Mutual Information Driven Equivariant Contrastive Learning for 3D Action Representation Learning

Lilang Lin, *Graduate Student Member, IEEE*, Jiahang Zhang, and Jiaying Liu, *Senior Member, IEEE*

***Abstract*— Self-supervised contrastive learning has proven to be successful for skeleton-based action recognition. For contrastive learning, data transformations are found to fundamentally affect the learned representation quality. However, traditional invariant contrastive learning is detrimental to the performance on the downstream task if the transformation carries important information for the task. In this sense, it limits the application of many data transformations in the current contrastive learning pipeline. To address these issues, we propose to utilize equivariant contrastive learning, which extends invariant contrastive learning and preserves important information. By integrating equivariant and invariant contrastive learning into a hybrid approach, the model can better leverage the motion patterns exposed by data transformations and obtain a more discriminative representation space. Specifically, a self-distillation loss is first proposed for transformed data of different intensities to fully utilize invariant transformations, especially strong invariant transformations. For equivariant transformations, we explore the potential of *skeleton mixing* and *temporal shuffling* for equivariant contrastive learning. Meanwhile, we analyze the impacts of different data transformations on the feature space in terms of two novel metrics proposed in this paper, namely, consistency and diversity. In particular, we demonstrate that equivariant learning boosts performance by alleviating the dimensional collapse problem. Experimental results on several benchmarks indicate that our method outperforms existing state-of-the-art methods.**

***Index Terms*— Self-supervised learning, skeleton-based action recognition, contrastive learning.**

I. Introduction

THE 3D skeleton is a highly efficient representation of the human body structure and motion, as it uses the 3D coordinates of key body joints to describe the human form. In comparison to RGB videos and depth data, skeletons are lightweight and protect the privacy of individuals. Due to their ease of analysis and discriminative nature, skeletons have become widely used in action recognition tasks.

Supervised skeleton-based action recognition methods [1], [2] have shown strong results, but they rely heavily on a large amount of labeled training data, which can be costly to obtain. To mitigate the need for full supervision, self-supervised learning [3], [4], [5], [6] has been applied to skeleton-based action recognition. Self-supervised learning-based approaches utilize a large amount of unlabeled data to model the spatial-temporal relationship and obtain a meaningful representation space through the use of various pretext tasks.

In the field of machine learning, there are two main paradigms for self-supervised learning: reconstruction-based and contrastive learning-based methods. Reconstruction-based approaches [5], [7], [8] employ an encoder-decoder model to reconstruct the input data from a latent representation, capturing the spatial-temporal correlations within the data. Contrastive learning-based methods [9], [10] aim to maximize the mutual information between different augmented views of the same data by comparing them with other views of unrelated data. This involves applying various transformations, such as rotation, scaling, and translation, to the input skeleton data to create augmented views of the original data. It enables the model to learn a discriminative representation of the data that is invariant to these transformations. Rao et al. [9] employed shearing and cropping for data transformation. Guo et al. [10] extended this approach by using additional augmentations, such as rotation, masking, and flipping, to improve the consistency of the contrastive learning method.

Data transformation has been proven to fundamentally affect the learned representation quality. Diverse data transformations provide rich movement patterns and greater semantic consistency. The model tends to discard the augmentation-related information to achieve invariant learning under data transformations. However, this can be detrimental to the performance of the downstream task if the transformation carries important information that is needed for the task. For example, temporal-domain information is often crucial for motion recognition tasks, as it captures the temporal dynamics of the action being performed. Previous methods [10], [11], [12], [13] based on contrastive 3D action representation learning solely focus on the invariant data transformations, *i.e.*, pursuing the consistency between two augmented views. However, this limits the wider application of skeleton data transformations in contrastive learning, because some transformations are found detrimental to the downstream tasks. For example, using temporal-domain shuffling as a transformation in contrastive learning destroys this temporal information,

Manuscript received 19 February 2023; revised 29 August 2023 and 28 November 2023; accepted 24 January 2024. Date of publication 7 March 2024; date of current version 12 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62172020; in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology); and in part by the State Key Laboratory of Media Audio and Video (Communication University of China), Ministry of Education, China. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yannick Berthoumieu. *(Lilang Lin and Jiahang Zhang contributed equally to this work.) (Corresponding author: Jiaying Liu.)*

The authors are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China (e-mail: linlilang@pku.edu.cn; zjh2020@pku.edu.cn; liujiaying@pku.edu.cn).

Digital Object Identifier 10.1109/TIP.2024.3372451

adversely affecting the downstream performance of the model. How to utilize the novel motion patterns exposed by these transformations, to improve the representation learning for 3D action recognition, remains an under-explored problem.

To address this issue, it is necessary to develop self-supervised learning methods that can preserve the transformation information for the use of data transformations. Therefore, we extend the previous contrastive learning approach, which is called invariant contrastive learning because invariant representations are learned, to equivariant contrastive learning. Invariant contrastive learning draws the data features closer after data transformation, while equivariant contrastive learning preserves the differences among the features of the transformed data, maintaining the augmentation-related information. By employing equivariant learning for certain transformations, the model can utilize relevant information about the transformations for the downstream task. We show that this approach, which combines equivariant and invariant contrastive learning, further improves the performance by better utilizing the relevant information about the transformations.

Specifically, for equivariant contrastive learning, we explore the potential of two data transformations: *skeleton mixing* and *temporal shuffling*. These transformations generate more diverse motion patterns and encourage the model to capture the augmentation-related information inherent in the data. Meanwhile, consistent learning becomes difficult when data transformations are enhanced. For better optimization of the feature consistency, we propose a distillation loss for transformed data of different intensities. We utilize knowledge gained from the basic transformations to guide the learning of the strong transformations.

To quantitatively assess the impacts of different data transformations on the feature space learned by the model, we analyze the distribution of the feature space from the perspective of mutual information. We decompose the mutual information into two novel metrics: *consistency* and *diversity*. These metrics enable us to measure the degree to which the transformed data maintain the underlying structure of the original data and the differences among the transformed data, respectively. Our experiments demonstrate that invariant data transformations prioritize consistency over diversity, while equivariant data transformations primarily optimize diversity.

In particular, we specifically examine the effects of different data transformations on the dimensional collapse phenomenon in the feature space. Dimensional collapse occurs when the learned representation collapses onto a lower-dimensional subspace, resulting in a loss of information for the downstream task. Our results show that consistency tends to exacerbate dimensional collapse, while diversity mitigates it. Based on these findings, we demonstrate that equivariant data transformations improve the representation quality by increasing the feature diversity and alleviating dimensional collapse.

Our main contributions are summarized as follows:

- In this paper, we propose to integrate equivariant data transformations with the existing invariant data transformations to improve the performance of skeleton-based representation learning for action recognition. The proposed equivariant skeleton data transformations assist the model in encoding more diverse motion patterns, while the invariant transformations lead to better semantic consistency between positive pairs.
- To further optimize the feature consistency, we employ a distillation loss for the transformed data of different intensities. This loss function enables us to transfer knowledge gained from the basic transformations to guide the learning of the strong transformations.
- We analyze the transformations in terms of mutual information and investigate the dimensional collapse phenomenon in the feature space. We demonstrate that equivariant data transformation improves the representation quality by increasing the feature diversity and alleviating dimensional collapse.

The remainder of the paper is organized as follows. In Sec. II, previous works on skeleton-based action recognition and self-supervised learning are reviewed. Sec. III delves into the details of the suggested self-supervised learning method and quantitative analysis. In Sec. IV, we present the results of our experiment and the model analysis. Sec. V presents the conclusions of this study.

II. Related Work

In this section, we first introduce the related work on skeleton-based action recognition and then briefly review contrastive learning.

