action understanding.

4. Dual-Pathway Graph Convolutional Networks (NeuroPath)

In this section, we provide an overview of the proposed NeuroPath framework and subsequently present its core architectural components.

4.1. Framework Overview

Fig. 3 illustrates the framework of NeuroPath, which is designed to explicitly exploit the complementary nature of spatial-aware and temporal-aware information. Given the input skeletal sequence $\mathbf{X}$ defined on the graph $\mathcal{G}$, NeuroPath first applies batch normalization (BN) and embedding (EB) to obtain a normalized feature representation $\tilde{\mathbf{X}} \in \mathbb{R}^{T\times N\times D}$, where D denotes the embedding dimension. Here, the embedding is implemented as a linear projection that maps the raw skeletal inputs into a unified feature space. The normalized features $\tilde{\mathbf{X}}$ are then processed by two parallel pathways that focus on spatial-aware and temporal-aware representations, respectively. Inspired by the functional segregation observed in the human visual system, the transformation

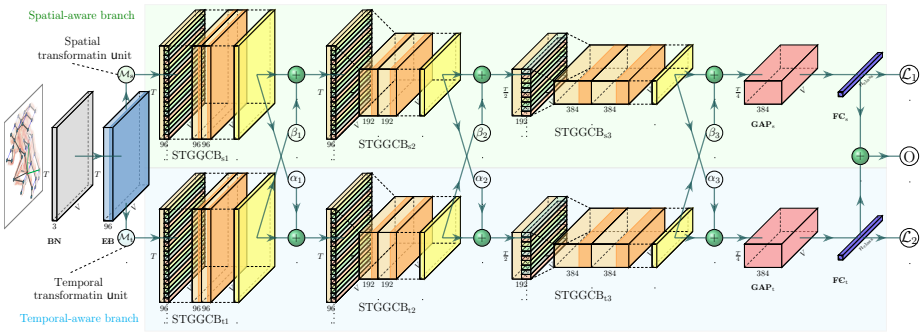


Figure 3: **Overview of the proposed NeuroPath framework.** NeuroPath adopts a brain-inspired dual-pathway architecture to explicitly model the complementary spatial-aware and temporal-aware information in skeletal data. The two pathways are conceptually motivated by the ventral and dorsal streams in the human visual system, which specialize in spatial structure and motion processing, respectively. In contrast to existing multi-stream networks that fuse modalities only at the output level, NeuroPath introduces dynamic inter-pathway fusion modules to enable effective interaction and coordination between spatial-aware and temporal-aware representations throughout the network.

units (see Sec. 4.2) decompose the input features into distinct skeletal modalities tailored to each pathway. Subsequently, the Spatial–Temporal Group Graph Convolution Block (STGGCB, see Sec. 4.3) captures coordinated body-part movements by integrating spatial structure with temporal cues and modeling their interdependencies. To facilitate effective information exchange between the two pathways, a dynamic fusion module (see Sec. 4.4) is introduced to explicitly model inter-pathway interactions and enhance spatial–temporal feature integration. Finally, the output features from both pathways are aggregated via a Global Average Pooling (GAP) layer and fed into a Fully Connected (FC) layer for action classification. The entire network is supervised using pathway-specific cross-entropy losses, denoted as $\mathcal{L}_1$ and $\mathcal{L}_2$, which are applied to the spatial-aware and temporal-aware pathways, respectively, enabling stable and collaborative learning of spatial–temporal representations.

4.2. Spatial/Temporal Transformation Unit

Design Rationale. As discussed in Sec. 3, existing STGCN-based approaches commonly exploit multiple skeletal modalities through independent streams and perform fusion at the score level [16, 10, 39, 9]. While this strategy empirically demonstrates modality complementarity, it also exposes two inherent limitations. First, different modalities exhibit pronounced performance imbalance, causing motion-dominated

streams to contribute marginally in isolation. Second, treating modalities independently prevents explicit coordination between spatial structure and temporal dynamics during representation learning, relegating their interaction to late-stage fusion. Our objective is fundamentally different. Rather than introducing additional streams or relying on ensemble strategies, we aim to explicitly coordinate complementary spatial and temporal information *within a single unified network*. To this end, we introduce a *spatial/temporal transformation unit*, which decomposes the input skeletal representation into spatial-aware and temporal-aware components and routes them to two dedicated pathways. This design directly addresses the observed imbalance by ensuring that each pathway receives modality-consistent information, while enabling subsequent layers to model their interaction explicitly.

Modality Transformation. Following the definition in Sec. 3, we consider four skeletal modalities, denoted by $\mathcal{X} = \{\mathbf{X}_{\text{jp}}, \mathbf{X}_{\text{bp}}, \mathbf{X}_{\text{jm}}, \mathbf{X}_{\text{bm}}\}$, corresponding to joint position, bone position, joint motion, and bone motion, respectively. All four modalities are computed beforehand from the joint-position sequence and are simultaneously available during training and inference. Joint and bone positions encode spatial structure, while joint and bone motion characterize temporal dynamics. Let $\tilde{\mathbf{X}}$ denote the normalized skeletal features and $\phi \in \mathcal{X}$ denote the *modality identifier* of the current stream setting. Importantly, ϕ does *not* indicate that one modality is derived from another. Instead, it specifies how the pre-computed modalities are *paired and routed* to the spatial-aware and temporal-aware pathways. We define two deterministic, parameter-free transformation functions, $\mathcal{M}_{\text{s}}(\cdot)$ and $\mathcal{M}_{\text{t}}(\cdot)$, which select the spatial-aware and temporal-aware representations, respectively. These functions do not perform any numerical transformation; they only determine which modality is assigned to each pathway. Accordingly, the spatial-aware representation is defined as

$$\mathbf{X}_{\text{s}} = \mathcal{M}_{\text{s}}(\phi) = \mathbb{I}_{\text{jp}}(\phi)\,\mathbf{X}_{\text{jp}} + \mathbb{I}_{\text{bp}}(\phi)\,\mathbf{X}_{\text{bp}}, \tag{2}$$

and the temporal-aware representation is defined as

$$\mathbf{X}_{\text{t}} = \mathcal{M}_{\text{t}}(\phi) = \mathbb{I}_{\text{jm}}(\phi)\,\mathbf{X}_{\text{jm}} + \mathbb{I}_{\text{bm}}(\phi)\,\mathbf{X}_{\text{bm}}, \tag{3}$$

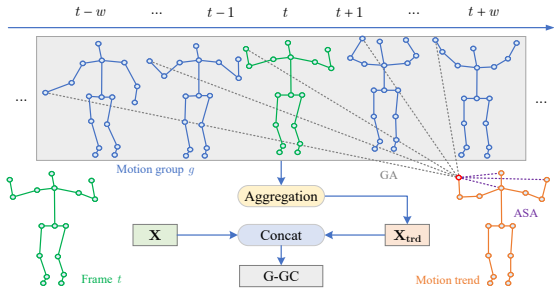


Figure 4: **Core design intuition of STGGCB**. Discriminative actions are characterized by coordinated groups of joints with correlated temporal dynamics. STGGCB first summarizes local temporal variation at the group level and then conditions spatial aggregation on this motion-aware context, enabling graph convolution to emphasize structurally and temporally relevant joint interactions.

where $\mathbb{I}_{(\cdot)}(\phi)$ is an indicator function that equals 1 if modality ϕ corresponds to the specified representation and 0 otherwise.

Discussion and Benefits. This routing strategy does not introduce additional parameters and does not increase model expressivity by itself. Instead, it provides a structured decomposition that exposes complementary spatial and temporal cues to different pathways, facilitating more effective coordination in subsequent modules. By explicitly separating spatial structure and temporal dynamics at the input stage, the proposed transformation unit alleviates the imbalance observed in conventional multi-stream settings and lays the foundation for coherent spatial-temporal modeling.

