 This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/tcds.2025.3561422, IEEE Transactions on Cognitive and Developmental Systems

Self-Supervised Hyperbolic Spectro-Temporal Graph Convolution Network for Early 3D Behavior Prediction

***Abstract*—3D human behavior is a highly nonlinear spatio-temporal interaction process. Therefore, early behavior prediction is a challenging task, especially prediction with low observation rates in unsupervised mode. To this end, we propose a novel self-supervised early 3D behavior prediction framework that learns graph structures on hyperbolic manifold. Firstly, we employ the sequence construction of multi-dynamic key information to enlarge the key details of spatio-temporal behavior sequences, addressing the high redundancy between frames of spatio-temporal interaction. Secondly, for capturing dependencies among long-distance joints, we explore a unique graph Laplacian on hyperbolic manifold to perceive the subtle local difference within frames. Finally, we leverage the learned spatio-temporal features under different observation rates for progressive contrast, forming self-supervised signals. This facilitates the extraction of more discriminative global and local spatio-temporal information from early behavior sequences in unsupervised mode. Extensive experiments on three behavior datasets have demonstrated the superiority of our approach at low to medium observation rates.**

***Index Terms*—Self-supervised learning, Early 3D behavior prediction, Hyperbolic manifold, Spatio-temporal interaction**

I. Introduction

UNLIKE behavior recognition, which classifies behaviors after they are completed, the primary goal of early behavior prediction [1] is to classify behaviors during their execution. It brings unique challenges for early behavior prediction: semantic ambiguity and uncertainty resulting from the incompleteness of actions. Although existing early behavior prediction methods have achieved high accuracy by leveraging labeled data, there is still one problem that is unavoidable: with the rise of visual big data, the prior knowledge and cost required for labeling data are rapidly increasing, and existing methods may struggle to exploit their advantages without labeled samples. Therefore, early behavior prediction without sample labels has become one of the key issues that needs to be addressed.

Behavior understanding can be divided into two types based on data: video and skeleton. Video, as the most common form of behavior sequences, is intuitive and contains rich environmental information. However, it is susceptible to various factors such as lighting and occlusion during collection and processing. In contrast, skeleton consists of a series of 3D coordinates, which can abstractly represent the ”core state” of human behavior with minimal storage and is not affected by environmental noise. It makes 3D behavior understanding more popular. Graph Convolutional Networks (GCNs) have emerged as a powerful tool for analyzing and understanding 3D human data across a multitude of applications. Their ability to capture spatial hierarchies and relationships within data has significantly advanced the state-of-the-art in areas like action recognition, where the temporal evolution of human poses is crucial [2]. GCNs are widely used in various fields to process 3D human data, such as human skeleton data representation learning [3], [4], human pose estimation and action Recognition [5], [6], self-supervised learning [7] and anomaly detection [8], interactive behavior analysis and crowd Analysis [9], etc.

However, these GCN-based methods have shortcomings. Firstly, they mainly rely on spatial graph convolution. Adjacency matrix is divided into multiple sub-matrices to aggregate spatial information within the first-order neighborhood of joints [10], [11]. However, various partitioning strategies cannot be directly applied to adjacency matrices, resulting in inefficiency in capturing spatial information between long-distance joints. Secondly, even if a node is removed, its information can still be conveyed through neighbors, potentially causing the network to over-rely on noise or irrelevant data during training. Lastly, the construction of temporal graphs by linking joint nodes across frames and using one-dimensional temporal convolution to aggregate temporal information might be straightforward [12]–[14]. However, given the high similarity often present between adjacent frames in human behavior, this direct aggregation can be disrupted by redundant information, impacting the accurate interpretation. All of those lead to the model not performing well when applied to intelligent surveillance and situations with unclear skeletal distributions. In surveillance scenarios, the movement range and speed of objects are affected by parameters such as frame rate and resolution, which makes it difficult for the model to learn the details in the video. Unclear skeletal distributions can interfere with the flow of information in graph convolutions, ultimately leading to a significant decline in performance.

Therefore, we propose a novel self-supervised network for early 3D behavior prediction. Firstly, we introduce a novel self-supervised network. Our approach employs trajectory functions to model behavior sequences, incorporating physical concepts such as displacement, velocity, and acceleration through Taylor expansions of behavior functions. This method is beneficial for constructing multi-dynamic key information sequences while enhancing the essential details in spatio-temporal interactions. Secondly, we introduce hyperbolic spectral graph convolution, which leverages the hierarchical structure of hyperbolic space. This convolutional approach is adept at capturing dependencies between long-distance joints and utilizes the hyperbolic Laplacian to significantly boost the representational power of graph data. Furthermore, we

progressively compare spatio-temporal features learned under various observation rates, and leverage these contrastive results as self-supervised signals. This enables the network to deeply understand intrinsic patterns and regularities within early behavior sequences without supervision, thereby extracting more discriminative global and local spatio-temporal information.

The main contributions of this paper lie in three aspects:

- We define Multi-dynamic key information sequences by employing trajectory functions to model behavior sequences and conducting Taylor expansion with displacement, velocity and acceleration, which not only eliminates high redundancy between frames, but also enhances core motion details and dynamic changes.
- We propose an innovative graph Laplacian on hyperbolic manifold to model the dependencies among long-distance joints. Specifically, by defining Fourier transform and spectral analysis on hyperbolic manifold, we implement hyperbolic spectral graph convolution to perceive subtle local difference information within the behavior sequences.
- We design a self-supervised early 3D behavior prediction framework to leverage graph structures on hyperbolic manifold. This framework generates self-supervised signals by progressively contrasting spatio-temporal features at different observation rates, revealing the hidden hierarchical structure and dynamic changes in behavior sequences.

II. Related Work

A. Early 3D Behavior Prediction

Compared to early behavior prediction from 2D videos [15]–[18], early 3D behavior prediction has gained the attention of researchers in recent years. Ke et al. [19] employed adversarial learning to minimize the difference between early sequences and complete sequences, thereby obtaining potential global information in early sequences. Weng et al. [20] leveraged the diversity information from different negative categories and employing category exclusion approach. Starting from early similar segments of different behaviors, Li et al. [21] designed a Hard Instance-Interference Class (HIIC) bank to record these similar segments and corresponding error categories, enabling the model to perceive details. To solve the same problem, Wang et al. [22] introduced a guided metric learning module for extracting category discrimination from initial sequences. This module minimizes within-class distances using a full-length guidance approach and maximizes between-class differences across varying observation rates. In order to fully leverage the information of early sequences, Chen et al. [23] developed a generative model to utilize early sequence data, enhancing motion tendency analysis. They also regulated the generation process recurrently to emphasize behavioral discriminative cues.

Li et al. [24] designed an adaptive graph convolution network with adversarial learning, which applied adversarial learning to make the features of early sequences as similar as possible to the features of complete sequences, thereby learning the potential global information in early sequences. Wang et al. [25] focused on the uncertainty and diversity of future sequences and proposed a diversified early action recognition network that is capable of outputting multiple reasonable action classes for each early sequence. Foo et al. [26] further considered the method of [21] and argued that not all samples should be used to train the network parameters. Therefore, they designed an Expert Retrieval and Assembly (ERA) module to generate a set of experts most specialized at using discriminative subtle differences, to distinguish samples with high similarity. Liu et al. [27] aimed at the problem of lack of discriminative information in the early stages of action sequences. They designed a graph convolution network suitable for 3D skeleton sequences, and developed an early attention module to encourage the model to focus more on the early parts of the motion. Wang et al. [28] analyzed the shortcomings of graph convolution-based prediction methods in graph construction and message passing, and proposed a dynamic dense graph convolution network. The network constructs a dense graph with 4D adjacency modeling and employed the designed dynamic message passing method to dynamically transfer information between multi-scale spatio-temporal skeletal joints.

Those methods focus on Euclidean feature extraction for action sequences, which may not capture dynamic changes due to nonlinear interactions. We use hyperbolic manifolds to analyze joint dependencies and enhance action analysis accuracy and efficiency with differential geometry.

B. Self-supervised 3D Behavior Recognition

In recent years, unsupervised learning has made significant progress in the field of artificial intelligence. Early works primarily involve training models through pretext tasks. Zheng et al. [29] firstly explored an unsupervised representation learning method to capture the long-term global motion dynamics in 3D sequences and proposed a learning framework consisting of encoder, decoder, and discriminator. Similarly, Su et al. [30] also tried to use an encoder-decoder structure while they used a strategy of weakening decoder to strengthen the learning ability of encoder, enabling the encoder to learn better skeleton feature representation.

Most subsequent works have adopted the concept of contrastive learning. This approach involves guiding sample features to be similar to corresponding positive sample features and far away from negative sample features to generate self-supervised signals. Li et al. [31] exploited multi-view information for mining positive samples and pursuing cross-view consistency in unsupervised contrastive learning. Yang et al. [32] introduced the MG-AL framework for self-supervised learning, using motion cues to guide attention and reduce dependence on large datasets or augmentation.

Hua et al. [33] proposed an attention-based contrastive learning framework SkeAttnCLR to learn the relationship between local and global features. The framework integrates local similarity and global features for skeleton-based action representations, and an attention mechanism is employed to highlight local salient features, thereby enhancing 3D action representation. Jin et al. [34] designed a self-supervised spatio-

temporal representation learning network, SSRL, to mine long-range semantic information with two inference tasks. The temporal inference task learns the temporal persistence through temporally incomplete sequences, while the spatial inference task learns the spatially coordinated nature through spatially partially sequences. Pang et al. [35] suggested that while retaining the human skeleton structure, capturing long-distance joint connections can enhance 3D behavior recognition. However, this method fails to fully capture the complex spatial relationships in human behavior. In contrast, spectral graph convolution can effectively extract spatial features between joint points in the irregular graph structure. Therefore, we employ spectral graph convolution to address this issue from a different perspective in this work.

