This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3318242

Skeleton-based Gesture Recognition with Learnable Paths and Signature Features

Jiale Cheng†, Dongzi Shi†, Chenyang Li, Yu Li, Hao Ni, Lianwen Jin, and Xin Zhang*

***Abstract*—For the skeleton-based gesture recognition, graph convolutional networks (GCNs) have achieved remarkable performance since the human skeleton is a natural graph. However, the biological structure might not be the crucial one for motion analysis. Also, spatial differential information like joint distance and angle between bones may be overlooked during the graph convolution. In this paper, we focus on obtaining meaningful joint groups and extracting their discriminative features by the path signature (PS) theory. Firstly, to characterize the constraints and dependencies of various joints, we propose three types of paths, i.e., spatial, temporal, and learnable path. Especially, a learnable path generation mechanism can group joints together that are not directly connected or far away, according to their kinematic characteristic. Secondly, to obtain informative and compact features, a deep integration of PS with few parameters are introduced. All the computational process is packed into two modules, i.e., spatial-temporal path signature module (ST-PSM) and learnable path signature module (L-PSM) for the convenience of utilization. They are plug-and-play modules available for any neural network like CNNs and GCNs to enhance the feature extraction ability. Extensive experiments have conducted on three mainstream datasets (ChaLearn 2013, ChaLearn 2016, and AUTSL). We achieved the state-of-the-art results with simpler framework and much smaller model size. By inserting our two modules into the several GCN-based networks, we can observe clear improvements demonstrating the great effectiveness of our proposed method.**

***Index Terms*—Gesture recognition, Graph convolutional network, Path signature features**

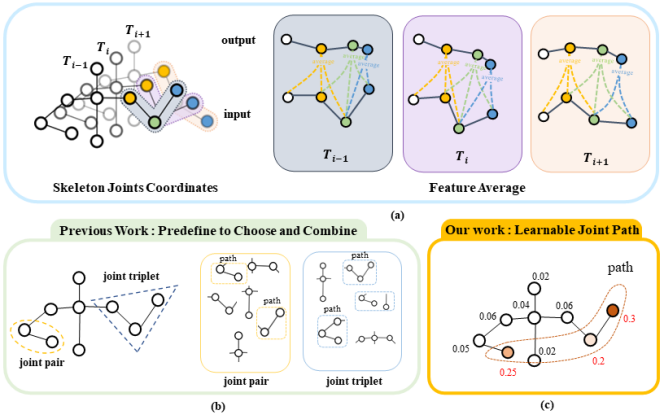

Fig. 1: The limits of GCNs and a comparison of previous work with our proposed Learnable Path Module.(a) On the left, the limited receptive field of GCNs is shown, which is constrained by the biological structure of the human body, as indicated in the figure. On the right, the disappearance of high-frequency features due to the smoothing effect of feature averaging is shown.(b) shows previous work has defined joint connections beyond natural body structures.(c) shows the Learnable Path is composed of the most relevant joints without any predefinition before.

I. Introduction

GESTURE recognition is an active research topic of computer vision in recent years for its wide range of applications, such as sign language translation [1] and human-computer interaction [2]. A vast of literature has been devoted to this field [3]–[5]. Compared with RGB-based methods [6]–[8], skeleton-based gesture recognition shows superior performance facing various challenges like noisy background, occlusion, camera view and illumination variations [9]–[12].

The skeleton sequence can be reshaped into a pseudo-image or a series of joint coordinates where CNNs [13]–[15] or RNNs [16]–[18] are employed to predict the action label. One of the main drawbacks is that the inherent skeleton and joint relationships are ignored. Since the skeleton can be interpreted as a graph, the graph convolutional network (GCN) [19] was adopted to replace CNNs and RNNs for feature extraction. Most recently, GCN-based algorithms have reached excellent results and gone mainstream in this area [20]–[26].

Even though, it is still a key unsolved problem to generate discriminative representation for describing the spatial body structure and dynamic motion pattern. We divide this problem into two folds and analyze them separately. The first issue is to define and group a set of joints (or known as region of interest) which are closely connected semantically and kinetically. The GCN-based methods typically transform the joint sequences into a series of graphs and apply graph convolution onto the data. Hence, regions of interest are defined based on the biological skeleton structure. Each region of interest is the

Manuscript received September 27, 2022; revised May 16, 2023. This project is sponsored by the National Key R&D Program of China (2022YFB4500600), Guangdong Key Laboratory of Human Digital Twin Technology (2022B1212010004), the Fundamental Research Funds for the Central Universities (2022ZYGXZR104), Zhuhai Industry Core and Key Technology Research Project (2220004002350), the Engineering and Physical Sciences Research Council (EPSRC) (EP/S026347/1), the Alan Turing Institute under the EPSRC grant (EP/N510129/1).

J. Cheng, D. Shi, C. Li, Y. Li and L. Jin are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China (e-mail: jialecheng100@gmail.com; eedongzishi@mail.scut.edu.cn; lichenyang.scut@foxmail.com; lyu.scut@qq.com; eelwjin@scut.edu.cn;).

X. Zhang is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China, and also with Pazhou Lab, Guangzhou, 510330, China (email: eexinzhang@scut.edu.cn).

H. Ni is with the Department of Mathematic, University College London, Gower Street, London WC1E 6BT, UK (e-mail: h.ni@ucl.ac.uk).

*Corresponding Author

†Equal Contribution

one-hop neighborhood of a node on the graph [26], which is straightforward. As shown in the left part of Figure 1(a), for the joint of elbow, the region of interest for GCNs is its adjacent connected shoulder and hand joints. These joint connections defined by the physical body are intuitive spatial constraints, but may not be the crucial ones to distinguish gestures or actions. For instance, it is ideal to have an instant information transformation between the left hand and right hand joints to evaluate their relative movement for the recognition of the gesture *Clapping*, while it is not accessible in a single GCN layer. Recent studies [23], [24], [27], [28] have tried to solve this issue by enlarging the local neighborhood from 1 hop to k hops or directly learning non-local joint connectivity, but they are still limited in the GCN framework. As shown in Fig. 1, in previous path signature based works, [29] manually defined combinations of any two or three joints to break the limitation of region of interest being constrained by natural body connections. Building upon this work, [9] and [10] selected the predefined joint combinations and kept the more informative ones, such as the left and right hand joints.

The second issue is how to extract and aggregate the critical spatial and temporal differential features, within the region of interest to form discriminative representation for classification. For the GCN based methods, within each region of interest, the message passing process is conducted spatially among nodes by averaging the joint features in the local neighborhood where information is aggregated and local dependency is described. This process is essentially a Laplacian smoothing filter in the spectral domain [30]. Specifically, in the gesture recognition, when the graph convolution is applied on the 3-dimension joint coordinates, the centroids of the local joint sets are computed as shown in Fig. 1, which may incur the loss of high-frequency information of the skeleton structure, such as the curvature of wrists and elbows. On the other hand, several researches observed that the temporal differential features are beneficial for describing the motion properties in video analysis [31]–[33]. Consequently, it has become a universal solution to design a multi-stream network in advanced researches [20], [21], [23]–[25], where the raw joint data and their spatial or temporal differential feature are fed into different streams and fused at last. It is in essence a model ensemble strategy which may be too heavy to be implemented on mobile devices.

Instead of designing a multi-stream neural network to process different information, we regard data sequence as the trajectory of the joint in spatial and temporal dimensions. In order to obtain the abundant geometrical and analytical properties of the trajectory, we introduce the concept of path and its signature features which come from rough path theory, a branch of stochastic analysis [34]. Through flexible path definition and its signature features, we can aggregate the critical spatial and temporal information without the excessive number of parameters caused by the ensemble strategy.