4.3. Spatial-Temporal Group Graph Convolution Block

Design Rationale. As illustrated in Fig. 4, a key challenge in skeleton-based action recognition is modeling coordinated body-part interactions that evolve over time. Discriminative actions are often characterized not by isolated joint motions, but by structured groups of joints exhibiting correlated temporal dynamics (e.g., arm–torso coordination in waving or punching). However, most existing (2+1)D STGCNs [15, 16, 17] either decouple spatial and temporal modeling or rely on sequential processing, which limits the ability of spatial aggregation to condition on temporal variation. To address this limitation, we design the Spatial-Temporal Group Graph Convolution Block (STGGCB, see Fig. 5) with two core objectives: (1) to explicitly inject temporal variation cues into spatial aggregation, so that graph convolution can prioritize

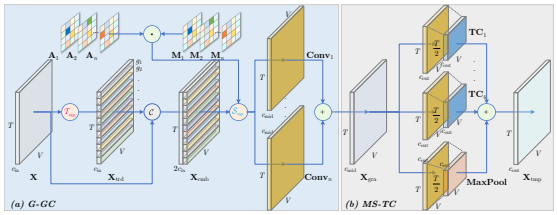


Figure 5: **The network architecture of STGGCB**: (a) Group Graph Convolution (G-GC) and (b) Multi-Scale Temporal Convolution (MS-TC). G-GC comprises group aggregation (see Eq. (4)) and adaptive structural aggregation (see Eq. (5)).

motion-relevant joint interactions; and (2) to capture multi-hop structural dependencies while avoiding redundant message mixing that commonly arises in dense spatial-temporal graphs. STGGCB integrates temporal-aware group aggregation with adaptive structural aggregation, followed by efficient temporal context modeling and spatial–temporal attention, thereby enabling coherent and scalable spatial–temporal feature learning within each pathway.

Group Aggregation. As shown in Fig. 5(a), group aggregation is designed to explicitly inject temporal evolution cues into spatial graph convolution, enabling motion-aware neighborhood aggregation. Instead of relying on a separate temporal convolution to capture motion patterns, we first compute a local temporal evolution feature that characterizes short-term motion trends of joints, and concatenate it with the original feature to form a temporally enhanced representation:

$$\mathbf{X}_{\text{cmb}}^{(l)} = \text{Concat}\Big(\mathbf{X}^{(l)}, \mathbf{X}_{\text{trd}}^{(l)}\Big), \quad \mathbf{X}_{\text{trd}}^{(l)} = \mathcal{T}_{\text{agg}}\Big(\mathbf{X}^{(l)}\Big), \tag{4}$$

where $\mathcal{T}_{\text{agg}}$ is implemented as a second-order temporal difference operator, which emphasizes motion transitions while suppressing constant-velocity patterns. By embedding temporal variation directly into the features used for graph convolution, group aggregation allows spatial message passing to prioritize motion-relevant neighbors, thereby achieving tighter spatial–temporal coupling than conventional (2+1)D designs that model spatial and temporal information sequentially.

Adaptive Structural Aggregation. To capture long-range spatial dependencies in a controlled manner, we introduce an adaptive structural aggregation that reweights

multi-hop message passing without modifying graph connectivity, as shown in Fig. 5(a). Given the combined features $\mathbf{X}_{\text{cmb}}^{(l)}$, aggregation over K-hop neighborhoods is formulated as

$$\mathbf{X}_{\text{gra}}^{(l)} = \sigma\left(\sum_{k=0}^{K} \bar{\mathbf{A}}_k \, \mathbf{X}_{\text{cmb}}^{(l)} \, \mathbf{W}_k^{(l)}\right), \tag{5}$$

$$\bar{\mathbf{A}}_k = \tilde{\mathbf{D}}^{-\frac{1}{2}}\left(\tilde{\mathbf{A}}_k \odot \mathbf{M}_k\right)\tilde{\mathbf{D}}^{-\frac{1}{2}}, \tag{6}$$

where $\tilde{\mathbf{A}}_k$ denotes the fixed *exact k-hop adjacency support*, i.e., $\tilde{\mathbf{A}}_k(i, j) = 1$ if and only if nodes i and j are connected by a path of length exactly k in the original skeleton graph, and 0 otherwise. This definition ensures that only relations at hop k are aggregated at the k-th term, avoiding redundant mixing across different neighborhood ranges. The matrix $\mathbf{M}_k$ is a learnable strength mask that adaptively reweights existing k-hop connections. The element-wise modulation $\odot$ adjusts the influence of existing multi-hop connections while preserving the original topology, thereby avoiding redundant shortcut edges that commonly arise from additive adjacency updates on dense k-hop graphs. The mask $\mathbf{M}_k$ is generated by a cross-attention gate,

$$\mathbf{M}_k = \mathbf{B}_k + \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right), \tag{7}$$

where $\mathbf{Q}$ and $\mathbf{K}$ are feature embeddings from $\mathbf{X}_{\text{cmb}}^{(l)}$ and $\mathbf{X}^{(l)}$, respectively, and $\mathbf{B}_k$ provides a hop-specific prior that stabilizes structural weighting when the normalized attention becomes diffuse. Unlike conventional higher-order topological message passing (see Fig. 6(c)) that explicitly alters graph connectivity, this formulation preserves the original multi-hop supports and enables adaptive long-range aggregation with controlled complexity and without increasing graph density (see Fig. 6(d)).

Multi-Scale Temporal Convolution. As shown in Fig. 5(b), we adopt a standard multi-scale temporal convolution to complement spatial–temporal aggregation. Specifically, the input channels are evenly divided into g groups, and each group is processed independently. The resulting features are concatenated as

$$\mathbf{X}_{\text{tmp}}^{(l)} = \text{Concat}\left(\text{Conv}_d\left(\mathbf{X}_{\text{gra}}^{(l)}\right), \text{MP}\left(\mathbf{X}_{\text{gra}}^{(l)}\right)\right), \tag{8}$$

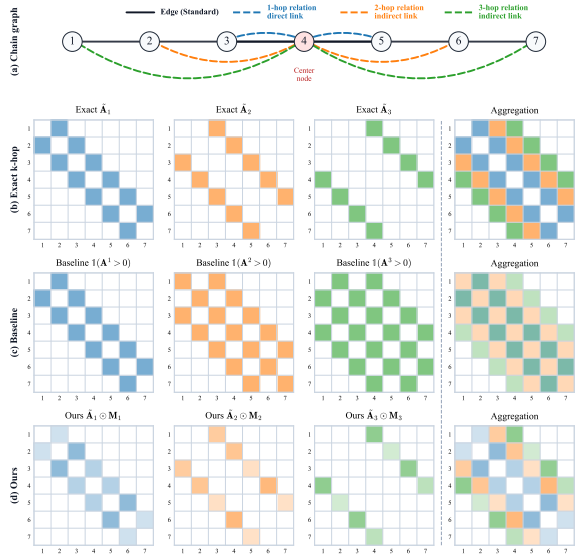


Figure 6: **Illustration of multi-hop aggregation strategies on a simple chain graph.** (a) A 7-node chain graph, where node 4 is selected as the center node and 1-hop, 2-hop, and 3-hop relations are shown by different colors. (b) Exact k-hop adjacency supports $\tilde{A}_k$, where each matrix contains only node pairs whose shortest-path distance is exactly k. (c) Conventional baselines using cumulative multi-hop aggregation $(A^1 + A^2 + A^3)$, which entangle different hop distances within a single representation. (d) Our method, which applies learnable masks M_k over exact k-hop supports, enabling adaptive reweighting while preserving explicit separation of neighborhood ranges.

where $\mathrm{Conv}_d(\cdot)$ denotes a dilated temporal convolution and $\mathrm{MP}(\cdot)$ denotes temporal max pooling. This module enlarges the temporal receptive field with low computational overhead and serves as a conventional temporal modeling component.