C. Self-Supervised Graph Neural Networks and Contrastive Learning

Using Graph Neural Networks for self-supervised learning leverages the inherent graph structure and positional information of skeleton data as a supervisory signal for learning [36], [37], thereby efficiently completing learning based on 3D skeleton data.

The application of contrastive learning on graph structures mainly involves these steps: using data augmentation to generate different views of skeleton data, encoding with Graph Neural Networks to obtain node feature vectors and designing contrastive loss functions, and designing effective contrastive learning supervision signals to assist model learning. Thus, research on Graph Neural Network contrastive learning in various fields centers around these three stages:

Focuses on different data augmentation methods. Guo et al. [38] proposed a method leveraging extreme augmentation for motion patterns, mixing various augmentations for rich unsupervised action info. Yet, training instability of extreme augmentation and risk of image/skeleton distortion remain challenges. Therefore, the subsequent work [39] employed a progressive augmentation strategy to create ordered positive pairs, ensuring representation consistency across views, buffering against semantic loss from direct strong augmentation.

Using a special Graph Neural Network encoder can effectively. Zeng et al. [40] proposed a hybrid network of Graph Neural Networks and MLPs for encoding, diversifying the distribution of negative samples, and using spatiotemporal occlusion to reduce information redundancy after data augmentation. Pang et al. [35] proposed a hybrid architecture of GCN and Transformer, trying to mix the spatial-temporal graph convolution stream and the spatial-temporal transformer stream in parallel.

Special unsupervised tasks can do a good job, Guo et al. [41] developed a graph representation supervision mechanism to enhance the intrinsic consistency between joint and skeletal information flows. Gao et al. [42] provided different spatiotemporal observation scenes and pulled them together in the embedding space to obtain action-specific features. This new type of pretext task has achieved good results under the condition of smaller model size and higher training efficiency.

The research trend of graph contrastive learning also includes exploring how to adapt to different graph data characteristics [43]. Sheng et al. [44] has constructed a triple-layer lncRNA-miRNA-disease heterogeneous graph, integrating the complex relationships among these entities. Work [45] enhanced the representation of user behavior sequences by constructing a global weighted item transition graph and introduces contrastive learning objectives to improve recommendation performance. Cai et al. [46] proposed that singular value decomposition can be used to process the adjacency matrix to obtain a new graph structure with low-rank approximation, emphasizing the main components of the graph and retaining the global collaborative signal. Different from the above work, we attempt to explore the graph structure characteristics in geometric space and construct an encoder that can effectively capture spatio-temporal relationships through knowledge of differential manifolds.

Researchers aim to develop more robust, interpretable and generalized self-supervised learning frameworks to address practical challenges. Self-supervised graph neural networks combined with contrastive learning offer an effective technical means for 3D early action recognition. By mining internal structure and attribute information, they learn feature representations for understanding complex 3D scenes and behaviors.

TABLE I
NOTATIONS AND DEFINITIONS

Notations	Definitions
X	complete sequence
r	the observation rate
T	the toal number of frames in complete sequence
X_t	3D coordinate matrix of complete sequence at frame t
f	the spatio-temporal feature in complete sequence
ϑ	the parameter of encoder
$\mathcal{L}$	contrast loss
ζ	the size of subsequence
F	the behavior implicit function
ξ	physical meanings, like dis, vel and acc
Φ_ξ^i	the ξ representation sequence under the i-th subsequence
A	the adjacency matrix of joint points
L	graph Laplacian matrix
$\mathbb{H}^{d,c}$	a d-dimensional hyperbolic manifold with curvature $-\frac{1}{c}$
$\mathcal{T}_x\mathbb{H}^{d,c}$	the tangent space at a point x in $\mathbb{H}^{d,c}$
$\otimes$	Kronecker product
$\oplus_c$	Möbius summation

III. PROPOSED METHOD

In this section, we present our proposed self-supervised hyperbolic spectro-temporal graph convolution network. Firstly, we provide an overview of our proposed network and the designed progressive contrastive self-supervised learning method. Next, we introduce the sequence construction of multi-dynamic key information and hyperbolic spectro-temporal graph convolution network, respectively. Table I shows the notations and their corresponding definitions.

A. Overview

In this paper, the proposed network uses Hyperbolic Spectro-Temporal Graph Convolution Network (HSTGCN) as encoder to enhance the feature aggregation process. At the same time, in order to remove the high similarity of consecutive frames, the behavior sequences are modeled via

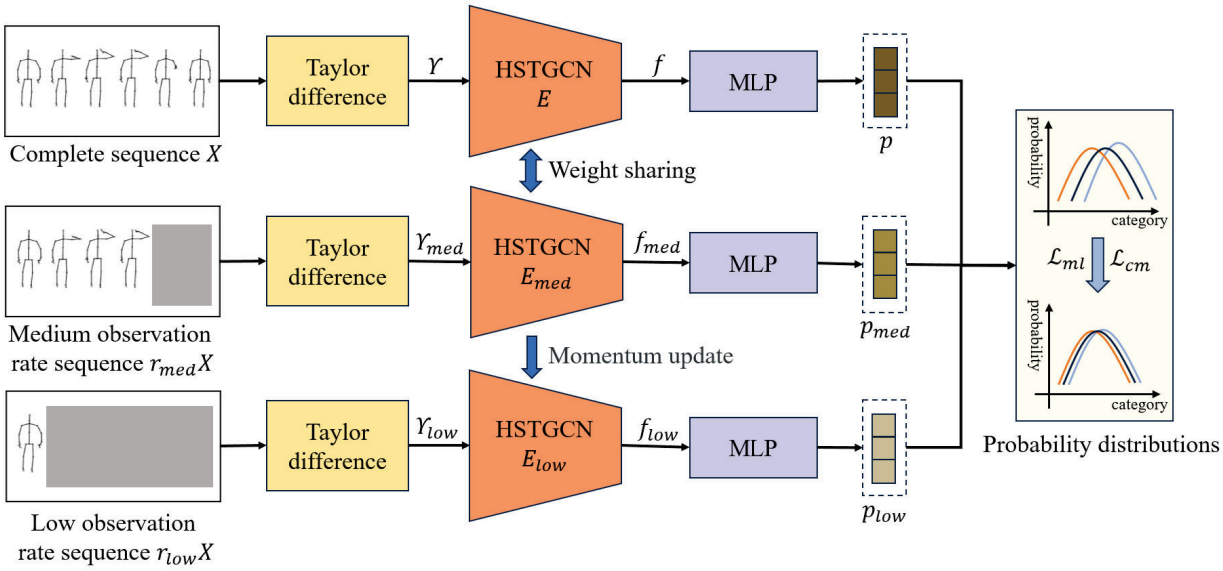


Fig. 1. The framework of self-supervised hyperbolic spectro-temporal graph convolution network (HSTGCN) for early 3D behavior prediciton. Firstly, Taylor difference operator is employed to transform input sequences into multi-dynamic key information sequences. Then, HSTGCN is used as an encoder to extract spatio-temporal features. HSTGCN converts spatial and spectral domains of spatio-temporal signals in hyperbolic manifold. Finally, a progressive contrast between behavior features with various observation rates is employed to bring their probability distributions closer.

implicit functions, and Taylor difference operator is employed to explore subtle changes with various physical meanings, thereby forming multi-dynamic key information sequences. Furthermore, the network adopts progressive contrast of 3D behavior sequences under different observation rates as self-supervised signals, avoiding errors caused by direct contrast between complete sequences and early sequences. The framework is illustrated in Figure 1.

Assuming $X = [X_1, X_2, \cdots, X_t, \cdots, X_T]$ represents the complete sequence, where T is the total number of frames in X, and X_t is the coordinate matrix of 3D behavior at frame t. $r \in (0, 1)$ represents the observation rate, so early behavior sequence can be expressed as rX. For these sequences under different observation rates, as shown in Figure 1, Taylor difference operator is firstly applied to extract behavior representation with various physical meanings, forming a multi-dynamic key information sequence. Subsequently, the proposed HSTGCN is employed to extract spatio-temporal information, resulting in corresponding spatio-temporal features denoted as f, f_r, respectively:

$$\begin{aligned} f &= HSTGCN(TD(X); \vartheta) \\ f_r &= HSTGCN(TD(rX); \vartheta_r) \end{aligned} \tag{1}$$

where, $TD(\cdot)$ denotes the transformation of input sequence into a multi-dynamic key information sequence through Taylor difference operator, and $HSTGCN(\cdot)$ represents the encoder, which leverages hyperbolic spectro-temporal graph convolution to extract rich spatio-temporal features contained in multi-dynamic key information sequence. Additionally, ϑ and ϑ_r are the parameters of these two encoders. Then we utilize Multi-layer Perceptrons (MLP) to map the features f, f_r into the output space, enabling us to obtain p, p_r. In order to ensure consistency between complete sequence and early sequence, we employ Mean Squared Error (MSE) loss function [34] to minimize the distance between them in feature space, defined as follows:

$$\mathcal{L} = \|p - p_r\|_2^2 = 2 - 2 \cdot \frac{< p, p_r >}{\|p\|_2 \cdot \|p_r\|_2} \tag{2}$$

where $\| \cdot \|_2$ denotes the operation of taking modulus, and $< \cdot, \cdot >$ represents the inner product between vectors.

Due to significant content differences between complete sequence and low observation rate sequence, direct contrast between them may lead to misleading results. Because low observation rate sequence may lacks crucial information present in complete sequence. Notably, medium observation rate sequence, falling between the two, retains some of the key information from complete sequence while providing a higher observation rate compared to low observation rate sequence. Therefore, medium observation rate sequence could serve as a better transitional solution. This also means that we will introduce two loss functions during training. The first loss function is employed to measure the feature variance between complete sequence and medium observation rate sequence, ensuring that model can extract enough information from complete sequence. The second loss function is used to evaluate the feature discrepancy between medium observation rate sequence and low observation rate sequence, aiming to enable model to make the accurate prediction even under low observation rates. So the total loss function is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{cm} + \alpha \mathcal{L}_{ml} \tag{3}$$

where, α represents the weight of $\mathcal{L}_{ml}$. $\mathcal{L}_{cm}$ and $\mathcal{L}_{ml}$ represent MSE loss between complete sequence and medium observation rate sequence, and between medium observation rate sequence and low observation rate sequence, respectively.