In previous path signature based methods [9], [10], [29], path is defined in a hand-crafted way. For the skeleton gesture data, these path definitions cannot reveal the joint relationship during the movement. Therefore, we leverage the self-attention mechanism [35] to enlarge the model capability, adaptively select the most correlative joints regardless of the biological skeleton structure, and group them into path. We propose to generate and compare three kinds of paths to extract features in diverse domains, i.e., the spatial path, temporal path, and learnable path. Further, a compact deep signature feature extraction block is designed by inserting several sets of parameters to improve the efficiency and effectiveness of PS calculation. We integrate these paths generation and signature features extraction procedures into two plug-and-play blocks, the spatial-temporal path signature module (ST-PSM) and learnable path signature module (L-PSM). These two blocks can not only obtain compact deep signature features, but also be inserted into other neural networks like CNNs and GCNs to provide the discriminative deep path signature features. Extensive experiments are conducted on three mainstream datasets (ChaLearn 2013 [36], ChaLearn 2016 [37], and AUTSL [38]) to evaluate the effectiveness and flexibility of our method. Using our path signature based modules and a simple CNN framework, we achieve the state-of-the-art results on ChaLearn 2013 and ChaLearn 2016 with much smaller model size. We also test our path signature based modules on AUTSL by inserting them into existing GCN frameworks to verify their compensatory expressive power and a universal boost in performance is observed. By visualizing learnable paths, we clearly show that joints are grouped differently according to motion dynamics.

In general, we conclude our contributions in three aspects:

- we adopt the path signature theory as a substitute for deep learning integration strategy, which was introduced to differentiate gesture movement trajectories. Compared to previous integration strategies, path signature requires fewer parameters and can be better integrated with the deep learning.
- We are the first to design learnable semantic-aware paths for the path signature based method. We overcome the limitation of the GCNs receptive field caused by the definition of biological skeleton domain and propose two plug-and-play feature extraction modules for skeleton-based gesture recognition.
- Extensive experiments conducted on three datasets to validate the effectiveness of the proposed plug-and-play path signature modules and demonstrate the potential of integrating path signatures with deep learning.

II. Related Work

A. Skeleton-Based Action Recognition

Recently, advanced approaches [20]–[26] have witnessed significant performance boost, with constructing spatial-temporal graphs and modeling the spatial correlation with GCNs directly, indicating the effectiveness of the inherent joint connectivity for action recognition. However, with a fixed adjacent matrix, the natural biological connection, the information of joints is restricted to flow into predefined directions, which greatly limits the expressive power of neural networks. Therefore, skeleton-based action recognition with GCNs is trying to be dynamic so that the network can learn more implicit information beyond predefined restrictions. Studies have been proposed to adjust the adjacent matrix based on the

data characteristics. In [24], researchers proposed to provide supplementary information of adjacent matrix by applying a set of learnable matrix masks. Similarly, [28], [27] proposed to add extra edges to the skeleton graph by MLP, LSTM [39], encoder-decoder leveraging the similarity of joints in spatial coordinates and temporal dynamics respectively. Note that these above mentioned methods mainly focus on the different approaches for adjustment of the region of interest (ROI), as known as the adjacent matrix, and ignore the plausibility of the aggregation method within the ROI. Most recently, Transformer [40] was proposed and demonstrated satisfying performance in information aggregation and feature extraction. Specifically, aggregated important nodes in space-time domain respectively to focus on the unconnected parts in space. [41] divided the action into different stages, learning the key information of each stage and the information association of different stages to describe the action. In fact, recognizing an action may only need a few joints. Self-attention is a little too blind, which brings an uneconomical computational cost. We found that the differential information discarded by the computation of graph convolution may be the crucial component for skeleton-based gesture and action recognition. Thereby, in this paper, we use the path signature to extract the complementary information of GCN.

B. Path Signature Method

Path signature is an infinite graded sequence of statistics known to characterize data streams (paths). It originated from Chen's study [42] in the form of iterated integrals to solve the differential equation of smooth paths. Lyons first extended it from paths of bounded variation [43] to paths of finite p-variation for any $p \geq 1$ [34]. As the success of deep learning, there is an emerging research area, which combines path signature features with deep learning to tackle various applications, such as financial data analysis [44], hand-written character recognition [45], writer identification [46], infant congnitive scores prediction [47] [48], gesture recognition [9], [10], and skeleton based action recognition [29]. We found that skeleton sequence data described by joints spatial positions is sampled at a higher frequency, and therefore, using path signatures allowed us to skip some noise in the data, enabling the extraction of effective motion trajectory features to describe the action. In previous related works based on skeleton action recognition based on path signature, [9], [29] are pioneer works constructing paths in both spatial and temporal domains and obtaining encouraging results. However, their paths are either pre-defined or randomly selected and signature features have a high dimensionality, which is not flexible and computational expensive. Most recently, Kidger [49] proposed that path signature transformation can be integrated into the network as a layer, which opened a new era of combining of PS and neural networks. Thanks to the work of [50], a fast and convenient path signature calculation is provided with back-propagation. In this paper, we further explore on the path generation, plug-in PS module design and data-driven signature transformation learning.

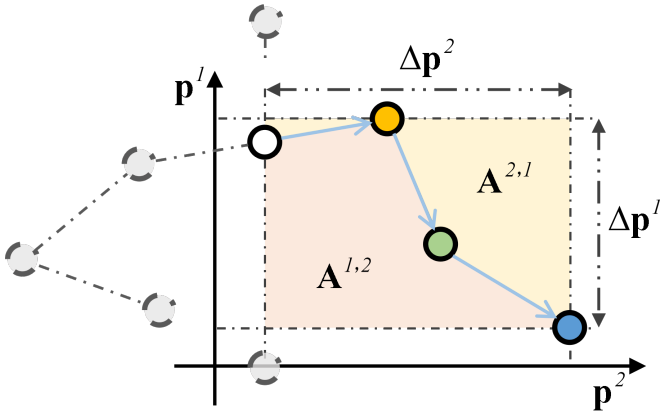

Fig. 2: The geometric illustration of the signature feature of path P on a 2-dimension plane. The path gets start at the white point when $t = a$ and stops at the blue point when $t = b$.

III. Path Signature Preliminaries

In this section, we will briefly introduce the calculation and properties of the signature features of paths. For interested readers, please refer to [51], [52] for more information.

Suppose $\mathbf{p} : [a, b] \to R^d$ is a d-dimensional path defined on the time interval $[a, b]$. Assuming t ∈ [a, b], and $\mathbf{p}_t \in R^d$ is a point on path $\mathbf{p}$. We can write $\mathbf{p}_t = \{p_t^1, p_t^2, \cdots, p_t^i, \cdots, p_t^d\}$, where p_t^i denotes the i^{th} coordinate of $\mathbf{p}_t$.

For a given path $\mathbf{p}$, we denote the 1^{st} order integral on the i^{th} coordinate as $Sig(\mathbf{p})_{a,b}^i$, i.e. $Sig(\mathbf{p})_{a,b}^i = \int_{a<t<b} dp_t^i$. Then the 1^{st} order integrals of $\mathbf{p}$ denoted by $S_1(\mathbf{p})_{a,b}$ is a collection of $Sig(\mathbf{p}_t)_{a,b}^i$ for all coordinates, i.e. $S_1(\mathbf{p})_{a,b} = \{Sig(\mathbf{p})^i\}_{i\in\{1,2,\cdots,d\}}$. For example, in Fig. 2, the 1^{st} fold iterated integral of path P is $S_1(P)_{a,b} = \{\Delta P^1, \Delta P^2\}$, where ΔP^i is the increment of the i^{th} coordinate of P.

Notably, $t_2 \to \int_{a<t_1<t_2} d\mathbf{p}_{t_1}$ is another d-dimensional valued path defined on the time interval $[a, b]$. Therefore, the 2^{nd} fold iterated integrals of $\mathbf{p}$ is the collection of the 2^{nd} order iterated integrals with all possible indices i_1, i_2, denoted by $Sig(\mathbf{p})_{a,b}^{i_1,i_2} = \int_{a<t_2<b} \int_{a<t_1<t_2} dp_{t_1}^{i_1} dp_{t_2}^{i_2}$. Specifically, in the example of Fig. 2, its 2^{nd} fold iterated integral contains four components:

$$\begin{aligned} S_2(P)_{a,b} &= \{Sig(P)_{a,b}^{1,1}, Sig(P)_{a,b}^{1,2}, Sig(P)_{a,b}^{2,1}, Sig(P)_{a,b}^{2,2}\} \\ &= \left\{\frac{(\Delta P^1)^2}{2!}, A^{1,2}, A^{2,1}, \frac{(\Delta P^2)^2}{2!}\right\}. \end{aligned}$$

where $A^{1,2}$, $A^{2,1}$ equals to the area enclosed by the twisting curve and different coordinates separately in this scheme.

