GCN-DevLSTM: Path Development for Skeleton-Based Action Recognition

Anonymous authors
Paper under double-blind review

Abstract

Skeleton-based action recognition (SAR) in videos is an important but challenging task in computer vision. The recent state-of-the-art (SOTA) models for SAR are primarily based on graph convolutional neural networks (GCNs), which are powerful in extracting the spatial information from skeleton data. However, it is not yet clear that such GCN-based models can effectively capture the temporal dynamics of human action sequences. To this end, we propose the G-Dev layer, which exploits the path development – a principled and parsimonious representation for sequential data by leveraging the Lie group structure. By integrating the G-Dev layer, the hybrid G-DevLSTM module enhances the traditional LSTM to reduce the time dimension while retaining high-frequency information. It can be conveniently applied to any temporal graph data, complementing existing advanced GCN-based models. Our empirical studies on the NTU-60, NTU-120 and Chalearn2013 datasets demonstrate that our proposed GCN-DevLSTM network consistently improves the strong GCN baseline models and achieves superior performance. [1]

1 Introduction

Action recognition has a wide range of applications in diverse fields such as human-computer interaction, performance assessment in athletics, virtual reality, and video monitoring Liu et al. (2019a). In comparison to video-based action recognition, skeleton-based action recognition (SAR) exhibits better robustness to the variety of light conditions and its efficiency in computation and storage Chen et al. (2021); Shi et al. (2019) while preserving data privacy. Skeleton data can be obtained either by localization of 2D/3D human joints using the depth camera Liu et al. (2019a) or pose estimation algorithms from the videos Cao et al. (2017).

Despite extensive literature devoted to SAR, the challenge still remains open. The key step towards accurate SAR models lies in designing an effective spatio-temporal representation of the skeleton sequence. Previously, Recurrent Neural Networks (RNNs) were commonly used due to their strength in capturing temporal dynamics. The advent of Spatial Temporal Graph Convolutional Networks (STGCN) has shifted the focus towards Graph Convolutional Networks (GCNs), which incorporate both spatial graph convolution and temporal convolution. The majority of subsequent GCN-based approaches Chen et al. (2021); Shi et al. (2020); Cheng et al. (2020b), following the STGCN framework, focus on the improvement of the spatial graph convolution, with minimal emphasis on improving temporal modules. These methods utilize 1-D temporal convolution operation to capture temporal features. However, whether convolution operation is most effective for capturing temporal dynamics is in question.

On the other hand, Rough path theory Lyons (1998) originated as a branch of stochastic analysis to make sense of controlled differential equations driven by highly oscillatory paths. More recently, the applications of rough path theory in sequential data analysis are emerging. In particular, the path signature, as a fundamental object of Rough path theory, has successful applications in various domains ranging from finance to healthcare, demonstrating its capacity as a principled and efficient feature of time series. By embedding time series into the continuous path space, the signature offers a unified treatment to handle variable length and irregular sampling. Signature-based algorithms show promise across diverse fields, including SAR task Liao

[1] The camera-ready version will contain a link to the code repository to ensure reproducibility.

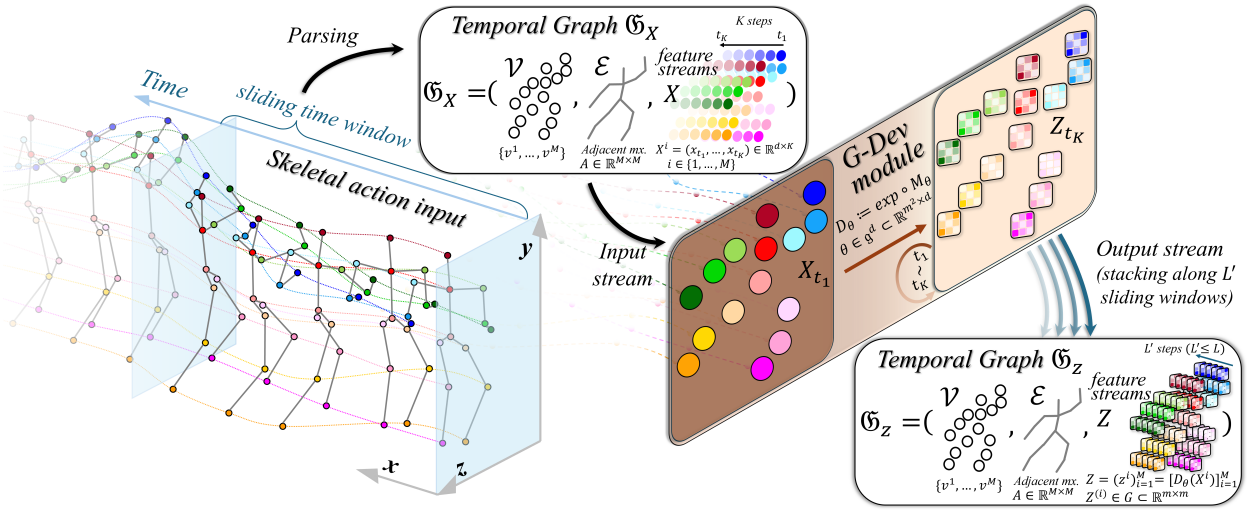


Figure 1: The work flow of G-Dev layer. The G-Dev layer takes a temporal graph (e.g., the skeleton sequence) as an input and outputs a temporal graph with a potentially significantly reduced time dimension. For each sub-time interval, we apply the G-Dev module (see Section 3.3) to generate a static graph representing the features within that window. We obtain the final temporal graph output by stacking these static graphs along the time dimension.

et al. (2021); Yang et al. (2022); Ahmad et al. (2019); Li et al. (2019a); Cheng et al. (2023). However, the path signature may face challenges such as the curse of dimensionality caused by high path dimension and lack of adaptability due to its deterministic nature Lou et al. (2022). To address these, Lou et al. Lou et al. (2022) propose the path development from rough path theory, as a trainable alternative to path signature. The path development enjoys the desirable properties of the signature while reducing feature dimension significantly. Lou et al. (2022) shows that combining the path development under the appropriate Lie group with LSTM Hochreiter & Schmidhuber (1997) alleviates the gradient vanishing of LSTMs, leading to significant enhancement across various sequential tasks.

