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: DOI10.1109/tip.2023.3328230, IEEE Transactions on Image Processing

Self-supervised 3D Behavior Representation Learning Based on Homotopic Hyperbolic Embedding

Jinghong Chen, Zhihao Jin, Qicong Wang, Hongying Meng, *Senior Member, IEEE*

***Abstract*—Behavior sequences are generated by a series of spatio-temporal interactions and have a high-dimensional nonlinear manifold structure. Therefore, it is difficult to learn 3D behavior representations without relying on supervised signals. To this end, self-supervised learning methods can be used to explore the rich information contained in the data itself. Context-context contrastive self-supervised methods construct the manifold embedded in Euclidean space by learning the distance relationship between data, and find the geometric distribution of data. However, traditional Euclidean space is difficult to express context joint features. In order to obtain an effective global representation from the relationship between data under unlabeled conditions, this paper adopts contrastive learning to compare global feature, and proposes a self-supervised learning method based on hyperbolic embedding to mine the nonlinear relationship of behavior trajectories. This method adopts the framework of discarding negative samples, which overcomes the shortcomings of the paradigm based on positive and negative samples that pull similar data away in the feature space. Meanwhile, the output of the network is embedded in a hyperbolic space, and a multi-layer perceptron is added to convert the entire module into a homotopic mapping by using the geometric properties of operations in the hyperbolic space, so as to obtain homotopy invariant knowledge. The proposed method combines the geometric properties of hyperbolic manifolds and the equivariance of homotopy groups to promote better supervised signals for the network, which improves the performance of unsupervised learning.**

***Index Terms*—spatio-temporal interaction, contrastive learning, Poincaré model, hyperbolic space, homotopic mapping.**

I. Introduction

3D behavior recognition is a challenging task in the field of computer vision, and many supervised methods have achieved good results. With the development of depth cameras, there are more and more large-scale datasets recording skeleton-based behavior sequences. However, labeling behavioral data requires sufficient prior knowledge, and manual labeling is often a heavy task. Recently, it is found that in self-supervised learning, the representation of data in the feature space often has a certain distribution law [1–8]. Therefore, learning discriminative representations from unlabeled skeleton sequence data through contrastive self-supervised learning is a possible solution.

J. Chen, Z. Jin and Q. Wang are with the Shenzhen Research Institute, Xiamen University, Shenzhen 518057, China, and the Department of Computer Science, Xiamen University, Xiamen 361005, China (e-mail: qcwang@xmu.edu.cn).

H. Meng is with the Department of Electronic and Electrical Engineering, Brunel University, London, UK, UB83PH (e-mail: hongying.meng@brunel.ac.uk).

Self-supervised learning is essentially a statistical method to capture valuable latent information from training samples with unknown labels. A behavior sequence records the spatial three-dimensional coordinate information corresponding to different times in the temporal domain, which can reflect the length of the movement time, the speed, the relative relationship between the spatial coordinates, the periodic change of the moving target, and the temporal context dependency. The data itself can provide more abundant knowledge than the label. Contrastive self-supervised learning have two learning methods: context-context self-supervision [1, 2, 9, 10], that uses a two-stream structure to fit two functions, where two function samples are mapped to two points in the feature space, and the distance is optimized through the contrast loss, so that similar samples are closer, and heterogeneous samples are far away; context-instance self-supervision [8, 11–14], where an auxiliary task is designed to generate transformed data from the original unlabeled data according to certain rules, and assign pseudo-labels as supervision signals. The network has certain prior knowledge and can be transferred to the target task to achieve better results. Self-supervision based on context-context comparison is a new method proposed in recent years. It finds a distribution that can effectively distinguish different types of data by directly comparing the representations between different samples, and only needs to enhance the data to allow the network to automatically extract immutability of data.

The context-context comparison method focuses on the comparison between global features of the data. It firstly maps the data into a unified feature space, and uses the loss function to optimize the distance between features to obtain an ideal distribution. The latent features of the data are automatically mined by neural networks. Most of the current works use Euclidean space as a metric space [1, 2, 9, 10]. However, for some high-dimensional data, in the structured input, different variables are not independent of each other. Skeleton sequence has typical nonlinear characteristics. It records the macroscopic response of target behavior, and contains information in both temporal and spatial dimensions. There is a strong correlation between different variables. This nonlinear relationship makes the data located in a manifold of a high-dimensional space. Since the manifold where the data is located is often not a vector space and does not satisfy the Euclidean axioms, the representation based on the Euclidean space cannot accurately capture this relationship of the data, so mining the internal non-Euclidean properties is of great significance to understand

the behavioral characteristics of the target [15–17]. Intuitively, behavior data has long dependencies on temporal domain. Our method in this paper conducts context-context contrastive self-supervision based on global representation of behavior sequences, and uses manifold as the representation space for global features of the data.

Hyperbolic manifold [18] is a Riemannian manifold with constant negative Gaussian curvature, it has the ability to efficiently model hierarchical structures [16, 17]. In this manifold space, an important intrinsic property is exponential growth, so data with tree-like hierarchical structure can be embedded into this space naturally with very low distortion. For skeleton sequences with topological structures, different joints may have tree-like relationships in space, and different skeleton sequences may contain many similar poses. This overlapping information within the data makes the data have spatio-temporal entailment relationship, which makes the relationship between features has a non-parallel hierarchical relationship, and hyperbolic manifolds can well mine such hierarchical relationships. To this end, we choose the hyperbolic space as the representation space, use the self-organizing ability of the hyperbolic space, organize the data in the hyperbolic space in a hierarchical structure according to the dependencies between the features, and guide the network with the idea of refined feature to perform comparative learning more efficiently.

In the comparative learning of behavior data, an encoder with spatio-temporal feature extraction ability can convert data with spatio-temporal interaction into trajectories in the feature space, and finally obtain the final representation through pooling [19, 20]. For this, the concept of homotopy is introduced in this paper to analyze the homotopy relationship between different spatio-temporal data. In the feature space, trajectories with homotopy relationship can be converted to each other through continuous functions, so the homotopy relationship can reflect the similarity between different trajectories.

In the self-supervised learning framework based on the comparison of positive and negative samples, since label knowledge is not introduced, similar samples are used as negative samples that will affect the learning effect. The BYOL framework [10] overcomes this difficulty to a certain extent. The framework discards the encoding of negative sample stored in the storage dictionary, and directly uses the nonlinear mapping with batch normalization to allow the network to implicitly compare the data without negative samples, so as to effectively distinguish the data. However, this makes the framework lose the ability to compare the representations of different samples. To this end, our method not only focuses on the hyperbolic representation of the global features of the sample, but also considers how to obtain a more reasonable representation in the hyperbolic space according to the homotopy relation. As an equivalence relation, the homotopy involves the knowledge of group theory. Equivalence relation can divide elements into multiple equivalence classes, and our method serves as a guide to express the global features of behavioral data in hyperbolic space, and their embedding points in hyperbolic space can be regarded as representations of equivalence classes so as to keep away from samples that do not have homotopy relations, and retain more homotopy invariant properties. In addition, we study the equivalence properties of homotopy trajectories and group combination that further supports the proposed homotopy view through theory.

The main contributions of this paper lie in three aspects:

- In order to fully exploit the non-parallel relationship of motion sequences with spatio-temporal interaction, we embed the target sample as a separate entity in the hyperbolic space, and use the Poincaré model to vectorized representation of samples that enable the distance between samples to reflect the similarity more accurately.
- We introduce the feature homotopy deformation module and use the Mobius addition in the Poincaré model to guide the network to construct a homotopy transformation. This module combines the hyperbolic manifold and homotopy transformation, infers the homotopy relationship of the samples by using the similarity of the spatio-temporal features in the temporal dimension, adaptively adjusts the distance between the sample pairs and mines the homotopic immutable knowledge among data.
- Experiments are conducted on three large human skeleton sequence datasets, and the results show that the proposed method outperforms the state-of-the-art methods on some evaluation metrics of the dataset without explicitly classifying the samples with positive and negative samples demonstrating its effectiveness.