Similarly, the k^{th} fold iterated integrals is defined as a collection of k^{th} order iterated integrals, i.e.

$$S_k(\mathbf{p})_{a,b} = \left\{ \int_{a<t_1<\cdots t_k<b} dp_{t_1}^{i_1} \cdots dp_{t_k}^{i_k} \right\}_{\substack{i_1,\cdots,i_k \\ \in\{1,\cdots,d\}}}. \quad (1)$$

The dimension of $S_k(\mathbf{p})_{a,b}$ is d^k. In general, the signature of a path is an infinite graded series containing its all folds

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3318242

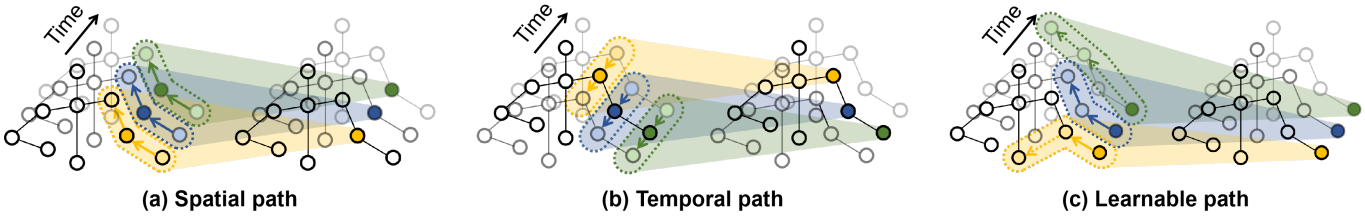

Fig. 3: Path overview. The dynamic skeleton sequence is taken as the input. Three different types of paths, i.e., the spatial, temporal, and learnable ones, are constructed for each joint. We extract their compact deep signature feature which is illustrated by the shadow surrounding the path. Specifically, we used dash lines in **(c)** to denote the learned connections which do not exist in the biological skeleton structure.

iterated integrals. In practice, one needs to truncate $S(\mathbf{p})$ up to the finite degree k to retain the first k fold iterated integrals

$$S(\mathbf{p})_{a,b}[0:k] = \{S_0(\mathbf{p})_{a,b}, S_1(\mathbf{p})_{a,b}, \cdots, S_k(\mathbf{p})_{a,b}\}, \quad (2)$$

where $S_0(\mathbf{p})_{a,b}$ is a scalar 1 by convention. Thus, the dimension of the signature of $\mathbf{p}$ up to degree k is $n_{PS} = \frac{d^{k+1}-1}{d-1}$.

The empirical time series is often discrete and hence, one may embed it into a piece-wise linear path interpolation to compute its signature. When $\mathbf{p}$ is a linear path, its k^{th} order iterated integral can be computed as:

$$Sig(\mathbf{p})^{i_1,i_2,\cdots,i_k}_{a,b} = \frac{1}{k!}\prod_{j=1}^{k}(p_b^{i_j} - p_a^{i_j}). \quad (3)$$

Moreover, the signature of piece-wise linear path can be computed by equation (3) and the Chen's identity [42]. The signature of a path has many algebraic and analytic properties. In the scheme of skeleton-based gesture recognition, it is noteworthy that if we regard the biological connected joints as a piece-wise linear path as illustrated in Fig. 2, its first fold iterated integral, i.e. $Sig(\mathbf{p})^{i}_{a,b}$, evaluates the direction and length of bones whose effectiveness has been proved by a variety of works [23]–[25]. The second fold iterated integral denotes for the area under or above the skeleton connections. It provides the information on Levy area, which is the area enclosed by the curve; for instance, the Levy area of P in Fig. 2 equals to $A_L = A^{1,2} - A^{2,1}$. Intuitively, the sign of A_L represents for the direction and concavity of a 4-node path. Additionally, the signature of a path uniquely determines the path up to time re-parameterization [34]. It is a universal feature implying that any continuous functions on the unparameterized path can be well approximated by the linear functional on the signature locally [53].

IV. Proposed Approach

The human body structure can be naturally represented by a graph, where nodes correspond to human joints and edges represent the connections between joints. However, as mentioned above, GCN is subject to two limitations. In the following, we firstly introduce path signature to enhance the expressive power of GCNs by designing paths and then extract key features from paths. Specifically, we propose two plug-and-play path signature modules. Compared to GCN, the path signature modules are more sensitive to variations in skeleton movements. Additionally, we propose a simple backbone to validate the effectiveness of the path signature modules.

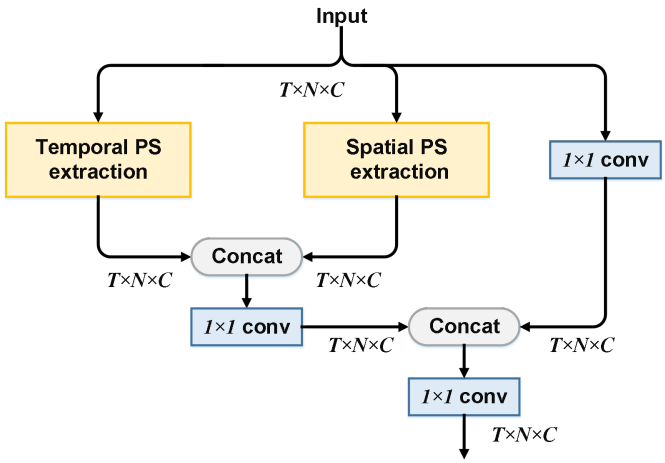

Fig. 4: The illustration of the spatial and temporal path signature module (ST-PSM), where "Concat" denotes for concatenation. The spatial and temporal PS features are linear combined with a 1×1 convolutional layer. The original information is inducted by a lateral connection.

A. Path Generation

Equation (4) is the basic function to describe the definition of path for every joint in $\mathcal{X}$, where $\mathcal{X} \in R^{T\times N\times C}$ is used to denote the skeleton-based sequence. T is the number of frames. N is the number of joints in each frame. C is the feature dimensionality, which is 3 or 2 for the input of neural network based on the spatial position of every joint.

$$\mathbf{p}_{n,t} = \omega_1 \cdot G(\mathbf{x}_{n,t}, \mathcal{X}). \quad (4)$$

The n^{th} joint in the t^{th} frame and its output of equation (4) are denoted as $\mathbf{x}_{n,t} \in R^C$ and $\mathbf{p}_{n,t} \in R^{L\times C}$ respectively. $\mathbf{p}_{n,t} = \{p_{1,t}, p_{2,t}, \cdots, p_{n,t}, \cdots, p_{L,t} \mid p_{n,t} \in R^C\}$ is a L-length path, while each $p_{n,t}$ is a joint selected from $\mathcal{X}$. We use function $G(\cdot, \mathcal{X})$ to denote the way we generate paths and a set of learnable parameters ω_1 to embed the path into a suitable feature space for the following deep signature feature extraction. Three kinds of path construction $G(\cdot) \in \{G_s(\cdot), G_t(\cdot), G_l(\cdot)\}$ will be discussed later, corresponding