This paper places a primary emphasis on enhancing the temporal module within our framework. To this end, we introduce a novel G-Dev layer (as shown in Figure 1) on the temporal graph, which refers to the graph with the node feature as time series. This layer utilizes the sliding window to summarize the local information via the principled path development, offering time dimension benefit and robustness to missing data and irregular sampling. Moreover, we propose the GCN-DevLSTM network, which retains the effectiveness of graph convolution for modeling spatial relationships among joints while leveraging the hybrid capacity of the G-Dev layer and LSTM to capture temporal dynamics. Our GCN-DevLSTM is flexible to incorporate with different GCN backbones. This combination captures discriminative temporal features and improves performance. Experimental results on three commonly used datasets demonstrate that when integrated with the CTRGC Chen et al. (2021) GCN backbone, our approach achieves state-of-the-art performance on the Chalearn2013 dataset and competitive results on NTU-60 and NTU-120, validating the effectiveness of the proposed temporal module. Furthermore, benefiting from the theory of path development, our method exhibits significantly enhanced robustness to missing or variable frames.

We summarize our main contributions as follows:

- We design a novel and principled graph development (**G-Dev**) layer for analyzing temporal graph data by extending the path development originally applied to vector-valued time series. As a generic network on graph data with time-varying node attributes, G-Dev exhibits efficacy and robustness to irregular sampling and missing data.

- We introduce the novel **GCN-DevLSTM** network for skeleton-based action recognition that effectively combines the G-Dev layer with LSTM to capture complex temporal dynamics.
- Experimental results demonstrate superior performance in terms of *accuracy* and *robustness*.

2 Related Work

2.1 Skeleton-based Action Recognition

The main challenge in SAR is extracting the effective representation that captures both spatial and temporal dependency of skeleton sequences. Earlier methods typically use convolutional neural networks (CNNs) Chéron et al. (2015); Liu et al. (2017b) and RNN Du et al. (2015); Liu et al. (2017a) to address this challenge without considering structural topology. STGCN Yan et al. (2018) is the first method to capture spatial and temporal relationships using GCN, taking the skeleton topology of the graph into consideration while this topology is not learnable, with only the spatial connection between adjacent joints being connected. As a result, some actions that rely on non-adjacent joints may lead to poor results. Liu et al. (2020) tackles this issue by introducing a learnable adjacency matrix alongside the original manually defined one, enabling connections between non-neighboring joints. Yet, this approach maintains the same graph for all action inputs, which might not be ideal, as different action classes may benefit from different graph topologies. To address this, Li et al. (2019b); Shi et al. (2019; 2020) propose adaptive GCN with the learnable and data-driven topology that shares among all channels. Cheng et al. (2020a); Chen et al. (2021) further propose the non-shared topology for each channel. Other recent studies improved model performance by enriching graph representations with higher-order features, such as angle-based information Schlegel et al. (2024). Despite these advances, current GCN-based methods mainly focus on the enhancement of the graph representation, with limited attention to temporal design. In particular, most GCN-based methods typically use CNNs to capture temporal dynamics. Although studies such as Si et al. (2019); Liao et al. (2021) attempt to incorporate LSTM with GCN for the SAR tasks, these methods still significantly underperform the sota GCN models that combine both GCN and CNN on large-scale action datasets.

2.2 Path signature & Path Development

The path signature is a core object of rough path theory, which provides a rigorous way of making sense of the solution to controlled differential equations driven by rough signals Lyons (1998). Playing a role as the non-commutative monomials on the path space, the path signature can be used as a faithful feature of time series Levin et al. (2013), which leads to applications for various sequential data learning tasks, such as handwritten recognition Xie et al. (2017), writer identification Yang et al. (2015), financial data analysis Lyons et al. (2014), time-series data generation Ni et al. (2021) and SAR Liao et al. (2021); Cheng et al. (2023); Yang et al. (2022); Ahmad et al. (2019); Li et al. (2019a). In the specific context of SAR, Ahmad et al. (2019) focuses on employing path signature solely for extracting spatial features while Yang et al. (2022); Li et al. (2019a) utilize path signature to obtain both temporal and spatial features. However, relying solely on path signature to capture spatial relationships may be suboptimal, as it is not data adaptive and only considers local spatial correlations. Liao et al. (2021) leverages GCN to extract spatial features and employ the log signature to capture temporal dynamics. However, the network tends to overfit rapidly when increasing its depth due to the curse of dimensionality of the signature for the high dimensional path.

More recently, the path development layer Lou et al. (2022) is proposed as a general sequential layer on time series data as a trainable alternative to the signature feature. It has useful applications in supervised learning on sequential data Lou et al. (2022), online signature verification Shi et al. (2025) and synthetic time generation Lou et al. (2024). This layer is inspired by the (Cantan) development of a path Driver (1995). The path development serves as a useful mathematical tool in rough path theory for proving the properties of the path signature (see e.g., Hambly & Lyons (2010); Chevyrev & Lyons (2016)). In particular,Chevyrev & Lyons (2016) proves that under an appropriate Lie group, the path development is a universal and characteristic feature of the path signature, which provides the theoretical support for the development layer Lou et al. (2022). Building upon Lou et al. (2022), our work extends the application of the path development from the vector-valued time series to that of the temporal graph.

3 G-Dev Layer: Development layer of a temporal graph

In this section, we start with a concise overview of the development of the path as the principled feature representation of d-dimensional time series. Subsequently, we proceed with introducing the path development layer proposed in Lou et al. (2022). For details of the path development, we refer readers to Lou et al. (2022). Building on this, we propose the G-Dev layer, a novel extension of the development layer applied to temporal graphs, wherein the features of nodes are represented by time series.