Spatial-Temporal Attention. To alleviate potential over-smoothing caused by repeated spatial–temporal aggregation, we incorporate a spatial–temporal attention module to adaptively reweight features. Given the input feature $\mathbf{X}_{\text{in}}$, we summarize global context along the spatial and temporal dimensions using both average pooling and max pooling. Specifically, we obtain pooled features $\boldsymbol{x}_{\text{ms}}, \boldsymbol{x}_{\text{as}} \in \mathbb{R}^{C_{\text{in}} \times T}$ and $\boldsymbol{x}_{\text{mt}}, \boldsymbol{x}_{\text{at}} \in \mathbb{R}^{C_{\text{in}} \times N}$, which capture complementary statistics of the feature distribution. These pooled features are projected by shared 1D convolutions to produce

$$\boldsymbol{h}_{\text{ms}}, \boldsymbol{h}_{\text{as}}, \boldsymbol{h}_{\text{mt}}, \boldsymbol{h}_{\text{at}} = \sigma(\mathrm{Conv}_{1D}(\boldsymbol{x}_{\text{ms}}, \boldsymbol{x}_{\text{as}}, \boldsymbol{x}_{\text{mt}}, \boldsymbol{x}_{\text{at}}))\,. \tag{9}$$

The decoded spatial and temporal attention signals are then combined to form a spa-

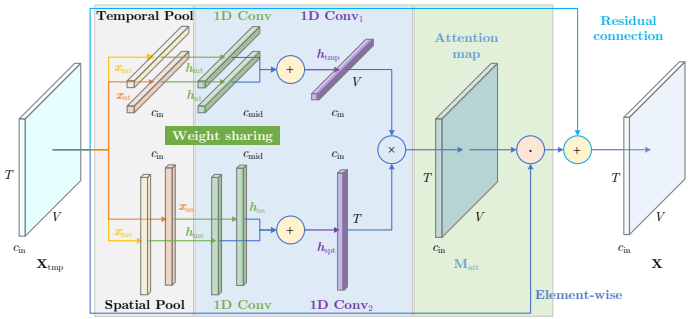


Figure 7: **The illustration of the spatial-temporal attention module.** '$\cdot$' and '$\times$' represent matrix dot-product and matrix multiplication operations, respectively.

tial–temporal attention map

$$\mathbf{M}_{\text{att}} = \kappa(\text{Conv}_{1D}(\boldsymbol{h}_{\text{ms}} + \boldsymbol{h}_{\text{as}})) \times \kappa(\text{Conv}_{1D}(\boldsymbol{h}_{\text{mt}} + \boldsymbol{h}_{\text{at}}))\,, \tag{10}$$

where $\kappa(\cdot)$ denotes the sigmoid function. Finally, the attention map is applied to refine the features via a residual formulation:

$$\mathbf{X}^{(l+1)} = \sigma\Big(\text{BN}\Big(\mathbf{X}_{\text{tmp}} \odot \mathbf{M}_{\text{att}} + \mathbf{X}_{\text{tmp}}\Big)\Big)\,. \tag{11}$$

This module follows common attention designs and serves as a lightweight feature reweighting mechanism to stabilize training and preserve discriminative information.

Discussion and Benefits. STGGCB enables effective long-range spatial-temporal modeling without modifying graph topology or increasing message-passing order. By injecting temporal variation into spatial aggregation and adaptively reweighting fixed multi-hop supports, the block captures global joint coordination while avoiding redundant connections and oversmoothing. Rather than increasing expressivity in theory, STGGCB provides a structured inductive bias that exposes complementary spatial and temporal cues, leading to more stable and efficient learning in practice.

4.4. Dual-Pathway Fusion

Design Rationale. Fig. 8 contrasts three paradigms. The single-branch STGCN (see Fig. 8(b)) [15, 17, 9] implicitly entangles spatial structure and temporal dynamics

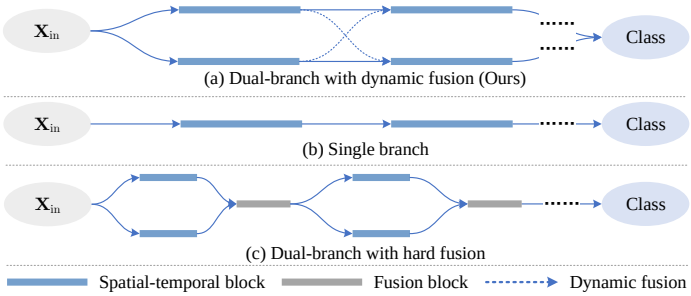


Figure 8: A comparison of STGCN network architectures: (a) our dual-pathway with dynamic fusion, (b) single pathway, (c) dual-pathway with hard fusion.

in a shared representation, which limits their distinct contributions. A dual-branch design with hard fusion (see Fig. 8(c)) [10, 21, 40] separates spatial and temporal processing but combines them using fixed operators, preventing adaptive interaction during representation learning. In contrast, our dual-branch architecture with dynamic fusion (see Fig. 8(a) explicitly separates spatial-aware and temporal-aware representations while enabling controlled, stage-wise interaction throughout the network. This mechanism fundamentally differs from conventional mid-level multimodal fusion, which typically merges heterogeneous features once at a predefined layer. Instead, NeuroPath enforces progressive coordination across the representation hierarchy, allowing spatial and temporal cues to influence each other during feature formation.

Dynamic Inter-Pathway Fusion. We perform *dynamic fusion* at multiple depths of the network. Specifically, after each STGGCB, features from the two pathways are softly exchanged, allowing complementary information to be progressively integrated while preserving pathway specialization. Formally, let $\mathbf{X}_{\text{s}}$ and $\mathbf{X}_{\text{t}}$ denote the spatial- and temporal-pathway features at a given stage. The fusion operation is defined as

$$\mathbf{X}'_{\text{s}} = \mathbf{X}_{\text{s}} + \alpha\,\mathbf{X}_{\text{t}}, \qquad \mathbf{X}'_{\text{t}} = \mathbf{X}_{\text{t}} + \beta\,\mathbf{X}_{\text{s}}, \tag{12}$$

where α and β are learnable scalar weights. This formulation implements a lightweight residual exchange mechanism, enabling each pathway to selectively absorb complementary cues from the other without collapsing them into a single representation. The fused features $\mathbf{X}'_{\text{s}}$ and $\mathbf{X}'_{\text{t}}$ are then propagated to the next block.

Discussion and Benefits. Compared with multi-stream ensemble methods [9], which fuse independently trained networks only at the score level, our fusion operates within a single network and during feature learning. This enables spatial and temporal cues to interact continuously across the representation hierarchy, rather than compensating for each other only at the final prediction stage. Importantly, the proposed fusion does not aim to increase expressivity through complex operators; instead, it introduces a structured coordination mechanism that explicitly exposes the imbalanced yet complementary nature of spatial and temporal information. As a result, NeuroPath achieves more effective and stable spatial–temporal coordination, reducing reliance on heavy multi-stream ensembles while improving accuracy and robustness in practice.

5. Experiments

This section presents the experimental setup followed by the experimental results.

5.1. Experimental Setting

Datasets. **Kinetics Skeleton 400** [41] is obtained from YouTube by OpenPose [42], containing 260,232 sequences with over 400 action types. The training and testing sets have 240,436 and 19,796 sequences, respectively. Each skeleton has 18 joints with 2D coordinates and prediction confidence as initial features. Two skeletons with higher overall confidence scores are preserved. Each sequence is padded to 300 by repeating it. For the Kinetics Skeleton 400 dataset, we report Top-1 and Top-5 accuracies for evaluation. **NTU RGB+D 60** [43] is collected by Microsoft Kinetics v2, including 56,578 sequences with over 60 action classes from 40 subjects and 3 camera views. Each skeleton contains 25 joints with 3D coordinates. The number of subjects in each frame ranges from 1 to 2. Two evaluation protocols are used: Cross-Subject (X-Sub) and Cross-View (X-View). In X-Sub, 40,091 samples from 20 subjects are used for training and 16,487 for testing. In X-View, 37,646 samples from camera 1 are used for training and 18,932 for testing. **NTU RGB+D 120** [44] extends the above with 57,367 samples with over 60 action classes, forming 113,945 samples of over 120 classes from 106 subjects and 32 camera setups. It replaces X-View with Cross-Setup (X-Set),

where 54,468 samples from 16 setups are used for training and 59,477 for testing. In X-Sub, 63,026 samples from 53 subjects are for training and 50,919 for testing.

Multi-Stream Evaluation Protocol. Following common practice [17, 9] in skeleton-based action recognition, we also report multi-stream results by combining predictions from multiple modality-specific inputs (e.g., joint, bone, joint motion, bone motion). These multi-stream results are used *only for evaluation purposes* to facilitate fair comparison with prior ensemble-based methods. Importantly, NeuroPath itself is trained end-to-end as a single unified model. The reported 4-stream results are obtained by averaging the prediction scores of independently evaluated NeuroPath models under different input modalities, and do not affect the training procedure.