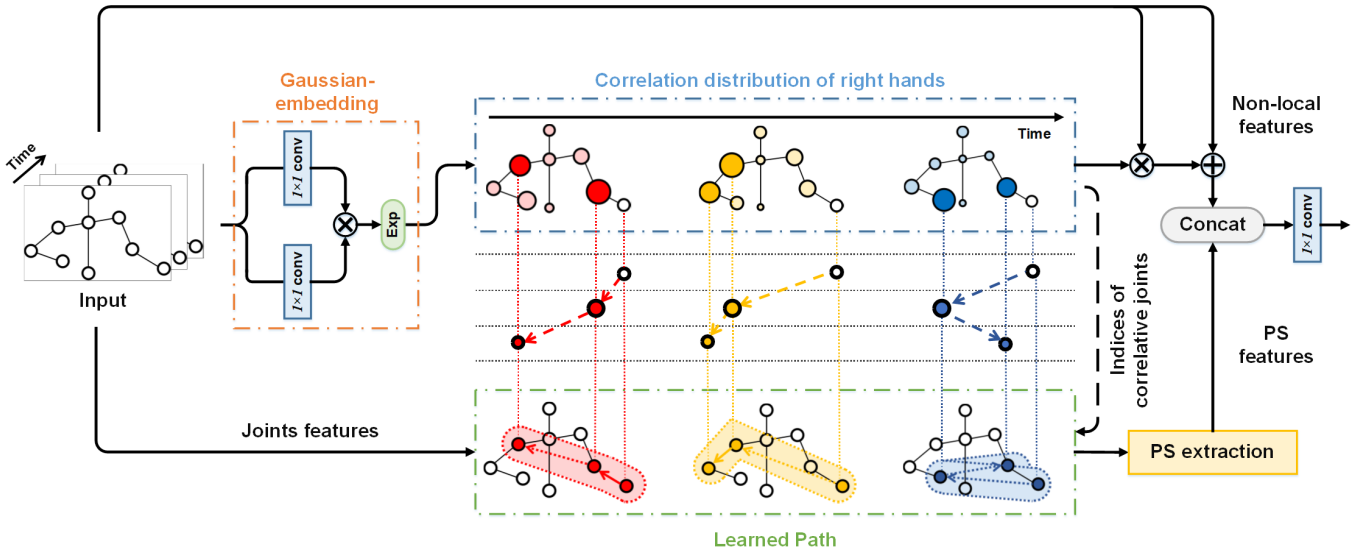

Fig. 5: The illustration of learnable path generation. A Gaussian-embedding function is applied for evaluating the correlative scores among pairs of joints in each frame. Paths are constructed with the root joints and its most correlative joints. Their compact deep signature features are then extracted and fused with the non-local features.

to spatial, temporal, and learnable paths respectively. It is noteworthy that various dependencies of the skeleton structure can be exploited as $G(\cdot, \mathcal{X})$, the way we define paths and changes.

B. Spatial and Temporal Path Signature Module (ST-PSM)

Effectiveness of spatial and temporal information has been proved by previous works [9], [29]. Accordingly, we define one path following the body structure and one path along the temporal sequence. Firstly, considering the biological body structure, we define the **spatial path**, $G_s(\cdot)$, as the physically neighboring joints. For example, as shown in Fig. 3(a), given the left elbow joint of frame t, $\mathbf{x}_{n,t}$, to generate a path of length $L = 3$, we select adjacent joints along the body structure, which are the left hand and the left shoulder joints. Therefore, the output of $G_s(\cdot)$ can be expressed as:

$$G_s(\mathbf{x}_{n,t}, \mathcal{X}) = \{\mathcal{N}^k(\mathbf{x}_{n,t}), \cdots, \mathcal{N}^1(\mathbf{x}_{n,t}), \mathbf{x}_{n,t}, \mathcal{N}^1(\mathbf{x}_{n,t}), \cdots, \mathcal{N}^k(\mathbf{x}_{n,t})\}, \tag{5}$$

where $k = \lfloor (L-1)/2 \rfloor$, $\mathcal{N}^k(\mathbf{x}_{n,t})$ denotes one of joints in the k-hop neighborhood of $\mathbf{x}_{n,t}$ in the t^{th} frame. Secondly, the moving trajectories of joints can be interpreted as a powerful indicator of dynamic patterns. Hence, we intuitively construct the **temporal path**, $G_t(\cdot)$, as the motion trajectory of each joint. Fig. 3(b) has shown three temporal paths regarding left shoulder, elbow and hand joint. For each given joint of the t^{th} frame $\mathbf{x}_{n,t}$, a time path of length L can be defined along the time dimension, where $k = \lfloor (L-1)/2 \rfloor$:

$$G_t(\mathbf{x}_{n,t}, \mathcal{X}) = \{\mathbf{x}_{n,t-k}, \cdots, \mathbf{x}_{n,t-1}, \mathbf{x}_{n,t}, \mathbf{x}_{n,t+1}, \cdots, \mathbf{x}_{n,t+k}\}. \tag{6}$$

By defining the spatial and temporal path, we can compute their compact deep signature features as defined in equation (3). Features from these two paths are further concatenated and fused together as shown in Fig. 4, which is denoted as the spatial and temporal path signature module (ST-PSM). Additionally, we adopt a one-layer linear transformation to obtain the original information beyond the path signature features. Finally, the signature and original features are concatenated together for further process.

C. Learnable Path Signature Module (L-PSM)

The biological connections among joints are not necessarily the most crucial ones for gesture recognition. The predefined paths may be restricted by the ambiguity of pairwise correlation among joints. It is vital to connect joints which are tightly connected kinetically. Thus, we design a self-attention based path generation mechanism $\mathbf{p}_{n,t} = G_l(\theta, \mathbf{x}_{n,t}, \mathcal{X})$ to automatically select the most correlative joints for the input root node $\mathbf{x}_{n,t}$ and construct paths in a data-driven manner as illustrated in Fig. 3(c).

The key problem is how to evaluate the joint correlation. In this work, we purpose to define paths with joints which are semantically similar. Inspired by [35], the learnable adjacent matrix A^t is thus introduced to evaluate the similarity between joints in the t^{th} frame in a data-driven manner. As shown in Fig. 5, the correlative scores between different joints with the right hand are illustrated.

It's a natural choice to calculate dot-product similarity to evaluate the correlation. Dot product measures the similarity between two vectors based on the angle between them, and the result is a scalar value. However, this can be limited, especially when dealing with high-dimensional data. For the construction of A^t, we empirically embed the root joint $\mathbf{x}_{n,t}$ and the other nodes $\mathbf{x}_{j,t}$ in the t^{th} frame differently with two sets of 1×1 kernels and apply the inner product with Gaussian function as the correlation metric:

$$A^t_{n,j} = \mathbf{e}^{(\theta_1 \cdot \mathbf{x}_{n,t})^T (\theta_2 \cdot \mathbf{x}_{j,t})}, \tag{7}$$

where $j \in \{1, 2, \cdots, N\}$. θ_1, θ_2 denote for the weights of linear transformation. The whole computational process in equation (7) is called as the Gaussian-embedding [54] as shown in Fig. 5. Gaussian embedding allows us to represent the similarity between two vectors as a continuous value, rather than a discrete value as in the case of dot product. On the other hand, Gaussian embedding represents vectors as probability distributions, and the similarity between them is calculated based on the overlap between their probability distributions. This approach provides a more nuanced measure of similarity and allows us to capture more complex relationships between vectors. Therefore, Gaussian embedding can be more effective in capturing subtle differences between gestures and provide a more accurate correlation score for hand gesture recognition tasks. The learnable parameters within $G_l(\theta, \cdot, \mathcal{X})$ is exactly $\theta = \{\theta_1, \theta_2\}$. In this example, a 3-node path $\mathbf{p}_{n,t}$ is constructed with the right hand serving as the root node $\mathbf{x}_{n,t}$. Its most correlative joints, i.e., the right elbow and left shoulder is then picked out and listed in $\mathbf{p}_{n,t}$ in the order of the correlative scores as follows:

$$\begin{aligned} &\mathbf{p}_{n,t} = \{\mathbf{x}_{n,t}, \mathbf{x}_{j_{(1)},t}, \mathbf{x}_{j_{(2)},t}, \cdots, \mathbf{x}_{j_{(L-1)},t}\}, \text{ where} \\ &j_{(\cdot)} = \mathrm{argsort}_{\mathrm{j}}\{\mathrm{A}^{\mathrm{t}}_{\mathrm{n,j}} / \sum_{\mathrm{j}} \mathrm{A}^{\mathrm{t}}_{\mathrm{n,j}} | \mathrm{j} = 1, 2, \cdots, \mathrm{N} \text{ and } \mathrm{j} \neq \mathrm{i}\}. \end{aligned} \tag{8}$$

With equation (7) and (8), we can adaptively select and group the most correlative joints in the semantic domain, which further improves the effectiveness and flexibility of path signature based methods on the non-sequential data. The network updates only the corresponding neurons in each iteration.