3.1 The development of a Path

Let $X \in V([0,T],\mathbb{R}^d)$, where $V([0,T],\mathbb{R}^d)$ represents the space of continuous paths of finite length in $\mathbb{R}^d$ defined on the time interval $[0,T]$. To fix the notations, let G denote a finite-dimensional Lie group and its associate Lie algebra $\mathfrak{g}$. Examples of the matrix Lie algebra associated with the special orthogonal group (denoted by $\mathfrak{so}$) and special Euclidean group ($\mathfrak{se}$) are given as follows:

$$\begin{aligned}
&\mathfrak{gl}(m;\mathbb{F}) = \{m \times m \text{ matrices over } \mathbb{F}\} \quad (\mathbb{F} = \mathbb{R} \text{ or } \mathbb{C});\\
&\mathfrak{so}(m,\mathbb{R}) := \{A \in \mathfrak{gl}(m;\mathbb{R}) : A^T + A = 0\};\\
&\mathfrak{se}(m) = \left\{ \begin{bmatrix} \omega & v \\ 0 & 0 \end{bmatrix} \mid \omega \in \mathfrak{so}(m,\mathbb{R}),\ v \in \mathbb{R}^m \right\}.
\end{aligned}$$

Definition 3.1 (Path Development) *Let $M : \mathbb{R}^d \to \mathfrak{g} \subset \mathfrak{gl}(m)$ be a linear map. The path development of $X \in V([0,T],\mathbb{R}^d)$ on G under M is the solution $Z_T \in G$ to the below equation*

$$dZ_t = Z_t \cdot M(dX_t) \quad \textit{for all } t \in [0,T] \quad \textit{with } Z_0 = Id_m, \tag{1}$$

where Id_m is the identity matrix and $\cdot$ denotes the matrix multiplication.

Example 3.1 (Linear path) *Let $X \in V([0,T],\mathbb{R}^d)$ be a* linear *path. For any linear map $M \in L(\mathbb{R}^d,\mathfrak{g})$, the development of the path under M admits the explicit formula and is simply*

$$D_M(X)_{0,T} = \exp\left(M(X_T - X_0)\right), \tag{2}$$

where exp *is the matrix exponential.*

Thanks to the multiplicative property of the path development (Lemma 2.4 in Lou et al. (2022)), the development of two paths $X \in V([0,s],\mathbb{R}^d)$ and $Y \in V([s,t],\mathbb{R}^d)$ under M is the same as the development of the concatenation of these two paths, in formula,

$$D_M(X * Y) = D_M(X) \cdot D_M(Y), \tag{3}$$

where $X * Y$ denotes the concatenation of X and Y and $\cdot$ is the matrix multiplication. Eq. equation 3 enables us to compute the development of any piecewise linear path explicitly.

The path development enjoys several desirable theoretical properties, making it a principled and efficient representation of sequential data. We summarize the following two properties, that are directly relevant to action recognition.

Invariance under Time-reparametrization Similar to the path signature, the path development is invariant under time reparametrization. This property ensures that the development of a path remains unchanged irrespective of the speed at which the path is traversed. It is well suited to action recognition, as the prediction of actions should not depend on the speed of the actions or the frame rate. Time-reparametrization invariance of path development may bring massive dimension reduction and enhance the robustness to variation in speed.

Uniqueness of the path development Intuitively, the path development effectively captures the order of events in time series, leveraging the non-commutative nature of matrix multiplication. Changing the order of events may lead to different development in general, highlighting its capacity to capture the temporal

dynamics of sequential data. It is proven in Chevyrev & Lyons (2016) the path development is a characteristic and universal feature of an un-parameterized path in $V_1([0,T],\mathbb{R}^d)$ under the appropriate Lie group. In other words, transforming a path of finite length to its development map $M \mapsto \mathrm{Dev}_M(X)$ can uniquely determine the path X up to translation. Thus, the path development combined with the starting point effectively summarizes the path without any loss of information.

3.2 Path Development Layer

Proposed in Lou et al. (2022), the path development layer is a novel neural network inspired by the path development by exploiting the representation of the matrix Lie group. More specifically, this layer utilizes the parameterized linear map $M \in L(\mathbb{R}^d, \mathfrak{g})$ by its linear coefficients $\theta \in \mathfrak{g}^d$.

Definition 3.2 (Path development layer) *The path development layer, denoted by D_θ, is defined as a mapping $\mathbb{R}^{d\times(N+1)} \to G$, which transforms any input time series $x = (x_1, x_2, ..., x_N)$ to the development $z_N \in G$ under the trainable linear map $M_\theta \in L(\mathbb{R}^d, \mathfrak{g})$. For $n \in \{0, ..., N-1\}$, we set*

$$z_{n+1} = z_n \exp(M_\theta(x_{n+1} - x_n)), z_0 = Id_m, \tag{4}$$

where $\theta \in \mathfrak{g}^d$ is the model parameter.

Eq. equation 4 is a direct consequence of Example 3.1 and the multiplicative property of the development (Eq. equation 3). Moreover, Eq. equation 4 shows the recurrence as an analogy to that of the RNNs, albeit in a much simpler manner, as depicted in Figure 1 (lower panel). Similar to RNNs, the path development can cope with time series input of variable length.

Dimension reduction The output of the path development has dimension m^2, where m is the matrix size. Note that its dimension does not depend on the path dimension d and time dimension T, resulting in the massive dimension reduction. This is in stark contrast to the signature feature of degree k, whose dimension $\left(\sum_{i=0}^{k} d^i = \frac{d^{k+1}-1}{d-1}\right)$ grows geometrically w.r.t d. Moreover, the path development is well suited to extract information from high-frequency time series or long time series without concern on the curse of dimensionality.

Robustness to irregular sampling As the path development can be interpreted as a solution to the differential equation (Eq. equation 1, it is a continuous time series model. Similar to other continuous time models such as signature and Neural CDEs Kidger et al. (2020), the path development exhibits robustness in handling time series data with irregular sampling.