II. Related work

Contrastive Self-Supervised Methods Based on Context-Context

The context-context contrastive self-supervised method mainly studies the relationship between the global representations of different samples. The current mainstream methods mainly use different views of a sample as positive samples, other samples as negative samples, and make positive samples closer each other while far away from negative samples in the feature space through the loss function. This method naturally has the problem that the solution space is easy to collapse. Therefore, a solution is to obtain the positive and negative samples by two parameter-sharing encoders [21]. A key problem in these methods is that each round of backpropagation causes the parameters of the network to change. This causes that different batches of learned negative samples are obtained from networks with different parameters, affecting the consistency of the negative sample features. In order to solve this problem, He et al. [1] proposed the MoCo framework, which adopts the momentum update method. Instead of updating the parameters of negative sample encoder by gradient propagation, it uses a storage dictionary to save its encoded results as a negative samples. In this way, it iterates at a very low speed. Each new representation entered into the queue is the output of the encoder updated in the previous step, which is as consistent as possible with the old representation. Subsequently, Chen et al. [22] proposed the MoCo V2 framework, adding the same nonlinear multilayer perceptron to the representation of the encoder during training, and using cosine decay instead of step decay to further improve performance. However, the

above methods simply treat different samples as different categories, and do not consider the correlation between them, which limits the performance. Pan et al. [23] used MoCo for unsupervised video representation to improve the temporal feature representation of MoCo from two perspectives. First, they introduced a generator to temporarily remove several frames from this sample. The discriminator is then learned to encode similar feature representations without regard to frame removal. Second, we use temporal decay to model the decay of keys in the memory queue when computing contrast loss.

For the representation learning based on positive and negative samples, it is difficult to obtain the optimal encoding of negative samples by using the momentum update method. Then siamese network is used for competitive learning, which can retain more information related to the data. Chen et al.[2] proposed the SimCLR framework, which uses the siamese network to encode the paired data obtained through data enhancement. To add more negative samples for calculation, SimCLR uses larger-scale batches to improve the effect of representation learning, but it also increases time and memory consumption. Subsequently, in order to solve this problem, the Facebook AI Research (FAIR) and the Institut National de Recherche en Informatique et en Automatique (INRIA) [3] launched a multi-view clustering exchange method. Instead of using a large number of negative samples, all kinds of samples are clustered, and then the clusters of each type are distinguished. However, in this method, artificially setting the clustering center lacks versatility. For different data, the optimal number of clusters are often different, and a large number of parameter adjustment experiments are required for different data. Therefore, there is still a huge room for improvement in self-supervised learning methods that discard positive and negative samples.

Grill [10] proposed a BYOL framework to guide the network's own potential. For the first time, they boldly abandoned the comparison between different data in the traditional method based on the dual-encoder structure, and improved the prediction result through the iteration of target network parameters. The framework proposes the concept of target network and online network, and updates the target encoder based on the momentum update method. Only by adding a nonlinear multi-layer perceptual layer, the encoding results are converted into features, and a batch normalization layer is added to the data. And they proved experimentally in [24] that batch normalization is not the key to BYOL success, not as providing implicit negative sample information. In the experiments when both encoder and Projector were without BN, SimCLR also failed, proving that BN is not providing an implicit negative sample, even if an explicit negative sample is given it is still not trained. Finally, it was agreed that the main role of BN is to improve the robustness of model training, resulting in no model collapse. The authors found that the BN in the encoder is crucial, and that the BN compensates for the effects of bad initialization. For this reason, the authors propose a new initialization method, which turns out to be much better than random, although not as good as the best results that can be achieved by BYOL. In conclusion, BN brings benefits only in terms of scaling parameters and stabilizing the training process, which is very important for BYOL. Subsequently, Chen et al. [25] continued the idea of guiding self-potential and conducted research on the Siamese network, and found that stopping the gradient backpropagation was the key to avoid the collapse of the solution space. The proposed simple Siamese network also achieved good results, but not yet surpassing the framework of self-potential guidance. In the past, self-supervised learning methods learned video representations by video playback speed prediction, however, the learned models may tend to focus on motion patterns and it is not easy to obtain accurate speed labels for videos. Chen et al. [26] propose a new approach to perceive playback speed and use the relative speed between two video clips as labels. In this way, the speed is well perceived and better motion features are learned.

Recently, the self-supervised training paradigm of masking-and-reconstruction has been successful in natural language processing and image understanding. Tong et al. [27] chose to use a sampling strategy with temporal interval for more efficient self-supervised pre-training of video and used a pipelined masking strategy with a very high mask ratio to obtain better video understanding network.

Self-Supervised Representation Learning Based on Skeleton Sequences

Compared to video sequences, human skeleton-based sequences increase the difficulty of self-supervised feature extraction due to the spatial topology. Though existing supervised learning methods [28–32] have been proposed and achieved satisfing results, considering the cost of labeled data the research on self-supervised methods became more and more significant. Su et al [4]. proposed a representation learning method based on prediction and clustering, which encodes and decodes the sequence with the GRU network. It encodes temporal features, and uses kNN clustering method to cluster the features obtained during the encoding process. Context-context contrastive self-supervised learning is a new method proposed in recent years applied to 3D skeletons. Some works use self-supervised frameworks to learn 3D skeleton sequences with spatio-temporal structure, and achieve good results. Rao et al. [19] used a self-supervised learning framework for human-based spatio-temporal interaction data for the first time. The method was based on a momentum contrastive learning framework, adopted a long short-term memory network as an encoder, and used data augmentation based on human skeletal motion data. These include rotation, shearing, inversion, Gaussian noise, Gaussian blur, joint mask, and channel mask. Experiments show that different data enhancements have a significant impact on performance. Li et al. [20] proposed a multi-view 3D behavior self-supervised learning method based on the MoCo V2 framework. The key idea is to use the complementary information of skeleton data under different views (joint, bone, motion) to obtain better self-supervision signal. In addition, the spatio-temporal graph convolution is adopted as the encoder, which greatly improves the spatio-temporal feature extraction ability of 3D skeleton data. The above methods have achieved competitive

results with supervised learning on human behavior datasets, reflecting the superiority of graph convolution in processing 3D skeleton sequences with spatial structure. The drawback of their momentum-contrastive learning framework is that it needs a queue to store a large number of coding results as negative samples, which is very unfavorable for narrowing the distance between similar samples. Therefore, we adopt the method based on contrastive learning, discarding the positive and negative samples, starting from the own characteristics of the data, encoding the representation that retains more spatio-temporal information, and improving the self-supervised learning effect of high-dimensional spatio-temporal interactive data. Yang et al. [6] represented the skeletal action sequences as 3D skeleton clouds and colored each point in the clouds according to the spatio-temporal order in the unlabeled skeletal sequences, using colored skeletal point clouds to effectively learn spatio-temporal features from the artificial color labels of skeletal joints. Kim et al. [33] devised a global and local attention mechanism in which global body movements and local joint movements attend to each other. A pre-training strategy for multi-zone posture displacement prediction is proposed to allow the model to estimate whole-body and joint movements at different time intervals and scales to learn global and local attention at different time scales. Guo et al. [34] introduced an extreme enhancement mechanism and an energy-based attention-guided descent module (EADM) to obtain a richer data enhancement strategy, further extended the positive sample with double-distribution divergence minimization loss and nearest neighbor mining, and finally obtained a high-quality action representation. Zhang et al. [35] designed a progressive growth augmentation strategy to generate multiple ordered positive pairs to achieve consistency in learning representations from different perspectives, and enhanced hierarchical consistency by directed clustering operations in the feature space to make the representation of the strongly augmented view closer to that in the weakly augmented view. Yang et al. [36] designed a two-steam pretraining network that utilizes both fine-grained and coarse-grained coloring to learn multi-scale spatio-temporal features. And the designed autoencoder framework is pre-trained to learn information representation through a masked skeleton cloud redrawing task.

However, the above work is limited to characterizing the data on the Euclidean space, ignoring the gestalt correlation and repetition exhibited by the action data.

Homotopic Feature Extraction Method

Some existing works [37, 38] have studied algorithms based on homotopic learning. The goal of these methods is to extract effective features that remain invariant to uncorrelated continuous deformations through homotopic transformation. The homotopy relationship provides a certain theoretical explanation for the relationship between trajectories, and the concept of homotopy equivalence has been applied to robot motion planning [39]. Since motion sequences are also high-dimensional spatio-temporal interaction data, it is natural to consider them as homotopic curves in the feature space, and the goal of self-supervised learning is to find a point as its equivalence class. The loss function is crucial for self-supervised contrastive learning, and homotopy analysis also plays a role in the definition of the loss function. Shit et al. [40] proposed a new topology-preserving loss function, that was achieved in the tubular structure segmentation task with excellent results. Therefore, homotopy analysis helps to define better measures of feature similarity. In addition, in the similarity calculation of contrastive learning, it is often necessary to firstly standardize the coding representation, so that it is in a high-dimensional unit sphere, and then calculate the similarity. The encoding representation and similarity can be used as the direction vector and norm, which are just embedded in the Poincaré model of the hyperbolic manifold.