Implementation Details. We implement NeuroPath using PyTorch and train it on two GeForce RTX 3090 GPUs. It is trained using SGD with a momentum of 0.9 and a batch size of 64. The initial learning rate is set to 0.1 for 70 epochs with step learning rate decay with a factor of 0.1 at epochs $\{30, 50\}$. Weight decay is set to 0.0001 and 0.0005 for Kinetics Skeleton 400 and NTU RGB 60 & 120. The processing of Kinetics Skeleton 400 is in line with [10], and NTU-RGB+D 60 and 120 are based on [9]. To speed up the training process on Kinetics Skeleton 400, 64 frames are randomly selected for the first 30 epochs and are successively increased to 300. For NTU RGB+D 60 and 120, 64 frames are selected, same as [9].

5.2. Comparisons with the State-of-the-Art

Baselines. We compare NeuroPath with representative state-of-the-art skeleton-based action recognition methods, including STGCN-based models [10, 26, 7, 9, 8], Transformer-based models [46, 30], and recent motion representation learning methods [2].We include MotionBERT [2] in Tab. 1 and directly compare it with our method under the same evaluation protocols. MotionBERT achieves strong performance due to its large-scale motion pretraining. Therefore, to provide a fair and comprehensive comparison, we report both its training-from-scratch and fine-tuned results. For other recent approaches [7, 24] that are not included in the main quantitative comparison, the primary reason is the inconsistency in input representation and evaluation pipeline with the standard coordinate-based ST-GCN setting [9, 26] adopted in this work. These

Table 1: Accuracy (%) comparison. The symbol "–" denotes unreported results. Underlined values indicate results taken from prior work [7]. Bold values indicate the best result in each column.

Ensemble stream	Method	Publication	Model type	Kinetics Skeleton 400		NTU RGB+D 60		NTU RGB+D 120	
				Top-1	Top-5	X-Sub	X-View	X-Sub	X-Setup
Joint-only	ST-GCN [15]	AAAI 2018	(2+1)D	30.7	52.8	81.5	88.3	70.7	73.2
	AGCN [17]	CVPR 2019	(2+1)D	34.8	56.5	86.8	94.2	77.9	78.5
	Shift-GCN [11]	CVPR 2020	(2+1)D	34.7	56.8	87.8	95.1	80.9	83.2
	CTR-GCN [9]	ICCV 2021	(2+1)D	36.5	58.5	89.0	94.8	83.6	84.5
	InfoGCN [8]	CVPR 2022	(2+1)D	36.7	58.7	90.4	95.0	84.1	85.2
	ML-STGNet [45]	TIP 2022	(2+1)D	–	–	89.8	95.0	84.9	86.5
	MS-G3D [10]	CVPR 2020	(2+1)D & 3D	36.5	58.6	89.4	95.0	83.1	85.0
	HetGCN [7]	TNNLS 2024	3D	**37.1**	59.2	90.5	95.3	84.3	85.7
	TranSkeleton [46]	TCSVT 2023	Transformer	–	–	90.1	95.4	84.9	86.3
	MotionBERT (scratch) [2]	ICCV 2023	Transformer	–	–	87.7	94.1	–	–
	MotionBERT (finetune) [2]	ICCV 2023	Transformer	–	–	**93.0**	**97.2**	–	–
	NeuroPath (Ours)	–	(2+1)D	37.0	**59.8**	90.8	95.8	**85.2**	**87.4**
Multi-stream	4s AAGCN [26]	TIP 2020	(2+1)D	37.8	61.0	90.0	96.2	–	–
	4s Shift-GCN [11]	CVPR 2020	(2+1)D	37.1	60.0	89.7	96.0	85.3	86.6
	3s Hyper-GNN [47]	TIP 2021	(2+1)D	37.1	60.0	89.5	95.7	–	–
	4s MST-GCN [22]	AAAI 2021	(2+1)D	37.8	60.3	91.1	96.4	87.0	88.3
	4s CTR-GCN [9]	ICCV 2021	(2+1)D	37.8	60.2	92.4	96.0	88.7	90.1
	6s InfoGCN [8]	CVPR 2022	(2+1)D	–	–	93.0	97.1	89.8	91.2
	2s ML-STGNet [45]	TIP 2023	(2+1)D	38.9	62.2	91.9	96.2	88.6	90.0
	4s KP-CTRGCN [39]	CVPR 2023	(2+1)D	–	–	92.9	96.8	90.0	91.3
	4s FRGCN [48]	CVPR 2023	(2+1)D	–	–	92.8	96.8	89.5	90.9
	4s JMDA [49]	TOMM 2025	(2+1)D	–	–	92.9	96.6	89.2	90.9
	4s LG-SGNet [29]	PR 2025	(2+1)D	–	–	**93.1**	96.7	89.4	91.0
	2s MS-G3D [10]	CVPR 2020	(2+1)D & 3D	38.0	60.9	91.5	96.2	86.9	88.4
	4s DualHead [50]	ACM MM 2021	(2+1)D & 3D	38.4	61.3	92.0	96.6	88.2	89.3
	2s HetGCN [7]	TNNLS 2024	3D	38.4	61.2	**93.1**	**97.2**	88.9	90.3
	4s TranSkeleton [46]	TCSVT 2023	Transformer	–	–	92.8	97.0	89.4	90.5
	4s THTFormer [30]	PR 2026	Transformer	–	–	93.0	96.8	**90.1**	91.3
	2s NeuroPath (Ours)	–	(2+1)D	38.9	61.9	92.4	97.0	89.5	91.0
	4s NeuroPath (Ours)	–	(2+1)D	**40.0**	**62.9**	93.0	**97.2**	89.9	**91.5**

Note: MotionBERT (finetune) [2] uses additional large-scale pretraining data (e.g., Human3.6M) before fine-tuning, while other methods follow standard NTU training protocols.

methods often rely on alternative pose representations, training pipelines, or preprocessing strategies, which makes direct comparison under identical experimental protocols non-trivial. For example, PoseConv3D [24] operates on heatmap-based pose representations using a 3D CNN architecture, which differs from coordinate-based skeleton inputs and graph convolution modeling. We thus treat it as a different pose representation paradigm rather than a directly comparable ST-GCN baseline.

Accuracy. Tab. 1 reports the accuracy comparison on multiple benchmarks. Here, "Js", "2s", and "4s" denote single-stream joint input, two-stream joint+bone inputs, and four-stream joint, bone, joint motion, and bone motion inputs, respectively. Overall, NeuroPath is primarily designed as a single-stream GCN-based framework that improves representation learning within each individual stream. The proposed design can also be extended to multi-stream settings as a plug-in module for evaluation. First, in the single-stream setting (Js), NeuroPath consistently achieves state-of-the-art performance among coordinate-based ST-GCN methods, demonstrating the effectiveness of the proposed intra-graph modeling. Importantly, the observed trends are consis-

tent with Fig. 2, where NeuroPath improves modality utilization and reduces imbalance across skeletal representations.Compared with MotionBERT, NeuroPath outperforms its training-from-scratch version on NTU RGB+D 60, while the 4s NeuroPath achieves the same reported accuracy as the finetuned MotionBERT on NTU RGB+D 60, i.e., 93.0% on X-Sub and 97.2% on X-View, without relying on large-scale motion pretraining. Second, extending to multi-stream configurations (2s and 4s) further improves performance in most cases, especially on Kinetics Skeleton 400 and NTU RGB+D 120. However, we observe that the benefit of multi-stream fusion is not always proportional, as naive aggregation of multiple streams may introduce redundancy or suboptimal cross-stream interactions. This suggests that while our method enhances per-stream representation quality, multi-stream performance also depends on the effectiveness of cross-stream fusion strategies, which is orthogonal to our main contribution. Finally, compared with DualHead [50], NeuroPath achieves clear improvements on NTU RGB+D 120, validating the effectiveness of the proposed single-stream enhancement mechanism.