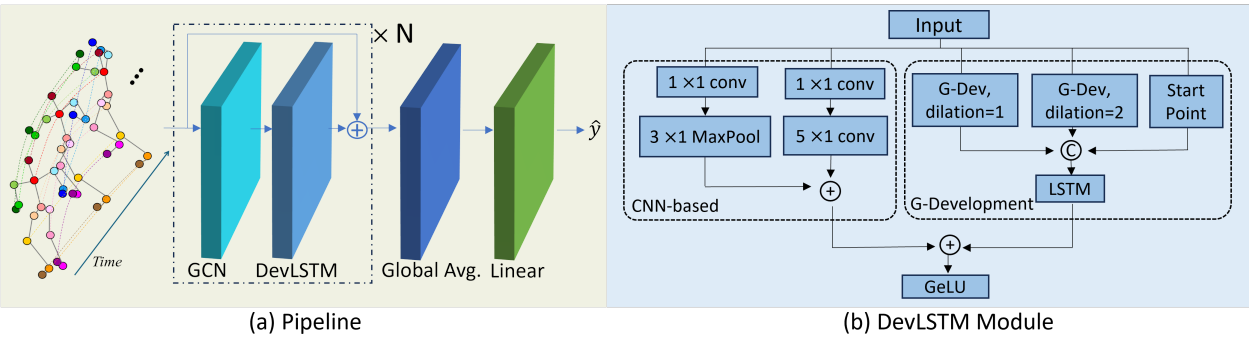
Computational Efficiency of analyzing online data Multiplicative property of the path development allows incremental computation of the development layer in streaming data, thereby reducing both computational load and memory requirements.

3.3 G-Dev Layer: Apply path development layer to a temporal graph

In this subsection, we propose the so-called G-Dev Sequence Layer, which extracts the local information of the temporal graph using the development layer in a rolling window fashion. The core component of the G-Dev Sequence layer serves as a generalization of the development layer from the d-dimensional time series to the temporal graph.

Throughout this paper, the temporal graph $\mathfrak{G}_X = (\mathcal{V}, \mathcal{E}, X)$ refers to the graph with each node associated with d-dimensional time series features X. Here $\mathcal{V} := \{v^1, \cdots, v^M\}$ denotes the vertex set and $\mathcal{E}$ can be represented by the adjacent matrix $A \in \mathbb{R}^{M\times M}$. We consider the general case where the time dimension of each node may vary across the nodes. More specifically, Let $X^i = (x_{t_1}, \cdots, x_{t_{L_i}}) \in \mathbb{R}^{d\times L_i}$ denote the features of the ith vertex, where d and L_i are the path and time dimension, respectively. The feature sets $X = (X^{(1)}, \cdots, X^{(M)})$ is obtained by stacking X^i for all $i \in \{1, \cdots, M\}$. Clearly, the skeleton sequence is an example of temporal graph data.

G-Dev module. Let us first define the G-Dev module, which is a trainable layer with model parameters $\theta \in \mathfrak{g}^d \subset \mathbb{R}^{m^2\times d}$, where m is the matrix order of development. It maps a temporal graph $\mathfrak{G}_X = (\mathcal{V}, \mathcal{E}, X)$ to



(a) Pipeline (b) DevLSTM Module

Figure 2: (a) The pipeline of our proposed approach consists of N blocks, with each block containing a GCN module and a DevLSTM module. (b) The detail of the DevLSTM module.

a static graph with matrix-valued features $\mathfrak{G}_Z = (\mathcal{V}, \mathcal{E}, Z)$ by applying the path development to the time-augmented transformation of each feature vector X_i of the vertex v_i, as shown in Figure 1. In formula, for each vertex index $i \in \{1, \cdots, M\}$, the output feature is $Z = (Z^{(i)})_{i=1}^{M}$, where $Z^{(i)} = D_\theta(X^{(i)}) \in G \subset \mathbb{R}^{m \times m}$.

Built on top of the G-Dev module, we employ a sliding window strategy to construct the following G-Dev Sequence layer, which can summarize (potentially long) time series over coarse time partitions to reduce the time dimension effectively while retaining the fine-grained information by using the low-dimensional representation.

Without loss of generality, let us focus on the case where the graph features X share the same time stamps $(t_i)_{i=0}^{L-1}$ across vertices[2]. Fix the window (kernel) size K and stride S. Let $L' = \lceil \frac{L-k}{S} \rceil$. By applying padding to Y, for any $Y \in \mathbb{R}^{d \times N}$, we define a sequence of time series $\tilde{Y}_j$:

$$\tilde{Y}_j = (Y_{j \times S}, \cdots Y_{j \times S + K - 1}) \in \mathbb{R}^{d \times K}, \forall j \in \{0, \cdots, L' - 1\}. \tag{5}$$

Definition 3.3 (G-Dev Sequence layer) *A G-Dev (Sequence) layer with the matrix degree m, window size K and stride S is defined as a trainable map that takes a temporal graph $\mathcal{G}_X = (\mathcal{V}, \mathcal{E}, X)$ with $X \in \mathbb{R}^{M \times L' \times d}$ as an input and outputs a temporal graph $\mathcal{G}_Z = (\mathcal{V}, \mathcal{E}, Z)$ with model parameter $\theta \in \mathbb{R}^{d \times m \times m}$. Here $Z = (z_j)_{j=1}^{L'}$ is a time series consisting of $z_j \in \mathbb{R}^{M \times m \times m}$. Each $z_j = \text{G-Dev}_\theta(\tilde{X}_j)$, where $\tilde{X}_j$ is obtained by applying sliding window by Eqn. 5.*

The G-Dev Sequence layer (Definition 3.3) can be easily extended to the moving window with the dilation parameter D, where we collect the path information at every D time steps during each moving window. To guarantee no information loss of the G-Dev sequence layer, it is crucial to choose S and K to ensure a minimum one-point overlap between two consecutive moving windows.