A. Skeleton-Based Action Recognition

Skeleton-based action recognition is a fundamental yet challenging field in computer vision research. Previous skeleton-based action recognition methods usually consider the geometric relationship among skeletal joints [14], [15]. The latest methods give more attention to deep networks. Skeleton-based action recognition methods can be divided into recurrent neural network (RNN)-based, convolutional neural network (CNN)-based, graph convolutional network (GCN)-based, and transformer-based methods. Du et al. [16] applied a hierarchical RNN to process body keypoints. Attention-based methods are further proposed for automatically selecting important spatial joints [17], [18], [19] and temporal frames [17], [18] to adaptively learn the simultaneous appearance of skeletal joints. For CNN-based methods, some works [20], [21] transform each skeleton sequence into image-like representations and apply a CNN model to extract spatial-temporal information. Recently, inspired by the natural topology graph of the human body, GCNs have aroused a surge of interest in skeleton-based action recognition. To extract both the spatial and temporal structural features from skeleton data, Yan et al. [22] proposed spatial-temporal graph convolutional networks. To make the graphic representation more flexible, attention mechanisms are applied in [1], [2], and [19] to adaptively capture discriminative features based on spatial composition and temporal dynamics. Chi et al. [23] designed an information bottleneck-based learning objective to guide the model to learn informative but compact latent representations. Duan et al. [24] implemented six different algorithms under a unified framework with both the latest and original good practices to ease the comparison of efficacy and efficiency. Meanwhile, transformer-based models [25] also show remarkable results by using the long-range temporal dependencies due to the use of an attention mechanism. To balance feature representation for cross-modal data, STAR-Transformer [26] effectively represents two cross-modal features as a recognizable vector. To close the performance gap between Transformers and GCNs, Zhou [27] proposed a new self-attention (SA) extension, named Hypergraph Self-Attention (HyperSA), to incorporate inherently higher-order relations into the model.

Generally, RNN-based methods can well model the temporal features while struggling to learn a good spatial representation. CNN-based methods can simultaneously model spatio-temporal information by reorganizing the skeleton. However, they rely heavily on the pre-defined skeleton representation, which is often designed heuristically and can be sub-optimal. Benefiting from the natural topological structure of the human body, GNNs are good at modeling the spatial-temporal features of the skeleton, while they are also found to be sensitive to the adjacent matrix. To get rid of the influence of artificially restricted adjacency matrix, the Transformer-based approaches can fully model the relationship between different joints by the attention mechanism. However, this often results in high computational complexity and possible over-fitting problems.

Most importantly, all these methods require lots of annotated data, which can be costly to achieve. To this end, we turn to the self-supervised learning paradigm, which learns representations from unlabeled data. Among the existing self-supervised learning methods for skeleton, contrastive learning has attracted lots of attention due to its effective ability to learn the discriminative feature space, and has shown superior performance advantage against other methods, *e.g.*, reconstruction-based [3], [7] and pseudo-label-based methods [8], [28]. Next, we will introduce and review the contrastive representation learning briefly.

B. Contrastive Representation Learning

Contrastive representation learning dates back to [29]. The key idea is to pull the positive pairs together while pushing away the negative pairs to learn a highly distinguishable feature space. Many works [30], [31], [32], [33], [34] have been presented in which representations are learned by contrasting positive pairs against negative pairs, which have achieved remarkable results. SimCLR, proposed by Chen et al. [35], uses a series of data augmentation methods, such as random cropping, Gaussian blurring and color distortion, to generate positive samples. He et al. [36] applied a memory module that adopts a queue to store negative samples, and the queue is constantly updated during training. In addition, many promising contrastive learning methods have been successively presented, such as BYOL [37], SimSiam [38], and SwAV [39].

In self-supervised skeleton-based action recognition, contrastive learning has also attracted the attention of numerous researchers. Rao et al. [9] applied MoCo for contrastive learning with a single stream. To utilize cross-stream knowledge, Li et al. [11] proposed a multiview contrastive learning method, and Thoker et al. [6] employed multiple models to learn from different skeleton representations. This work proposes performing knowledge distillation among the different modalities, *i.e.*., *joint*, *bone*, and *motion*. Recently, data transformations have been proven to significantly affect representation quality [40], [41], and many works have focused on constructing positive/negative pairs via data transformation. Rao et al. [9] first proposed a series of augmentations and applied the MoCo framework. Guo et al. [10] proposed using more extreme augmentations to boost contrastive learning. Zhang [13] further utilized strong augmentations by hierarchical consistency learning. Huang et al. [42] proposed a graph contrastive learning framework for skeleton-based action recognition to explore the global context across all sequences. Zhou et al. [43] proposed a Partial Spatio-Temporal Learning (PSTL) framework to exploit the local relationship from a partial skeleton sequences built by a unique spatio-temporal masking strategy. However, most works have only employed invariant contrastive learning, and the exploration of equivariant contrastive learning is still insufficient. Recently, E-SSL [44] was proposed, which applies a rotation transformation to images to perform equivariant contrastive learning, and showed notable results. Inspired by this, we explore the potential of combining equivariant and invariant contrastive learning for skeleton data in this paper. Meanwhile, we present a detailed analysis of the improvements achieved by equivariant contrastive learning.

III. Method

A. Invariant and Equivariant Representation Learning

In this part, the proposed invariant and equivariant contrastive learning method is introduced for skeleton-based action representation learning. We first present the definitions of invariance and equivariance. Formally, we let $\mathcal{T}$ be the transformation set and $f(\cdot)$ be the encoder. For any given input $\mathbf{x}$, invariant contrastive learning aims to obtain the representation space invariant to $\mathcal{T}$:

$$\forall t \in \mathcal{T},\ \ f(t(\mathbf{x})) = f(\mathbf{x}). \tag{1}$$

In addition, more recent works [35], [36], [38] utilize the two different transformed views and consider another invariant form of derivation, which is also used in our method:

$$\forall t_1,\ t_2 \in \mathcal{T},\ f(t_1(\mathbf{x})) = f(t_2(\mathbf{x})). \tag{2}$$

For the learning of equivariance, we follow the definition of previous works [44], [45]:

$$\forall t \in \mathcal{T},\ \exists p_t(\cdot),\ \ f(t(\mathbf{x})) = p_t(f(\mathbf{x})), \tag{3}$$

where $p_t(\cdot)$ is the transformation in the encoder space, parameterized by the transformation t. If $p_t(\cdot)$ is the identity transformation, then Eq. 3 is reduced to Eq. 1. In other words, invariance can be regarded as a specific trivial case of equivariance. From this definition, we know that the model is encouraged to encode augmentation-related information to achieve equivariant learning, and the different augmented views are projected to different embeddings.

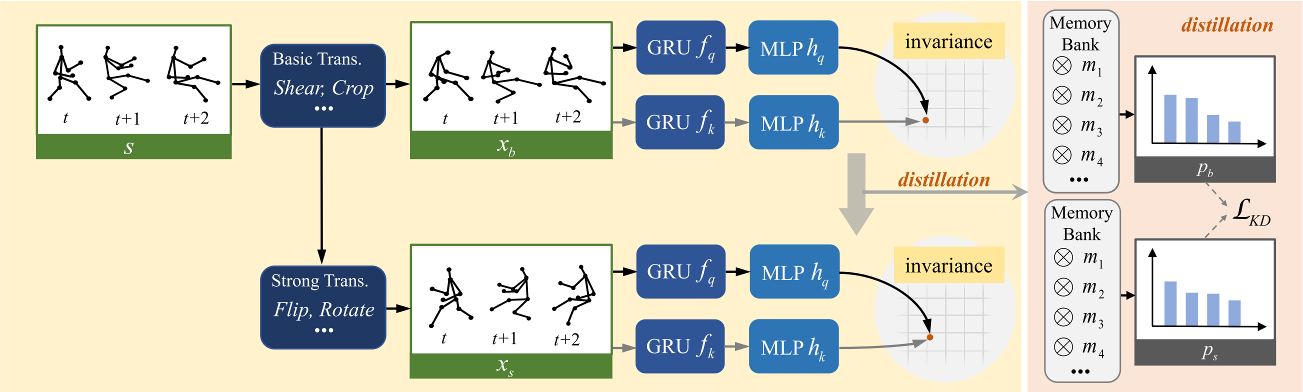


Fig. 1. Pipeline of invariant contrastive learning. The top stream has a basic data transformation set, which includes mainly the *cropping, shearing* and *joint jittering* data transformations. The bottom stream corresponds to a strong data transformation set, adding the *flipping, rotation* and *Gaussian noise* transformations. In addition, we propose a distillation objective, which utilizes the knowledge of the basic data transformation to guide invariant learning under strong data transformations. The features obtained under the strong data transformations are encouraged to be close to those obtained under the basic transformations in terms of the similarity distribution.