Efficiency. The computational efficiency of NeuroPath under the X-Sub protocol of NTU RGB+D 60 is illustrated in Fig. 1, where the horizontal axis denotes GFLOPs, the vertical axis denotes accuracy, and the bubble size represents the number of parameters. As shown in Fig. 1, NeuroPath achieves a favorable accuracy–efficiency tradeoff, attaining high recognition accuracy with relatively low computational cost and compact model size. For the joint-stream setting, NeuroPath requires only 2.8M parameters and 4.17 GFLOPs, which is substantially more efficient than MS-G3D (3.2M parameters, 24.44 GFLOPs) and DualHead (3.0M parameters). This efficiency advantage stems from our unified spatial–temporal modeling within each pathway, which avoids the heavy two-stage spatial–temporal convolution commonly used in STGCN variants such as CTR-GCN [9]. Although 3D convolution-based approaches [10, 7] attempt to address this limitation, they typically incur significantly higher computational and parameter costs. Overall, NeuroPath demonstrates that explicitly coordinated dual-pathway modeling can achieve strong performance without sacrificing efficiency.

5.3. Ablation Study

To investigate the contribution of each component of NeuroPath, we conducted an ablation study using the Js NeuroPath on NTU RGB+D 60 in the X-Sub setting.

Effectiveness of Group Aggregation. Tab. 2 reports the results of the proposed group aggregation strategies. We evaluate two temporal grouping schemes: an *overlapping* strategy, where adjacent groups share temporal frames, and a *non-overlapping* strategy, where groups are disjoint in time. Compared with the baseline without group aggregation, both strategies bring clear performance gains (2.1% and 1.5%, respectively), demonstrating the effectiveness of group-level temporal modeling. The overlapping strategy consistently outperforms the non-overlapping one, as shared frames preserve temporal continuity across groups and enable the model to capture motion correspondence between consecutive segments, whereas non-overlapping grouping processes each segment independently and thus miss such cross-group dependencies.

Table 2: Accuracy (%) with various grouping strategies. The changes are relative to the first row.

Configuration	**Param. (M)**	**Acc. (%)**
Baseline w/o Group	2.51	88.7
Baseline w/ Overlapping	2.80 $^{\uparrow 0.29}$	**90.8** $^{\uparrow 2.1}$
Baseline w/ Non-Overlapping	2.80 $^{\uparrow 0.29}$	90.2 $^{\uparrow 1.5}$

We further investigate the impact of different overlapping strategies on the performance of NeuroPath, as summarized in Tab. 3. Here, "B1", "B2", and "B3" denote the three STGGCBs in the network, and "✗" indicates that the corresponding block is not grouped. The results show that grouping the first two blocks yields the best performance, while extending grouping to all three blocks introduces an additional 0.85M parameters compared to grouping only the first two blocks, increasing model complexity and leading to overfitting and degraded accuracy. We also observe that the temporal window size plays a critical role: both excessively small and overly large windows hurt performance. For instance, a window size of 3 results in a 0.4% accuracy drop compared to a window size of 7. Accordingly, we set the window size to 7 for all experiments, as it provides a better balance between temporal receptive field and sensitivity to local motion cues, i.e., larger windows tend to smooth out local dynamics,

while smaller windows lack sufficient temporal context.

Table 3: Accuracy (%) with various group configurations. The changes are relative to the first row.

Configuration	Window Size			Param. (M)	Acc. (%)
	B1	B2	B3		
w/o Group	✗	✗	✗	2.51	88.7
w/ Group	✓	✗	✗	2.65 $^{\uparrow 0.14}$	90.4 $^{\uparrow 1.7}$
	✓	✓	✗	2.80 $^{\uparrow 0.29}$	**90.8** $^{\uparrow 2.1}$
	✓	✓	✓	3.36 $^{\uparrow 0.85}$	90.4 $^{\uparrow 1.7}$
	✗	✓	✗	2.66 $^{\uparrow 0.15}$	90.1 $^{\uparrow 1.4}$
	✗	✗	✓	3.07 $^{\uparrow 0.56}$	89.9 $^{\uparrow 1.2}$
	✓	✗	✓	3.21 $^{\uparrow 0.70}$	90.3 $^{\uparrow 1.6}$
	✗	✓	✓	3.22 $^{\uparrow 0.71}$	90.2 $^{\uparrow 1.5}$

Additionally, Tab. 4 reports the plug-and-play results of integrating the proposed modules into AGCN [26]. Introducing the attention mask $\mathbf{M}_{\text{att}}$ yields a moderate accuracy improvement with a small parameter increase, while incorporating the proposed group aggregation module leads to a substantially larger performance gain. Combining both components further improves accuracy, demonstrating their complementary effects. These results indicate that the proposed group aggregation module is not only compatible with existing STGCN architectures, but also consistently improves their representation capability under a modest parameter overhead.

Table 4: Plug-and-play accuracy (%) of proposed modules. The changes are relative to the first row.

Configuration	Param. (M)	Acc. (%)
AGCN (Js) [26]	3.47	87.4
AGCN (Js) w/ $\mathbf{M}_{\text{att}}$	3.87 $^{\uparrow 0.40}$	88.2 $^{\uparrow 0.8}$
AGCN (Js) w/ Group	4.45 $^{\uparrow 0.98}$	89.1 $^{\uparrow 1.7}$
AGCN (Js) w/ Group + $\mathbf{M}_{\text{att}}$	4.85 $^{\uparrow 1.38}$	89.7 $^{\uparrow 2.3}$

Effectiveness of Adaptive Structural Aggregation. Tab. 6 shows the results of various structural aggregation strategies, among which the element-wise product ($\odot$) outperforms the additive strategy (+) proposed in the previous work [10] by 0.6%. This demonstrates the effectiveness of our adaptive structural aggregation. In addition, Tab. 5 presents the results of different multi-hop aggregation strategies, where 'T', 'M', and 'B' correspond to the top, middle, and bottom aggregation strategies in Fig. 6. NeuroPath achieves the best performance among them, highlighting its effectiveness.

By combining the results from Tabs. 5 and 6, it confirms that the importance-unaware aggregation strategy (90.6%) is inflexible, and adding residual links (90.2%) introduces redundancy, thereby limiting action recognition performance.

Table 5: Accuracy (%) in various spatial and temporal configurations. The changes are relative to the third row.

Configuration					Param. (M)	Acc. (%)
T	M	B	MK	MD		
✓	✗	✗	✗	✓	2.80	90.2
✗	✓	✗	✗	✓	2.80	90.6
✗	✗	✓	✗	✓	2.80	**90.8**
✗	✗	✓	✓	✗	3.04 $^{\uparrow 0.24}$	90.5 $^{\downarrow 0.3}$

Effectiveness of Multi-Scale Temporal Convolution. We investigate multi-kernel (MK) and multi-dilation (MD) temporal convolutions, as summarized in Tab. 5. Compared to MK, MD achieves higher accuracy with fewer parameters, indicating superior parameter efficiency. When combined with the proposed aggregation strategy, MD further improves performance while keeping the model compact. These results suggest that MD provides a more effective temporal modeling mechanism and works particularly well with our spatial aggregation design, offering a favorable balance between temporal receptive field and model efficiency for action recognition.

Effectiveness of Spatial-Temporal Attention. As shown in the sixth row of Tab. 6, removing the spatial–temporal attention module ($\mathbf{M}_{\text{att}}$) results in a 0.9% accuracy drop, indicating its clear contribution to performance. In addition, the plug-and-play results in Tab. 4 show that incorporating $\mathbf{M}_{\text{att}}$ into AGCN leads to a consistent 0.8% accuracy improvement. Together, these results demonstrate that spatial–temporal attention provides effective adaptive weighting of joint interactions over space and time, contributing to more discriminative representations for action recognition.

Effectiveness of Dual-pathway Architecture. The results of different architectural designs are reported in the first four rows of Tab. 6. Both dual-pathway variants (see Fig. 8(a) and Fig. 8(c)) outperform the single-pathway baseline (see Fig. 8(b)), demonstrating the effectiveness of separating spatial-aware and temporal-aware processing. Moreover, our proposed architecture (see Fig. 8(a)) achieves a 0.3% accuracy

Table 6: Accuracy (%) with various configurations. The changes are relative to the first row.