TABLE I
NOTATIONS AND DEFINITIONS

Notations	Definitions
B_θ	input data of the online network
B_ξ	input data of the target network
f_θ	the encoder of the online network
f_ξ	the encoder of the target network
$H(\cdot,\cdot)$	the homotopy function
x_θ	encoded representation after spatial pooling
x_ξ	target representation after spatial pooling
y_θ	encoded representation after global pooling
y_ξ	target representation after global pooling
$\mathrm{sim}(\cdot,\cdot)$	similarity of two feature
$\mathrm{mean}(\cdot)$	average
$g_\theta(\cdot)$	projection function of online network
$g_\xi(\cdot)$	projection function of target network
$g'_\xi(\cdot)$	projection function of homotopy deformation
z_θ	projection result of online network
z_ξ	projection result of target network
z'_ξ	homotopic projection result
p	similarity between target and online representation
p'	similarity between homotopy and online representation
z^h_ξ	hyperbolic embedding result of target network
$\mathcal{L}$	contrast loss of target and online representation
$\mathcal{L}'$	contrast loss of homotopy and online representation

III. PROPOSED METHOD

This session firstly describes the theoretical basis, namely the hierarchical relational feature representation based on the Poincaré model, and then describes and explains the added innovative modules, namely, homotopy projection learning based on hyperbolic embeddings, and groups invariance analysis based on homotopy equivalence, and finally describes the entire network, that is, a high-dimensional motion trajectory self-supervised network based on homotopy equivalence classes. The overall framework of our network is shown in Fig.1. For a given skeleton sequence, a pair of input data $B_\theta, B_\xi \in \mathcal{R}^{C\times V\ \times T}$ is obtained through data enhancement, where C, V, and T represent the number of channels, joints and frames of the sample respectively. The human joint sequence is treated as a spatio-temporal graph, and a spatio-temporal graph convolutional network is used to fuse features in each frame. For the temporal dimension, ordinary convolutions are used to fuse adjacent frames. Table I shows the notations and corresponding definitions.

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: DOI10.1109/tip.2023.3328230, IEEE Transactions on Image Processing

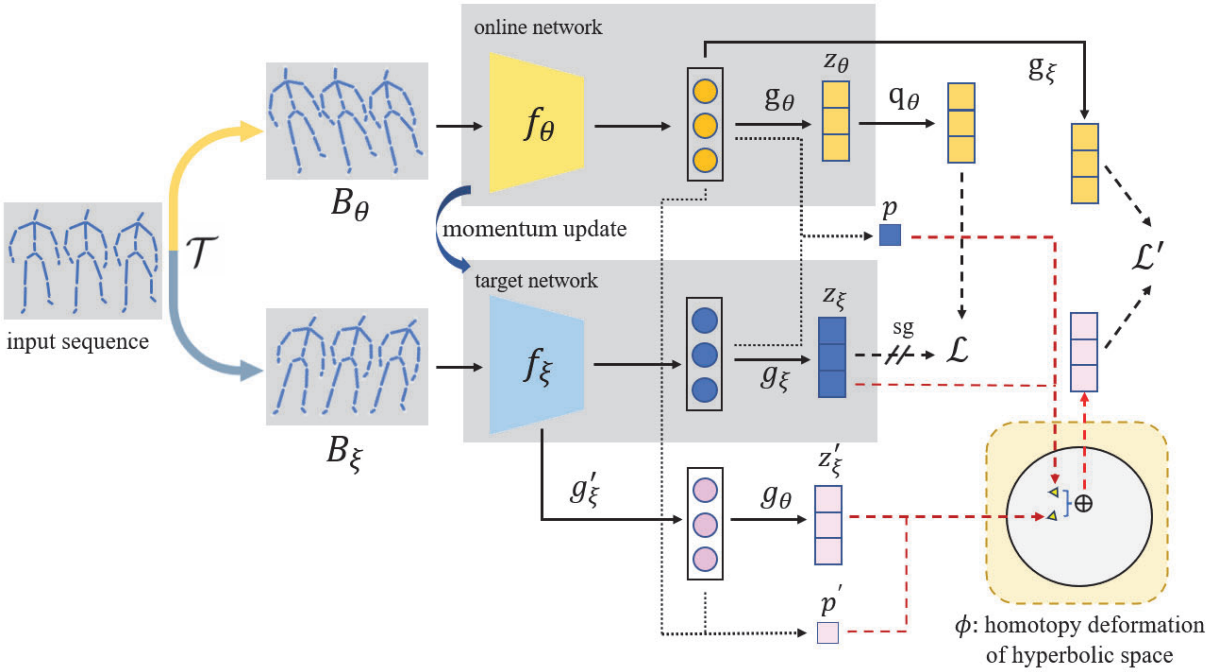


Fig. 1. The framework of self-supervised behavior recognition network based on homotopy hyperbolic embedding. The model uses a siamese structure with ST-GCN as backbone network. Augmented samples are fed into the prediction network (top) and the target network (bottom) to obtain two representations, which are output to the same feature space for prediction after a projection transformation. The prediction network updates the parameters by back-propagation, while the target network is updated by momentum. In the target network, we additionally introduce a network to homotopy the original representation. The homotopy deformation is based on hyperbolic embedding. Using the normalized output of the spatio-temporal features as directions, we compute similarities on a frame-by-frame basis for the sequence representations, taking the average as the parametric, and embedding the output features into hyperbolic space.

A. Behavior sequence self-supervised learning network

In order to make the distribution of samples in the projection space contain more temporal information, we design a self-supervised learning network based on high-dimensional motion sequence of homotopy equivalent classes, as shown in Fig.1.

The goal of the self-supervised network is to find a mapping function without relying on the real labels of the data, and use this function to map different high-dimensional motion sequences into a certain feature space. The obtained space points can represent the base points of different homotopy equivalence classes, so as to improve the performance of downstream classification tasks.

We use a contrastive learning framework to learn the mapping function, which consists of two sub-networks, an online network that updates parameters through back propagation, and a target network that discards gradients. After the gradient is updated, the parameters of the target network are updated with the new parameters through momentum. ST-GCN[30] is used as the backbone network, in which f_θ is used for the encoder branch, f_ξ is used for the target branch. For the obtained output, the base point of the homotopy equivalent class is obtained by calculating the average value. These two points in the feature space are transformed to the same space by projection layers (MLP) $g_\theta(\cdot)$, $g_\xi(\cdot)$. The online network is connected to a multilayer perceptron $q_\theta(\cdot)$ to make the output approximate the target network, and MSE loss is used to learn the network:

$$\mathcal{L}_{\theta,\xi} = \left\| \frac{q_\theta\left(z_\theta\right)}{\left\|q_\theta\left(z_\theta\right)\right\|_2} - \frac{z'_\xi}{\left\|z'_\xi\right\|_2} \right\|_2^2 = 2 - 2 \cdot \frac{\left\langle q_\theta\left(z_\theta\right), z'_\xi \right\rangle}{\left\|q_\theta\left(z_\theta\right)\right\|_2 \cdot \left\|z'_\xi\right\|_2} \tag{1}$$

In order to force the model to learn deeper information, we introduced high-dimensional spatio-temporal sequence homology relations, explored the retention of different sequence homology relations in the learning process of manifold representation by spatio-temporal interaction data, and proposed a hyperbolic manifold embedding method based on homology mapping to achieve the extraction of homology invariant features of spatio-temporal sequences. After the output of the encoder is averagely pooled, the temporal dimension information is lost. In order to utilize the similar information of each frame of two outputs, we use a multi-layer perceptron and a average function $g'_\xi(\cdot)$ to learn the homotopy transformation, and project the features to the Poincaré model using the temporal similarity of the two representations as a norm. The high-dimensional motion sequence is not linear in space, and the base points obtained by the average calculation cannot accurately represent the equivalent class. The Mobius summation of the Poincaré model is used to represent the representative point of the sequence in the hyperbolic space:

$$z_\xi^h = \frac{\lambda \cdot g_\xi\left(y_\xi\right)}{\left\|g_\xi\left(y_\xi\right)\right\|_2} \oplus \frac{\lambda' \cdot g_\theta \circ g'_\xi\left(y_\xi\right)}{\left\|g_\theta \circ g'_\xi\left(y_\xi\right)\right\|_2} \tag{2}$$

Among them, λ and λ' are the similarity between the homotopy and the encoding end, respectively. We use the projection layer of the target network to project the output of the encoder to obtain y_θ^h, and use MSE to calculate the loss [10]:

$$\mathcal{L}'_{\theta,\xi} = 2 - 2 \cdot \frac{\left\langle z_\theta^h, z_\xi^h \right\rangle}{\left\| z_\theta^h \right\|_2 \cdot \left\| z_\theta^h \right\|_2} \tag{3}$$

Adding $\mathcal{L}_{\theta,\xi}$ (the MSE loss in Euclidean space) to $\mathcal{L}'_{\theta,\xi}$ (the MSE loss in hyperbolic space), the final loss function is defined as:

$$\mathcal{L} \triangleq \mathcal{L}_{\theta,\xi} + \tilde{\mathcal{L}}_{\theta,\xi} + \lambda(\mathcal{L}'_{\theta,\xi} + \tilde{\mathcal{L}}'_{\theta,\xi}) \tag{4}$$

where λ controls the loss proportion of the hyperbolic manifold and $\tilde{\mathcal{L}}$ represents the symmetric form of the loss function, i.e., the cross predicts the output at the other end. In the model prediction stage, only f_θ on the encoder side is reserved for representation learning.