In the following, we will elaborate on the pipelines of the proposed invariant contrastive learning and equivariant contrastive learning separately and provide an overview of the full model.

1) Invariant Contrastive Learning: Traditional contrastive learning frameworks encourage the model to learn the invariants under different data transformations, which often represent the semantic information. Formally, such a framework is usually comprised of the following components:

- **Data augmentation module**, which employs data transformations $\mathcal{T}$ to generate different views of the original data, which are regarded as positive pairs;
- **Query/Key encoder** $f(\cdot)$ for mapping the input into the latent feature space;
- **Embedding projector** $h(\cdot)$ for mapping the latent feature into an embedding space where the following self-supervised loss is applied;
- **Self-supervised loss**, which is used to perform the feature clustering operation in the embedding space. The key idea is to pull the positive pairs together and push the negative pairs away, obtaining a distinguishable embedding space.

Specifically, our framework, as shown in Fig. 1, is based on MoCo v2 [36]. Given a skeleton sequence $\mathbf{s}$, the positive pair $(\mathbf{x}, \mathbf{x}')$ is constructed via the transformation set $\mathcal{T}$. Subsequently, we can obtain the corresponding feature representations $(\mathbf{z}, \mathbf{z}')$ via the query/key encoder $f_q(\cdot)/f_k(\cdot)$ and embedding projector $h_q(\cdot)/h_k(\cdot)$. Meanwhile, a memory queue is maintained to store the negative samples according to a first-in, first-out strategy. The model optimizes the InfoNCE [46] loss to perform invariant contrastive learning:

$$\mathcal{L}_{Info} = -\log \frac{\exp(\mathbf{z} \cdot \mathbf{z}'/\tau)}{\exp(\mathbf{z} \cdot \mathbf{z}'/\tau) + \sum_{i=1}^{\mathbf{M}} \exp(\mathbf{z} \cdot \mathbf{m}_i/\tau)}, \tag{4}$$

where $\mathbf{m}_i$ is the i_{th} negative sample feature in $\mathbf{M}$, M is the number of negative features and τ is the temperature hyperparameter. The key encoder is a momentum-updated version of the query encoder in MoCo v2.

Following recent works [6], [12], we adopt three data transformations as the basic transformation set, namely, *temporal cropping, shearing, and joint jittering*, which are used in our baseline algorithm for contrastive learning. Meanwhile, to further boost the invariant learning of the model, we introduce more data transformations as the strong transformation set. In addition to the transformations above, three more spatial transformations are adopted in the set: *flipping, rotation, and Gaussian noise*. Therefore, we optimize $\mathcal{L}^b_{Info}$ and $\mathcal{L}^s_{Info}$ using the basic and strong data transformation sets, respectively, to generate the positive pairs and learn the invariance under different transformations.

However, strong data transformations can lead to semantic information loss due to severe distortion [13], [41], [47], and hence, directly optimizing the $\mathcal{L}^s_{Info}$ term is difficult. The basically transformed views naturally provide a clue for the invariant learning of strongly transformed views. Therefore, we propose transferring the knowledge learned from the basically transformed data to the strongly transformed data for better guidance of invariant learning under strong transformations. Specifically, we adopt the self-distillation design for the relational knowledge [48], and the following proposed objective is optimized jointly with the InfoNCE loss:

$$\begin{aligned} \mathcal{L}_{KD} &= -p\left(\mathbf{z}'_b, \tau_b\right) \log p\left(\mathbf{z}_s, \tau_s\right), \\ p_i\left(\mathbf{z}, \tau\right) &= \frac{\exp(\mathbf{z} \cdot \mathbf{m}_i/\tau)}{\sum_{i=1}^{M} \exp(\mathbf{z} \cdot \mathbf{m}_i/\tau)}, \end{aligned} \tag{5}$$

where $\mathbf{z}'_b = f_k(\mathbf{x}_b)$; $\mathbf{z}_s = f_q(\mathbf{x}_s)$; $\mathbf{x}_b$ and $\mathbf{x}_s$ represent the basically and strongly augmented data, respectively; and τ_b and τ_s are the temperature hyperparameters. The gradient of $\mathbf{z}'_b$, which is the learning target, is stopped. This is because the knowledge from the basically transformed data is more confident and accurate, so we only use basic views to supervise the strong views but not vice versa. This self-distillation objective can further boost the invariant learning under strong data transformations using the knowledge of the basic transformation set, giving rise to better representation.

Overall, the optimization objective for invariant contrastive learning is as follows:

$$\mathcal{L}_{inv} = \mathcal{L}^b_{Info} + \mathcal{L}^s_{Info} + \mathcal{L}_{KD}. \tag{6}$$

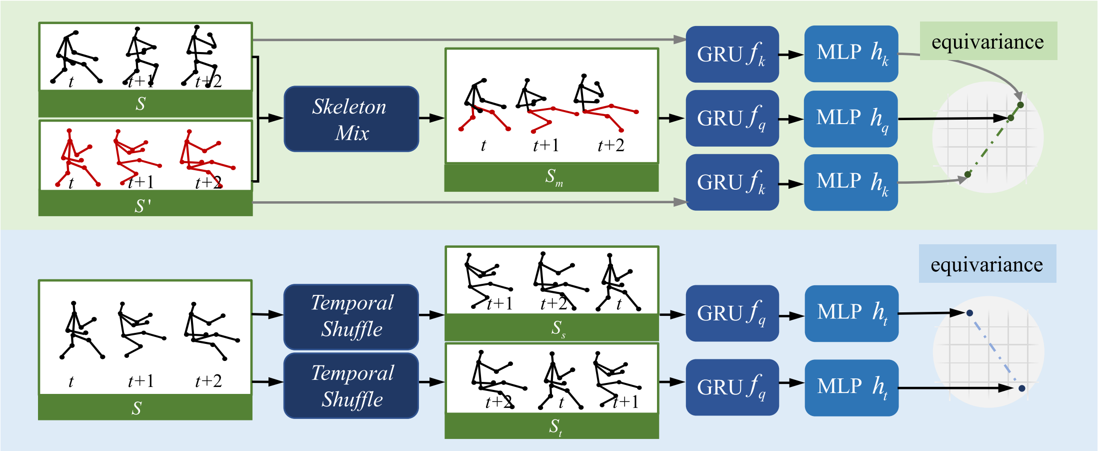


Fig. 2. Pipeline of equivariant contrastive learning. The green stream corresponds to *skeleton mixing*. We maximize the similarity between the projected feature of the mixed data and the constructed feature as the learning target, which is obtained according to the mixing ratio λ. The blue stream corresponds to *temporal shuffling*. A predictor $h_t(\cdot)$ is attached to predict the shuffling label. In this way, different shuffled data are encoded into different features, encouraging the encoder to extract temporal information.

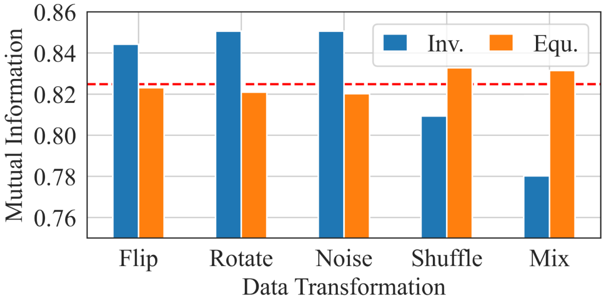


Fig. 3. Histogram of the relationship between data transformation and mutual information. The red line is the baseline result. "Inv." denotes invariance and "Equ." denotes equivariance.

2) Equivariant Contrastive Learning: To better utilize the novel patterns exposed by data transformations, we integrate the equivariant data transformations with the aforementioned invariant data transformations. We show that this hybrid approach can further improve the performance. Specifically, we utilize two data transformations, namely, *skeleton mixing* for the spatial dimension and *temporal shuffling* for the temporal dimension, in equivariant contrastive learning, as shown in Fig. 2. Intuitively, these transformations may cause semantic loss or changes, so we regard them as equivariant transformations, which is also supported by the results of the subsequent experiments shown in Fig. 3. Next, we introduce these data transformations and the task design.