Configuration	**Param. (M)**	**Acc. (%)**
Fig. 8(a) w/ $\odot\mathbf{M}_k$ (see Eq. (7))	2.80	**90.8**
Fig. 8(a) w/ $+\mathbf{M}_k$ (see Eq. (7))	2.80 $^{\uparrow 0.00}$	90.2 $^{\downarrow 0.6}$
Fig. 8(b) w/ $\odot\mathbf{M}_k$ (see Eq. (7))	1.40 $^{\downarrow 1.40}$	88.5 $^{\downarrow 2.3}$
Fig. 8(c) w/ $\odot\mathbf{M}_k$ (see Eq. (7))	2.80 $^{\uparrow 0.00}$	90.5 $^{\downarrow 0.3}$
Fig. 8(a) w/o $\mathcal{M}_\mathrm{s}, \mathcal{M}_\mathrm{t}$	2.80 $^{\uparrow 0.00}$	89.7 $^{\downarrow 1.1}$
Fig. 8(a) w/o $\mathbf{M}_\mathrm{att}$ (see Eq. (10))	2.23 $^{\downarrow 0.57}$	89.9 $^{\downarrow 0.9}$
Fig. 8(a) w/o Dynamic Fusion	2.80 $^{\uparrow 0.00}$	90.4 $^{\downarrow 0.4}$
Fig. 8(a) w/o α_1, β_1	2.80 $^{\uparrow 0.00}$	90.3 $^{\downarrow 0.5}$
Fig. 8(a) w/o α_2, β_2	2.80 $^{\uparrow 0.00}$	90.5 $^{\downarrow 0.3}$
Fig. 8(a) w/o α_3, β_3	2.80 $^{\uparrow 0.00}$	90.4 $^{\downarrow 0.4}$

improvement over (c), indicating that dynamic fusion enables more effective interaction between the two pathways than hard fusion.

Effectiveness of Modality Transformation. We further evaluate the contribution of the modality transformation modules $\mathcal{M}_\mathrm{s}$ and $\mathcal{M}_\mathrm{t}$ by removing them in Tab. 6. Without these modules, both pathways receive identical inputs, which leads to a 1.1% accuracy drop. This result highlights the importance of explicitly assigning complementary modalities to the spatial-aware and temporal-aware pathways. In addition, compared with prior work that processes a single modality in a single-stream network [39], where the bone-motion stream achieves 88.0% accuracy, our method reaches 90.7%. This comparison further demonstrates the benefit of jointly leveraging complementary spatial-aware and temporal-aware information for action recognition.

Effectiveness of Dynamic Fusion. As shown in Tab. 6, removing the dynamic fusion module results in a 0.4% accuracy drop, confirming its positive contribution. We further analyze the effect of fusion at different network stages by individually removing the fusion weights in each block, which leads to varying degrees of performance degradation. These results indicate that dynamic fusion consistently contributes across multiple stages of the network. While most state-of-the-art methods in Tab. 1 rely on late fusion strategies, earlier fusion approaches [24] often introduce additional complexity or require richer input modalities. In contrast, our dynamic fusion mechanism enables effective stage-wise interaction within a single model, providing a more bal-

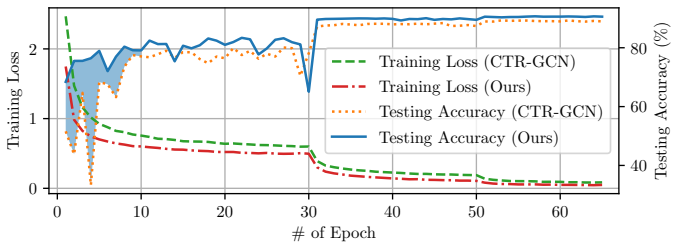


Figure 9: A comparison of the training loss and testing accuracy between CTR-GCN [9] and our method. The results demonstrate the effectiveness of our proposed model, as it achieves lower training loss and higher testing accuracy compared to CTR-GCN.

anced trade-off between performance and model complexity.

Effectiveness of Multi-Stream Skeletal Modeling. To evaluate the effectiveness of multi-stream modeling and the benefit of jointly using joint- and bone-based features, we conduct an ablation study on different skeletal modality combinations, as shown in Fig. 2(d). The evaluated settings include single-modality inputs, all pairwise combinations, and the full four-stream (4s) configuration. Single-modality inputs achieve comparable but limited performance (e.g., 90.82% for joint and 90.85% for bone), whereas combining joint and bone features yields a clear improvement to 92.40%, demonstrating their complementarity. Further gains are consistently obtained when incorporating motion cues, and the full 4s setting achieves the best performance (93.0%). These results confirm that multi-stream modeling effectively exploits complementary structural and motion information, leading to more robust representations.

5.4. Qualitative and Quantitative Results

We show more qualitative and quantitative results. All results are based on Js NeuroPath in the X-Sub setting of NTU RGB+D 60.

Fast Convergence. In Fig. 9, the training loss and testing accuracy curves for NeuroPath and CTR-GCN [9] are presented. The blue shading in the first 10 epochs represents the training phase. Our method demonstrates faster convergence than CTR-GCN, and around the 30th epoch, the learning rate changes noticeably. The results show that NeuroPath achieves lower training loss and higher testing accuracy than CTR-GCN, indicating its efficacy in modeling spatial-temporal features for action recognition. Furthermore, NeuroPath converges faster in the early stages, delivering superior

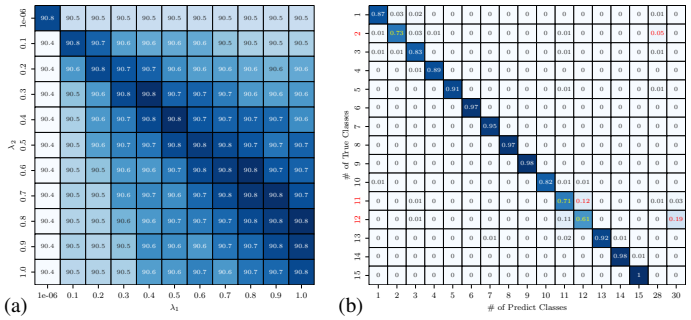
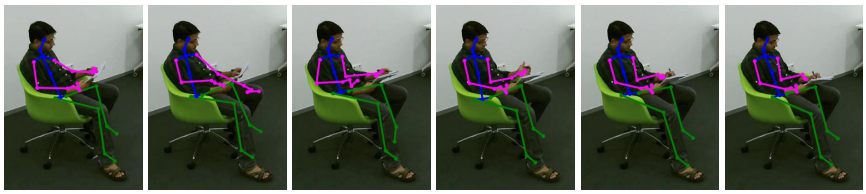


Figure 10: Heatmaps of (a) the fusion matrix and (b) the confusion matrix. λ_1 and λ_2 are coefficient weights.

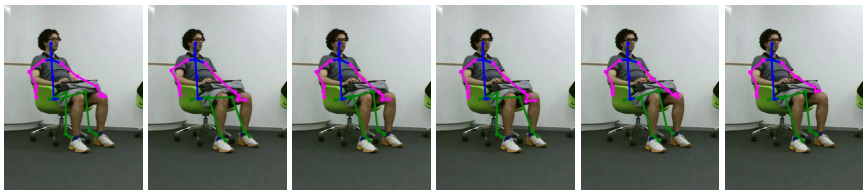
performance with fewer training epochs. This shows the superiority of NeuroPath.

Effective Fusion. Fig. 10(a) visualizes the confusion matrices under different fusion settings. We observe that using either spatial or temporal pathway alone yields inferior performance compared to their fused counterparts, indicating that neither pathway is sufficient in isolation. In contrast, dual-pathway fusion consistently improves recognition accuracy, validating the benefit of jointly modeling spatial structure and temporal dynamics. In all experiments, the fusion coefficients are set to $\lambda_1 = \lambda_2 = 1$. Notably, the performance gains are robust across different fusion configurations, suggesting that NeuroPath does not rely on a specific weighting scheme. This robustness stems from the dynamic fusion mechanism, which enables progressive and bidirectional information exchange between the two pathways while preserving pathway-specific representations. These results demonstrate that effective coordination between spatial-aware and temporal-aware pathways, rather than the dominance of a single pathway, is the key factor underlying the performance improvements of NeuroPath.