Here, we define r_{med} as medium observation rate and r_{low} as low observation rate. Therefore, the corresponding behavior sequences are $r_{med}X$ and $r_{low}X$. According to the similar process, the spatio-temporal features of them are denoted as $f_{r_{med}}$ and $f_{r_{low}}$. By substituting them into Eq.2, we can obtain $\mathcal{L}_{cm}$ and $\mathcal{L}_{ml}$.

$$\begin{aligned} \mathcal{L}_{cm} &= \|p - p_{r_{med}}\|_2^2 \\ \mathcal{L}_{ml} &= \|p_{r_{med}} - p_{r_{low}}\|_2^2 \end{aligned} \tag{4}$$

For the parameter update methods of these three encoders, we ensure the parameter sharing between encoder corresponding to f and encoder corresponding to $f_{r_{med}}$, allowing them to extract and represent features consistently. The parameter sharing ensures that these two encoders perform consistently in feature extraction and representation learning. During training, the model adjusts parameters based on the common features of X and $r_{med}X$, enabling these two encoders to adapt to changes in both sequences simultaneously. This approach allows the encoder parameters trained on medium observation rate and complete sequences to be used for extracting useful features and enhancing the prediction accuracy when processing early 3D behavior sequences. As for the parameters in encoder corresponding to $f_{r_{low}}$, they are updated via momentum update. Because momentum update can prevent network from collapsing during the learning process. Therefore, the parameter update method of three encoders can be expressed as:

$$\begin{aligned} \vartheta_{r_{med}} &= \vartheta \\ \vartheta_{r_{low}} &\leftarrow \beta\vartheta_{r_{low}} + (1-\beta)\vartheta \end{aligned} \tag{5}$$

where, ϑ, ϑ_{med} and ϑ_{low} represent the parameters of three encoders respectively, and $\beta \in [0, 1]$ is the momentum factor.

In the subsequent evaluation stage, we introduce classifiers behind encoders corresponding to $f_{r_{med}}$ and $f_{r_{low}}$. These classifiers aim to predict the category of behavior sequence based on the features extracted by encoders and output the category probabilities. Considering that prediction result only relying on a single encoder and classifier might have some limitations and biases, we mitigate it by averaging the category probabilities obtained from both encoders, yielding a more robust and reliable final prediction classification result.

B. Sequence Construction of Multi-Dynamic Key Information

Human behavior, as a natural and continuous process, contains both diverse and unique attributes. Its diversity shows various expressions due to individual differences, environmental dynamics, and temporal evolution. Conversely, its uniqueness lies in fundamental physical concepts such as displacement, velocity, and acceleration. These concepts together constitute the essential characteristics. With these properties, it is possible to model human behavior using trajectory function. Trajectory functions can capture the changes in behavior across temporal and spatial domains simultaneously, providing a comprehensive understanding of the dynamics of behavior. By employing trajectory functions, we can accurately describe and predict nonlinear and complex motion patterns inherent in human behavior.

Algorithm 1 Training Algorithm for HSTGCN

Input: Set of skeleton data D,
encoder network E, projector q,
target encoder network E_{t}, target projector q_{t},
momentum hyper-parameter β, weight parameters α ,
number of optimization steps K and batch size N
Output: Trained encoder E_t
Randomly initialize $\vartheta_{r_{med}}$ and copy to $\vartheta_{r_{low}}$
for each epoch e from 1 to K **do**
  for all data points X in D **do**
    Obtain samples at different observation rates $X, r_{med}X, r_{low}X$
    Compute $\Upsilon, \Upsilon_{\mathrm{med}}, \Upsilon_{\mathrm{low}}$ from $X, r_{\mathrm{med}}X, r_{\mathrm{low}}X$ respectively
    $p = q(E(\Upsilon; \vartheta_{r_{med}}))$
    $p_{med} = q(E(\Upsilon_{med}; \vartheta_{r_{med}}))$
    $p_{low} = q_t(E_t(\Upsilon_{low}; \vartheta_{r_{low}}))$
    $\mathcal{L}_{cm} = \|p - p_{r_{med}}\|_2^2$
    $\mathcal{L}_{ml} = \|p_{r_{med}} - p_{r_{low}}\|_2^2$
    $\mathcal{L}_{total} = \mathcal{L}_{cm} + \alpha\mathcal{L}_{ml}$
  end for
  Update $\vartheta_{r_{med}}$ by back-propagation
  Update the $\vartheta_{r_{low}}$ with momentum β
end for

To represent a behavior sequence as an trajectory function, we adopt a method starting with the generation of individual subsequences. It is achieved by applying a sliding window of size ζ and step size 1 on the original sequence X. As a result, we obtain $K = T - \zeta + 1$ subsequences, denoted as $\{X^1, X^2, \cdots, X^i, \cdots, X^K\}$, where each X^i represents a subsequence of length ζ, i.e., $X^i = [X_1^i, X_2^i, \cdots, X_\zeta^i]$. Next, we construct the trajectory function of the behavior sequence F. The purpose of F is to establish a mapping from input space to feature space, allowing us to indirectly understand and describe the nature of behavior by analyzing features. Given a subsequence, our primary goal is to enhance the key information within it. In order to accomplish this, we employ Taylor series to expand the trajectory function:

$$\begin{aligned} F(X_\zeta^i) = F(X_1^i) &+ \frac{F^{'}(X_1^i)}{1}(X_\zeta^i - X_1^i) \\ &+ \frac{F^{''}(X_1^i)}{2}(X_\zeta^i - X_1^i)^2 + o(X_\zeta^i - X_1^i)^2 \end{aligned} \tag{6}$$

The calculation of this formula covers motion over the entire subsequence, focusing mainly on long-range motion. However, to fully understand the motion dynamics, it is not enough to consider only long-range motion. The short-range motion within subsequence also plays a crucial role. So we replace X_ζ^i in Eq.6 with $X_\tau^i(\tau < \zeta)$ to capture short-range motion details by analyzing the motion between X_τ^i and X_1^i. The long-range and short-range motions are then averaged to provide a more comprehensive and balanced description of behavior. This process can be described as

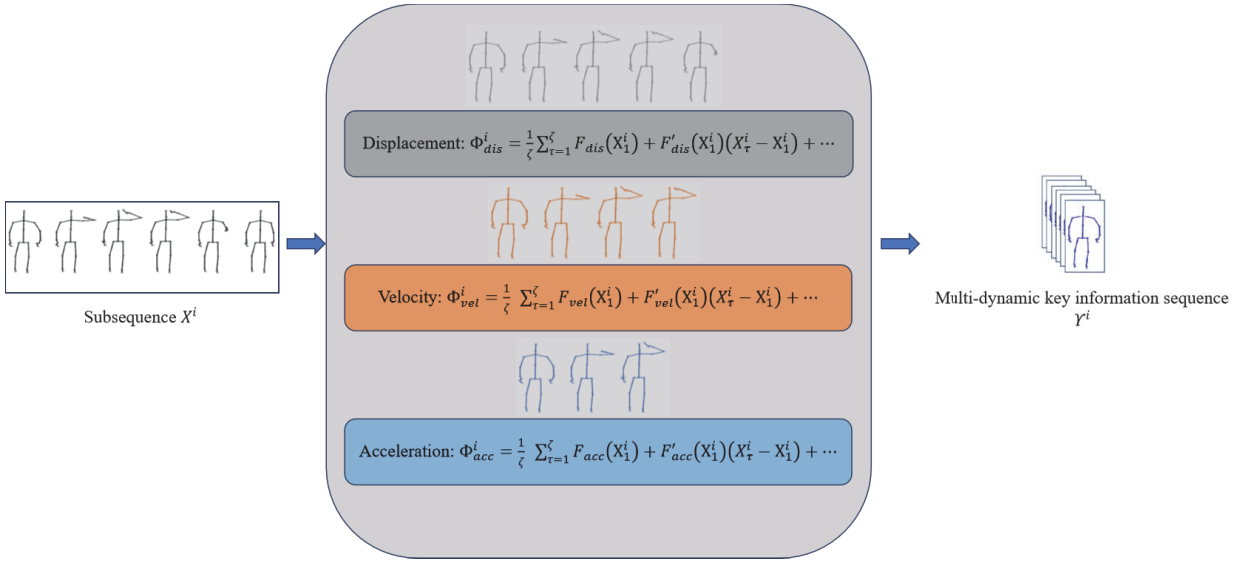


Fig. 2. The generation process of multiple dynamic key information sequences. The subsequence is Taylor expanded under three different physical concepts, and then the generated new sequences are fused to obtain their dynamic key information.

$$\Phi^i_\xi = \frac{1}{\zeta}\left(F(X^i_1)+F(X^i_2)+\cdots+F(X^i_\zeta)\right) = \frac{1}{\zeta}\sum_{\tau=1}^{\zeta} F(X^i_\tau) \quad (7)$$

Here, Φ^i_ξ represents the key motion in a subsequence starting with X^i_1 and ending with X^i_ζ. The subscript ξ denotes a specific physical concept. By defining different ξ, the new subsequences under various motion concepts can be obtained.