Additionally, based on the observation that the relationship between pairs of joints is a process of dynamic evolution with their instant moving patterns, the learnable adjacent matrix A^t is generated differently for each frame. Thereby, the path we construct for the same joint also changes dynamically. We name the whole process in Fig. 5 as the learnable path signature module (L-PSM).

D. Compact Deep Signature Feature Extraction

Given a non-retraceable path $\mathbf{p}_{i,t}$, its compact deep signature features can be computed as follows:

$$\mathbf{s}_{n,t} = \mathrm{ReLU}(\omega_2 \cdot \mathrm{S}(\mathbf{p}_{\mathrm{n,t}})[0 : \mathrm{k}]). \tag{9}$$

Function $S(\cdot)$ denotes for the mathematical calculation for path signature corresponding to equation (2) and (3). The original signature features have high dimensionality, which motivates us to process data along the path with a set of learnable parameters ω_2 for removing redundancy and fusing the multi-fold iterated integrals. As stated in section III, in practice, we often truncate the path signature up to a certain degree to avoid an overlarge dimension. To reduce the dimensionality of features and avoid redundancy, we set k=2. Therefore,$S(\cdot)$ can be represented as:

$$\begin{aligned} &S(\mathbf{p}_{n,t})[0:2] = \{S_0(\mathbf{p}_{n,t}), S_1(\mathbf{p}_{n,t}), S_2(\mathbf{p}_{n,t})\}, \\ &S_k(\mathbf{p}_{n,t}) = \{Sig(\mathbf{p}_{n,t})^{\overbrace{1, \cdots, 1}^{k}}, Sig(\mathbf{p}_{n,t})^{\overbrace{1, \cdots, k}^{k}}, \cdots, \\ &\qquad Sig(\mathbf{p}_{n,t})^{\overbrace{k, \cdots, k}^{k}}\}. \end{aligned} \tag{10}$$

The truncation operation may lead to the loss of some critical features hidden in the high-order iterated integrals. By introducing ω_2, the information of the high-order iterated integrals can be revealed by the non-linear combination of the low-order iterated integrals with the shuffle product property of the path signature method [55]. Further, it is noteworthy that $\mathrm{ReLU}(\cdot)$ is used in equation (9) for non-linearity, while the path signature $S(\mathbf{p}_{n,t})$ itself is a set of non-linear properties of data sequence, which explains the absence of the non-linear function in equation (4).

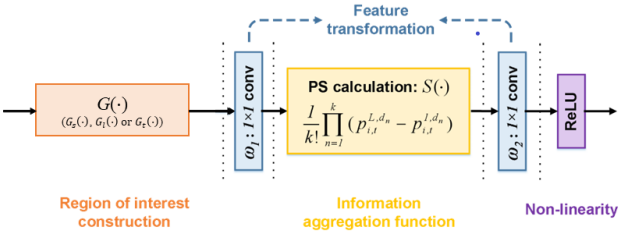

Fig. 6: The illustration of the process corresponding to equation (12). One of three $G(\cdot)$ is used for an implementation of PS feature extraction. Notably, we use "$\cdots$" to separate different components.

E. Comparison with GCN

On the graph-structure input $\mathcal{X}$ and its adjacent matrix A, the layer-wise update function for graph convolution network at layer n can be defined as:

$$\mathcal{X}^{n+1} = \sigma(\underbrace{\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\mathcal{X}^n}_{\text{Aggregation}} W^n), \tag{11}$$

where $\tilde{A} = A + I$, I is the identity matrix introduced for maintaining local dependency, $\tilde{D}$ is the degree matrix of $\tilde{A}$ introduced for normalization. We can roughly divide the equation (11) into three components, i.e., the information aggregation mechanism, feature transformation function and non-linearity as shown in Fig. 6. Specifically, GCN aggregates information from the 1-hop neighborhood by matrix muplication between $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ and $\mathcal{X}$, which is essentially the averaging computation. Further, the learnable matrix W and sigmoid function σ are introduced respectively for feature embedding and inducting non-linearity.

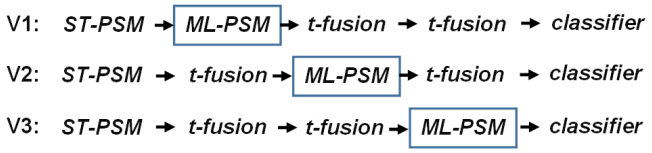

Fig. 7: We build three variants of a shallow network by inserting the L-PSM at different places. The *t-fusion* layer is the combination of one temporal convolution layer and one temporal pooling layer. We abbreviate the *multi-head L-PSM* as *ML-PSM* for convenience.

Similarly, we summarize the computation of path signature based modules as below by integrating equation (4) and (9):

$$\mathbf{s}_{n,t} = PS(\omega_1, \omega_2; \mathbf{x}_{n,t}, \mathcal{X}) = \mathrm{ReLU}(\omega_2 \cdot \underbrace{S(\omega_1 \cdot G(\mathbf{x}_{n,t}, \mathcal{X}))}_{\text{Aggregation}}), \tag{12}$$

where we use function $S(\cdot)$ for information aggregation, ω_1 and ω_2 for feature transformation. Both $\mathrm{ReLU}(\cdot)$ and path signature calculation are non-linear functions. Compared to GCN, which aggregates information within the 1-hop neighborhood, our module can go beyond the fixed biological connections and combine joints which are most semantically relevant to the action when aggregating information. This is because we have designed multiple ways $G_s(\cdot)$, $G_l(\cdot)$ and $G_t(\cdot)$ to generate paths.

F. Design of the Whole Framework

Remarkably, both ST-PSM and L-PSM are plug-and-play modules which do not change the data structure of input. Hence, it is feasible to insert our path signature based modules into existing methods.

First, we try to build a shallow network to test the performance of our PS based modules without the graph convolutional layers. In the task of skeleton based gesture recognition, the number of frames in each video clip is far more larger than the number of joints. Hence, we construct a simple baseline with temporal convolution and pooling layers to enlarge the temporal receptive field as shown in Fig. 7. The ST-PSM is applied on the top of the network for extracting features with abundant physical contextual information, while the L-PSM is inserted into the backbone at different places to adaptively generate diverse paths for describing the skeleton structures and testing its sensitivity against the temporal dependency. A multi-head L-PSM is utilized to capture different moving patterns of the gesture skeleton.

On the other hand, we try to combine our path signature module and the advanced GCNs to validate our assumptions above. The basic strategy is to apply our PS based modules on the GCN frameworks to provide the differential properties, which is similar to the model ensemble strategy in [20], [21], [23]–[25], but with much fewer parameters.

V. Experiments

A. Datasets

ChaLearn 2013. ChaLearn 2013 multimodel gesture dataset [36], which provides RGB, depth, foreground segmentation and skeleton data, contains 20 Italian gestures performed by 27 different persons. Each sequence lasts 1-2 minutes and includes 8-20 gesture instances. This dataset is split into training, validation and testing sets, containing 6850, 3454 and 3579 samples respectively. We only use skeleton data for gesture recognition [56].

ChaLearn 2016. ChaLearn 2016 dataset [37] is the largest public gesture recognition dataset currently. The whole dataset contains 47933 gesture samples and 249 types of gestures collected by 21 volunteers. It consists of two parts, the Isolated Gesture Dataset (IsoGD) and the Continuous Gesture Dataset (ConGD). We only use the IsoGD part in our experiments. ChaLearn 2016 only provides RGB and depth image sequences, so we use Openpose [57] to estimate the skeleton joints as [58] did.

AUTSL. AUTSL dataset [38] is a large scale multi-modal Turkish sign language dataset providing 38336 video clips collected from 43 signers with 20 backgrounds and 226 different Turkish sign actions. Each sample contains multiple modalities such as color image (RGB), depth and skeleton. Following [59], we use a pretrained HRNet [60] pose estimator provided by MMPose [61] to estimate the 133-point whole-body keypoints from the RGB videos and preprocess the skeleton-based data.