Skeleton Mixing. Mixing transformations are widely used in self-supervised learning [49], [50], [51], [52]. Here, we explain and analyze them from the perspective of equivariant learning. Specifically, *skeleton mixing* contains three mixing methods designed for the skeleton data: cut mix [53], resize mix [54], and mix up [55]. In the cut mix and resize mix methods, we first divide the skeletal joints into multiple subsets according to different body parts of the human topology, *i.e.*, the *trunk, right hand, left hand, right leg* and *left leg*. Then, we randomly select another skeleton sequence $\mathbf{s}'$ and mix it with the original data $\mathbf{s}$ at the level of a human body part to generate the transformed data $\mathbf{s}_m$. The mixing ratio λ is defined as the ratio of the total number of replaced joints to the total number of joints in the skeleton $\mathbf{s}$. For the mix-up method, we simply mix all the joints of $\mathbf{s}$ and $\mathbf{s}'$ at a certain mixing ratio λ to generate the transformed data $\mathbf{s}_m$:

$$\mathbf{s}_m = (1-\lambda)\mathbf{s} + \lambda\mathbf{s}'. \tag{7}$$

In the implementation, we randomly select a mixing method from the above three methods and apply it to the skeleton data. *Skeleton mixing* is viewed as a transformation of the original data $\mathbf{s}$. In this sense, it can be unreasonable to constrain the model to learn the invariance under the transformation because semantic preservation is not guaranteed when applying it. Therefore, inspired by the mixed labels widely used in previous works [53], [54], [55], we manually construct $\mathbf{z}'_m$ as the target feature to be learned according to the mixing ratio:

$$\mathbf{z}'_m = (1-\lambda) f_k(\mathbf{s}) + \lambda f_k(\mathbf{s}'). \tag{8}$$

We find that the equivariant learning target based on this prior can achieve a remarkable improvement in performance. Specifically, the model is constrained to optimize the following InfoNCE loss:

$$\mathcal{L}_m = -\log \frac{\exp(\mathbf{z}_m \cdot \mathbf{z}'_m/\tau)}{\exp(\mathbf{z}_m \cdot \mathbf{z}'_m/\tau) + \sum_{i=1}^{M} \exp(\mathbf{z}_m \cdot \mathbf{m}_i/\tau)}, \tag{9}$$

where $\mathbf{z}_m = f_q(\mathbf{s}_m)$. $\mathbf{s}_m$ is mixed skeleton data, $f_q(\cdot)$ is the online encoder. In this way, different anchors are constructed corresponding to different augmentations, and the model learns the equivariance by modeling the mapping relationship between the transformed data and the anchors.

Temporal Shuffling. For a given skeleton sequence $\mathbf{s}$, we chunk it into C clips along the temporal dimension. Then, we randomly shuffle the clips and reassemble them into time-out-of-order data. There are $C!$ ways to shuffle the clips. We deliberately choose a specific permutation for shuffling, and the shuffling label corresponds to the associated label. We constrain the model to be sensitive to the shuffling transformation and learn the equivariance in the encoder latent space.

Inspired by E-SSL [44], we apply a classifier $h_t(\cdot)$ to take the latent feature as input and predict the shuffling label. In turn, the representation space learned by the model is encouraged to contain augmentation-related information, such as the temporal order. To ensure the preservation of temporal relationships and prevent the model from learning incorrect information, we propose the incorporation of equivariant contrastive learning. Specifically, our strategy involves training the network to predict temporal order shuffling patterns, thereby enabling the network to effectively capture temporal-domain information. This innovative approach serves to enhance action classification performance by enhancing the model's temporal understanding.

Specifically, denoting the data transformed using *temporal shuffling* as $\mathbf{s}_t$, the objective to be optimized is as follows:

$$\mathcal{L}_t = -\log \frac{\exp(\mathbf{z_t} \cdot \mathbf{w_c}/\tau)}{\exp(\mathbf{z_t} \cdot \mathbf{w_c}/\tau) + \sum_{i=1,i\neq c}^{C!} \exp(\mathbf{z_t} \cdot \mathbf{w}_i/\tau)}, \quad (10)$$

where $\mathbf{z_t} = f_q(\mathbf{s_t})$ and c is the shuffling pseudolabel of $\mathbf{s}_t$. $\mathbf{w}_i$ corresponds to the i_{th} column vector in the weight of predictor $h_t(\cdot)$, which is a bank containing the learnable positive/negative samples. Positive pairs refer to sequences that undergo the same shuffling method, signifying that their shuffling orders match. Conversely, negative pairs encompass sequences with different permutation orders.

For equivariant contrastive learning, the optimization objective of the model is summarized as follows:

$$\mathcal{L}_{equ} = \mathcal{L}_m + \mathcal{L}_t. \quad (11)$$

3) Full Model: We employ the above invariant and equivariant contrastive learning design to train the encoder $f(\cdot)$, which is then used in the downstream task after pretraining. The model is trained in a multitasking manner, and the objective is

$$\mathcal{L} = \mathcal{L}_{inv} + \lambda_{equ}\mathcal{L}_{equ}, \quad (12)$$

where λ_{equ} is a hyperparameter.

B. Consistency and Diversity Analysis of Self-Supervised Representations

In this part, we present a detailed analysis of the relation between data transformations and the feature space obtained by pretraining. Two new metrics are proposed for evaluating the consistency and diversity within the feature space based on mutual information between transformed samples. These metrics enable us to measure the quality of the feature space.

1) Mutual Information Estimation: Mutual information is a measure of the dependence between two random variables, $\mathbf{X}$ and $\mathbf{Z}$. The mutual information between $\mathbf{Z}$ and $\mathbf{Z}'$ can be regarded as the Kullback-Leibler (KL-) divergence between the joint, $\mathcal{P}_{\mathbf{ZZ}'}$, and the product of the marginals $\mathcal{P}_{\mathbf{Z}} \otimes \mathcal{P}_{\mathbf{Z}'}$:

$$I(\mathbf{Z};\mathbf{Z}') = D_{\mathrm{KL}}(\mathcal{P}_{\mathbf{ZZ}'}||\mathcal{P}_{\mathbf{Z}} \otimes \mathcal{P}_{\mathbf{Z}'}), \quad (13)$$

where D_{KL} is the KL divergence. The strength of the dependence between two random variables, $\mathbf{Z}$ and $\mathbf{Z}'$, can be quantified using mutual information, which is a measure of the divergence between the joint distribution of $\mathbf{Z}'$ and $\mathbf{Z}'$ and the product of their individual marginals. The large divergence indicates that there is a strong relationship between $\mathbf{Z}$ and $\mathbf{Z}'$ and that knowledge of the value of one variable can provide significant information about the other.

To calculate the mutual information, we utilize the Donsker–Varadhan representation. A representation of the KL divergence is given by the following theorem.

Theorem 1 (Donsker-Varadhan Representation): The KL divergence is estimated by a variational method by maximizing the dual problem:

$$D_{KL}(\mathcal{P}||\mathcal{Q}) = \sup_{T\in\mathcal{T}} E_{\mathcal{P}}[T] - \log(E_{\mathcal{Q}}[e^T]), \quad (14)$$

where the supremum is taken over all fuctions T in $\mathcal{T}$.

Therefore, the estimation of the mutual information of transformed samples can be expressed as

$$\begin{aligned} I(\mathbf{Z};\mathbf{Z}') &= D_{\mathrm{KL}}\left(\mathcal{P}_{\mathbf{ZZ}'}||\mathcal{P}_{\mathbf{Z}} \otimes \mathcal{P}_{\mathbf{Z}'}\right) \\ &= \sup_{T\in\mathcal{T}} E_{\mathcal{P}_{\mathbf{ZZ}'}}[T] - \log\left(E_{\mathcal{P}_{\mathbf{Z}}\otimes\mathcal{P}_{\mathbf{Z}'}}[e^T]\right) \\ &= \sup_{g\in\mathcal{G}} E_{(\mathbf{z},\mathbf{z}')\sim\mathcal{P}_{\mathrm{pos}}}\left[k(g(\mathbf{z}),g(\mathbf{z}'))\right] \\ &\quad - \log\left(E_{(\mathbf{z},\mathbf{z}')\sim\mathcal{P}_{\mathrm{data}}\otimes\mathcal{P}_{\mathrm{data}}}\left[e^{k(g(\mathbf{z}),g(\mathbf{z}'))}\right]\right), \end{aligned} \quad (15)$$

where $(\mathbf{z},\mathbf{z}') \sim \mathcal{P}_{\mathrm{pos}}$ are features extracted by the pretrained model of positive pairs. $\mathcal{P}_{\mathrm{data}} = \int \mathcal{P}_{\mathrm{pos}}(\mathbf{z},\mathbf{z}')\mathrm{d}\mathbf{z}'$ is the marginal distribution in the feature manifold. The function $g(\cdot)$ maps $\mathbf{z}$ and $\mathbf{z}'$ to a d-dimensional normalized feature embedding on the hypersphere $\mathbb{S}^{d-1}$. The function $k(\cdot)$ is a cosine similarity measure that quantifies the similarity between the two embeddings produced by $g(\cdot)$. To investigate the feature space properties of different pretrained encoders, we fix the pretrained encoder weights and then train the function $g(\cdot)$ to maximize the equation for estimating the mutual information. Therefore, we regard the former item as consistency and the latter as diversity.