Displacement, velocity and acceleration are three common physical concepts in time series analysis. Together, they provide a comprehensive description of the motion process, offering detailed information from the overall motion trajectory to local dynamic changes. By combining these three concepts, we can more accurately capture and analyze the complexity of motion. We define the formulas for displacement, velocity, and acceleration as shown in Eq.8. $F^{'}$ and $F^{''}$ of these can be obtained by calculating the differences between adjacent terms. We notice that $o(X^i_\zeta - X^i_1)^2$ might have impact on the overall calculation, therefore, we adopt a small, learnable value ϵ to balance this effect.

$$\begin{aligned} F_{dis}(X^i_k) &= dis(X^i_k) = X^i_{k+1} - X^i_k \\ F_{vel}(X^i_k) &= vel(X^i_k) = dis(X^i_{k+1}) - dis(X^i_k) \\ F_{acc}(X^i_k) &= acc(X^i_k) = vel(X^i_{k+1}) - vel(X^i_k) \end{aligned} \quad (8)$$

Displacement denotes the position change of a joint, characterized by both magnitude and direction. According to Eq.7, we can get the dynamic key information sequence about displacement:

$$\begin{aligned} \Phi^i_{dis} = \frac{1}{\zeta}\sum_{\tau=1}^{\zeta} F_{dis}(X^i_1) + \frac{F'_{dis}(X^i_1)}{1}(X^i_\zeta - X^i_1) \\ + \frac{F''_{dis}(X^i_1)}{2}(X^i_\zeta - X^i_1)^2 + \epsilon_{dis} \end{aligned} \quad (9)$$

where, F_{dis} is the trajectory function of displacement. In a similar manner, Φ^i_{vel} and Φ^i_{acc} can be obtained through the trajectory function of velocity F_{vel} and the trajectory function of acceleration F_{acc}. Φ^i_{vel} and Φ^i_{acc} share a similar structure with Φ^i_{dis}.

Therefore, given a subsequence, we can calculate three motion representation sequences: Φ^i_{dis}, Φ^i_{vel} and Φ^i_{acc}. These three new sequences are then concatenated, and weighting matrix are learned using a 3D convolution operation ϱ. It allows for effective extraction of key information from the sequences in different physical concepts, resulting in the generation of multi-dynamic key information sequences. More specifically, we use *concat* and batch normalization to combine these three sequences:

$$\Phi^i_{concat} = BN(concat(\Phi^i_{dis}, \Phi^i_{vel}, \Phi^i_{acc})) \quad (10)$$

Then, a weight matrix is calculated to merge sequences of different physical concepts in an optimal way:

$$W^i_\Phi = softmax(LeakyRELU(\varrho(\Phi^i_{concat}))) \quad (11)$$

where, W^i_Φ is responsible for calculating the appropriate weights in units of frames to accurately merge sequences in different physical senses. Finally, the output multi-dynamic key information sequence Υ^i is calculated as follows:

$$\Upsilon^i = W^i_\Phi \otimes \Phi^i_{concat} \quad (12)$$

By generating multi-dynamic key information sequences of all subsequences in a similar way, a final multi-dynamic key information sequence $\Upsilon = [\Upsilon^1, \Upsilon^2, \cdots, \Upsilon^i, \cdots, \Upsilon^K]$ can be constructed. Furthermore, by inputting Υ into feature encoder, in-depth learning and understanding of behavior sequences can be achieved.

C. Hyperbolic Spectro-Temporal Graph Convolution Network

In this section, we propose a novel encoder, the Hyperbolic Spectro-Temporal Graph Convolutional Network (HSTGCN). The specific structure is shown in Figure 3. The architecture of HSTGCN is composed of multiple HSTGCN blocks, which are designed to progressively obtain high-level abstract representations of the input skeletal sequence features. After stacking the HSTGCN blocks, we incorporate a global pooling layer and a fully connected layer to summarize the output and generate the final representation. Within the HSTGCN block, we employ a parallel structure, divided into a CNN stream and a GCN stream. The CNN stream captures local features within the skeletal representation through two-dimensional convolutions and adjusts the spatial distribution of the representation using Batch Normalization. On the other hand, the GCN stream acquires global node representations with hyperbolic graph topological structure information through Hyperbolic Spectro-Temporal Graph Convolution. We add the local feature representation from the CNN stream to the global node representation from the GCN stream and gradually fuse them through a Temporal Convolutional Network (TCN) to obtain the temporal characteristics of the skeletal information. The output is a representation with excellent spatiotemporal properties as the output of the HSTGCN block.

In ST-GCN, the most important operation is spatial graph convolution, which aggregates features by computing the weighted average of node features with the neighborhood of each node. Let the feature of l-th layer be denoted as $f^l \in \mathbb{R}^{C\times T\times N}$, and the feature of $l+1$-th layer be denoted as $f^{l+1} \in \mathbb{R}^{C\times T\times N}$. Here, C is the channel dimension, and N is the number of human joints. Then the spatial graph convolution operation can be expressed as:

$$f^{l+1} = \mathcal{D}^{-\frac{1}{2}}(A+I)\mathcal{D}^{-\frac{1}{2}}f^l W^l \tag{13}$$

Among Eq.13, A is the adjacency matrix of joint points. I is the identity matrix with same size as A, representing self-connection between each joint. $\mathcal{D}$ is the degree matrix of $A+I$, and it is also a diagonal matrix defined as $\mathcal{D}^{pp} = \sum_q (A^{pq} + I^{pq})$. p and q represent two joint points of human body. W^l is the weight matrix selected according to partitioning strategy during convolution process.

The partitioning strategy employed in ST-GCN is spatial configuration partitioning strategy, which divides the node neighborhood into three parts: the root node, nodes adjacent to the root node and closer to the center of gravity, and nodes adjacent to the root node but farther from the center of gravity. However, this strategy divides A into multiple adjacency submatrices, then conducts different convolution operations to learn spatial features under each submatrices, and finally aggregates these features to obtain the result of spatial graph convolution. While the approach calculates within these three categories, it fails to establish relationships between them, and ignores spatial information of long-distance joints. Futhermore, it is important to note that spatial graph convolution is a first-order local approximation of spectral graph convolution. This paper attempts to better aggregate spatial features of body joints through first-order spectral graph convolution.

First, spectral graph convolution can be defined as the product of input vector and filter in Fourier domain. This process is typically implemented by Fourier transformation of Laplacian matrix. Specifically, we define graph Laplacian matrix as $L = D - A$, and the normalized Laplacian matrix as $L^{sym} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = U\Lambda U^T$. Here, D is the degree matrix of A. U is the eigenvector matrix, which can be obtained by performing eigendecomposition on L^{sym}, and Λ is the eigenvalue matrix corresponding to U. Then, spectral graph convolution can be expressed as:

$$g_\theta \star x = U g_\theta U^T x \tag{14}$$

where, g_θ represents a parameterized filter, θ denotes the set of these parameters, and x is the input vector. The symbol $\star$ represents the convolution operation. In Eq.14, x is transformed from spatial domain to spectral domain through graph Fourier transform $U^T x$. Subsequently, $g_\theta U^T x$ evaluates the computations in spectral domain. And finally, through the inverse Fourier transform, $U g_\theta U^T x$ is converted back to spatial domain. Considering that this graph structure corresponds to human joints, the feature decomposition of L^{sym} requires a large amount of computational resources. To address this, Chebyshev polynomials are introduced to approximate g_θ, transforming the above formula into:

$$g_\theta \star x = U g_\theta U^T x \approx \sum_{w=0}^{\mathcal{W}} \gamma_w Q_w(L) x \tag{15}$$

Among these parameters, γ_w is the sequence of coefficients for Chebyshev polynomial, $\mathcal{W}$ is the number of terms of polynomial, and $Q_w(L)$ is the w-order polynomial of L. Additionally, for simplicity, only the first-order Chebyshev inequality is considered, as follows:

$$g_\theta \star x \approx \theta(D^{-\frac{1}{2}}AD^{-\frac{1}{2}} + I)x \tag{16}$$

This paper considers transforming Eq.16 into a potential residual structure. In general, it involves directly adding the input vector x to result through a shortcut connection without multiplying by θ, which can be rewritten as:

$$g_\theta \star x \approx \theta D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x + x \tag{17}$$

Finally, the above process is extended to the feature f^l, as follows

$$f^{l+1} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}f^l W^l + f^l \tag{18}$$

Combined with one-dimensional temporal convolution, we have implemented spectro-temporal graph convolution. Next, we aim to extend this process to hyperbolic manifold, leveraging the geometric properties in hyperbolic manifold to enhance the representation of spatio-temporal relationships.

The hyperbolic manifold is a Riemannian manifold with constant negative curvature. It possesses numerous straight lines that are parallel to a given line and pass through a common point, offering a powerful representation of implication

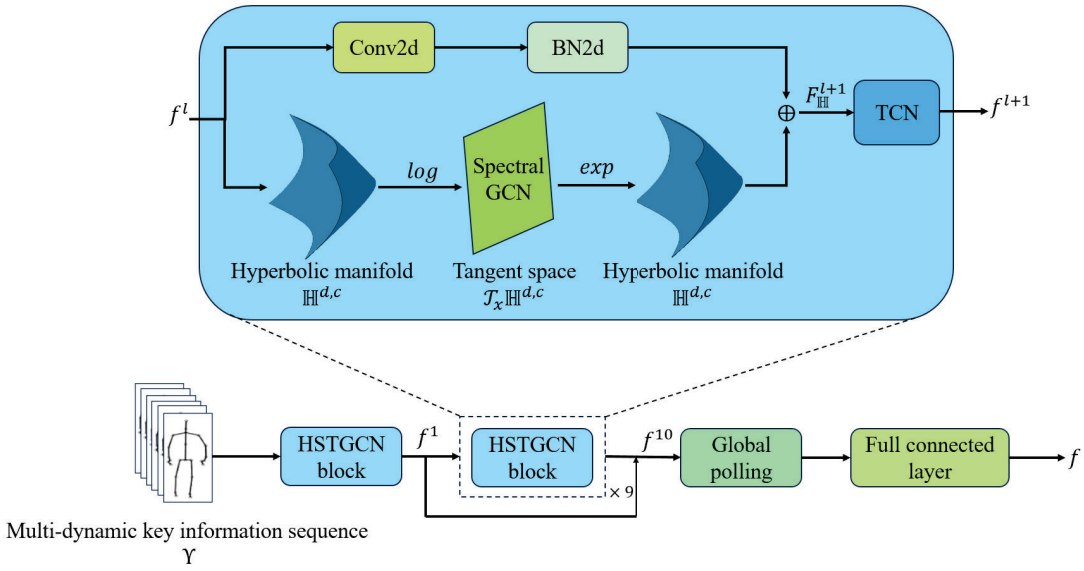


Fig. 3. The framework of hyperbolic spectro-temporal graph convolution network. By transforming the hyperbolic space and its tangent space, we extend spectral graph convolution to the hyperbolic space, thereby realizing hyperbolic spectro-temporal graph convolution network.