B. Implementation Details

In terms of training parameters, SGD with momentum is used as the optimizer. The learning rate is updated between 1e-5 to 1e-2 with a step of 1060. The weight decay coefficient is 1e-5, and the batch size is 64. The network is implemented in the Pytorch framework and trained on a GeForce GTX 1080 GPU. An open-source package, signatory[1], is used to provide efficient path signature computation with backpropagation.

We conducted experiments on different datasets to test the effectiveness of our proposed backbone mentioned in section IV F and compared it with other methods. We also inserted our modules into existing GCNs. Specifically, we first pass the skeleton sequence through the ST-PSM and L-PSM, and after feature fusion, the output channel is 64. Then, we set the input channel of the GCNs to 64 when inserting our module into existing GCNs, and we mostly use the default parameters in the source code except for changing the input channel.

C. Ablation Study

1) Evaluation on the configuration of L-PSM: As shown in Fig. 5, we apply the self-attention mechanism to calculate the correlative scores for each pair of joints and then we select the most correlated joints adaptively to form a path. Therefore, the correlation metric equation (7) plays an important role in L-PSM, which motivates to compare different correlation metrics in Table I.

Besides, in the literature, researchers found that the path length in path signature theory and number of heads in multi-head attention are two sensitive hyper-parameters which have a huge impact on the model performance [29], [35]. Hence, we compare the performance of L-PSM with different hyper-parameter settings in Fig. 8(a) and Fig. 8(b). We follow the

Evaluation Metrics	Acc (%)
Dot product	92.46
Gaussian	90.24
Gaussian-embedding	**92.77**
Concatenation	92.57

TABLE I: Evaluation of correlation metric.

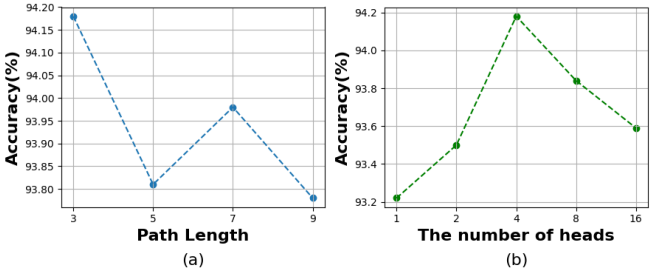

Fig. 8: Evaluations of the (a) path length and (b) the number of nodes. Paths composed of different nodes exhibit distinct connection modes, which correspond to different ranges of gesture changes. Each head focuses on distinct patterns and variations within the gestures, influencing the composition of paths.

Models	Params	Acc (%)
V1	1.27M	93.76
V2	**1.27M**	**94.18**
V3	2.51M	93.36
w/o ω_1 & ω_2	6.00M	94.09
w/o ST-PSM	1.26M (↓ 0.01M)	92.60 (↓ 1.58)
w/o L-PSM	1.24M (↓ 0.03M)	93.28 (↓ 0.90)

TABLE II: Evaluation of the components of the path signature based modules on ChaLearn 2013. The increase and decrease compared to V2 is shown.

TABLE III: Evaluation of pluggable modules ST-PSM and L-PSM on ChaLearn 2013 dataset.

Methods	Params(M)	infer(ms)	Acc (%)
ST-GCN [26]	3.07	3.59	93.11
–with ST-PSM	3.09	6.71	93.56
–with L-PSM	**3.10**(↑ 0.03)	22.72	**93.64**(↑ 0.53)
–with ST-PSM and L-PSM	3.11	26.02	93.59
AS-GCN [24]	6.88	21.46	92.66
–with ST-PSM	6.91	24.85	93.59
–with L-PSM	6.97	47.48	94.18
–with ST-PSM and L-PSM	**6.98**(↑ 0.10)	49.52	**94.35** (↑ 1.69)
MS-G3D [23]	4.56	22.93	94.71
–with ST-PSM	4.57	26.64	94.46
–with L-PSM	4.67	49.46	95.11
–with ST-PSM and L-PSM	**4.68**(↑ 0.12)	50.94	**95.47** (↑ 0.76)

experiment setting in [54] for the comparison of evaluation metrics and empirically select the Gaussian-embedding function with its superior performance. Furthermore, it is observed that the L-PSM does not exhibit improved performance with longer paths and an increased number of attention heads. One potential explanation is attributed to the limited diversity within the ChaLearn 2013 dataset, which comprises only 20 distinct gestures accompanied by an 11-node skeleton. As a result, the model requires a relatively small number of nodes in the path and attention heads to accurately capture and characterize the motion patterns.

2) Evaluation of the internal components and position of path signature based modules: As stated in section IV-F, we try to insert the L-PSM into the backbone at different places to generate diverse paths for describing the skeleton structures and testing the sensitivity of L-PSM against the temporal dependency. After comparison, it can be observed that a medium length (7 of 39 frames, about 0.2× length of the video clip) temporal receptive field is preferred to capture the moving patterns of hands. The learned paths at different time points will be shown later. Further, to verify the effectiveness and efficiency of our modifications on the path signature features calculation and path signature modules, an ablation study is conducted in Table II. According to section III, the dimensionality of path signature grows exponentially with the dimensional of the path, which builds a great challenge to integrate the path signature theory and deep learning. Obviously, as shown in Table II, by introducing two sets of parameters, we greatly decrease the model size of the path signature based network (from 6.00M to 1.27M), while the model capability is remained. We also try to insert our modules in other advanced backbones in Table III. The best performance is achieved with our modules inserted. Clearly, the L-PSM and ST-PSM steadily improve the accuracy at a low cost.

D. Comparison against the State-of-the-art

Based on ST-PSM and L-PSM, we can construct a simple classification model as illustrated in Fig. 7. It only contains several basic modules, including ST-PSM, L-PSM, 1×3 convolution and pooling layers. It is a non-GCN-based framework, and does not require a complex network structure. In this section, we compare the proposed model with other advanced methods, including both the non-GCN-based ones and GCN-based ones, and test their performance on ChaLearn 2013 and ChaLearn 2016 datasets. The best model configuration according to the ablation study is applied.

In recent years, several methods have been proposed for gesture recognition and achieved good results on ChaLearn 2013 dataset. Some of them are mainly benefited from the powerful characterization ability of CNN and LSTM models, such as HiVideoDarwin [62], VideoDarwin [63], D-LSDA [64], CNN for Skeleton [65], Two-stream LSTM [56] and Multi-path CNN [10]. Among them, 3s_net_TTM [9] uses PS of predefined paths to improve the ability of feature extraction. However, it constructs the spatial path by handcraft and extracts the PS before feeding into the neural networks overlooking the possibility of a further integration between the PS theory and machine learning. Results of all these methods and our proposed method are shown in Table IV. Our proposed model gets superior results benefiting from the strong feature extraction capability of PS and the flexibility of learnable paths.

ChaLearn 2016 dataset contains different kinds of data, such as RGB, depth, optical flow and skeleton data. Some advances have validated the effectiveness of the fusion with multi-modalities [59], [69], while we concentrate on the skeleton

TABLE IV: Classification accuracy comparison against state-of-the-art methods on the ChaLearn 2013 dataset.

Methods	Params	infer(ms)	Acc (%)
HiVideoDarwin [62]	-	-	74.90
VideoDarwin [63]	-	-	75.30
D-LSDA [64]	-	-	76.80
CNN for Skeleton [65]	-	-	91.20
Two-stream LSTM [56]	-	-	91.70
3s_net_TTM [9]	-	-	92.80
Multi-path CNN [10]	-	-	93.13
STFFormer [41]	5.5M	8.42	92.77
ST-TR [40]	19.4M	75.89	93.50
Shift-GCN [21]	0.6M	13.96	90.86
AS-GCN [24]	6.9M	21.46	92.66
CTR-GCN [66]	1.4M	14.63	92.82
GCN-Logsig-RNN [67]	13.0M	-	92.86
ST-GCN [26]	3.1M	3.59	93.11
MS-G3D [23]	4.6M	50.93	94.71
Proposed-V2	**1.3M**	59.74	**94.18**

TABLE V: Classification accuracy comparison against state-of-the-art methods on the ChaLearn 2016 dataset.