2) Consistency and Diversity Analysis: We quantitatively evaluate the feature quality of different pretrained encoders for a given data transformation using consistency and diversity. By examining the consistency and diversity of the feature representations produced by different encoders, we can identify which encoders are most effective at capturing the relevant information and relationships within the data.

- **Consistency**: Consistency refers to the degree of similarity between positive samples. A high degree of consistency indicates that the information related to the data transformation has not been extracted by the encoder, and as a result, positive samples are mapped to the same features. This means that the features are not affected by the data transformation, as they are representative of the inherent characteristics of the samples rather than the transformation. On the other hand, a low degree of consistency indicates that the extracted features contain information about the data transformation, resulting in the features of the same data being different under different transformations. Formally, we define this as follows:

$$\mathcal{C} = E_{(\mathbf{z},\mathbf{z}')\sim\mathcal{P}_{\mathrm{pos}}}\left[k(g(\mathbf{z}),g(\mathbf{z}'))\right]. \quad (16)$$

- **Diversity**: Diversity is the similarity of features between two random samples. Higher diversity means that the feature

TABLE I

ACCURACY OF DOWNSTREAM TASKS WITH DIFFERENT DATA TRANSFORMATIONS USING EQUIVARIANT CONTRASTIVE LEARNING AND INVARIANT CONTRASTIVE LEARNING

Transformation	Invariant	Equivariant
Baseline	84.7	
Flip	**86.7**	84.0
Rotate	**86.1**	84.1
Noise	**86.3**	83.2
Shuffle	83.2	**85.7**
Mix	83.6	**85.2**

space encodes more information. Lower diversity indicates a collapse of the feature space, making some of the sample features indistinguishable. Formally, we define this as follows:

$$\mathcal{D} = \log\left(E_{(\mathbf{z},\mathbf{z}')\sim\mathcal{P}_{\text{data}}\otimes\mathcal{P}_{\text{data}}}\left[e^{k(g(\mathbf{z}),g(\mathbf{z}'))}\right]\right). \tag{17}$$

3) Invariant and Equivariant Representation Learning: Fig. 3 and Table I show the mutual information and action recognition accuracy for different data transformations under invariant contrastive learning and equivariant contrastive learning. The baseline method adopts the basic transformation set for invariant learning only. The first three transformations, *flipping, rotation and Gaussian noise*, obtain better performance under invariant contrastive learning, while the performance decreases under equivariant contrastive learning. The other two transformations, in contrast, show performance improvement under equivariant contrastive learning instead of invariant contrastive learning. By comparing the mutual information metrics of the feature spaces obtained by invariant and equivariant learning under these data transformations, we can unsupervisedly distinguish between the two types of transformations. The mutual information of the first three data transformations is better under invariant contrastive learning than under equivariant contrastive learning. This shows that these three transformations are suitable for invariant contrastive learning. The mutual information metrics of the latter two data transformations are compromised under invariant contrastive learning. This indicates that consistent learning of these transformations introduces more noise and degradation into the feature space.

C. Relation to Dimensional Collapse

In self-supervised learning, dimensional collapse is a phenomenon that occurs when the representation learned by the model becomes constant or collapses onto a lower-dimensional subspace. This can occur when the model is trained to minimize the distances between embedding vectors of augmented samples. One way to prevent complete collapse is through the use of contrastive methods. However, it has been observed that even with the use of contrastive methods, dimensional collapse can still occur, where the embedding vectors occupy a subspace that is of lower dimensionality than their original space. This can limit the effectiveness of the self-supervised learning model and should be taken into consideration when designing and training these models.

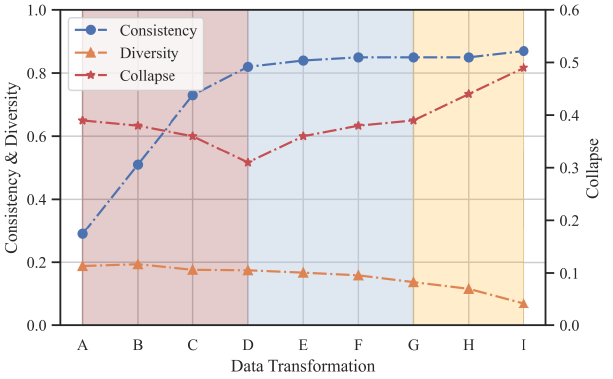


Fig. 4. Curves of consistency, diversity and dimensional collapse with data transformation. A-I are continuously enhanced data transformations. A is *shearing, clipping*. B adds *joint jitter* for 5 joints. C applies it to 10 joints. D applies it to 15 joints. E, F, and G add the data transformations *flipping*, *rotation* and *noise*, respectively, based on D. H and I add *shuffling* and *mixing*, respectively. A-G are invariant contrastive learning. H and I add equivariant contrastive learning. Simultaneously, we have amplified the diversity by a factor of five, ensuring that distinctions are more prominently highlighted.

To investigate the dimensionality of the feature space, we calculate the covariance matrix as follows:

$$\mathbf{M} = \sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T, \tag{18}$$

where $\mathbf{M}$ is the covariance matrix. And $\mathbf{z}_i = f(\mathbf{x}_i)$ is the extracted feature of skeleton data $\mathbf{x}_i$. $\bar{\mathbf{z}} = \sum_{i=1}^{N}\mathbf{z}_i/N$.

Then, we compute the singular value decomposition of the matrix. We observe that some of the singular values are close to 0, *i.e.,* , the dimensional collapse phenomenon occurs. The collapse of the matrix $\mathbf{M}$ is harmful because it leads to a reduction in the amount of information encoded.

To quantify the degree of collapse quantitatively, we propose metrics for measuring the degree of variation in the distribution of singular values:

$$\mathcal{P}(\{\sigma_i\},\{a_i\},\{b_i\}) = \frac{\sum_{i=1}^{d} a_i\sigma_i}{\sum_{i=1}^{d} b_i\sigma_i}, \tag{19}$$

where $\{\sigma_i\}$ are the singular values arranged from largest to smallest, $\{a_i\}$ is a given nondecreasing sequence, and $\{b_i\}$ is a given nonascending sequence. Here, we set $\{a_i\}$ to be from 1 to d and $\{b_i\}$ to be from d to 1. When the difference between the singular values is large, the indicator is close to 0. When the singular values are equal, the indicator is 1.

Next, we observe the effects of consistency and diversity on dimensional collapse.

Consistency, Diversity and Dimensional Collapse with Growing Augmentations: Fig. 4 illustrates the trends of the consistency and diversity measures as the number of data transformations is progressively increased. In this particular scenario, the first three data transformations are incorporated into the training using an invariant contrastive learning approach, while the last two transformations are incorporated using equivariant contrastive learning.

We observe that the curve can be divided into three stages. In the first stage, the model focuses on optimizing consistency, which results in dimensional collapse. This is a common issue

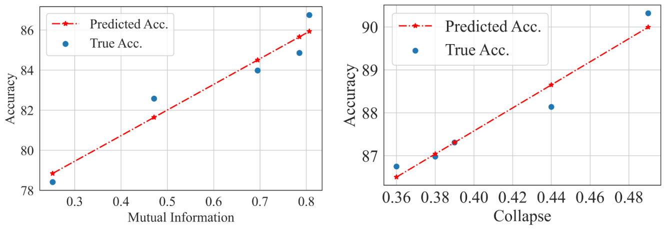


Fig. 5. Curves between mutual information and dimensional collapse with the downstream task accuracy. The mutual information predicts the accuracy under data transformations A-E, and the dimensional collapse predicts the accuracy under E-I in Fig. 4.