relationships. We denote $\mathbb{H}^{d,c}$ as a d-dimensional hyperbolic manifold with constant negative curvature $-\frac{1}{c}(c>0)$, and its tangent space at a point x is expressed as $\mathcal{T}_x\mathbb{H}^{d,c}$. The tangent space is a linear space and exhibits Euclidean geometric properties, enabling the transformation of operations that are challenging on hyperbolic manifolds into its tangent space. The transformation between these two spaces is achieved through exponential mapping and logarithmic mapping:

$$\begin{aligned}\exp_x(z) &= x \oplus_c (\tanh(\sqrt{c}\frac{\lambda\|z\|_2}{2})\frac{z}{\sqrt{c}\|z\|_2}2)\\ \log_x(y) &= \frac{2}{\sqrt{c}\lambda}\tanh^{-1}(\sqrt{c}\|x\oplus_c y\|_2)\frac{-x\oplus_c y}{\|-x\oplus_c y\|_2}\end{aligned} \tag{19}$$

where, $x, y \in \mathbb{H}^{d,c}$ and $z \in \mathcal{T}_x\mathbb{H}^{d,c}$. $\exp_x(z)$ maps point z in tangent space to hyperbolic space, while $\log_x(y)$ maps point y in hyperbolic space to tangent space. $\oplus_c$ denotes addition in hyperbolic space, known as Möbius summation, defined as

$$v_1 \oplus_c v_2 = \frac{(1+2c<v_1,v_2>+c\|v_2\|_2^2)v_1+(1-c\|v_1\|_2^2)v_2}{1+2c<v_1,v_2>+c^2\|v_1\|_2^2\|v_2\|_2^2} \tag{20}$$

v_1 and v_2 are two points in hyperbolic space, $\lambda_x = \frac{2}{1+c\|x\|_2^2}$ serves as a conformal factor, used to describe the metric structure on hyperbolic manifold.

Based on these two mappings, we apply spectral graph convolution to tangent space, as shown in Figure 3. First, we map the feature f^l into the hyperbolic space.

$$f_{\mathbb{H}}^l = \exp_o(f^l) \tag{21}$$

Define the linear operation function $F_{\mathbb{H}}$, which is an operation performed in hyperbolic space and can be expressed as

$$F_{\mathbb{H}}^o(f_{\mathbb{H}}^l, X) = \exp_o(X log_o(f_{\mathbb{H}}^l)) \tag{22}$$

This operation involves mapping features $f_{\mathbb{H}}^l$ to tangent space for computation $F(.)$ and then mapping them back to the original hyperbolic space. $o = \{\sqrt{c}, 0, 0, \cdots, 0\}$ denotes the origin in hyperbolic space. This operation function will aid in understanding the subsequent formulas. For example, we perform linearly transforming of $f_{\mathbb{H}}^l$ first, which is hyperbolic Fourier transform and expressed as

$$\hat{f_{\mathbb{H}}^l} = F_{\mathbb{H}}^o(f_{\mathbb{H}}^l, W^l) \tag{23}$$

Next, we apply spectral filtering operation to the result. When calculating the adjacency matrix, we can leverage hyperbolic distance to measure the relationship between two nodes, and modify the weight based on the distance, given by:

$$A_{\mathbb{H}}(p,q) = \exp(-\frac{d_{\mathbb{H}}^2(p,q)}{\delta}) \tag{24}$$

where, δ is an artificially specified radial range parameter, and $d_{\mathbb{H}}(p,q)$ represents the distance between two nodes p and q in hyperbolic manifold, defined as

$$d_{\mathbb{H}}(p,q) = (\frac{2}{\sqrt{c}})\tanh^{-1}(\sqrt{c}\|-p\oplus_c q\|_2) \tag{25}$$

Subsequently, we calculate the degree matrix $D_{\mathbb{H}}$ using a method similar to that in Euclidean space, enabling spectral filtering operation on the linear transformation results:

$$f_{\mathbb{H}}^{l+1} = F_{\mathbb{H}}^{f_{\mathbb{H}}^l}(\hat{f_{\mathbb{H}}^l}, D_{\mathbb{H}}^{-\frac{1}{2}}A_{\mathbb{H}}D_{\mathbb{H}}^{-\frac{1}{2}}) \oplus_c f_{\mathbb{H}}^l \tag{26}$$

Here, $f_{\mathbb{H}}^{l+1}$ is the feature in hyperbolic manifold of layer $l+1$. Finally, we have completed the extension of spectro-temporal graph convolution to hyperbolic space.

However, hyperbolic spectro graph convolution can obtain global representations with skeletal topological information. We need to further employ convolutional layers to learn the local information of the samples. Specifically, we use batch normalization layers and convolutional layers for feature processing.

$$\hat{f^l} = BatchNorm(Conv2d(f^l)) \tag{27}$$

We map $f_{\mathbb{H}}^{l+1}$ back to the flat space and add it with $\hat{f}^l$, obtaining a representation that contains both local and global information. To further capture the temporal relationships present in the sequence, we utilize Temporal Convolutional Networks (TCNs) as the final part of our model, as follows:

$$f^{l+1} = TCN(\hat{f}^l + log_{f_{\mathbb{H}}^l}(f_{\mathbb{H}}^{l+1})) \tag{28}$$

With this, we have completed the construction of the Hyperbolic Spectro-Temporal Graph Convolution Network.

IV. Experiments and Analysis

A. Datasets

In order to verify the effectiveness of our proposed method on early 3D behavior prediction task, experiments are carried out on three human-based 3D behavior datasets, namely NTU RGB+D 60 dataset, NTU RGB+D 120 dataset and PKU-MMD dataset. These three datasets are the largest benchmark datasets, and the experimental results on them are more convincing.

NTU RGB+D 60 [47]: This dataset contains RGB+D videos and skeleton data for human behavior. The behavior data is captured by 3 Microsoft Kinect V2 cameras from 40 human subjects, with a total of 56,880 samples containing 60 categories totaling 4 million frames, where the maximum frame for all samples is 300. 25 joints are recorded for each body skeleton. The dataset provides two original settings, namely two evaluation protocols, Cross-Subject (Xsub) and Cross-View (Xview). In Xsub protocol, the training set contains 40,320 samples from 20 subjects, and the remaining 16,560 samples are used for testing. In Xview protocol, 37,920 samples captured by cameras 2 and 3 are used for training, and camera 1 is used for training. The remaining 18960 samples were used for testing. We follow these two settings and report the Top-1 accuracy of experimental results.

NTU RGB+D 120 [48]: This dataset is an extended version of NTU RGB+D 60, adding 57,367 skeleton sequences in 60 additional action categories, totaling 113,945 samples, 120 action category categories, captured from 106 different subjects and 32 different cameras. Two evaluation protocols are used: Cross-subject (Xsub) and Cross-setting (Xset). In Xsub protocol, 63,026 samples from half of the participating subjects were used for training, while the remaining 50919 samples were used for testing. In Xset protocol, 54468 samples taken from half of the camera devices are used for training and the remaining 59477 samples are used for testing.

PKU-MMD [49]: This dataset covers a wide range of complex humanactivity categories, collecting 1,076 long video sequences with 51 action categories. These sequences were captured by 66 participating subjects from various perspectives using three Kinect V2 cameras, totaling 21,545 behavior instances across 5.4 million frames. The label of each long sequence marks the behavior category of each action instance, the start frame, endframe and label confidence of the action. Additionally, the dataset offers two different settings, Part I and Part II. We conduct experiments under the cross subject protocol on Part I.

TABLE II
Ablation of different modules on NTU-60 and PKU-MMD

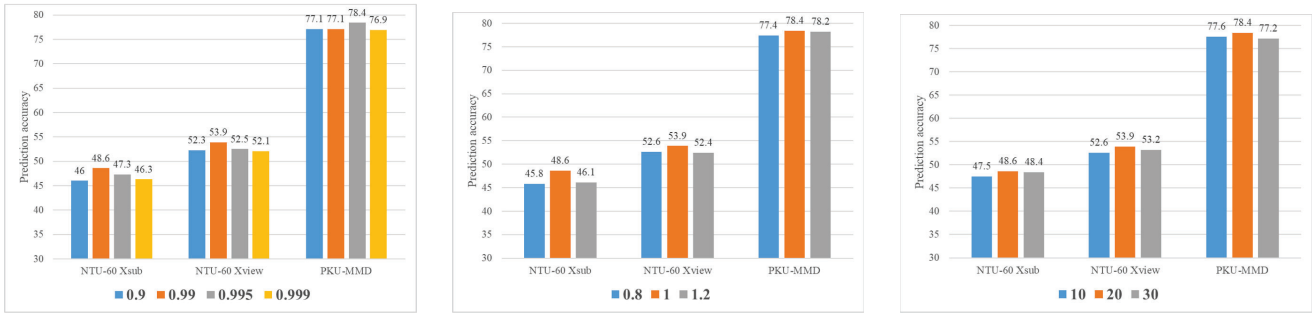
MS	SC	SL	HM	NTU-60		PKU-MMD
				Xsub	Xview	
				38.8	41.2	70.5
✓				44.5	49.3	73.8
✓	✓			46.2	51.5	74.3
✓	✓	✓		46.9	52.6	76.5
✓	✓	✓	✓	**48.6**	**53.9**	**78.4**

B. Experimental Setup

All experiments are performed using the PyTorch framework [50]. Following the standard methods in existing behavior prediction tasks [19], [24], [51], early 3D behavior sequences are generated. For each complete 3D behavior sequence within dataset, partial sequences are obtained at observation rates ranging from 0.1 to 0.9 and these partial sequences collectively form a 3D behavior prediction dataset. The sample size in this dataset is 9 times larger than that in original dataset, and we unify the total frame number of all samples to 50 frames. For sequences with more than 50 frames, downsampling is used to reduce the number of frames. For sequences with less than 50 frames, fill the blank frames with 3D joint point coordinates of the last frame. The network is trained on an NVIDIA RTX 3090 GPU with a batch size of 128. Throughout all datasets and evaluation protocols, we only report the Top-1 accuracy.