B. Hierarchical relation feature representation of Poincaré model

Hyperbolic manifold is a Riemannian manifold with constant negative Gaussian curvature, and five isometric models are given in literature [18]. We choose the Poincaré model as the isometric model and embed each sample as an instance in this space. Without loss of generality, a space with a curvature of -1 is choosed, and the n-dimensional Poincaré sphere model is used as the feature space.

The two enhanced skeleton sequences are encoded into high-dimensional feature vectors $z_\theta, z_\xi \in \mathcal{R}^n$. In order to calculate a direction vector r and a norm λ, the feature vector is firstly unitized:

$$\bar{z}_\theta = \frac{z_\theta}{\left\| z_\theta \right\|_2}, \bar{z}_\xi = \frac{z_\xi}{\left\| z_\xi \right\|_2} \tag{5}$$

After obtaining the unit vector, we define $\bar{z}_\xi$ as the norm, calculate its similarity in the form of inner product. At this time, the sphere center distance of the feature in the Poincaré model reflects the similarity of the two output vectors, and the embedding method is as follows:

$$\begin{cases} \lambda = \langle \bar{z}_\theta, \bar{z}_\xi \rangle \\ r = \bar{z}_\xi \end{cases} \tag{6}$$

where $\lambda \in [0, 1]$, the parametric Poincaré space embedding is defined as $x = \lambda r$ in the n-dimensional unit sphere $\mathcal{D}^n$.

It can be seen from the parameterization method that if the output is more similar to the output of the other encoder, its norm is closer to 1. In self-supervised learning, we hope that the network pays more attention to the samples that are difficult to approach, and the similarity of the samples is determined by the direction vector. Therefore, it is possible to correct the direction of the feature vector of the samples that are easy to zoom in, so as to zoom in on the samples that have the same semantics, but are far apart. Defining $\oplus$ as the Möbius summation on the Poincaré model [15], letting $\eta, \xi \in \mathcal{D}^n$, then the summation operation is defined as:

$$\eta \oplus \xi = \frac{\left(1 + 2\langle \eta, \xi \rangle + \|\xi\|^2\right)\eta + \left(1 - \|\eta\|^2\right)\xi}{1 + 2\langle \eta, \xi \rangle + \|\eta\|^2\|\xi\|^2} \tag{7}$$

It can be seen from this definition that the operation does not satisfy the commutative law. Since the Poincaré model is a conformal model with conformal properties, the proposed method calculates the similarity after uniting it, so its norm can be ignored. In this formula, the two coefficients are respectively $1 + 2\langle \eta, \xi \rangle + \|\xi\|^2$ and $1 - \|\eta\|^2$. Therefore, if the norm of η and ξ is closer to 1, the degree of similarity is higher, the summation result is closer to η, on the contrary, it is closer to ξ. Using this property, in contrastive learning, the similarity between the vectorized output of one enhanced sample and that of another sample is calculated, and its value is used as norm of the feature in the embedding space. In equation 3, η is the vector that is expected to be close, and ξ is the vector that is expected to be far away, the result $\eta \oplus \xi$ can adaptively adjust the original direction.

C. Homotopic projection learning based on hyperbolic em-bedding

1) The basic concept of homotopic projection: In real-world scenarios, samples exhibiting high similarity are often indicative of belonging to the same underlying category. However, due to the absence of explicit category labels as supervised signals, the network will pull the distance between samples belonging to the same category. Consequently, this can pose challenges in effectively clustering samples of the same category within the representation space solely based on their category attributes, leading to potential performance degradation in unsupervised learning tasks. In contrast, homogeneous deformation is a continuous deformation that does not destroy the topology of the data, and if the network can mine invariants from homogeneous deformation, it can find the prior knowledge which is provided by the data itself. In this regard, this paper investigates the ability to obtain a good representation through the self-improvement capability of the network without using negative samples, and to make the output representations available as representative elements for homotopy classification of the data, using the ability of the network to fit continuous functions, transforming the feature sequences into the same space using homotopy map-ping, and finally embedding them in hyperbolic space. Since the features are transformed into the same space, different representations have isoren relations, thus forming different equivalence classes. Using this property can make the network adaptively establish the association between different repre-sentations to obtain a better approximation of the sequence isoren class of representative elements, thus achieving the purpose of improving the unsupervised learning performance. In order to choose robust features with obvious distinguishing significance, that is, features are easy to extract, invariant to relative deformation, and are insensitive to noise, we use the homotopy trajectory curve to find the homotopy invariant features of the sample. For the skeleton sequence dataset, each serialized sample is composed of different poses. These poses

are described by coordinate positions, which are often in a low-dimensional submanifold of a high-dimensional space, and the entire sample can be regarded as a discrete curve in manifold space.

Firstly, the definition of homotopy is given here. X and Y are assumed two topological spaces, f and g are continuous mappings of X to Y. If there is a continuous mapping $\mathrm{H}: X \times I \rightarrow Y$ such that $\forall x \in X$, $\mathrm{H}(x,0) = f(x)$, $\mathrm{H}(x,1) = g(x)$, then H is called to be a homotopic mapping that transforms f into g, where $I = [0,1]$. The skeleton sequence data is described as a time series composed of three-dimensional coordinates of joints, where each frame describes an action pose, and the constraint relationship between the coordinates makes the vector representation of each frame of high-dimensional data in a manifold of a high-dimensional space, so each sample can be approximately treated as a discrete curve in a manifold space. The input of the network is two trajectory curves, which are assumed to be continuous functions of the manifold space defined on the temporal dimension, which are projected as points in the manifold space by the encoder f_θ, f_ξ, and then projected to the same space using the projection function g_θ, g_ξ. Among them, g_ξ and f_ξ use momentum to update parameters. g_ξ approximately converts nonlinear g_θ into multiple linear problems. Therefore, the function fitted by the projected multilayer perceptron (MLP) at the target side is very dependent on the encoder side. We study the method of overcoming this dependency to approximate the high-order approximation of the projection function through homotopy analysis, so that the network can learn homotopy invariant features and improve the accuracy of the solution. The proposed method intends to construct a function $\mathrm{Hy}(x;p)$, and use the multilayer perceptron to learn this function.

If $p \in \mathcal{D}^n$ is the hyperbolic embedded variable, where $\|p\| \in [0, 1)$, construct the real function $\mathrm{Hy}(x; p)$ to be the homotopy function of the hyperbolic space [41], and the homotopy function is the relationship between the hyperbolic space and the projection function

$$\mathrm{Hy}(x;p) \triangleq \|p\| g_\theta + (1 - \|p\|) g_\xi \tag{8}$$

When $\|p\| = 0$, $\mathrm{Hy}(x;0) = g_\xi$, and $\lim_{\|p\| \rightarrow 1} \mathrm{Hy}(x;p) = g_\theta$. When $\|p\|$ changes from 0 to 1, the function $\mathrm{Hy}(x;p)$ changes continuously from g_ξ to g_θ. This continuous change is called homotopy, which is represented as $\mathrm{Hy}(x;p) : g_\theta \sim g_\xi$, that is, the solution of function g_θ and function g_ξ is homotopy.

2) Constructing homotopic functions based on hyperbolic embedding of behavior sequences: In order to extract more temporal information of the homotopy data, the output representation of the target end is input into the homotopy module proposed here, as shown in Fig.1. This module embeds the target end output into the hyperbolic space based on the similarity of the pose of each frame. In this process, a homotopy deformation of the target side multilayer perceptron is constructed. The input data of these two networks $B_\theta, B_\xi \in \mathcal{R}^{C \times V \times T}$ are spatially pooled to obtain $x_\theta, x_\xi \in \mathcal{R}^{C \times T}$, and the global pooling is performed to obtain $y_\theta, y_\xi \in \mathcal{R}^{C}$, then the feature representations are L_2 unitized in the channel dimension to obtain $\bar{y}_\theta$, $\bar{y}_\xi$ and $\bar{x}_\theta$, $\overline{\mathcal{X}}_\theta$.