Methods	Top-1(%)	Top-5(%)
SkeLSTM [58]	35.39	-
3s_net_TTM [9]	39.95	-
Multi-path CNN [10]	43.82	-
MS-G3D [23]	48.05	60.69
ST-GCN [26]	22.08	28.62
–with GPS [68]	23.26	29.42
–with ST-PSM	25.25	35.89
–with L-PSM	26.71	34.14
–with ST-PSM and L-PSM	**27.23**	35.32
Shift-GCN [21]	21.08	26.22
–with GPS [68]	21.35	26.36
–with ST-PSM	21.63	27.44
–with L-PSM	22.21	26.23
–with ST-PSM and L-PSM	**23.24**	26.72
Proposed-V2	**51.60**	**77.04**

data in this work. There are several skeleton based methods experiment on it, including SkeLSTM [58], 3s_net_TTM [9] and Multi-path CNN [10]. As shown in Table V, our proposed method also performs the best among these methods, showing clear improvement.

In the field of skeleton-based action recognition, several excellent GCN-based methods have been proposed, such as ST-GCN [26], AS-GCN [24], Shift-GCN [21] and MS-G3D [23]. These methods perform well in the skeleton-based action recognition. We tested these methods on both ChaLearn 2013 and ChaLearn 2016, and compared their performance with our proposed model. As shown in Table IV and Table V, our proposed method shows relatively good results with fewer parameters. In Table IV, Shift-GCN has the least parameters but works worst, MS-G3D is a little better than our proposed model with more than 3 times parameter than ours. Our method works the best considering both effectiveness and efficiency. In Table V, the accuracy of the GCN-based methods are inferior to the non-GCN-based models. It is mainly because the precision of joint position is affected by the drastic background and illumination changes, which has a huge impact on static hand gesture recognition. MS-G3D outperforms with its superior ability in extracting multi-scale spatial-temporal features, while our method performs slightly better with the combination of PS features and self-attention mechanism. Additionally, we compared Global Position Self-attention [68], with the modules we proposed. The results, shown in Table V, indicate that the ST-PSM and L-PSM captured local details of the actions and learned more discriminative representations.

E. Comparison against the Ensemble Strategy

As discussed before, PS focuses on multi-scale differential information within the dynamic skeleton structure, including the motion features (temporal differential) and bone features (spatial differential). In this part, ST-PSM and L-PSM are considered pluggable modules and inserted into the GCN-based methods to compare with the ensemble strategy. For each method, four streams are trained with different input, including joint stream (Js), joint motion stream (Ms), bone stream (Bs), and bone motion stream. We ensemble the output of these streams following the setting in [59]. Experiments are carried out on AUTSL dataset, and the results are shown in Table VI. It is noteworthy that our proposed path signature modules achieve comparable results against the two-stream ensemble with a slight increment on the model size indicating the great potential of the path signature modules.

TABLE VI: Classification accuracy comparison against the ensemble strategy on AUTSL. We abbreviate path signature modules as PSMs here for convenience.

Methods	Params	Top-1 (%)	Top-10 (%)
ST-GCN [26]			
–with Js	3.14M	94.25	99.28
–with Js & Bs	6.28M (2×)	94.75	99.21
–with Js & Ms	6.28M (2×)	94.45	**99.46**
–with all streams	12.56M (4×)	**95.06**	99.34
–with Js & PSMs	3.19M (1.02×)	94.93(↑ 0.68)	99.18
AS-GCN [24]			
–with Js	3.52M	94.84	99.18
–with Js & Bs	7.04M (2×)	94.95	99.28
–with Js & Ms	7.04M (2×)	95.11	**99.50**
–with all streams	14.08M (4×)	**95.29**	99.41
–with Js & PSMs	3.56M (1.01×)	95.04(↑ 0.20)	99.34
Shift-GCN [21]			
–with Js	0.74M	94.86	99.12
–with Js & Bs	1.48M (2×)	95.56	99.41
–with Js & Ms	1.48M (2×)	95.02	99.46
–with all streams	2.96M (4×)	**95.88**	**99.48**
–with Js & PSMs	0.80M (1.08×)	95.16(↑ 0.30)	99.25
SL-GCN [59]			
–with Js	3.84M	94.84	99.21
–with Js & Bs	7.68M (2×)	95.04	99.34
–with Js & Ms	7.68M (2×)	94.95	99.23
–with all streams	15.36M (4×)	**95.56**	**99.41**
–with Js & PSMs	3.89M (1.01×)	95.09(↑ 0.25)	99.30

F. Illustration of Learned Paths

Compared with the PS of predefined paths, our L-PSM module constructs learnable paths to capture features more flexibly. L-PSM is not limited by the natural biological structure of human body and can automatically find the path associated with the current action. We obtained the paths generated by the L-PSM module corresponding to equation (8), which are shown in Fig. 9. Compared with predefined paths, our learned paths change dynamically to adapt to gestures, and capture

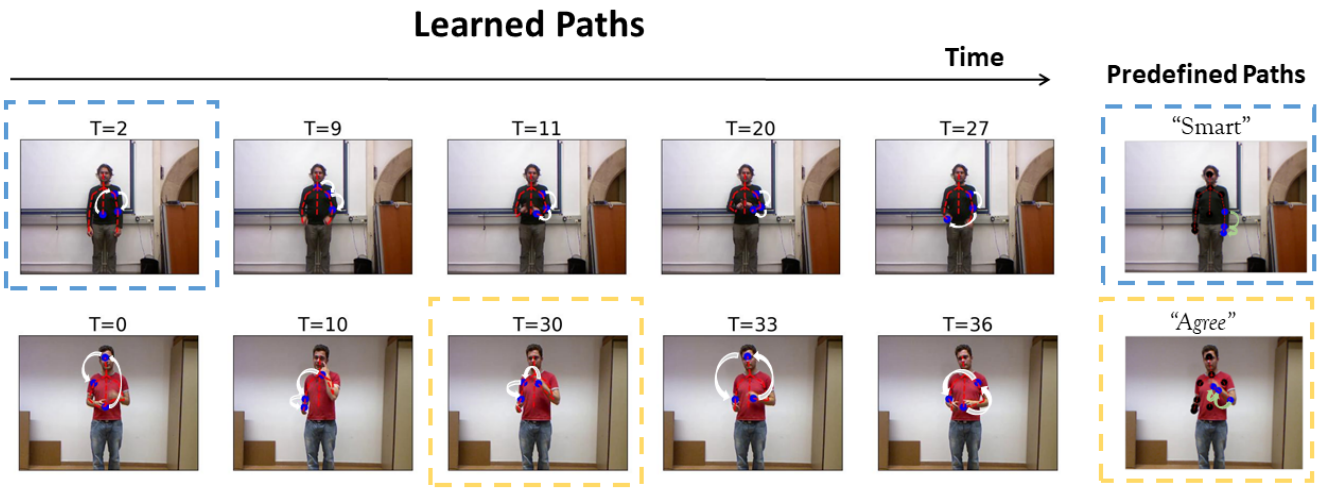

Fig. 9: Illustration of learned paths and predefined paths. We selected keyframes that reflect the process of action changes from a video comprising 39 frames, where dashed boxes represent the same frame. Our learnable path (white) dynamically changes to adapt to gestures, associating important related joint points in each frame, while predefined paths (green) can only associate adjacent joints on the body structure.

important characteristics between the relevant joints in each frame.

To be specific, the sub-figure above shows a "Smart" gesture, where the man extends his index fingers and folds them together. At the beginning, L-PSM can notice the relative relationship between the raised shoulder and elbow and the relatively static joints ($T = 2, 9$). When two hands are close, it focuses on the critical path composed of the whole arm ($T = 11, 20$), while when the hands are down, it focuses on the relationship between the hands and arms on both sides ($T = 27$). However, the predefined path only focuses on on one side of the shoulder, elbow and hand.

And in the sub-figure below, a man raises his left hand and point it towards his eyes then putting it down, meaning "Agree". L-PSM still pays attention to the relative motion of non adjacent joints on both sides ($T = 10, 30, 36$) and the motion of the hand relative to the head ($T = 33$), while the predefined path can only focus on one side.