Self-supervised Pre-training: We feed complete sequence, sequence with a random observation rate greater than 0.5, and sequence with a random observation rate less than 0.5 into the proposed network. Throughout the optimization process, we employ Stochastic Gradient Descent (SGD), with a momentum of 0.9 and a weight decay of 0.0001. Additionally, the momentum factor β is set to 0.99. The model is trained for 300 epochs with a fixed learning rate of 0.1, and the learning rate remains unchanged throughout the training process.

Linear Evaluation: We evaluate model through linear evaluation on self-supervised 3D behavior recognition methods [52]–[54], where the encoder weights are kept frozen during testing. Based on this evaluation, two linear classifiers, each consisting of a fully connected layer and a softmax layer, are appended to the encoders corresponding to medium observation rate and low observation rate sequences. The classifiers are trained for 100 epochs with a learning rate of 3.0, which is multiplied by 0.1 at 80-th epoch. Typically, existing supervised early 3D behavior prediction methods evaluate prediction accuracy under the observation rates of 0.2, 0.4, 0.6, and 0.8. We also compute the prediction accuracy under these four observation rates and compare these results with some self-supervised 3D behavior recognition methods for early 3D behavior prediction and some supervised 3D behavior prediction. This comparison is particularly important given the lack of research on unsupervised early 3D behavior prediction. As part of our experiments, we also computed the average accuracy at all observation rates as an outcome of ablation experiments.



(a) Momentum Factor (b) Loss Function Weight (c) Subsequence Length

Fig. 4. Ablation expermental results of main parameters on NTU-60 and PKU-MMD datasets.

C. Ablation Experiments

In this section, we perform ablation experiments on two protocols of NTU-60 dataset and PKU-MMD dataset to assess the effectiveness of each proposed module. Additionally, we explore the optimal configuration of main parameters, and the experimental results are thoroughly analyzed and discussed. In Table II, "MS" refers to medium observation rate sequence, "SC" denotes the sequence construction of multi-dynamic key information, "SL" represents spectral graph convolution based on laplacian operator, and "HM" indicates hyperbolic manifold.

We also employ the t-SNE algorithm to visualize the impact of our proposed method on classification effects. Specifically, we use the distance between cluster centers and the degree of aggregation of samples within the same category to judge the specific classification effects. We have conducted targeted research on the influence of hyperbolic mapping and the sequence construction of multi-dynamic key information on our algorithm. As shown in the Fig. 7, the impact of hyperbolic mapping on the overall data distribution is not significant. This is because we used the exponential and logarithmic methods during the hyperbolic mapping to ensure that data can be correctly mapped between the hyperbolic space and other spaces. Since we use the geodesic distance in the hyperbolic space as a reference, our method can effectively identify some outliers and bring them closer to the cluster centers. The method of sequence construction of multi-dynamic key information changes the overall distribution of the input data, which also has a significant impact on the final visualization results. It can be seen that this method can effectively prevent the cluster centers from being too close to each other. In contrast, the visualization results of our method exhibit more distinct cluster centers, and the distribution of these centers is more dispersed. , proving the effectiveness of our approach.

The effectiveness of the proposed modules: As shown in Table II, the prediction performance is notably lower in the absence of medium observation rate sequence input compared to other methods. This suggests that the spatio-temporal information provided by medium observation rate sequence is critical for early 3D behavior prediction. Additionally, the sequence construction of multi-dynamic key information, applied on two evaluation protocols of NTU-60, contributes to nearly a 2% improvement in prediction accuracy. The similar results are observed on PKU-MMD. This operator aims at reducing redundancy between adjacent frames and more accurately capturing frame-to-frame differences to represent dynamic changes in behavior sequence. Furthermore, with the introduction of Laplacian-based spectral graph convolution and hyperbolic manifold, the network achieves its highest performance. It further validates the effectiveness of our designed approach. The spectral graph convolution can establish the dependencies of long-distance joints, while hyperbolic manifold is suitable for extracting the complex relationships within human 3D behavior sequences.

The effects of main parameters: Figure 7 shows our ablation experiments on the main parameters. It can be seen from Figure 4a, the proposed network exhibits a high sensitivity to the choice of momentum factor. A small coefficient leads to unstable learning of encoders during training, thus reducing representation quality. Conversely, when the parameter is close to 1, network learns minimal changes from contrastive learning. Regarding the contrast loss weight α, we perform ablation experiments as shown in Figure 4b. Excessively large values cause model to overly focus on the feature differences between medium and low observation rate, potentially leading to overfitting to the features of medium observation rate sequence while ignoring global information from complete sequence. Conversely, excessively small values make model insensitive to the feature differences and unable to effectively leverage this information. Therefore, the experimental data indicates that an appropriate balance point is achieved at 1.0. In addition, the length of subsequence ζ is also ablated and the results of Figure 4c show that the performance of model is optimal when $\zeta = 20$. This may be because in shorter subsequence, i.e., 10 frames, model mainly encodes the dominant motion information, thereby reducing the interference of noise. However, this setup also has some limitations, namely that model cannot fully capture the long-range dynamic features of behavior. In contrast, extending the subsequence's length to 30 frames may enhance the model's ability to capture long-range motion information, but it may also introduce unnecessary noise, which may have a negative impact on the performance of model.

TABLE III
COMPARISONS OF THE EARLY 3D BEHAVIOR PREDICTION ACCURACY WITH VARIOUS METHODS ON NTU-60

Method	Backbone	Xsub					Xview				
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
Suspervised											
Local+LGCN (TIP 2019) [19]	CNN	32.1	63.8	77.0	82.5	83.2	-	-	-	-	
CEL (TCSVT 2020) [20]	RNN	35.6	54.6	67.1	72.9	75.5	37.2	57.2	69.9	75.4	78.0
HARD-Net (ECCV 2020) [21]	GCN	42.4	72.2	83.0	**86.8**	**87.5**	**53.2**	**82.9**	91.3	93.7	94.0
Local+AGCN-AL (TCDS 2021) [24]	GCN	38.2	71.2	82.3	86.3	87.2	-	-	-	-	-
GA-Net (TMM 2021) [22]	GCN	**42.5**	**72.6**	**83.1**	**86.8**	87.2	49.8	80.2	**91.6**	**94.0**	**94.2**
Dear-Net (TMM 2023) [25]	CNN	32.7	69.7	80.2	83.5	-	-	-	-	-	-
Self-supervised											
SkeletonCLR (CVPR 2021) [31]	GCN	19.1	45.6	59.9	65.9	65.3	21.0	48.1	64.1	70.0	68.4
AimCLR (AAAI 2022) [38]	GCN	16.2	40.6	55.2	62.2	63.2	23.0	52.7	**69.4**	75.3	74.5
HiCLR (AAAI 2023) [39]	GCN	21.2	50.1	65.7	**71.9**	72.3	24.3	54.6	70.3	**76.3**	76.0
SkeAttnCLR (IJCAI 2023) [33]	GCN	20.2	48.0	63.7	70.6	71.1	18.2	46.8	62.5	70.0	67.6
SSRL (TCSVT 2024) [34]	GCN	21.8	50.1	62.5	68.9	68.7	24.1	53.5	66.3	70.4	69.8
Ours	**GCN**	**21.9** ±0.65	**51.8** ±0.24	**66.4** ±0.44	71.5 ±0.52	**72.8** ±0.61	**25.8** ±1.58	**55.6** ±0.72	69.3 ±0.61	75.7 ±0.74	**76.2** ±0.46

TABLE IV
COMPARISONS OF THE EARLY 3D BEHAVIOR PREDICTION ACCURACY WITH VARIOUS METHODS ON NTU-120

Method	Backbone	Xsub					Xset				
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
Self-supervised											
SkeletonCLR (CVPR 2021) [31]	GCN	5.3	15.7	28.2	32.0	33.7	7.4	19.0	31.2	36.5	38.0
AimCLR (AAAI 2022) [38]	GCN	4.6	14.5	24.7	30.0	46.4	8.5	22.1	37.0	42.4	38.4
HiCLR (AAAI 2023) [39]	GCN	**6.3**	19.9	33.3	38.8	51.7	8.5	23.6	38.2	44.0	48.4
SkeAttnCLR (IJCAI 2023) [33]	GCN	5.0	16.6	32.8	**42.9**	**57.6**	7.4	20.1	36.6	**46.0**	**57.5**
SSRL (TCSVT 2024) [34]	GCN	4.3	11.5	17.8	21.3	32.2	6.1	15.9	24.2	26.6	28.5
Ours	**GCN**	6.2 ±0.95	**20.5** ±1.26	**34.6** ±0.82	42.7 ±0.63	55.6 ±0.65	**8.9** ±1.02	**24.1** ±1.45	**38.3** ±0.77	45.5 ±0.78	56.0 ±0.61

TABLE V
COMPARISONS OF THE EARLY 3D BEHAVIOR PREDICTION ACCURACY WITH VARIOUS METHODS ON PKU-MMD

Method	Backbone	Xsub				
		0.2	0.4	0.6	0.8	1.0
Self-supervised						
SkeletonCLR (CVPR 2021) [31]	GCN	50.3	69.2	74.5	71.7	72.3
AimCLR (AAAI 2022) [38]	GCN	61.7	74.5	80.4	79.3	76.7
HiCLR (AAAI 2023) [39]	GCN	63.9	78.0	83.0	84.7	81.7
SkeAttnCLR (IJCAI 2023) [33]	GCN	64.7	80.5	83.5	84.0	80.7
SSRL (TCSVT 2024) [34]	GCN	73.8	85.1	**88.4**	**89.1**	87.8
Ours	**GCN**	**74.1** ±0.61	**85.7** ±0.70	87.3 ±0.45	88.0 ±0.51	**87.9** ±0.45