4 GCN-DevLSTM Network for SAR

A skeleton sequence can be modeled by a temporal graph, where the vertex set corresponds to M possible joints, with each $v^i \in \mathbb{R}^2$ or $\mathbb{R}^3$ indicating 2 or $3D$ coordinates of the ith joint, respectively. The edge set $\mathcal{E}$ denotes the set of all the joint pairs connected by a bone.

In Figure 2(a), we introduce the proposed GCN-DevLSTM Network, a novel framework for SAR tasks that integrates the GCN and DevLSTM modules into a single block. The GCN module is designed to detect spatial correlations among vertices, while the DevLSTM module handles the temporal dynamics across video frames. Further details on these modules are discussed in Sections 4.1 and 4.2, respectively. Our network is designed to capture discriminative spatial-temporal features with N such blocks, where $N = 9$ in this paper. Additionally, within each block, we build a residual path that establishes a shortcut connection with the input to mitigate gradient vanishing/ explosion issues, enhancing the training stability.

[2]Otherwise, we apply linear interpolation for data imputation, which would not affect the development feature thanks to invariance to time re-parameterization.

Following these GCN-DevLSTM blocks, we employ global average pooling to effectively reduce both temporal and spatial dimensions to 1. After that, a linear function is applied to adjust the feature dimension to match the number of action classes, thereby facilitating the classification process. Cross-entropy loss is employed at the end to compare the predicted action class with the ground truth.

4.1 GCN Module

In the GCN module, we basically follow the network of CTR-GC Chen et al. (2021). However, we depart from the use of three initial adjacency matrices representing self, inward, and outward correlations. Instead, we propose a Hops-CTRGC to enhance the model by incorporating various hops of information to construct three adjacency matrices denoted as A_0, A_1, and A_2.

A_0 remains an identical (self) matrix. For A_1, the elements a_{ij} are set to 1 if vertex i and vertex j are physically connected in the skeleton graph; otherwise, a_{ij} in A_1 is set to 0. A_1 captures the direct connections in the graph. To enrich spatial information, we introduce $A_2 = A_1^2$, enabling a vertex to connect with its neighbour's neighbour. A_2 allows us to extend our understanding beyond direct connections to include relationships involving two hops in the graph. Since GCN module is not our main focus in this paper, readers can find more details about the network of CTR-GC in Chen et al. (2021) or in our appendix A.

4.2 G-DevLSTM Module

The proposed DevLSTM module, illustrated in Figure 2(b) is mainly composed of the CNN-based branch and the G-development-based branch. The output of DevLSTM module is given by applying GeLU activation function to the summation of the output of both branches.

G-Development branch. Inspired by the design of multi-scale TCN Liu et al. (2020), we also design a multi-scale network to capture temporal dynamics. We utilize two G-dev layers with the dilation parameter $D = 1$ and $D = 2$ to enhance its ability to capture temporal patterns across different scales in the input sequence.

The path development combined with the starting point can determine the path uniquely. Hence, we append the location of the starting points to the output of the path development layer along the feature dimension over each sub-time interval. The resulting features are then fed into an LSTM network, which is able to effectively aggregate the local development summary and learn deep features to extract the long-term dependencies in time series data. We use the sequential output of LSTM, which keeps the time dimension the same as the input.

CNN-based branch. MaxPool and a convolutional layer with 5×1 kernel size have consistently demonstrated superior performance within the temporal module of other GCN-based studies. Introducing both MaxPool and the 5×1 convolutional operation to our DevLSTM temporal module has the potential to enhance model performance even further. We can reasonably design kernel size, padding, stride and dilation to ensure the output size of Maxpool, LSTM and 5×1 convolution remain the same.

5 Numerical Experiments

To validate the effectiveness of the proposed approach, we carry out numerical experiments on the following popular benchmarking datasets for skeleton-based action and gesture recognition. These datasets are chosen for their diversity in data size and task complexity, providing a comprehensive evaluation of the model's performance. (1) **Chalearn2013** Escalera et al. (2013) is a public dataset for gesture recognition, consisting of 11,116 skeleton examples and 20 gesture categories, filmed by 27 people. Each frame has 20 joints. (2) **NTU RGB+D 60** (NTU-60) Shahroudy et al. (2016) consists of 56,880 action samples and 60 action classes, filmed by 40 people using three Kinect cameras in different camera views. 3D skeletal data in the dataset have different sequence lengths, with each owning 25 joints. (3) **NTU RGB+D 120** (NTU-120) Liu et al. (2019b) extends upon NTU RGB+D 60, consisting of 114,480 samples and 120 action classes, filmed by 106 people in three camera views. We consider the Cross-subject (X-Sub) and Cross-view (X-View)

evaluation benchmarks for both NTU-60 and NTU-120 as proposed by Shahroudy et al. (2016). We use only skeleton data from these datasets for our experiments.

Experiments are conducted using a single Nvidia A6000 graphics card, where a batch of 16 examples is fed to the network. Stochastic gradient descent optimiser Bottou (2012) is employed, with a Nesterov momentum = 0.9 and a weight decay = 0.0003. The training epoch is set to 65 and we apply the warm-up strategy He et al. (2016) to the first 5 epochs. The initial learning rate is 0.02 and a step learning rate decay is applied with a factor of 0.1 at epoch 35 and 55, respectively. Data processing on NTU datasets mainly follows the procedure in Chen et al. (2021), with the length of all input data sequences resizing to 64. We use $\mathfrak{sc}$ Lie algebra in this paper, with a corresponding matrix size of $m = 10$. A detailed network architecture can be found in the supplementary. The replicate padding is used in our experiments.

5.1 Fusion results from different data streams

Most of the recent approaches that we are set to compare combine predicted scores from different input data streams for enhanced performance. The most commonly used data streams are joint (J), bone (B), joint motion (JM) and bone motion (BM) where details can be found in Shi et al. (2019; 2020). Apart from the above-mentioned data streams, we introduce a novel stream: the dual graph (D). In a dual graph, vertices represent bones, and edges establish connections between these bones. More details about dual graph can be found at Appendix B.