TABLE II
LINEAR RELATIONSHIP BETWEEN MUTUAL INFORMATION AND DIMENSIONAL COLLAPSE AND DOWNSTREAM TASK ACCURACY

	Mutual Information	Dimensional Collapse
R^2	0.93	0.94
coeff.	12.9	26.8
intercept.	75.6	76.8

in deep learning, where the model may prioritize minimizing the similarity loss over capturing the underlying structure of the data. Stronger transformations make the discrimination task nontrivial, constraining the model to extract meaningful motion patterns from the skeleton data. However, this increased consistency of positive pairs also leads to the aggregation of nearby features, resulting in the collapse of the dimensionality of the feature space. This can lead to the loss of information inherent in the data.

In the second stage, we introduce more challenging data transformations into the model. More novel motion patterns are exposed, improving the diversity of the model, in addition to the consistency. This helps mitigate the dimensional collapse because more difficult data transformations lead to richer motion patterns and representations. The consistency gradually saturates, so the model mainly optimizes the diversity, which increases the dimensionality of the feature space.

Finally, in the third stage, we incorporate equivariant contrastive learning into the model, which further enhances the optimization of the diversity. Because equivariant contrastive learning extracts diverse features encoding the augmentation-related information, the dimensional collapse problem is further corrected. Thus, the model is able to effectively encode and learn richer and more meaningful representations for the downstream task.

Overall, our proposed method effectively alleviates dimensional collapse by introducing equivariant contrastive learning, giving rise to better performance on the downstream task.

D. Relation to Downstream Tasks

To further explore the mutual information and dimensional collapse in relation to downstream tasks, we employ these two metrics to predict the accuracy of downstream action recognition. Fig. 5 and Table II show the correlations of these two metrics with the action recognition performance. The accuracy of action recognition continues to increase as the mutual information increases. After the mutual information stabilizes, the accuracy continues to improve as the dimensional collapse is alleviated.

We next analyze the performance of these two metrics in relation to downstream tasks from a theoretical perspective. The Bayes error rate P_e represents the lowest error that can be achieved by any classifier that is trained on the given data representations. It is considered to be the benchmark for the performance of a classifier. We let $\mathbf{z}$ and $\mathbf{z}'$ be the extracted features and $\hat{\mathbf{y}}$ be the prediction for labels $\mathbf{Y}$ given $\mathbf{z}$. The Bayes error rate P_e is defined as $P_e = 1 - E_{\mathbf{z}\sim\mathbf{Z}}[\max_{y\in\mathbf{Y}} p(\hat{\mathbf{y}} = \mathbf{y}|\mathbf{z})]$.

Theorem 2 (Bayes Error Rate of Representations): For a given feature distribution $\mathbf{Z}$ *of skeleton data and a distribution* $\mathbf{Z}'$ *of its transformed data, the Bayes error can be estimated as*:

$$\begin{aligned} P_e &\leq 1 - e^{-(H(\mathbf{Y}) - I(\mathbf{Z},\mathbf{Y}))} \\ &= 1 - e^{-(H(\mathbf{Y}) - I(\mathbf{Z},\mathbf{Y}|\mathbf{Z}') - I(\mathbf{Z},\mathbf{Z}',\mathbf{Y}))}. \end{aligned} \tag{20}$$

Therefore, to reduce the error rate, we need to increase two types of mutual information: $I(\mathbf{Z},\mathbf{Y}|\mathbf{Z}')$ and $I(\mathbf{Z},\mathbf{Z}',\mathbf{Y})$. Moreover, $I(\mathbf{Z},\mathbf{Z}',\mathbf{Y}) = I(\mathbf{Z},\mathbf{Z}') - I(\mathbf{Z},\mathbf{Z}'|\mathbf{Y})$. We increase the mutual information $I(\mathbf{Z},\mathbf{Z}')$ when optimizing invariant contrastive learning, so the mutual information $I(\mathbf{Z},\mathbf{Z}',\mathbf{Y})$ will keep increasing. However, the previous method using only invariant contrastive learning extracts only this part of the mutual information. $I(\mathbf{Z},\mathbf{Y}|\mathbf{Z}')$ is ignored, resulting in a suboptimal final result. This information is then extracted by the encoder in equivariant contrastive learning.

Next, we show that alleviating dimensional collapse can boost $I(\mathbf{Z},\mathbf{Y}|\mathbf{Z}')$. Since it is less than the entropy of the feature space $H(\mathbf{Z})$, increasing the entropy of the feature space can raise the upper bound of this part of the mutual information. Therefore, we show that alleviating dimensional collapse can increase the entropy of the feature space.

The covariance matrix is rewritten as

$$\mathbf{M} = \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T, \tag{21}$$

where $\tilde{\mathbf{Z}} = [\mathbf{z}_1 - \bar{\mathbf{z}}, \ldots, \mathbf{z}_N - \bar{\mathbf{z}}]$ is randomly sampled from the centered feature distribution $\mathbf{Z}$. For simplicity of analysis, we assume that each item of the feature distribution obeys the same Gaussian distribution $\mathbf{z}_i - \bar{\mathbf{z}} \sim \mathrm{N}(0, \sigma^2\mathbf{I})$. We estimate their eigenvalues using the Marcenko–Pastur theorem.

Theorem 3 (Marcenko-Pastur Law for Wishart Matrices): Let $\rho(x)$ *be the empirical spectral measure of the random matrix* $\mathbf{M} = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T$*, where* $\tilde{\mathbf{Z}}$ *is a* $d \times N$ *random matrix of i.i.d. Gaussian random variables* $\mathrm{N}(0, \sigma^2)$*. When* N *tends to infinity, the* $\rho(x)$ *converges weakly almost surely to the Marcenko-Pastur law defined by*:

$$\rho(x) \propto \frac{1}{2\pi u \sigma^2} \frac{\sqrt{(a_+ - x)(x - a_-)}}{x}, \tag{22}$$

where $a_\pm = \sigma^2(1 \pm \sqrt{u})^2$ *and* $u = \frac{d}{N}$.

This theorem states that the eigenvalues of the matrix are distributed in an interval $[a_-, a_+]$. The smallest eigenvalue is proportional to the variance σ^2. Alleviating dimensional collapse increases the minimum eigenvalue. Thus, the mitigation of dimensional collapse increases the variance of the feature distribution.

However, the entropy of the Gaussian distribution is

$$H(\mathbf{Z}) = \frac{d}{2}(\ln 2\pi + 1) + d \ln \sigma. \tag{23}$$

As the variance increases, the entropy of the feature space also increases. Hence, we show that by introducing equivariant contrastive learning, we alleviate the dimensional collapse and, thus, increase the entropy of the feature space. Therefore, $I(\mathbf{Z}, \mathbf{Y}|\mathbf{Z}')$ is increased.

IV. Experiment Results

A. Dataset

To fully demonstrate the effectiveness of our method, we evaluate the model on three large-scale datasets:

1) NTU RGB+D 60 Dataset (NTU 60) [61]: There are 56,578 videos in this dataset, with 60 annotations and 25 joints in each frame. These samples were recorded by Microsoft Kinect v2 cameras. We adopt the following two evaluation protocols: a) cross-subject (xsub): the data for training and testing are collected from 40 different subjects; and b) cross-view (xview): the data for training and testing are captured in 3 different views: front view and 45-degree views for the left side and right side.

2) NTU RGB+D 120 Dataset (NTU 120) [62]: This is an extension of NTU 60. It contains a total of 114,480 videos, with 120 action categories. Two recommended protocols are adopted: a) cross-subject (xsub): the data for training and testing are collected from 106 different subjects; and b) cross-setup (xset): the data for training and testing are collected from 32 different setups with different camera locations.

3) PKU Multimodality Dataset (PKUMMD) [63]: PKU-MMD covers a range of detailed information about human activities and a multimodality 3D understanding of human actions. The actions are organized into 51 action categories and include almost 20,000 instances. There are 25 joints in each sample. PKUMMD is divided into two versions, Part I and Part II. In Part II, action recognition is more difficult because the large view variation and heavy occlusion cause more skeleton noise. Experiments are conducted according to a cross-subject protocol and on the two subsets.