TABLE VI
COMPARISONS OF THE 3D ACTION RECOGNITION ACCURACY ON DIFFERENT DATASETS

Methods	Encoder	NTU-60		NTU-120		PKU-MMD
		Xsub	Xview	Xsub	Xset	
Self-supervised						
P&C (CVPR 2020) [30]	GRU	50.7	76.3	42.7	41.7	59.9
MG-AL (TCSVT 2022) [32]	GCN	64.7	68.0	46.2	49.5	-
ST-CL (TMM 2023) [42]	GCN	68.1	69.4	54.2	55.6	-
HiCLR (AAAI 2023) [39]	GCN	77.6	82.0	66.8	66.1	-
SDS-CL (TNNLS 2024) [55]	DSTA [56]	73.6	78.9	50.6	55.6	-
IKEM (ICASSP 2024) [57]	GCN	75.5	81.8	-	-	-
SSRL (TCSVT 2024) [34]	GCN	**80.4**	82.0	**68.0**	68.6	89.9
Ours	**GCN**	78.7 ±0.23	**83.6** ±0.41	67.7 ±0.21	**70.2** ±0.41	**90.3** ±0.36

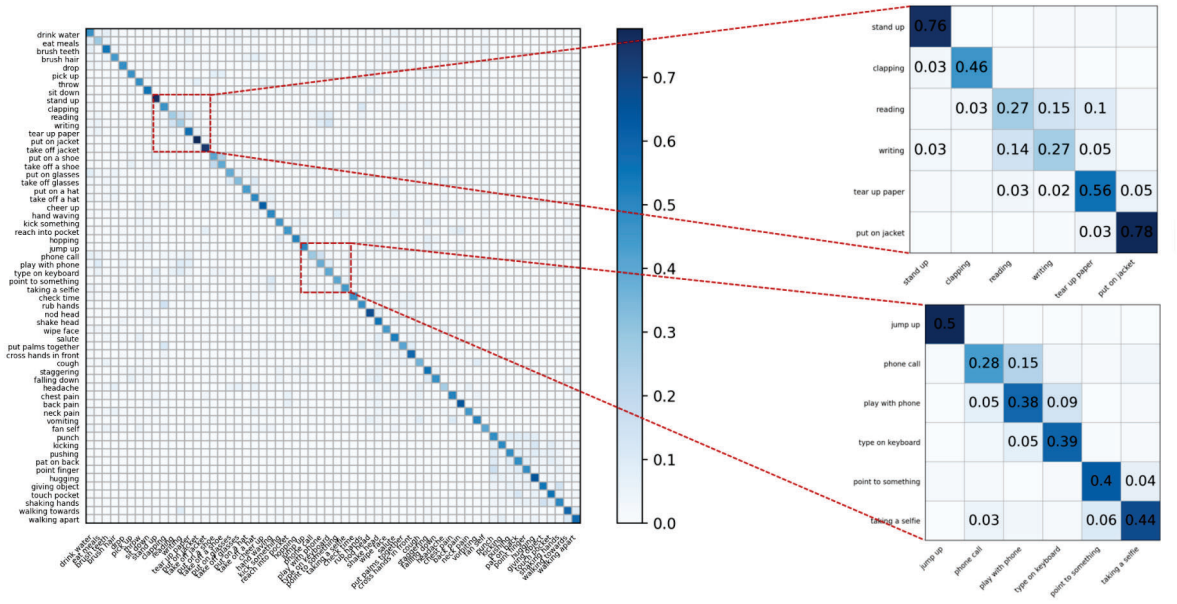


Fig. 5. Confusion matrix visualization results of NTU-60 dataset under Xsub protocol.

D. Comparison with Existing Methods

Since there are few studies on self-supervised learning in early 3D behavior prediction, in order to make an effective comparison, this paper simply modifies five common self-supervised 3D behavior recognition methods to make them suitable for early 3D behavior prediction task, including SkeletonCLR [31], AimCLR [38], HiCLR [39], SkeAttnCLR [33] and SSRL [34].

Table III presents a comparison of the proposed method on two protocols of NTU-60 dataset, while Table IV illustrates the comparison on two protocols of NTU-120 dataset, and Table V shows the comparison on PKU-MMD. According to these two tables, it is evident that the proposed method generally achieves higher prediction accuracy compared to other self-supervised methods under the listed observation rates. This fully demonstrates the benefits of the introduction of medium observation rate sequences and the design of HSTGCN. At an observation rate of 0.2, our method is only 0.1% higher than SSRL [34], but at the rest of observation rates, SSRL lags behind the proposed method. This shows that the proposed method has some advantages under medium to high observation rates while maintaining high accuracy at low observation rates. However, it is worth noting that the performance gap between our method and others gradually narrows with the increase of observation rate, even worse than HiCLR [39] and SkeAttnCLR [33] at 0.8 observation rate. This trend may arise because our method focuses more on capturing features or patterns at the very beginning of behavior, which may become less representative at high observation rates. At the same time, our method also has good results on another large dataset NTU-120, generally leading at four observation rates. This illustrates the effectiveness and robustness of the proposed self-supervised hyperbolic spectro-temporal graph convolution network when dealing with large-scale datasets. At the same time, the proposed method also achieves good results under low observation rates on PKU-MMD, which further shows the effectiveness of this approach in handling early behavior sequences. This advantage is mainly due to the utilization of the medium observation rate sequences and hyperbolic spectro-temporal convolution.

Furthermore, we compare our proposed method with some supervised early 3D behavior prediction methods. Since NTU-120 is a large-scale dataset, we have not yet found the prediction results of supervised methods. Therefore, we only list some supervised methods in Table III. The results indicate that our method is not much different from some supervised methods under the four observation rates. This further validates the feasibility of employing multi-dynamic key information sequences and highlights the powerful feature extraction capability of HSTGCN. Moreover, it also shows that our method can effectively capture and leverage the potential information in early 3D behavior sequences to obtain the accurate prediction results. Our method also achieves excellent results in action recognition, as shown in Table VI, where it outperforms most datasets and shows significant improvements over previous methods. Overall, our approach demonstrates advanced performance in early action prediction and action recognition.

Additionally, we visualize the prediction probability distributions under two protocols of NTU-60 using confusion matrix, as shown in Figure 5 and Figure 6. These matrices provide a more intuitive understanding of the performance of self-supervised early 3D behavior prediction network proposed in this paper. Notably, the network has better prediction ability

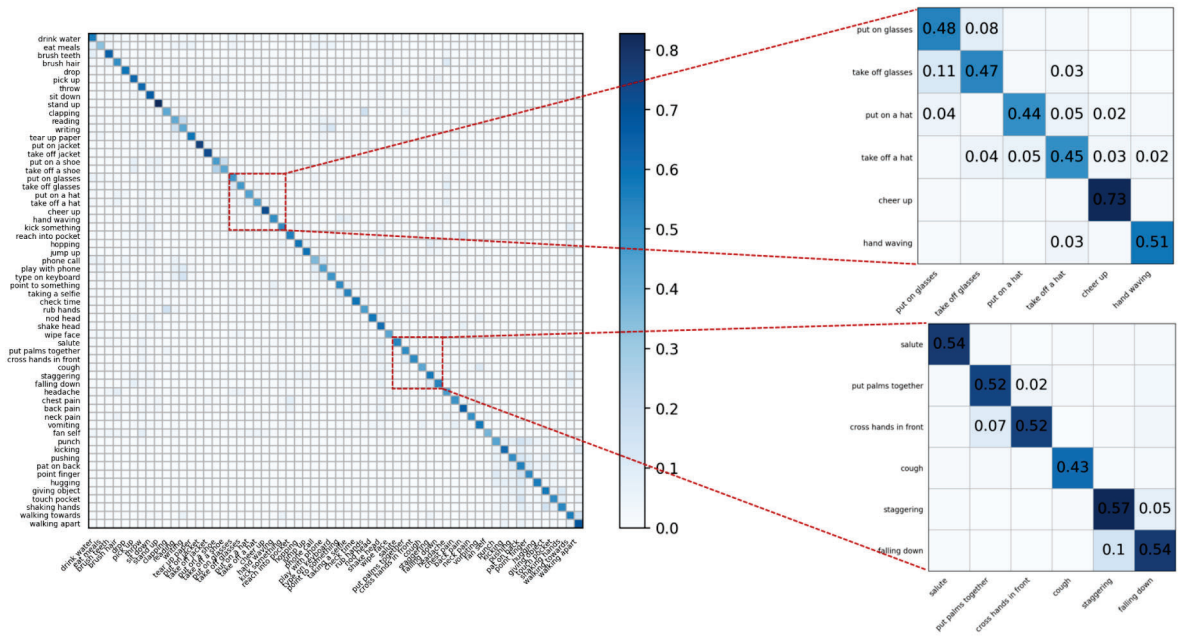
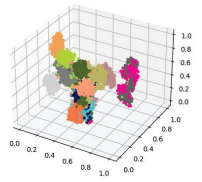
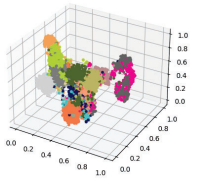


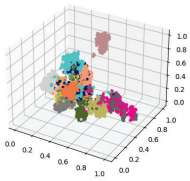
Fig. 6. Confusion matrix visualization results of NTU-60 dataset under Xview protocol.