5.2 Comparison with state-of-the-art methods

Table 1: Results on Chalearn2013 dataset.

	Acc (%)
Two-streamLSTM Wang & Wang (2017)	91.70
ST-LSTM + Trust Gate Liu et al. (2018)	92.00
Multi-path CNN Liao et al. (2019)	93.13
3s_net_TTM Li et al. (2019a)	92.08
GCN-Logsig-RNN Liao et al. (2021)	92.86
ST-PSM+L-PSM Cheng et al. (2023)	94.18
AGGP Shi et al. (2024)	95.18
Our method	**95.61**

Table 1 shows the comparison results between our result and other state-of-the-art methods on Chalearn2013 dataset. We only employ the joint stream for fair comparison as other approaches only use the joint representation on Chalearn2013 dataset. It is evident that our model yields the best result on the small-scale gesture dataset, surpassing the second-best method by 0.42%, proving its effectiveness on the small dataset.

We compare our approach with other state-of-the-art methods on both NTU-60 and NTU-120 datasets, as shown in Table 2. Since our GCN module is following CTR-GCN Chen et al. (2021) with the major difference lying in the temporal module, we show more results on CTR-GCN for a more comprehensive comparison. As we can see all of our results, no matter whether using only joint or using ensembled streams combining joint and bone, outperform CTR-GCN with a large margin on all benchmarks. Our method's limited performance gain with extra joint or bone motion, compared to CTR-GCN, may stem from our path development layer exploiting position differences between consecutive time points. As a result, streams using joint and bone motion provide identical information at the beginning of the network, limiting their help. Our model, integrating five data streams, achieves the best result compared to other state-of-the-art approaches across three out of four benchmarks. It is important to highlight that our network structure remains the same on all datasets, demonstrating its strong adaptability for different dataset sizes.

Table 2: Experimental results on NTU-60 and NTU-120 dataset. The best results for each benchmark are bolded.

						NTU-60		NTU-120	
						X-sub (%)	X-view (%)	X-sub (%)	X-view (%)
ST-GCN Yan et al. (2018)						81.5	88.3	70.7	73.2
AGC-LSTM Si et al. (2019)						89.2	95.0	-	-
2S-AGCN Shi et al. (2019)						88.5	95.1	82.5	84.2
MS-AAGCN Shi et al. (2020)						90.0	96.2	-	-
Shift-GCN Cheng et al. (2020b)						90.7	96.5	85.9	87.6
Logsig-RNN Liao et al. (2021)						-	-	75.8	78.0
CTR-GCN(J) Chen et al. (2021)						89.8	94.8	84.9	86.7
CTR-GCN(J+B) Chen et al. (2021)						92.0	96.3	88.7	90.1
CTR-GCN Chen et al. (2021)						92.4	**96.8**	88.9	90.6
EfficientGCN Song et al. (2022)						92.1	96.1	88.7	88.9
Ta-CNN Xu et al. (2022)						90.4	94.8	85.4	86.8
FR-Head Zhou et al. (2023)						92.8	**96.8**	89.5	90.9
SPIANet Yin et al. (2024)						92.8	**96.8**	89.2	90.4
FD-GCN Huo et al. (2026)						92.8	96.7	89.4	90.7
	Joint	Bone	Dual	Joint motion	Bone motion				
Our methods	✓					90.8	95.7	86.2	87.7
		✓				91.1	95.2	87.0	88.6
			✓			91.2	95.3	87.1	88.7
	✓	✓				92.4	96.4	89.0	90.4
	✓	✓		✓	✓	92.6	96.6	89.3	90.7
	✓	✓	✓	✓	✓	**92.9**	96.7	**89.7**	**91.2**

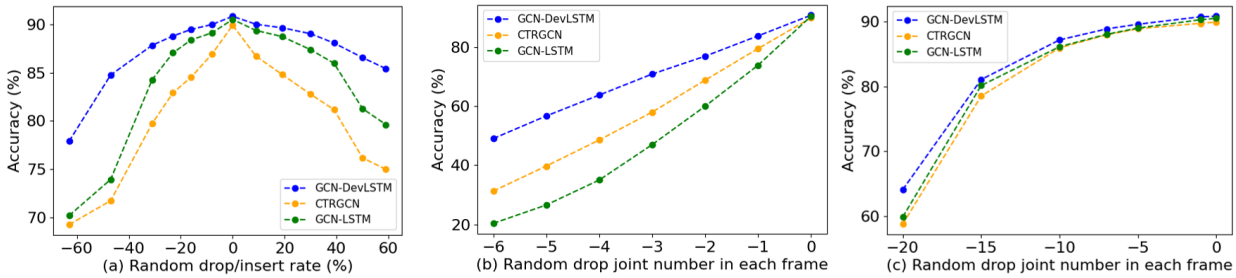


Figure 3: Robustness analysis on NTU60 X-sub benchmark. Figure 3a shows the results of randomly dropping or inserting a certain frames to original time series. Figure 3b and 3c shows the results of randomly dropping a number of joints in each frame with zero-filling and linear interpolation respectively.

5.3 Comparison with signature-based models

Since path development layer can be seen as a data-adaptive alternative to the signature, we further compare our method with other signature-based models in SAR task. Compared to the Logsig-RNN Liao et al. (2021), one of the recent signature-based action recognition methods (using the joint stream only), our model (J) outperforms 10.4% and 9.7% respectively on both benchmarks of NTU-120 dataset, and 2.75% on Chalearn2013 dataset. Compared to the latest signature-based action recognition approach Cheng et al. (2023), we still outperform 1.43% on Chalearn2013 dataset while unfortunately, they didn't report their results on NTU dataset, leaving no way to compare. These results support the advantages of using path development that we claimed before, compared with path signature.