In this paper, we use a multi-layer perceptron $h(\cdot)$ that has the same structure as g_ξ and learn parameters through gradient propagation. We perform the same transformation on each frame of output x_ξ to obtain output x'_ξ, and perform L_2 unitization to get $\bar{x}'_\xi$. At this time, the target end has two unitized output representations compared to the representations of the encoding end. The similarity is calculated [2] and averaged, and two results p and p' are obtained:

$$\begin{cases} p = \mathrm{sim}\left(\bar{x}_\theta, \bar{x}_\xi\right) = \mathrm{mean}\left(\bar{x}_\theta^T \cdot \bar{x}_\xi\right) \\ p' = \mathrm{sim}\left(\bar{x}_\theta, \bar{x}'_\xi\right) = \mathrm{mean}\left(\bar{x}_\theta^T \cdot \bar{x}'_\xi\right) \end{cases} \tag{9}$$

Similarly, the symmetrical forms $\tilde{p}$ and $\tilde{p}'$ can be obtained by exchanging the input enhanced samples. Here, the multi-layer perceptron h is regarded as a function that converts the target representation to the encoding representation, and the projection layer after the two networks is exchanged at the same time. We average x'_ξ in the time dimension to obtain y'_ξ, use $g_\theta(\cdot)$ to project and unitize y'_ξ to obtain the final output $\bar{z}'_\xi$, and use the similarity to embed the output into the hyperbolic space. The similarity p' is the comparison result between the changed target end representation and the encoding end, where it can be taken as the norm, and multiplied by $\bar{z}'_\xi$ to obtain the embedding of the hyperbolic space:

$$s' = p' \bar{z}'_\xi \tag{10}$$

At the same time, taking the comparison result p outputted by the original target end and the encoding end as the norm, and the normalization of the original projection result as the direction vector, we get:

$$s = p \bar{z}_\xi \tag{11}$$

Finally, as shown in the lower right corner of Fig.1, the embeddings of the two Poincaré models are summed through a Möbius to obtain the final hyperbolic embedding:

$$z_\xi^h = s' \oplus s \tag{12}$$

Similarly, when another enhanced sample is input to this end, its symmetric form $\tilde{z}_\xi^h$ can be obtained. At this time, the original input of the other end becomes the current input, and the two hyperbolic embedding norms are symmetric forms of $\tilde{p}$ and $\tilde{p}'$.

From the definition of the summation symbol, it can be concluded that if the homotopy result is highly similar to the other end, that is, $p \rightarrow 1$, $p' \rightarrow 1$, the output direction is closer to s', and the target representation is inputted into the multi-layer perceptron h and averaged, the process is regarded as a function g'_ξ, the homotopy module approximates the composite form $g_\theta \circ g'_\xi(\cdot)$ of the function and the end-to-end projection function of the encoding. Otherwise, the output result is close to S_θ^h. At this time, the homotopy module is equivalent to an identity transformation, and the projection function remains unchanged, thus the whole homotopy module is the homotopy function between $g_\theta \circ g'_\xi(\cdot)$ and g_ξ.

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: DOI10.1109/tip.2023.3328230, IEEE Transactions on Image Processing

D. Group invariance analysis based on homotopy equivalent class

The purpose of homotopic mapping is to find homotopy invariant features of samples. This invariance often enables homotopy transformations to have good properties of groups, which can be automatically learned by the network through gradient propagation. Since the proposed network architecture adds an additional multi-layer perceptron $q_\theta(\cdot)$ to predict the projection result of the target end after the projection transformation at the encoding end, it also plays the role of dispersing the sample points, so that the data is more evenly distributed on the surface of the unit sphere. For the representation obtained by the encoding end, $q_\theta(\cdot)$ is also used, then the transformation from the encoder representation to the final output prediction is a composite form of $q_\theta \circ g_\theta$. For the loss function, the comparison with the original target projection is retained, that is, the solution:

$$\arg \max_{q_\theta, g_\theta, y_\theta} \cos \left\langle g_\xi \left(y_\xi\right), q_\theta \circ g_\theta \left(y_\theta\right)\right\rangle \tag{13}$$

Based on the loss function of hyperbolic homotopy, the required solution is as follows:

$$\arg \max_{y_\theta, s} \cos \left\langle g_\xi \left(y_\theta\right), s' \oplus s\right\rangle \tag{14}$$

where the direction vector of s' is from $g_\theta \circ h\left(y_\xi\right)$ and the norm is from $p' = \cos\left\langle x_\theta, x'_\xi \right\rangle$, while the direction vector of s is from y_ξ and the norm is from $p = \cos\left\langle x_\theta, x_\xi\right\rangle$. At this time, if the two projection functions are very similar, the representations at both ends need to be in similar spaces. It can be known from the property of Möbius summation that the above formula is equivalent to

$$\arg \max_{g_\theta, y_\theta} \cos \left\langle g_\theta \left(y_\theta\right), g_\theta \circ g'_\xi \left(y_\xi\right)\right\rangle \tag{15}$$

At this time, if a function g'_ξ can be found, and the transformation of each frame of the target representation can transform y_ξ into an approximation of y_ξ, then p and p' are both large values, so the network will automatically learn a homotopic map. If the representation similarity of both ends is very low, then solve:

$$\arg \max_{y_\theta} \cos \left\langle g_\xi \left(y_\theta\right), g_\xi \left(y_\xi\right)\right\rangle \tag{16}$$

That is, the network will automatically learn a y_θ that is close to y_ξ, which is equivalent to undergoing an identity transformation.

We define the inverse of the trajectory $\alpha : I \rightarrow X(I = [0, 1])$ as $\bar{\alpha}$ and specify it as $\bar{\alpha}(t) = \alpha(1-t)$. For two paths α and β on X, if $\alpha(1) = \beta(0)$ is satisfied, the product $\alpha\beta$ of them is stipulated as [41]:

$$\alpha\beta(t) = \alpha(2t), 0 \leq t \leq \frac{1}{2}, \alpha\beta(t) = \beta(2t-1), \frac{1}{2} \leq t \leq 1 \tag{17}$$

Therefore, for a closed loop passing through the base point, the loop that can be reduced to the base point is regarded as the unit element. Since the homotopy relation is an equivalence relation, it is obvious that the equivalence classes of closed loops passing through the base point form a group under the defined product.

For the spatio-temporal trajectory in the feature space, selecting a point in the space as the representative element, and the equivalence classes of all closed loops passing through the base point under the homotopy relation constitute a group structure.

IV. EXPERIMENTS AND ANALYSIS

A. Datasets

In order to verify the effectiveness of the proposed method, experiments are carried out on three large-scale human-based 3D skeleton datasets, namely NTU RGB+D 60 dataset, NTU RGB+D 120 dataset and PKU-MMD dataset, that are three largest datasets in the field of behavior recognition, on which it is more convincing to compare experimental results.

NTU RGB+D 60 [42]: This dataset is currently one of the largest 3D behavior recognition datasets, containing RGB+D videos and skeleton data for human action recognition. The data was captured from 40 human subjects by 3 Microsoft Kinect V2 cameras. There are 56880 samples with 4 million frames in 60 categories, and the maximum number of frames in all samples is 300. Each body skeleton records 25 joints. The original benchmark provides two evaluation methods, namely Cross-Subject (Xsub) and Cross-View (Xview) evaluation. In Xsub evaluation, the training set contains 40,320 videos from 20 subjects, and the remaining 16,560 videos are used for testing. In Xview evaluation, 37920 videos captured from No. 2 and No. 3 cameras were used for training, and the remaining 18,960 videos from No. 1 camera were used for testing. We follow these two benchmarks and report of Top-1 accuracy.

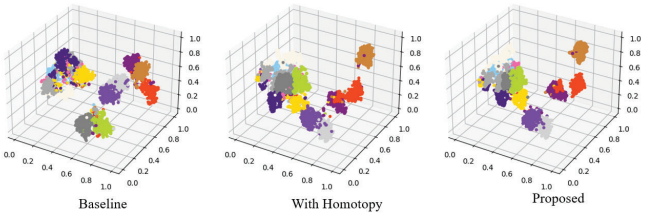
NTU RGB+D 120: This dataset is an extended version based on NTU RGB+D 60, adding 57,367 skeleton sequences in additional 60 action categories, totaling 113,945 samples, 120 action category categories, captured from 106 different subjects and 32 different cameras. Two evaluation criteria are used: Cross-subject (Xsub) and Cross-setting (Xset). In Xsub protocol, 63,026 samples from half of the participating subjects were used for training, while the remaining 50,919 samples were used for testing. In Xset evaluation, 54,468 samples taken from half of the camera devices are used for training and the remaining 59,477 samples are used for testing.