Failure Cases. The partial inter-class confusion matrix shown in Fig. 10(b) reveals that some actions share similarities in terms of body movements, leading to high confusion rates between them. For example, the 12-th *writing*, has been confused with the 30-th *typing on a keyboard*, in 19% of the samples. We provide visual examples of these two confused actions in Fig. 11, where different limbs are highlighted with different colors. In Fig. 11(a), the 311-th sample of *writing* is misclassified as *typing on a keyboard*, while in Fig. 11(b), the 3326-th sample of *typing on a keyboard* is mis-



(a) Writing (311-th sample) → typing on a keyboard



(b) Typing on a keyboard (3326-th sample) → writing

Figure 11: Selected frames with their human skeletons of two failure cases.

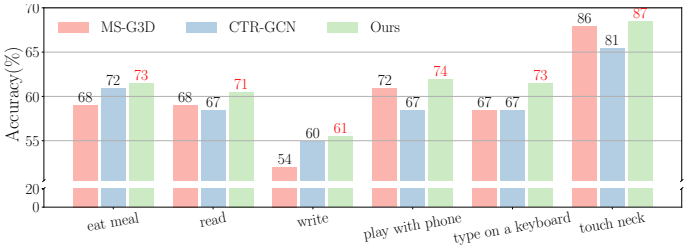


Figure 12: An accuracy comparison of six challenging classes in the X-Sub setting on NTU RGB+D 60.

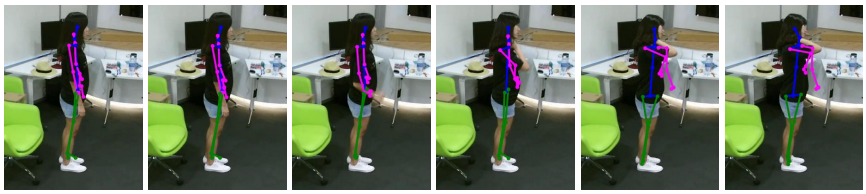
classified as *writing*. However, the hand and forearm movements are too subtle for NeuroPath to recognize. Our findings underscore the challenge of recognizing similar actions and the importance of developing robust features to capture subtle differences between them. It is noted that NeuroPath outperforms MS-G3D [10] in recognizing *writing*, achieving an accuracy of 61% compared to that of 54%. Furthermore, NeuroPath exhibits a lower misclassification rate of *writing* as *typing on a keyboard* at 19%, whereas MS-G3D has a higher rate of 23%. These results further demonstrate the superiority of NeuroPath over existing state-of-the-art approaches.

Effectiveness of Long-Range Dependencies. To provide further insights into the effectiveness of NeuroPath in recognizing challenging actions, we have selected six challenging actions in Fig. 13, and the result is shown in Fig. 12. Additionally, Tab. 7 reports the predicted scores of CTR-GCN [9] and our NeuroPath. Ours outperforms

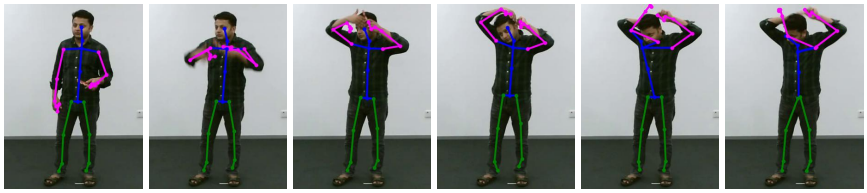
CTR-GCN in all the actions, demonstrating the effectiveness of incorporating long-range dependency information in NeuroPath. For instance, *touching neck* in Fig. 13(a) and *brushing hair* in Fig. 13(b) are hard to distinguish as they both involve hand movement close to the head. However, the subtle difference is mainly in the interaction of different parts of the human body rather than the local hand movements. NeuroPath achieves a correct recognition rate of 98%, demonstrating the effectiveness of incorporating long-range dependencies in distinguishing the two actions.

Table 7: Prediction scores of CTR-GCN [9] and NeuroPath.

#	Action	CTR-GCN [9]	Ours
1	Reading (GT)	0.21	0.82
	Typing on a keyboard	0.53	0.05
2	Typing on a keyboard (GT)	0.12	0.59
	Writing	0.47	0.28
3	Touching neck (GT)	0.02	0.88
	Brushing hair	0.96	0.10



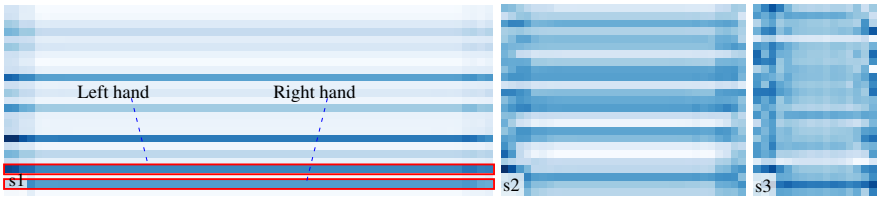
(a) Touching neck



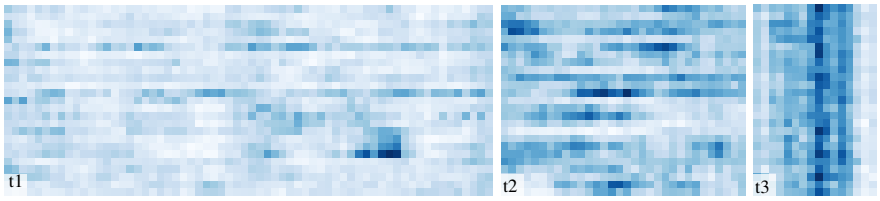
(b) Brushing hair

Figure 13: Selected challenging and well-recognized samples.

Attention Maps. Fig. 14 shows attention maps, specifically for a sample in Fig. 13(b) of *brushing hair*. The vertical axis denotes the joint index while the horizontal axis denotes the frame index. Darker colors indicate active joints with more discriminative features. Notably, the hand joints exhibit stronger attention in Fig. 14(a), ensuring the preservation of crucial information. Additionally, Fig. 14(a) attends to more variations in the spatial domain across joints, while Fig. 14(b) attends to more



(a) Spatial-aware pathway



(b) Temporal-aware pathway

Figure 14: Attention maps in the first STGGCB: (a) 's1', 's2', and 's3' are heatmaps of three STGGCBs in the spatial-aware pathway, (b) 't1', 't2', and 't3' are heatmaps of three STGGCBs in the temporal-aware pathway. The darker color denotes the higher attention weight for the corresponding joint.

variations in the temporal domain across frames, indicating the presence of unique information in each pathway. These findings emphasize NeuroPath's effectiveness in enhancing the discriminative features and improving action recognition accuracy.

6. Conclusion

This paper revisits skeleton-based action recognition from a modality-aware perspective. While STGCNs have achieved remarkable progress, we identify two persistent limitations: inefficient spatial–temporal cohesion within single branches and inadequate exploitation of complementary skeletal modalities. To address these issues, we propose *NeuroPath*, a brain-inspired dual-pathway framework that explicitly separates and coordinates spatial-aware and temporal-aware representations. NeuroPath integrates a spatial–temporal group graph convolution block for coherent intra-pathway modeling, a modality transformation unit for structured input decomposition, and dynamic inter-pathway fusion for progressive feature interaction. Extensive experiments on three large-scale benchmarks demonstrate that NeuroPath consistently achieves state-of-the-art performance with improved computational efficiency. The results validate that explicitly modeling the imbalanced yet complementary nature of

skeletal modalities leads to more effective and compact representations, reducing reliance on heavy multi-stream ensembles.