(a) HSTGCN on NTU-60 Xsub



(b) w/o hyperbolic mapping



(c) w/o sequence construction of multi-dynamic key information

Fig. 7. Ablation expermental results of main parameters on NTU-60 Xsub protocol

when faced with behaviors exhibiting obvious discriminative features, such as ”stand up” and ”cheer up”. For some behaviors that are initially similar, this method also has certain discriminative capabilities, but there are still some limitations, such as ”reading” and ”writing”. This may be because HST-GCN does not take into account the interaction information with interactive items. Overall, the proposed method effectively addresses the challenges associated with self-supervised learning in early 3D behavior prediction tasks.

V. Conclusion

In this paper, we propose a progressive contrastive self-supervised framework for early 3D behavior prediction. This framework leverages the contrast between spatio-temporal sequences' features under various observation rates to optimize model learning. Additionally, we design the sequence construction of multi-dynamic key information and hyperbolic spectro-temporal graph convolution network. On one hand, Taylor difference operator forms multi-dynamic key information sequences by calculating motion sequences, which have different physical meanings, enabling the capture of subtle but important changes in motion sequence. On the other hand, hyperbolic graph Laplacian operator employs the geometric properties of hyperbolic manifold and the long-distance feature aggregation of spectral graph convolution to handle complex relationships within graph structure. The proposed self-supervised prediction framework is evaluated on three 3D behavior datasets, and the experimental results fully demonstrate the effectiveness of each part of the proposed method.

References

[1] M. S. Ryoo, “Human activity prediction: Early recognition of ongoing activities from streaming videos,” in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 1036–1043.
[2] 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*, 2020, pp. 536–553.
[3] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, “Revisiting skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2969–2978.
[4] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 7444–7452.

[5] F. Lino, C. Santiago, and M. Marques, "3d human pose estimation with occlusions: Introducing blendmimic3d dataset and gcn refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4646–4656.

[6] S. Mehraban, V. Adeli, and B. Taati, "Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6920–6930.

[7] G. Wang, M. Liu, H. Liu, P. Guo, T. Wang, J. Guo, and R. Fan, "Augmented skeleton sequences with hypergraph network for self-supervised group activity recognition," *Pattern Recognition*, vol. 152, p. 110478, 2024.

[8] A. Flaborea, G. M. D. di Melendugno, S. D'arrigo, M. A. Sterpa, A. Sampieri, and F. Galasso, "Contracting skeletal kinematics for human-related video anomaly detection," *Pattern Recognition*, p. 110817, 2024.

[9] B. Ganga, B. Lata, and K. Venugopal, "Object detection and crowd analysis using deep learning techniques: Comprehensive review and future directions," *Neurocomputing*, p. 127932, 2024.

[10] F. M. Thoker, H. Doughty, and C. G. Snoek, "Skeleton-contrastive 3d action representation learning," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 1655–1663.

[11] S. Yang, J. Liu, S. Lu, E. M. Hwa, Y. Hu, and A. C. Kot, "Self-supervised 3d action representation learning with skeleton cloud colorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 509–524, 2024.

[12] B. Xu, X. Shu, J. Zhang, G. Dai, and Y. Song, "Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[13] X. Shu, B. Xu, L. Zhang, and J. Tang, "Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7559–7576, 2023.

[14] B. Xu, X. Shu, and Y. Song, "X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 3852–3867, 2022.

[15] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng, "Progressive teacher-student learning for early action prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3551–3560.

[16] Y. Cai, H. Li, J.-F. Hu, and W.-S. Zheng, "Action knowledge transfer for action prediction with partial videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 8118–8125.

[17] X. Liu, J. Yin, D. Guo, and H. Liu, "Rich action-semantic consistent knowledge for early action prediction," *IEEE Transactions on Image Processing*, vol. 33, pp. 479–492, 2023.

[18] W. Guan, X. Song, K. Wang, H. Wen, H. Ni, Y. Wang, and X. Chang, "Egocentric early action prediction via multimodal transformer-based dual action prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4472–4483, 2023.

[19] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Learning latent global network for skeleton-based action prediction," *IEEE Transactions on Image Processing*, vol. 29, pp. 959–970, 2019.

[20] J. Weng, X. Jiang, W.-L. Zheng, and J. Yuan, "Early action recognition with category exclusion using policy-based reinforcement learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4626–4638, 2020.

[21] T. Li, J. Liu, W. Zhang, and L. Duan, "Hard-net: Hardness-aware discrimination network for 3d early activity prediction," in *Proceedings of Computer Vision*, 2020, pp. 420–436.

[22] W. Wang, F. Chang, C. Liu, G. Li, and B. Wang, "Ga-net: a guidance aware network for skeleton-based early activity recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 1061–1073, 2021.

[23] L. Chen, J. Lu, Z. Song, and J. Zhou, "Recurrent semantic preserving generation for action prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 231–245, 2020.

[24] G. Li, N. Li, F. Chang, and C. Liu, "Adaptive graph convolutional network with adversarial learning for skeleton-based action prediction," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 1258–1269, 2021.

[25] R. Wang, J. Liu, Q. Ke, D. Peng, and Y. Lei, "Dear-net: Learning diversities for skeleton-based early action recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 1175–1189, 2021.

[26] L. G. Foo, T. Li, H. Rahmani, Q. Ke, and J. Liu, "Era: Expert retrieval and assembly for early action prediction," in *Proceedings of Computer Vision*, 2022, pp. 670–688.

[27] C. Liu, X. Zhao, Z. Yan, Y. Jiang, and X. Shi, "A graph convolutional network with early attention module for skeleton-based action prediction," in *Proceedings of the International Conference on Pattern Recognition*, 2022, pp. 1266–1272.

[28] X. Wang, W. Zhang, C. Wang, Y. Gao, and M. Liu, "Dynamic dense graph convolutional network for skeleton-based human motion prediction," *IEEE Transactions on Image Processing*, vol. 33, pp. 1–15, 2023.

[29] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised representation learning with long-term dynamics for skeleton based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[30] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9628–9637.

[31] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3d human action representation learning via cross-view consistency pursuit," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4739–4748.

[32] Y. Yang, G. Liu, and X. Gao, "Motion guided attention learning for self-supervised 3d human action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8623–8634, 2022.

[33] Y. Hua, W. Wu, C. Zheng, A. Lu, M. Liu, C. Chen, and S. Wu, "Part aware contrastive learning for self-supervised action recognition," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 855–863.

[34] Z. Jin, Y. Wang, Q. Wang, Y. Shen, and H. Meng, "Ssrl: Self-supervised spatial-temporal representation learning for 3d action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 274–285, 2024.

[35] C. Pang, X. Lu, and L. Lyu, "Skeleton-based action recognition through contrasting two-stream spatial-temporal networks," *IEEE Transactions on Multimedia*, vol. 25, pp. 8699–8711, 2023.

[36] M. Liu, K. Liang, Y. Zhao, W. Tu, S. Zhou, X. Gan, X. Liu, and K. He, "Self-supervised temporal graph learning with temporal and structural intensity alignment," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[37] Y. Mo, Y. Chen, Y. Lei, L. Peng, X. Shi, C. Yuan, and X. Zhu, "Multiplex graph representation learning via dual correlation reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12 814–12 827, 2023.

[38] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding, "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 762–770.

[39] J. Zhang, L. Lin, and J. Liu, "Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3427–3435.

[40] Q. Zeng, C. Liu, M. Liu, and Q. Chen, "Contrastive 3d human skeleton action representation learning via crossmoco with spatiotemporal occlusion mask data augmentation," *IEEE Transactions on Multimedia*, vol. 25, pp. 1564–1574, 2023.

[41] R. Guo, J. Sun, C. Zhang, and X. Qian, "A self-supervised metric learning framework for the arising-from-chair assessment of parkinsonians with graph convolutional networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6461–6471, 2022.

[42] X. Gao, Y. Yang, Y. Zhang, M. Li, J.-G. Yu, and S. Du, "Efficient spatio-temporal contrastive learning for skeleton-based 3-d action recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 405–417, 2021.

[43] Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan, "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 1819–1831, 2022.

[44] N. Sheng, Y. Wang, L. Huang, L. Gao, Y. Cao, X. Xie, and Y. Fu, "Multi-task prediction-based graph contrastive learning for inferring the relationship among lncrnas, mirnas and diseases," *Briefings in bioinformatics*, vol. 24, no. 5, p. bbad276, 2023.

[45] Y. Zhang, Y. Liu, Y. Xu, H. Xiong, C. Lei, W. He, L. Cui, and C. Miao, "Enhancing sequential recommendation with graph contrastive learning," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022, pp. 2398–2405.

[46] X. Cai, C. Huang, L. Xia, and X. Ren, "LightGCL: Simple yet effective graph contrastive learning for recommendation," in *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

[47] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010–1019.
[48] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.
[49] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," *arXiv preprint arXiv:1703.07475*, 2017.
[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.
[51] Q. Ke, J. Liu, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Global regularizer and temporal-aware cross-entropy for skeleton-based early action recognition," in *Proceedings of Asian Conference on Computer Vision*, 2019, pp. 729–745.
[52] Y.-B. Cheng, X. Chen, J. Chen, P. Wei, D. Zhang, and L. Lin, "Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2021, pp. 1–6.
[53] J. Chen, Z. Jin, Q. Wang, and H. Meng, "Self-supervised 3d behavior representation learning based on homotopic hyperbolic embedding," *IEEE Transactions on Image Processing*, vol. 32, pp. 6061–6074, 2023.
[54] W. Wu, Y. Hua, C. Zheng, S. Wu, C. Chen, and A. Lu, "Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops*, 2023, pp. 224–229.
[55] B. Xu, X. Shu, J. Zhang, G. Dai, and Y. Song, "Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 11 035–11 048, 2024.
[56] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition," in *Proceedings of the Asian conference on computer vision*, 2020.
[57] Y. Wei, K. Peng, A. Roitberg, J. Zhang, J. Zheng, R. Liu, Y. Chen, K. Yang, and R. Stiefelhagen, "Elevating skeleton-based action recognition with efficient multi-modality self-supervision," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6040–6044.