PKU-MMD [43]: This dataset is a large-scale multimodal 3D human behavior recognition dataset, which covers a wide range of complex human activity categories and has been manually annotated with labels. The dataset collected 1,076 long video sequences with 51 action categories, completed by 66 participating subjects under different perspectives of three Kinect V2 cameras, containing 21,545 behavior instances and a total of 5.4 million frames. The label of each long sequence marks the behavior category of each action instance, the start frame, end frame and label confidence of the action. The 51 behavior categories collected in this dataset can be divided into two types: 41 single-person behaviors and 10 two-person interaction behaviors. We choose the skeleton data of this dataset, and each frame consists of the 3D coordinates of

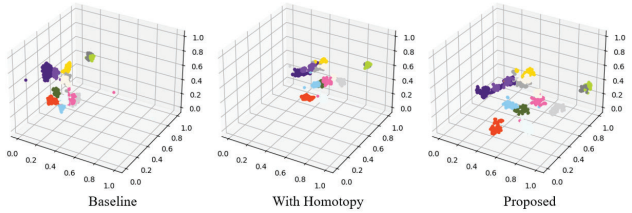
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/tip.2023.3328230, IEEE Transactions on Image Processing

TABLE II
ABLATION EXPERIMENT RESULTS OF DIFFERENT COMPONENTS

Method	Params	linear evaluation				fine-tune evaluation			
		NTU-60		PKU-MMD		NTU-60		PKU-MMD	
		Xsub	Xview	part I	part II	Xsub	Xview	part I	part II
Baseline	2.01M	75.9	77.9	80.4	40.6	82.2	89.0	87.0	52.8
With Homotopy	2.04M	77.0	78.8	85.1	48.1	83.8	90.6	90.5	54.9
Proposed	2.14M	**78.9**	**82.3**	**88.5**	**51.7**	**84.9**	**91.5**	**92.6**	**56.7**



(a) NTU-60 Xsub



(b) PKU-MMD part I

Fig. 2. The t-SNE visualization of embeddings of ablation experiments.

the 25 main body joints of the behavior participant. In the behavior recognition task, this dataset provides data of two different parts, among which part II introduces more skeleton noise, which brings more challenges.

B. Experimental Setup

The proposed method is implemented through the PyTorch deep learning framework. For data preprocessing, this paper follows the approach adopted in SkeletonCLR [20], no new data augmentation method is used, and the batch size is set to 128 during training.

The entire network framework adopts ST-GCN as the encoder. In order to ensure the fairness of the comparison, the conditions in the experiment are consistent, and the network structure and parameters of the encoder are consistent with the SkeletonCLR, in which the number of GCN layers is 5. For the problem of inconsistent frame numbers of action sequences, linear interpolation is used to unify all sample sequences into 50 frames. Stochastic Gradient Descent with momentum 0.9 and weight decay 0.0001 were used on network optimization. The model was trained for 400 epochs with a learning rate of 0.1, and no learning rate decay strategy was used during the learning process.

Linear Evaluation The model is validated by linear evaluation on an behavior recognition task. Specifically, a linear classifier (a fully connected layer followed by a softmax layer) is trained, supervised by a fixed encoder, and the final Top-1 classification accuracy is compared.

TABLE III
ABLATION EXPERIMENT RESULTS OF λ IN LOSS FUNCTION

λ	NTU-60		PKU-MMD	
	Xsub	Xview	part I	part II
0.05	78.2	81.8	87.8	51.2
0.1	**78.9**	82.3	88.5	**51.7**
0.2	78.0	**82.9**	**89.0**	48.8

Semi-Supervised Evaluation We pre-train the encoder with all the data, then fine-tune the entire model with five protocols of the randomly selected labeled data.

Fine-tune Evaluation We append a linear classifier to the trained encoder, then train the entire model and compare it to supervised methods.

C. Ablation Study

In this section, we verifies the effectiveness of different components in our proposed method through ablation experiments. In order to evaluate the hyperbolic space embedding module and the homotopy deformation module respectively, the following experiments are carried out for comparison: directly using the contrastive learning framework without additional modules (Baseline), adding the method of homotopy deformation (With Homotopy), and our proposed method combining homotopy deformation and hyperbolic embedding. The results under linear evaluation protocol and fine-tune protocol are shown in Table II.

It can be seen from the experimental results that the direct use of the BYOL framework without additional modules has a higher accuracy. When the homotopy deformation module is added, the effect of Xsub and Xview based on linear evaluation protocol is improved by 1.1% and 0.9% respectively, indicating that the homotopy relationship is closely related to the similarity of global features of the samples. The samples with homotopy relationship can be brought closer together in space, that can better gather similar samples and improve the classification effect. The method of combining homotopy deformation and hyperbolic embedding proposed in this paper embeds the global representation of samples into hyperbolic space for analysis on the basis of homotopy, which further improves the classification effect. The improvement is 1.9% and 3.5%, indicating that the representation of behavior sequence data satisfies the hypothesis of hyperbolic manifold in the feature space to a certain extent, that is, there is a certain hierarchical relationship, that is suitable for modeling with hyperbolic space to mine data. The hierarchical structure of the system can better improve its classification effect. Fig.2 shows the distribution of embeddings on NTU-60 xsub and PKU-MMD part I using the t-SNE algorithm [44]. It

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: DOI10.1109/tip.2023.3328230, IEEE Transactions on Image Processing

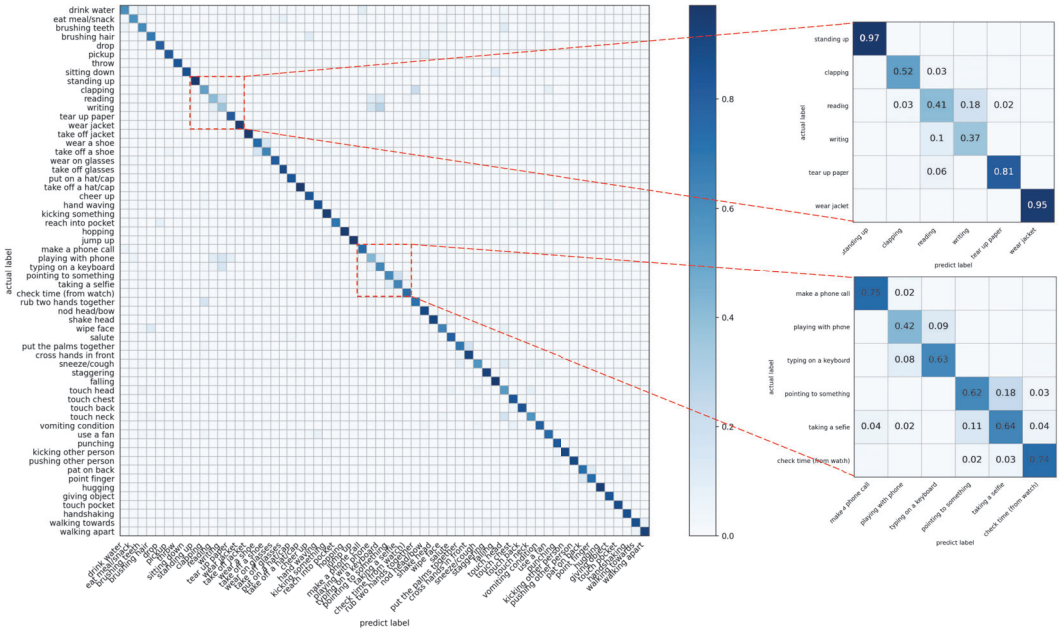


Fig. 3. Confusion matrix on NTU-60 under xsub setting.

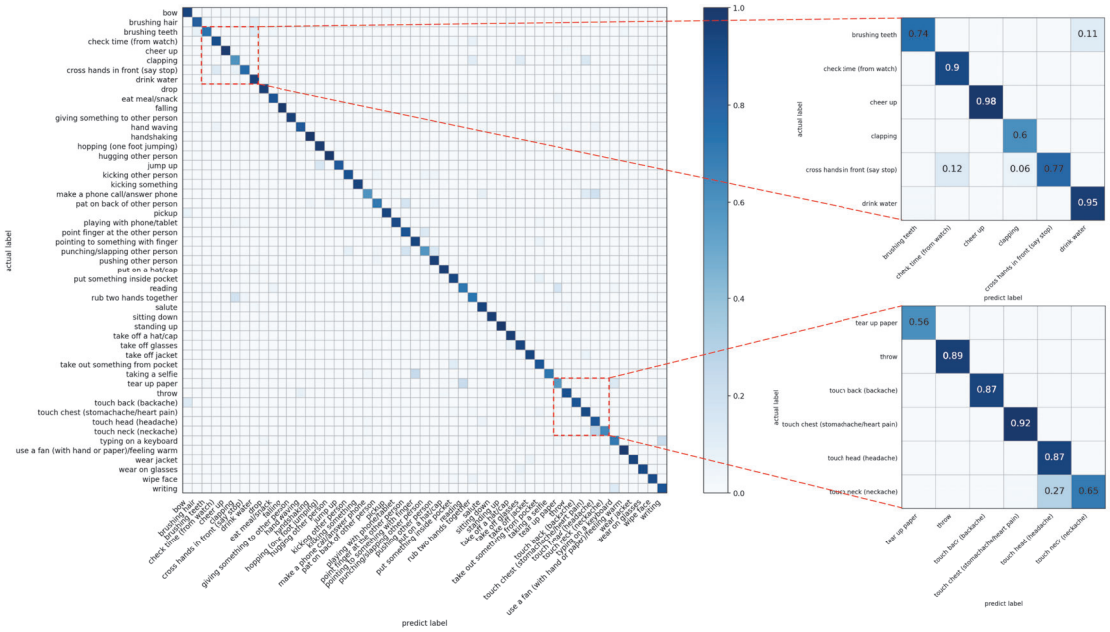


Fig. 4. Confusion matrix on PKU-MMD part I under xsub setting.

can be seen from the visualization results that the skeleton representations learned by the homotopy deformation and combining homotopy deformation and hyperbolic embedding have certain distinguishability in the feature space.