VI. Conclusion

In this work, we introduce path signature, a powerful mathematical tool for extracting the differential features, which is essential in skeleton-based gesture recognition, and modify its calculation to make it more efficient to be applied in neural networks. Then, ST-PSM is designed to calculate spatial and temporal path characteristics for the dynamic skeleton structure. To address the limitation of predefined paths in ST-PSM, we proposed L-PSM, which automatically generates paths through the self-attention mechanism. It can construct different paths according to different inputs so that more meaningful features can be acquired. Base on ST-PSM and L-PSM, we construct a classification model that performs better than other non-GCN-based models, and it works the best among GCN-based models considering both effectiveness and efficiency. Besides, being considered as plug-and-play modules, ST-PSM and L-PSM are inserted into several methods to improve their performance at a low cost illustrating their great potential in future researches.

References

[1] W. Wu, C. Li, Z. Cheng, X. Zhang, and L. Jin, "Yolse: Egocentric fingertip detection from single rgb images," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 623–630, 2017.

[2] C. Mou and X. Zhang, "Attention based dual branches fingertip detection network and virtual key system," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2159–2165.

[3] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, pp. 976–990, 2010.

[4] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation and segmentation and recognition," *Comput. Vis. Image Underst.*, vol. 115, pp. 224–241, 2011.

[5] H. Zhang, Y. Zhang, B. Zhong, Q. Lei, L. Yang, J. Du, and D. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, p. 1005, 2019.

[6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," vol. abs/1406.2199, 2014.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2015.

[8] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4305–4314, 2015.

[9] C. Li, X. Zhang, L. Liao, L. Jin, and W. Yang, "Skeleton-based gesture recognition using several fully connected layers with path signature features and temporal transformer module," vol. 33, 2019, pp. 8585–8593.

[10] L. Liao, X. Zhang, and C. Li, "Multi-path convolutional neural network based on rectangular kernel with path signature features for gesture recognition," in *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2019, pp. 1–4.

[11] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 148–157, 2017.

[12] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," *ArXiv*, vol. abs/1603.07772, 2016.

[13] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaïd, "A new representation of skeleton sequences for 3d action recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4570–4579, 2017.

[14] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn," *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 601–604, 2017.

[15] H. Liu, J. Tu, and M. Liu, "Two-stream 3d convolutional neural network for skeleton-based action recognition," *ArXiv*, vol. abs/1705.08106, 2017.

[16] L. Li, W. Zheng, Z. Zhang, Y. Huang, and L. Wang, "Skeleton-based relational modeling for action recognition," *ArXiv*, vol. abs/1805.02556, 2018.

[17] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," in *ECCV*, vol. 40, no. 12, 2018, pp. 3007–3021.

[18] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: a large scale dataset for 3d human activity analysis," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010–1019, 2016.

[19] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ArXiv*, vol. abs/1609.02907, 2016.

[20] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling gcn with dropgraph module for skeleton-based action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 1–18.

[21] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and 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.

[22] M. Korban and X. Li, "Dggcn: a dynamic firected graph convolutional network for action recognition," in *European Conference on Computer Vision (ECCV)*, 2020.

[23] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and 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.

[24] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 026–12 035.

[25] S. Lei, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7904–7913, 2019.

[26] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *32nd AAAI Conference on Artificial Intelligence*, pp. 7444–7452, 2018.

[27] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3590–3598, 2019.

[28] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," 2019, pp. 8561–8568.

[29] W. Yang, T. Lyons, H. Ni, C. Schmid, and L. 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*. Springer, 2022, pp. 431–464.

[30] N. Hoang and T. Maehara, "Revisiting graph neural networks: all we have is low-pass filters," *ArXiv*, vol. abs/1905.09550, 2019.

[31] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[32] L. Wang, Z. Tong, B. Ji, and G. Wu, "Tdn: Temporal difference networks for efficient action recognition," *arXiv preprint arXiv:2012.10071*, 2020.

[33] Y. Zhao, Y. Xiong, and D. Lin, "Recognize actions by disentangling components of dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6566–6575.

[34] H. Boedihardjo, X. Geng, T. Lyons, and D. Yang, "The signature of a rough path: uniqueness," *ArXiv*, vol. abs/1406.7871, 2014.

[35] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

[36] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, I. Guyon, V. Athitsos, H. Escalante, L. Sigal, A. Argyros, C. Sminchisescu *et al.*, "Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 365–368.

[37] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 56–64.

[38] O. M. Sincan and H. Y. Keles, "Autsl: A large scale multi-modal turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181 340–181 355, 2020.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[40] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding*, vol. 208, p. 103219, 2021.

[41] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal tuples transformer for skeleton-based action recognition," *ArXiv*, vol. abs/2201.02849, 2022.

[42] K. tsai Chen, "Integration of paths-a faithful representation pf paths by noncommutative formal power series," *Transactions of the American Mathematical Society*, vol. 89, pp. 395–407, 1958.

[43] B. Hambly and T. Lyons, "Uniqueness for the signature of a path of bounded variation and the reduced path group," *Annals of Mathematics*, vol. 171, pp. 109–167, 2010.

[44] T. Lyons, H. Ni, and H. Oberhauser, "A feature set for streams and an application to high-frequency financial tick data," in *ACM International Conference Proceeding Series*, 2014, pp. 1–8.

[45] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1903–1917, 2018.

[46] S. Lai, Y. Zhu, and L. Jin, "Encoding pathlet and sift features with bagged vlad for historical writer identification," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3553–3566, 2020.

[47] X. Zhang, J. Cheng, H. Ni, C. Li, X. Xu, Z. Wu, L. Wang, W. Lin, D. Shen, and G. Li, "Infant cognitive scores prediction with multi-stream attention-based temporal path signature features," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*. Springer, 2020, pp. 134–144.

[48] J. Cheng, X. Zhang, H. Ni, C. Li, X. Xu, Z. Wu, L. Wang, W. Lin, and G. Li, "Path signature neural network of cortical features for prediction of infant cognitive scores," *IEEE transactions on medical imaging*, vol. 41, no. 7, pp. 1665–1676, 2022.

[49] P. Kidger, P. Bonnier, I. P. Arribas, C. Salvi, and T. Lyons, "Deep signature transforms," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1–11, 2019.

[50] P. Kidger and T. Lyons, "Signatory: differentiable computations of the signature and logsignature transforms, on both cpu and gpu," *arXiv:2001.00706*, 2020.

[51] B. Graham, "Sparse arrays of signatures for online character recognition," *ArXiv*, vol. abs/1308.0371, 2013.

[52] J. Reizenstein and B. Graham, "The iisignature library: efficient calculation of iterated-integral signatures and log signatures," *CoRR*, vol. abs/1802.08252, 2018.

[53] D. Levin, T. Lyons, and H. Ni, "Learning from the past and predicting the statistics for the future and learning an evolving system," *arXiv: Statistical Finance*, vol. abs/1712.02757, 2013.

[54] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[55] I. Chevyrev and A. Kormilitzin, "A primer on the signature method in machine learning," *ArXiv*, vol. abs/1603.03788, 2016.

[56] H. Wang and L. 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*, 2017, pp. 499–508.

[57] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[58] C. Lin, J. Wan, Y. Liang, and S. Z. Li, "Large-scale isolated gesture recognition using a refined fused model based on masked res-c3d network and skeleton lstm," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 52–58.

[59] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3413–3423.

[60] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[61] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[62] H. Wang, W. Wang, and L. Wang, "Hierarchical motion evolution for action recognition," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 574–578.

[63] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5378–5387.

[64] B. Su, X. Ding, H. Wang, and Y. Wu, "Discriminative dimensionality reduction for multi-dimensional sequences," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 1, pp. 77–91, 2017.

[65] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 579–583.

[66] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and 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. 13 359–13 368.

[67] S. Liao, T. Lyons, W. Yang, K. Schlegel, and H. Ni, "Logsig-rnn: a novel network for robust and efficient skeleton-based action recognition," in *BMVC 2021-32nd British Machine Vision Conference*, 2021.

[68] J. Kim and J. Lee, "Global positional self-attention for skeleton-based action recognition," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 3355–3361.

[69] P. Narayana, R. Beveridge, and B. A. Draper, "Gesture recognition: Focus on the hands," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5235–5244.