5.4 Robustness analysis

We comprehensively compare the robustness of our full model (GCN-DevLSTM), CTRGCN Chen et al. (2021) and our model without path development layer (GCN-LSTM) in terms of missing frames, missing joints and variable sequence length. To this end, we firstly create a new test dataset by randomly dropping

or inserting a certain number of frames out of 64 frames to test the robustness of missing frames and variable length. Figure 3a illustrates that our full model with the path development layer consistently outperforms the other methods in terms of accuracy, regardless of the proportion of frames missing or inserting, proving its robustness to irregular sampling and variable sequence length. Notably, at a 47% frame-drop rate, the accuracy of our full model decreases by only 6.07%, while CTR-GCN drops by 18.18% and GCN-LSTM by 16.56%. A similar trend is observed when frames are inserted.

Additionally, to validate the robustness of missing joints due to occlusions in real life, we create another dataset by randomly dropping a certain number of joints in each frame. To fill in these missing joints, two techniques are adopted: zero-filling (in Figure 3b) and linear interpolation (in Figure 3c). Both figures show that our model with G-Dev layer, is consistently more robust against missing joints than others. More details are in the supplementary material. We credit the superior robustness of our full model to the theory foundation provided by the path development layer, mentioned in 3.2.

5.5 Comparison with MS-TCN

To demonstrate the flexible integration of our DevLSTM temporal module with various GCN modules in a plug-and-play fashion, we employ the adaptive graph convolution (AGC) graph in 2S-AGCN Shi et al. (2019) and a fixed graph without learnable parameters, respectively as an additional GCN module backbone in our network. In addition, we compare our DevLSTM with multi-scale temporal convolution network (MS-TCN) Chen et al. (2021), commonly used as the temporal module in other recent works across the different GCN backbones. Table 3 illustrates that the proposed DevLSTM module consistently outperforms MS-TCN with different GCN backbones by a large margin, indicating that our designed temporal module is effective in enhancing temporal modeling across various GCN backbones. An overview of adaptive graph convolution and fixed graph can be found in the supplementary material.

Table 3: Comparison between our DevLSTM module with MS-TCN across various GCN modules on NTU60 dataset.

GCN Module	Temporal module	X-sub	X-view
CTR-GC	MS-TCN (J)	90.3	94.9
	DevLSTM (J)	**90.8**	**95.7**
AGC	MS-TCN (J)	89.8	94.5
	DevLSTM (J)	**90.6**	**95.6**
Fixed graph	MS-TCN (J)	88.7	93.1
	DevLSTM (J)	**89.5**	**94.5**

5.6 Ablation studies

We further investigate the importance of residual path design to connect the input and output of each block and each component in our temporal module. As evident from Table 4, residual path plays an important role to enhance the model performance since it can help alleviate gradient vanishing/explosion in the deep neural network. For temporal components in our DevLSTM module, the path development layer emerges as the primary contributor to the overall performance improvement since removing it causes the largest performance degradation, with other components also contributing to enhanced performance.

Table 4: Ablation studies of DevLSTM module on NTU-60 dataset. w/o means without.

Components	X-sub	X-view
with all components (J)	90.8	95.7
w/o residual path (J)	89.9	95.2
w/o 1× 1 conv & 5×1 conv (J)	90.7	95.3
w/o 1× 1 conv & 3× 1 maxpool (J)	90.6	95.4
w/o path dev (J)	90.5	95.2

Table 5: Model complexity analysis. Compare the model parameters between our approach with the corresponding accuracy on NTU-60 X-view dataset using joint stream only.

	ST-GCN Yan et al. (2018)	2s-AGCN Shi et al. (2019)	CTRGCN Chen et al. (2021)	**Ours**								
Blocks	-	-	-	9	8	7	6	5	4	3	2	1
Params (M)	1.22	1.55	1.46	4.61	3.29	1.78	1.38	0.98	0.53	0.40	0.26	0.13
Acc (%)	88.3	94.0	94.8	95.7	95.5	95.4	95.1	95.0	94.0	93.7	92.7	89.1

5.7 Model complexity analysis

Efficacy of compact GCN-DevLSTM networks in Table 5 shows our trade-off between the accuracy and model complexity using various block numbers. Increasing the block number from 1 to 9 leads to boosting the accuracy from 89.1% to 95.7% while increasing the model parameter number from 0.13M to 4.61M. However, even with fewer blocks, our results are competitive with much less model parameters in comparison with other stoa methods. As illustrated in Table 5, adding more blocks to our networks can further improve accuracy, albeit with increased parameter number and computational complexity. It is worthy noting that our model with 5 blocks only has 68% of model parameters of the strong baseline CTRGCN, reducing 0.49 million parameters (from 1.46M to 0.98M) while slightly outperforming it. Besides, our model with a single block surpasses ST-GCN, despite the differences in preprocessing or training strategies between ST-GCN and our method.

5.8 Lie Group & Hidden Size Selection

Table 6: Evaluation of accuracy (%) across different hidden sizes in special Euclidean (SE) and special orthogonal (SO) Lie group on NTU60-Xview benchmark with the joint stream.

Lie Group	Hidden Size				
	8	9	10	11	12
SE	95.6	95.4	95.7	95.5	95.5
SO	95.5	95.5	95.4	95.6	95.4

Given that path development is key to our methodology, the choice of a suitable Lie Group, along with its corresponding hidden size, influences the ultimate performance of the model. Table 6 presents results for two commonly used Lie groups with varying hidden sizes. The results indicate that a larger hidden size does not necessarily ensure an enhancement in the model and the best result is obtained by utilizing the special Euclidean group with a hidden size of 10. These important hyperparameters should be tuned based on numerical results in practice.

6 Conclusion and future work

In summary, we propose a novel and generic GCN-DevLSTM network for analyzing temporal graph data by leveraging the path development layer. This approach not only offers flexibility when combined with various GCN modules but also significantly enhances performance by effectively capturing the temporal dependencies within data. The effectiveness of the proposed method is validated by the superior results on three datasets in SAR. Moreover, it demonstrates strong robustness to missing data and variable frame rates.