Limitations and Future Work. While NeuroPath focuses on improving single-stream representation learning and achieves strong performance under both single-stream and multi-stream evaluation settings, an interesting future direction is to explore more advanced multi-stream fusion strategies. In particular, modeling cross-stream interactions among joint, bone, and motion modalities in a more adaptive manner may further improve performance beyond simple aggregation. We consider such extensions to be orthogonal to the proposed single-stream enhancement framework. Beyond multi-stream fusion, another promising direction is to investigate more structured grouping and segmentation strategies, which may better capture fine-grained action semantics and further strengthen spatial–temporal coordination. Beyond the standard skeleton-based setting, the proposed dual-pathway framework may be extended to richer skeletal representations, such as heatmap-based pose encodings, as well as to multi-modal inputs including RGB, depth, or audio cues. At a higher level, leveraging pretrained backbone models and fine-tuning them on target domains [2, 51] remain a promising direction for advancing skeleton-based action understanding. Overall, NeuroPath provides an effective framework for unified spatial–temporal modeling, and offers useful insights for developing robust and interpretable action recognition systems.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62272019) and the China Postdoctoral Science Foundation (No. 2025M781489).

References

[1] K. Zhou, H. P. H. Shum, F. W. B. Li, X. Zhang, X. Liang, Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction, IEEE Transactions on Image Processing 34 (2025) 3718–3732.

[2] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, Y. Wang, Motionbert: A unified perspective on learning human motion representations, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 15085–15099.

[3] L. Zhao, X. Lu, Q. Bao, M. Wang, In-place gestures classification via long-term memory augmented network, in: 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2022, pp. 224–233.

[4] L. Zhao, X. Lu, M. Zhao, M. Wang, Classifying in-place gestures with end-to-end point cloud learning, in: 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2021, pp. 229–238.

[5] K. Zhou, R. Cai, L. Wang, H. P. H. Shum, X. Liang, A comprehensive survey of action quality assessment: Method and benchmark, Pattern Recognition 179 (2026) 113933.

[6] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, J. Chen, Memory attention networks for skeleton-based action recognition, IEEE Transactions on Neural Networks and Learning Systems 33 (9) (2022) 4800–4814.

[7] X. Gao, Y. Yang, Y. Wu, S. Du, Learning heterogeneous spatial-temporal context for skeleton-based action recognition, IEEE Transactions on Neural Networks and Learning Systems 35 (9) (2024) 12130–12141.

[8] 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.

[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] 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.

[11] 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.

[12] B. Ren, M. Liu, R. Ding, H. Liu, A survey on 3d skeleton-based action recognition using learning method, Cyborg and Bionic Systems 5 (2024) 0100.

[13] K. Zhou, Y. Ma, H. P. Shum, X. Liang, Hierarchical graph convolutional networks for action quality assessment, IEEE Transactions on Circuits and Systems for Video Technology 33 (12) (2023) 7749–7763.

[14] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017.
URL https://openreview.net/forum?id=SJU4ayYgl

[15] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Thirty-second AAAI conference on arti-

ficial intelligence, 2018.

[16] L. Shi, Y. Zhang, J. Cheng, Lu, Hanqing, Skeleton-based action recognition with directed graph neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 7912–7921.

[17] 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.

[18] V. Van Polanen, M. Davare, Interactions between dorsal and ventral streams for controlling skilled grasp, Neuropsychologia 79 (2015) 186–191.

[19] K. Zhou, H. P. Shum, F. W. Li, X. Liang, Multi-task spatial-temporal graph auto-encoder for hand motion denoising, IEEE Transactions on Visualization and Computer Graphics 30 (10) (2023) 6754–6769.

[20] K. Zhou, Z. Cheng, H. P. Shum, F. W. Li, X. Liang, Stgae: Spatial-temporal graph auto-encoder for hand motion denoising, in: 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2021, pp. 41–49.

[21] Z. Huang, X. Shen, X. Tian, H. Li, J. Huang, X.-S. Hua, Spatio-temporal inception graph convolutional networks for skeleton-based action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2122–2130.

[22] 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, Vol. 35, 2021, pp. 1113–1122.

[23] L. Huang, Y. Huang, W. Ouyang, L. Wang, Part-level graph convolutional network for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11045–11052.

[24] 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.

[25] 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.

[26] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Transactions on Image Processing 29 (2020) 9532–9545.

[27] Y.-F. Song, Z. Zhang, C. Shan, L. Wang, Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1625–1633.

[28] Z. Yang, J. Li, H. Zhang, D. Zhao, B. Wei, Y. Xu, Restore-rwkv: Efficient and

effective medical image restoration with rwkv, IEEE Journal of Biomedical and Health Informatics 30 (1) (2025).

[29] 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 Recognition 159 (2025) 111106.

[30] N. Ma, G. Xu, Y. Han, B. Sun, Thtformer: Topology-adaptive hypergraph transformer network for skeleton-based action recognition, Pattern Recognition (2026) 112125.

[31] Y. Chen, C. Lin, Y. Qiao, Dped: Bio-inspired dual-pathway network for edge detection, Frontiers in Bioengineering and Biotechnology 10 (2022) 1876.

[32] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Advances in neural information processing systems 27 (2014).

[33] Z. Yang, Y. Zhou, H. Chen, H. Zhang, D. Zhao, B. Wei, Y. Xu, Unipet: a universal network for high-quality pet image denoising across varied dose reduction factors, Medical Image Analysis (2026) 104059.

[34] W. Shi, T. Xu, H. Yang, Y. Xi, Y. Du, J. Li, J. Li, Attention gate based dual-pathway network for vertebra segmentation of x-ray spine images, IEEE Journal of Biomedical and Health Informatics 26 (8) (2022) 3976–3987.

[35] Z. Yang, H. Chen, Z. Qian, Y. Yi, H. Zhang, D. Zhao, B. Wei, Y. Xu, All-in-one medical image restoration via task-adaptive routing, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2024, pp. 67–77.

[36] Y. Wu, Y. Zhou, J. Saiyin, B. Wei, M. Lai, J. Shou, Y. Xu, Attriprompter: Auto-prompting with attribute semantics for zero-shot nuclei detection via visual-language pre-trained models, IEEE Transactions on Medical Imaging 44 (2) (2024) 982–993.

[37] X. Gao, W. Hu, J. Tang, J. Liu, Z. Guo, Optimized skeleton-based action recognition via sparsified graph regression, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 601–610.

[38] M. A. Goodale, A. D. Milner, Separate visual pathways for perception and action, Trends in neurosciences 15 (1) (1992) 20–25.

[39] X. Wang, X. Xu, Y. Mu, Neural koopman pooling: Control-inspired temporal dynamics encoding for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10597–10607.

[40] W. Zhang, Z. Lin, J. Cheng, C. Ma, X. Deng, H. Wang, Sta-gcn: two-stream graph convolutional network with spatial–temporal attention for hand gesture recognition, The Visual Computer 36 (10) (2020) 2433–2444.

[41] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the

kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[42] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: realtime multi-person 2d pose estimation using part affinity fields, IEEE transactions on pattern analysis and machine intelligence 43 (1) (2019) 172–186.

[43] 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.

[44] 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 transactions on pattern analysis and machine intelligence 42 (10) (2019) 2684–2701.

[45] Y. Zhu, H. Shuai, G. Liu, Q. Liu, Multilevel spatial-temporal excited graph network for skeleton-based action recognition, IEEE Transactions on Image Processing 32 (2023) 496–508.

[46] H. Liu, Y. Liu, Y. Chen, C. Yuan, B. Li, W. Hu, Transkeleton: Hierarchical spatial-temporal transformer for skeleton-based action recognition, IEEE Transactions on Circuits and Systems for Video Technology (2023).

[47] X. Hao, J. Li, Y. Guo, T. Jiang, M. Yu, Hypergraph neural network for skeleton-based action recognition, IEEE Transactions on Image Processing 30 (2021) 2263–2275.

[48] 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.

[49] L. Xiang, Z. Wang, Joint mixing data augmentation for skeleton-based action recognition, ACM Transactions on Multimedia Computing, Communications and Applications 21 (4) (2025) 1–24.

[50] T. Chen, D. Zhou, J. Wang, S. Wang, Y. Guan, X. He, E. Ding, Learning multi-granular spatio-temporal graph network for skeleton-based action recognition, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, p. 4334–4342.

[51] Y. Zhou, Y. Wu, J. Saiyin, B. Wei, Y. Xu, Sdpt: Synchronous dual prompt tuning for visual-language pre-trained models, IEEE Transactions on Pattern Analysis and Machine Intelligence (2026).