At the same time, we conduct ablation experiments with different proportions of hyperbolic loss in the loss function, as shown in Table III. It can be clearly seen from the experimental results that different values of λ have a certain impact on the accuracy. Here, this paper chooses the ratio of $\lambda = 0.1$ to carry out subsequent experiments.

D. Comparison with Existing Methods

In this section, the proposed method is experimentally compared to related existing methods to evaluate its effectiveness. The evaluation method follows three protocols: linear evaluation protocol, semi-supervised protocol, and fine-tune protocol.

1) Comparison of results on linear evaluation: In order to evaluate whether the learned representations contain sufficient discriminative information, this section compares them on linear classification tasks with the results shown in Table IV.

TABLE IV
COMPARING THE PERFORMANCE WITH THE CURRENT METHOD UNDER THE TWO VALIDATION BENCHMARKS ON NTU 60 DATASET

Methods	Year	Accuracy(%)	
		Xsub	Xview
LongT GAN[5]	2018	39.1	48.1
MS2L[45]	2020	52.6	-
P&C[4]	2020	50.7	76.3
SeBiReNet[7]	2020	-	79.7
AS-CAL[19]	2021	58.5	64.8
SkeletonCLR[20]	2021	68.3	76.4
CrosSCLR[20]	2021	72.9	79.9
'TS' Colorization[6]	2021	71.6	79.9
GL-Transformer[33]	2022	76.3	**83.8**
AimCLR[34]	2022	74.3	79.7
HiCLR[35]	2023	77.6	82.0
Proposed	-	**78.9**	82.3

TABLE V
COMPARING THE PERFORMANCE WITH THE CURRENT METHOD UNDER THE TWO VALIDATION BENCHMARKS ON NTU 120 DATASET

Methods	Year	Accuracy(%)	
		Xsub	Xset
P&C[4]	2020	42.7	41.7
AS-CAL[19]	2021	48.6	49.2
SkeletonCLR[20]	2021	56.8	55.9
ISC[46]	2021	**67.9**	67.1
GL-Transformer[33]	2022	66	***68.7***
AimCLR[34]	2022	63.4	63.4
Proposed	-	68.4	67.3

Under the two verification benchmarks, the proposed method achieves the best performance under the Xsub proto-col. Compared to HiCLR, the proposed method improves the accuracy by 1.3%. In HiCLR, the authors use gradual growing data enhancement methods to provide more information to the network, that is, a multi-view method is used for the skeleton data in different views to provide the network with richer supervision signals. The complementarity of information improves the network performance, while our proposed method does not add additional enhanced samples, and adopts the BYOL framework that discards the comparison of positive and negative samples. We show confusion matrix on NTU-60 in Fig.3. It can be seen that recognition mistakes are concentrated in action such as clapping, reading, writing, using a mobile phone and using a keyboard, which contains subtle movements with little body variation, and our proposed method achieves good accuracy in the remaining categories.

Fig.5 shows the comparison of the linear evaluation among the models trained with different epochs. It can be seen that the proposed method can achieve high accuracy when the number of rounds is small, and it has always maintained leading results under different rounds.

The proposed method is further compared on NTU 120, a larger-scale dataset of more classes, and the results are shown in Table V. The proposed method outperforms most of the methods, ranking second on both Xsub protocol and Xview protocol. Compared to the best method ISC on Xsub protocol, the proposed method is only 0.3% behind and hava an advantages under the Xview protocol. Similarly, the situation is opposite to the previous one compared to GL-TransFormer. These prove that the method is also competitive on multi-category large-scale datasets.

TABLE VI
COMPARING THE PERFORMANCE WITH THE CURRENT METHOD UNDER THE TWO VALIDATION BENCHMARKS ON PKU-MMD DATASET

Methods	Year	Accuracy(%)	
		Part I	Part II
ST-GCN(supervised)[30]	2018	84.1	48.2
VA-LSTM(supervised)[47]	2019	84.1	50.0
LongT GAN[5]	2018	67.7	26.0
MS2L[45]	2020	64.9	27.6
3s-CrosSCLR[20]	2021	84.9	21.2
ISC[46]	2021	80.9	36.0
AimCLR[34]	2022	83.4	-
Proposed	-	**88.5**	**51.7**

TABLE VII
SEMI-SUPERVISED RESULTS ON PKU-MMD DATASET

Methods	Year	1%		10%	
		Part I	Part II	Part I	Part II
LongT GAN[5]	2018	35.8	12.4	*69.5*	25.7
MS2L[45]	2020	36.4	13.0	33.1	-
ISC[46]	2021	37.7	-	72.1	-
Proposed	-	**55.4**	**24.8**	**83.5**	**37.8**
3s-CrosSCLR[20]	2021	49.7	10.2	82.9	28.6
3s-AimCLR[34]	2022	57.5	15.1	86.1	33.4
3s-Proposed	-	**64.5**	**26.7**	**86.9**	**39.0**

The comparison results on the dataset PKU-MMD are shown in Table VI. There are two different parts in this dataset. Part II is more challenging because of the view changes introduced by more skeleton noise. On Part I, our proposed method outperforms other unsupervised methods and outperforms some supervised methods, in which the network ST-GCN is used as the encoder of the proposed method, which proves the excellent representation learning ability. It can be seen from the table that the effect of the method 3s-CrosSCLR drops sharply in part II, while the proposed method can achieve good results in both parts, which proves that the proposed method has a strong ability to deal with skeleton noise, the extracted features are more robust. We plot the confusion matrix results for the PKU-MMD Part I in Fig.4. It can be seen from the figure, the most easily confused categories are mainly actions with small movements, such as brushing teeth and tear up papers. For actions with more obvious amplitudes, our proposed method achieves satisfing results. At the same time, 3s-CrosSCLR uses data from different views of the skeleton, namely joint data, bone data and motion data, while the proposed method only uses joint data and achieves higher accuracy, which further reflects the superiority.

2) Comparison on semi-supervised evaluation: In order to evaluate the classification effect of the proposed method where only a small number of labels are input for training, experiments were carried out on PKU-MMD and NTU-60 datasets with different labeled data. The results are shown in Table VII and Table VIII respectively.

Through experiments, it can be seen that the proposed method achieves high performance under the two protocols of PKU-MMD. Our method achieves state-of-the-art results on both single-stream and three-stream. And compared to 3s-AimCLR, our method of single stream also surpasses 9.7% and 4.4% on part II. It's not too far behind on the single

TABLE VIII
SEMI-SUPERVISED RESULTS ON NTU-60 DATASET

Methods	Year	1%		5%		10%		20%		40%	
		Xsub	Xview	Xsub	Xview	Xsub	Xview	Xsub	Xview	Xsub	Xview
LongT GAN[5]	2018	35.2	-	-	-	62.0	-	-	-	-	-
MS2L[45]	2020	33.1	-	-	-	65.2	-	-	-	-	-
ISC[46]	2021	35.7	38.1	59.6	65.7	65.9	72.5	70.8	78.2	-	-
'TS' Colorization[6]	2021	42.9	46.3	60.1	63.9	66.1	73.3	72.0	77.9	75.9	**82.7**
GL-Transformer[33]	2022	-	-	64.5	68.5	68.6	74.9	-	-	-	-
Proposed	-	**59.9**	**56.5**	**70.0**	**71.7**	**72.7**	**75.5**	**75.0**	**78.5**	**77.0**	80.6
3s-CrosSCLR[20]	2021	51.1	50.0	-	-	74.4	77.8	-	-	-	-
3s-Colorization[6]	2021	48.3	52.5	65.8	70.3	71.7	78.9	76.4	**82.7**	**79.8**	**86.8**
3s-AimCLR[34]	2022	54.8	54.3	-	-	78.2	81.6	-	-	-	-
3s-HiCLR[35]	2023	58.5	58.3	-	-	**79.6**	**84**	-	-	-	-
3s-Proposed	-	**65.7**	**58.4**	**74.6**	**75.0**	75.9	77.5	**78.9**	81.3	79.5	82.8

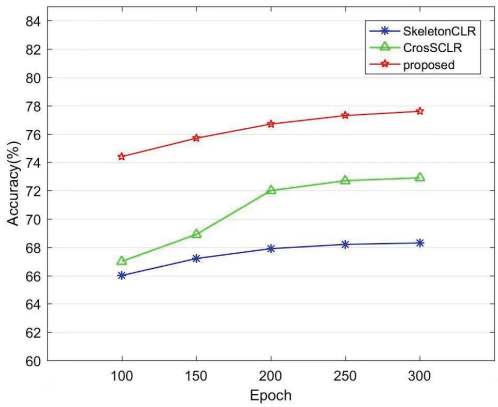


Fig. 5. Comparison of linear evaluation results of models obtained from different rounds of training.