The current G-Dev Sequence layer relies on other GCNs to extract spatial information, which limits its standalone utility. Future enhancements could include integrating the features of adjacent nodes directly into the input of the path development layer to better capture spatio-temporal dependencies.

Our proposed G-Dev Sequence layer provides a generic model for temporal graph data, opening up possibilities for its applications across various fields, including urban planning, social network analysis and others.

References

Tasweer Ahmad, Lianwen Jin, Jialuo Feng, and Guozhi Tang. Human action recognition in unconstrained trimmed videos using residual attention network and joints path signature. *IEEE Access*, 7:121212–121222, 2019.

Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition*, pp. 421–436. Springer, 2012.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *IEEE International Conference on Computer Vision*, pp. 13359–13368, 2021.

Jiale Cheng, Dongzi Shi, Chenyang Li, Yu Li, Hao Ni, Lianwen Jin, and Xin Zhang. Skeleton-based gesture recognition with learnable paths and signature features. *IEEE Transactions on Multimedia*, 2023.

Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *European Conference on Computer Vision*, pp. 536–553, 2020a.

Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 183–192, 2020b.

Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *IEEE International Conference on Computer Vision*, pp. 3218–3226, 2015.

Ilya Chevyrev and Terry J Lyons. Characteristic functions of measures on geometric rough paths. *Annals of probability: An official journal of the Institute of Mathematical Statistics*, 44(6):4049–4082, 2016.

BRUCE K Driver. A primer on riemannian geometry and stochastic analysis on path spaces. 1995.

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

Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Isabelle Guyon, Vassilis Athitsos, Hugo Escalante, Leonid Sigal, Antonis Argyros, Cristian Sminchisescu, et al. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In *15th ACM on International conference on multimodal interaction*, pp. 365–368, 2013.

Ben Hambly and Terry Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, pp. 109–167, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

Jinze Huo, Haibin Cai, and Qinggang Meng. Fast distance-enhanced graph convolutional network for skeleton-based action recognition. *Pattern Recognition*, 169:111827, 2026.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33:6696–6707, 2020.

Daniel Levin, Terry Lyons, and Hao Ni. Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*, 2013.

Chenyang Li, Xin Zhang, Lufan Liao, Lianwen Jin, and Weixin Yang. Skeleton-based gesture recognition using several fully connected layers with path signature features and temporal transformer module. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 8585–8593, 2019a.

Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3595–3603, 2019b.

Lufan Liao, Xin Zhang, and Chenyang Li. Multi-path convolutional neural network based on rectangular kernel with path signature features for gesture recognition. In *IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4. IEEE, 2019.

Shujian Liao, Terry Lyons, Weixin Yang, Kevin Schlegel, and Hao Ni. Logsig-rnn: a novel network for robust and efficient skeleton-based action recognition. *British Machine Vision Conference*, 2021.

Bangli Liu, Haibin Cai, Zhaojie Ju, and Honghai Liu. Rgb-d sensing based human action and interaction analysis: A survey. *Pattern Recognition*, 94:1–12, 2019a.

Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1647–1656, 2017a.

Jun Liu, Amir Shahroudy, Dong Xu, Alex C Kot, and Gang Wang. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):3007–3021, 2018.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex 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):2684–2701, 2019b.

Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017b.

Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 143–152, 2020.

Hang Lou, Siran Li, and Hao Ni. Path development network with finite-dimensional lie group representation. *arXiv preprint arXiv:2204.00740*, 2022.

Hang Lou, Siran Li, and Hao Ni. Pcf-gan: generating sequential data via the characteristic function of measures on the path space. *Advances in Neural Information Processing Systems*, 36, 2024.

Terry Lyons, Hao Ni, and Harald Oberhauser. A feature set for streams and an application to high-frequency financial tick data. In *2014 International Conference on Big Data Science and Computing*, pp. 1–8, 2014.

Terry J Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2): 215–310, 1998.

Hao Ni, Lukasz Szpruch, Marc Sabate-Vidales, Baoren Xiao, Magnus Wiese, and Shujian Liao. Sig-wasserstein gans for time series generation. In *Second ACM International Conference on AI in Finance*, pp. 1–8, 2021.

Kevin Schlegel, Lei Jiang, and Hao Ni. Using joint angles based on the international biomechanical standards for human action recognition and related tasks. *arXiv preprint arXiv:2406.17443*, 2024.

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

Dongzi Shi, Xin Zhang, Jiale Cheng, Tong Xiong, and Hao Ni. Adaptive global gesture paths and signature features for skeleton-based gesture recognition. In *International Conference on Pattern Recognition*, pp. 278–292. Springer, 2024.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035, 2019.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020.

Yilin Shi, Lei Jiang, Hao Ni, and Lianwen Jin. Devinsight: Weaving path development into online signature verification. In *International Conference on Document Analysis and Recognition*, pp. 398–415. Springer, 2025.

Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *IEEE International Conference on Computer Vision*, pp. 1227–1236, 2019.

Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (2):1474–1488, 2022.

Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 499–508, 2017.

Zecheng Xie, Zenghui Sun, Lianwen Jin, Hao Ni, and Terry Lyons. Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1903–1917, 2017.

Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 2866–2874, 2022.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Weixin Yang, Lianwen Jin, and Manfei Liu. Chinese character-level writer identification using path signature feature, dropstroke and deep cnn. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 546–550. IEEE, 2015.

Weixin Yang, Terry Lyons, Hao Ni, Cordelia Schmid, and Lianwen Jin. Developing the path signature methodology and its application to landmark-based human action recognition. In *Stochastic Analysis, Filtering, and Stochastic Optimization: A Commemorative Volume to Honor Mark HA Davis's Contributions*, pp. 431–464. Springer, 2022.

Xinpeng Yin, Jianqi Zhong, Deliang Lian, and Wenming Cao. Spatiotemporal progressive inward-outward aggregation network for skeleton-based action recognition. *Pattern Recognition*, 150:110262, 2024.

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