B. Implementation Details

1) Data Preprocessing and Training Strategy: For a fair comparison, we follow the experimental settings of recent works [6] [12]. For data preprocessing, first, all sequences of skeletons are downsampled to 300 frames. Then, in every forward pass, we crop and resize the skeleton sequences to 64 frames via the *temporal cropping* transformation to train the model.

We adopt a 3-layer Bi-GRU as the backbone, of which the hidden dimension is set to 1024. MLPs are used as the projection heads $h_q(\cdot)$ and $h_k(\cdot)$, which project the features into the embedding with 128 dimensions. Similarly, we use an MLP as the classifier $h_t(\cdot)$ for temporal shuffling prediction.

During self-supervised pretraining, the model is trained for 450 epochs in total, with a batch size of 128. The initial learning rate is 0.02 and is reduced to 0.002 at 350_{th} epochs. We employ the SGD optimizer with a momentum of 0.9, and the weight decay is 0.0001. The size of the memory bank M is set to 16384. The temperature τ in the InfoNCE loss is set to 0.07, and the temperatures τ_b and τ_s are set as 0.05 and 0.1, respectively. For temporal shuffling transformation, we set the number of clips to be shuffled to 4. λ_{equ} is 1.0.

For a fair comparison, we obtain the fusion results of multiple streams by summing the prediction scores after the fully connected layer following previous works. Specifically, we adopt the following three streams to obtain the ensemble results, denoted *3s-*:

- *Joint*: The 3D positions of the human body joints.
- *Bone*: The differences between the adjacent joints in the same frame according to the human body topology.
- *Motion*: The differences between the coordinates of the same joint in adjacent frames.

2) Data Transformation: Next, we describe the adopted transformations in detail. We select the following transformations as the basic transformation set, which have been widely adopted in previous works [6], [12]:

- *Temporal Cropping:* This transformation randomly selects a subsequence of the original data by sampling a starting frame and a length ratio. Then, the selected sequence is resampled to a fixed length of 64 frames.
- *Shearing:* This transformation slants the human body 3D coordinates to a random angle by using a shear transformation matrix:

$$\mathbf{S} = \begin{bmatrix} 1 & a_{12} & a_{13} \\ a_{21} & 1 & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix}, \tag{24}$$

where a_{ij} is the shear factor randomly sampled from $[-1, 1]$.

- *Joint Jittering:* This transformation randomly selects j of the 25 joints in the skeleton data. Then, these selected joints are masked to zero or perturbed by a uniformly distributed random matrix. We set j to 15 by default.

For the strong transformation set, these strategies are used in our implementation:

- *Flipping:* Considering that the skeleton of the human body is symmetrical, we exchange the positions of the left and right subskeletons based on a probability $p = 0.5$.
- *Rotation:* For all joint coordinate sequences, the *rotation* transformation randomly selects a main rotation axis $\mathbf{A} \in \{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ and chooses a random rotation angle $[0, \pi/6]$. For the remaining two rotation axes, the rotation angle is randomly sampled from $[0, \pi/180]$. This mimics the variability of the camera position in the real world.
- *Gaussian Noise:* Gaussian noise $\mathcal{N}(0, 0.01)$ is added to the skeleton data based on a probability $p = 0.5$.

C. Evaluation and Comparison

In this section, we investigate the quality of the features extracted by our self-supervised model for action recognition. To this end, we evaluate our approach under a variety of evaluation protocols, including unsupervised, semisupervised, KNN, and transfer learning approaches. This provides a

TABLE III

Comparison of Action Recognition Results Under Linear Evaluation on NTU Datasets. Ensemble Represents the Fusion Results of the Joint, Bone, and Motion Streams

Method	Year	Stream	Backbone	NTU 60		NTU 120		PKUMMD I
				xsub (%)	xview (%)	xsub (%)	xset (%)	xsub (%)
Long TGAN [3]	AAAI 2018	Joint	GRU-based	39.1	48.1	-	-	67.7
MS^{2}L [4]	ACM MM 2020			52.6	-	-	-	64.9
P&C [5]	CVPR 2020			50.7	76.3	42.7	41.7	59.9
PCRP [56]	TMM 2021			54.9	63.4	43.0	44.6	-
CRRL [57]	TIP 2022			67.6	73.8	56.2	57.0	-
ISC [6]	ACM MM 2021			76.3	85.2	67.1	67.9	80.9
CMD [12]	ECCV 2022			79.8	86.9	70.3	71.5	-
SkeletonCLR [11]	CVPR 2021	Joint	GCN-based	68.3	76.4	56.8	55.9	-
AimCLR [10]	AAAI 2022			74.3	79.7	-	-	-
HiCLR [13]	AAAI 2023			77.6	82.0	66.8	66.1	-
SkeleMixCLR [52]	arXiv 2022			80.7	85.5	69.0	68.2	-
CPM [58]	ECCV 2022			78.7	84.9	68.7	69.6	88.8
H-Transformer [59]	ICME 2021	Joint	Transformer-based	69.3	72.8	-	-	-
GL-Transformer [8]	ECCV 2022			76.3	83.8	66.0	68.7	-
Ours	-	Joint	GRU-based	**83.9**	**90.3**	**75.7**	**77.2**	**89.7**
3s-HiCo [60]	AAAI 2023	Ensemble	GRU-based	82.6	90.8	75.9	77.3	-
3s-CMD [12]	ECCV 2022			84.1	90.9	74.7	76.1	-
3s-CrosSCLR [11]	CVPR 2021	Ensemble	GCN-based	77.8	83.4	67.9	66.7	84.9
3s-AimCLR [10]	AAAI 2022			78.9	83.8	68.2	68.8	87.4
3s-HiCLR [13]	AAAI 2023			80.4	85.5	70.0	70.4	-
3s-CPM [58]	ECCV 2022			83.2	87.0	73.0	74.0	-
3s-Ours	-	Ensemble	GRU-based	**87.0**	**92.9**	**79.4**	**81.2**	**91.7**

comprehensive comparison with other state-of-the-art methods. We report the top-1 accuracy for all datasets.

1) Unsupervised Learning Approaches: In the unsupervised setting, we use a linear evaluation mechanism to evaluate the quality of the learned feature representation by the encoder $f(\cdot)$. The encoder $f(\cdot)$ is not fine-tuned during the linear evaluation protocol, and the linear classifier is finetuned for the downstream task. Specifically, we use a fully connected layer together with a softmax layer as the classifier. The SGD optimizer is utilized with an initial learning rate of 0.1. The learning rate is reduced to 0.01, 0.001, and 0.0001 at the 20_{th}, 50_{th} and 70_{th} epochs, respectively. The classifier is trained for 100 epochs in total.

Table III shows the results under linear evaluation on NTU and PKUMMD benchmarks. We compare our method with other methods, including GRU-based, GCN-based, and transformer-based methods. Due to the knowledge exposed by invariant and equivariant data transformations, our method achieves a significant improvement and obtains state-of-the-art scores under different protocols. Compared with CMD, the enhancements observed in our proposed method stem from two primary factors. Firstly, the introduction of equivariant contrastive learning serves to mitigate dimensionality collapse, thereby enriching the learned representations. Secondly, the strategic utilization of data transformations and training strategies also contributes to the observed improvements. Compared with HiCLR, we further investigate and successfully introduce the equivariant contrastive learning for another form of strong data augmentation, *i.e.*, the equivariant augmentation. Remarkably, our method with only a single joint stream can perform on par with or better than many multistream-fusion methods, demonstrating the great advantage of our approach. Meanwhile, with the fusion of three streams, the performance of our method further improves. Notably, the proposed method outperforms 3s-CMD by 4.7% and 5.1% on the NTU 120 xsub and xset protocols, respectively. Our method also obtains significant improvements on the PKUMMD dataset, which illustrates the generalization ability of our method.