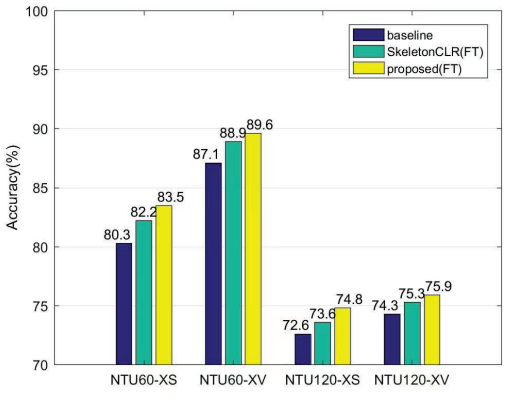


Fig. 6. Comparison of the results of the three models on fine-tune evaluation.

stream of part I either. That means the proposed method also has great advantages compared to methods that do not utilize multi-view data.

Under the five protocols of NTU 60, the proposed method has achieved good results using only joint data, outperforming almost the previous methods. This method has a certain improvement compared to 'TS' Colorization under almost every protocol. At the same time, our method also achieves competitive results when using data from different views of the skeleton. The method achieves the best results on 1% and 5% labels of Xsub protocol, respectively. Compared to the 3s-Colorization, the proposed method lags behind by only 0.3%on 40% label of Xsub protocol, and has advantages on the other labels.

3) Comparison on fine-tune evaluation: For fair comparison, the spatio-temporal graph convolution used in the proposed method has the same network structure and parameters as the existing method, and on large datasets NTU 60 and NTU 120, the fine-tune results of the proposed network outperform the baseline methods ST-GCN and method CrosCLR, as shown in Fig.6.

The results show that the proposed method achieves the highest accuracy under different validation benchmarks of NTU 60 and NTU 120, where "FT" indicates that the model is obtained by fine-tuning, and the baseline method is the encoder used in the experiment does not perform self-supervised learning and directly connects with full connections for classification. It can be seen that the proposed model has a significant improvement compared to the baseline model, and certain improvement compared to SkeletonCLR in different datasets.

V. Conclusion

This paper proposes a self-supervised learning method based on hyperbolic homotopy embedding, which adopts contrastive learning framework to learn different equivalent classes from spatio-temporal sequences through homotopic mapping, and maps the extracted high dimensional spatio-temporal interaction features to hyperbolic space through homotopy functions. In order to prevent similar samples from being pushed far in the feature space, the proposed method adopts a self-supervision framework that discards negative samples, and uses the potential of the network to guide itself to automatically obtain better supervision signals. At the same time, hyperbolic embedding can capture the contextual correlation of high-dimensional spatio-temporal sequences. The proposed method uses the Poincaré model to quantify the global features, and uses the similarity between the two stream output features of the contrast network as the embedding norm, and leverages the Poincaré model to represent the global feature vectorization. Homotopic mapping is used to represent the relationship of equivalent classes between data, so that the mapping satisfies the property of group invariance. The proposed method combines the geometric properties of hyper-

bolic manifolds and the equivariance of homotopy groups to promote better supervised signals for the network and improve the performance of unsupervised learning.

REFERENCES

[1] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[3] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[4] 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. 9631–9640.

[5] 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.

[6] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Skeleton cloud colorization for unsupervised 3d action representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 423–13 433.

[7] Q. Nie, Z. Liu, and Y. Liu, "Unsupervised 3d human pose representation with viewpoint and pose disentanglement," in *European Conference on Computer Vision*. Springer, 2020, pp. 102–118.

[8] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[9] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European conference on computer vision*. Springer, 2020, pp. 776–794.

[10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[11] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Learning image representations by completing damaged jigsaw puzzles," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 793–802.

[12] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*. Springer, 2016, pp. 69–84.

[13] C. Wei, L. Xie, X. Ren, Y. Xia, C. Su, J. Liu, Q. Tian, and A. L. Yuille, "Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1910–1919.

[14] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[15] O.-E. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," *arXiv preprint arXiv:1805.09112*, 2018.

[16] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," *Proceedings of the International Conference on Neural Information Processing Systems*, vol. 30, pp. 6338–6347, 2017.

[17] F. Sala, C. De Sa, A. Gu, and C. Ré, "Representation tradeoffs for hyperbolic embeddings," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2018, pp. 4460–4469.

[18] J. W. Cannon, W. J. Floyd, R. Kenyon, W. R. Parry *et al.*, "Hyperbolic geometry," *Flavors of geometry*, vol. 31, no. 59-115, p. 2, 1997.

[19] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition," *Information Sciences*, vol. 569, pp. 90–109, 2021.

[20] 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. 4741–4750.

[21] Bardes, Adrien and Ponce, Jean and LeCun, Yann, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[22] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[23] Pan, Tian and Song, Yibing and Yang, Tianyu and Jiang, Wenhao and Liu, Wei, "Videomoco: Contrastive video representation learning with temporally adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 205–11 214.

[24] Richemond, Pierre H and Grill, Jean-Bastien and Altché, Florent and Tallec, Corentin and Strub, Florian and Brock, Andrew and Smith, Samuel and De, Soham and Pascanu, Razvan and Piot, Bilal and others, "Byol works even without batch statistics," *arXiv preprint arXiv:2010.10241*, 2020.

[25] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

[26] Chen, Peihao and Huang, Deng and He, Dongliang

and Long, Xiang and Zeng, Runhao and Wen, Shilei and Tan, Mingkui and Gan, Chuang, "Rspnet: Relative speed perception for unsupervised video representation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1045–1053.

[27] Tong, Zhan and Song, Yibing and Wang, Jue and Wang, Limin, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *arXiv preprint arXiv:2203.12602*, 2022.

[28] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3010–3022, 2016.

[29] Q. Nie, J. Wang, X. Wang, and Y. Liu, "View-invariant human action recognition based on a 3d bio-constrained skeleton model," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3959–3972, 2019.

[30] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[31] K. Cheng, Y. Zhang, X. He, J. Cheng, and H. Lu, "Extremely lightweight skeleton-based action recognition with shiftgcn++," *IEEE Transactions on Image Processing*, vol. 30, pp. 7333–7348, 2021.

[32] 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/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.

[33] Kim, Boeun and Chang, Hyung Jin and Kim, Jungho and Choi, Jin Young, "Global-local motion transformer for unsupervised skeleton-based action learning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer, 2022, pp. 209–225.

[34] Guo, Tianyu and Liu, Hong and Chen, Zhan and Liu, Mengyuan and Wang, Tao and Ding, Runwei, "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.

[35] Zhang, Jiahang and Lin, Lilang and Liu, Jiaying, "Hierarchical Consistent Contrastive Learning for Skeleton-Based Action Recognition with Growing Augmentations," 2022.

[36] Yang, Siyuan and Liu, Jun and Lu, Shijian and Hwa, Er Meng and Hu, Yongjian and Kot, Alex C, "Self-Supervised 3D Action Representation Learning with Skeleton Cloud Colorization," *arXiv preprint arXiv:2304.08799*, 2023.

[37] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Homotopy continuation for sparse signal representation," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5. IEEE, 2005, pp. v–733.

[38] Y. Shinagawa, "Homotopic image pseudo-invariants for openset object recognition and image retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 11, pp. 1891–1901, 2008.

[39] Bhattacharya, Subhrajit, "Search-based path planning with homotopy class constraints," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 24, no. 1, 2010, pp. 1230–1237.

[40] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. Pluim, U. Bauer, and B. H. Menze, "cldice-a novel topology-preserving loss function for tubular structure segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 560–16 569.

[41] Hatcher, Allen, *Algebraic Topology*. Cambridge University Press, 2002.

[42] 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.

[43] 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.

[44] Van der Maaten, Laurens and Hinton, Geoffrey, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[45] L. Lin, S. Song, W. Yang, and J. Liu, "Ms2l: Multi-task self-supervised learning for skeleton based action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2490–2498.

[46] F. M. Thoker, H. Doughty, and C. G. Snoek, "Skeleton-contrastive 3d action representation learning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1655–1663.

[47] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1963–1978, 2019.