2) Semi-Supervised Learning Approaches: In semi-supervised settings, we first utilize all data to train the encoder $f(\cdot)$ in a self-supervised manner and then apply only partial training data to finetune the model for the downstream task. The results reflect the generalization performance in scenarios with less labeled data available, where the model tends to face severe overfitting problems. In the implementation, we jointly train the linear classifier and the encoder $f(\cdot)$ for 50 epochs. Subsets of the labeled data, *i.e.*, 1%, 5%, 10%, and 20% of the data, are randomly selected as the training data for evaluation. The initial learning rate is set to 0.01 and is reduced to 0.001 and 0.0001 at the 30_{th} and 50_{th} epochs, respectively.

As shown in Table IV, our method remarkably outperforms other methods on NTU 60. Compared with the latest GRU-based method, CMD, our method achieves significant performance improvements with very small amounts of training data, *i.e.*, 1%, and 5%, which proves the effectiveness of the learned representations. It is noted that the work [58] achieves better performance using 1% data. This is because it adopts GCN as the backbone while our method is based on GRU, which has more parameters and is more vulnerable to

TABLE IV

PERFORMANCE COMPARISON ON NTU 60 IN TERMS OF THE SEMI-SUPERVISED EVALUATION PROTOCOL

Method	NTU 60							
	xview				xsub			
	(1% data)	(5% data)	(10% data)	(20% data)	(1% data)	(5% data)	(10% data)	(20% data)
LongT GAN [3]	-	-	-	-	35.2	-	62.0	-
MS^{2}L [4]	-	-	-	-	33.1	-	65.1	-
ASSL [64]	-	63.6	69.8	74.7	-	57.3	64.3	68.0
ISC [6]	38.1	65.7	72.5	78.2	35.7	59.6	65.9	70.8
CPM [58]	**57.5**	-	77.1	-	**56.7**	-	73.0	-
CMD [12]	53.0	75.3	80.2	84.3	50.6	71.0	75.4	78.7
Ours	54.9	**78.6**	**82.9**	**86.3**	55.2	**74.9**	**78.9**	**81.6**
3s-CrosSCLR [11]	50.0	-	77.8	-	51.1	-	74.4	-
3s-AimCLR [10]	54.3	-	81.6	-	54.8	-	78.2	-
X-CAR [65]	-	70.0	78.2	85.7	-	67.3	76.1	79.4
MAC-Learning [66]	-	70.4	78.5	84.6	-	63.3	74.2	78.4
3s-CMD [12]	55.5	77.2	82.4	86.6	55.6	74.3	79.0	81.8
3s-Ours	**60.7**	**82.0**	**86.0**	**89.0**	**60.8**	**78.5**	**81.7**	**84.7**

TABLE V

KNN EVALUATION RESULTS ON NTU DATASETS

Method	NTU 60		NTU 120	
	xsub	xview	xsub	xset
LongT GAN [3]	39.1	48.1	31.5	35.5
ISC [6]	62.5	82.6	50.6	52.3
CRRL [57]	60.7	75.2	-	-
HiCLR [13]	60.6	73.1	46.0	46.0
CMD [12]	70.6	85.4	58.3	60.9
Ours	**72.3**	**89.3**	**61.1**	**64.0**

TABLE VI

COMPARISON OF THE TRANSFER LEARNING PERFORMANCE ON PKUMMD WITH PRETRAINING ON NTU DATASETS

Method	To PKU II xsub	
	NTU 60 (%)	NTU 120 (%)
LongT GAN [3]	44.8	-
MS^{2}L [4]	45.8	-
ISC [6]	51.1	52.3
CRRL [57]	48.5	-
CMD [12]	56.0	57.0
Ours	**56.5**	**57.8**

the over-fitting problem when the labeled data is extremely little. And when more labeled data is available, *i.e.*, 10%, our approach shows a significant performance improvement. We also compare our method with the latest semi-supervised learning methods, X-CAR and MAC-Learning. Although the self-supervised pretraining method does not rely on any labels, our method can still achieve remarkable performance compared with the semi-supervised learning methods, indicating the strong generalization ability of our method.

3) Supervised Learning Approaches: We apply a K-nearest neighbor (KNN) classifier, which is a nonparametric supervised learning method. It directly evaluates the quality of the feature space learned by the encoder in the self-supervised pretraining stage. We set K=1 to assign the label according to the cosine similarity distance, following a recent approach, CMD [12].

Table V shows the KNN evaluation results on NTU datasets. Our method notably outperforms other methods, including reconstruction-based methods and contrastive learning methods. Compared with CMD, our method achieves 3.9% improvement on NTU 60 xview and 3.1% improvement on NTU 120 xset, verifying the effectiveness of the introduced invariant and equivariant contrastive learning. Our method obtains a highly distinguishable feature space by modeling the different data transformations.

4) Transfer Learning Approaches: We study the transfer representation learning of our method to explore whether it can acquire general knowledge across datasets. Specifically, we first pretrain the encoder on the source dataset and then finetune a linear classifier jointly with the encoder on the target dataset. We choose the NTU datasets as the source datasets and the PKUMMD dataset as the target dataset. All results reported are under the cross-subject protocol. As shown in Table VI, our method gives the best results under the transfer learning evaluation. These results indicate that our method obtains more transferrable knowledge than existing models.

D. Ablation Study

To present a detailed analysis of the proposed method, we show the ablation experiment results in this part. All experiments are conducted using linear evaluation with the cross-view protocol on NTU 60 by default.

1) Invariant and Equivariant Transformations: We first analyze the impact of the proposed invariant and equivariant transformations. As shown in Table VII, different combinations of the transformations are evaluated, as well as the effect of the knowledge distillation objective $\mathcal{L}_{KD}$ for the strong transformation set.

First, it is observed that the strong transformation set can bring a 2+% improvement in performance. This indicates that these transformations can further benefit invariant learning, which supports our claim that they are invariant transformations.

The equivariant transformations, namely, *skeleton mixing* and *temporal shuffling*, are both proven to be beneficial for the model performance. Taking advantage of the novel

TABLE VII

ABLATION STUDY ON THE DIFFERENT TRANSFORMATIONS. *Shuffle* AND *Mix* REPRESENT THE *temporal Shuffling* AND *skeleton Mixing* TRANSFORMATIONS, RESPECTIVELY

Invariant Trans.		Equivariant Trans.		Loss	Top-1
Basic	Strong	*Mix*	*Shuffle*	$\mathcal{L}_{KD}$	Accuracy
✓					84.7%
✓	✓			✓	87.3%
✓	✓	✓		✓	89.2%
✓	✓		✓	✓	88.1%
✓	✓	✓	✓	✓	**90.3%**
✓	✓				86.7%
✓	✓	✓			88.5%
✓	✓	✓	✓		89.7%

TABLE VIII

ABLATION STUDY ON THE WEIGHT λ_{equ}

Weight λ_{equ}	Acc. (%)
$\lambda_{equ} = 0$	87.3
$\lambda_{equ} = 0.5$	89.4
$\lambda_{equ} = 1.0$	**90.3**
$\lambda_{equ} = 1.5$	89.2
$\lambda_{equ} = 2.0$	88.8

motion patterns exposed by the transformations, the model further optimizes the diversity of the representation space. We integrate the invariant and equivariant transformations into our method and obtain the best results, as shown in Table VII.

In addition, we analyze the knowledge distillation design for the strong transformation set. We advocate transferring the knowledge of the basic transformation set to facilitate the learning of a strong transformation set. As shown in Table VII, $\mathcal{L}_{KD}$ can further boost the performance consistently.

2) Weight for Equivariant Contrastive Learning: The equivariant contrastive learning relies on the two equivariant transformations, *i.e.*, *skeleton mixing* and *temporal shuffling*. λ_{equ} controls the loss weight of equivariant contrastive learning, which is highly related to consistency and diversity in the feature space. The effect of λ_{equ} is shown in Table VIII. When λ_{equ} is too large, the model tends to optimize the diversity, and satisfactory consistency cannot be obtained. Instead, when λ_{equ} is too small, the performance of the model degrades because the motion patterns generated by the equivariant transformations are not well utilized and the diversity of the feature space is insufficient. The best performance is achieved when $\lambda_{equ} = 1.0$. It is also observed that training with equivariant contrastive learning always improves the performance.

3) Temporal Shuffling Configurations: In this section, we report ablation studies on the *temporal shuffling* configurations with respect to the loss weight and number of clips C.

- **Loss weight for temporal shuffling**. We explore the impact of the loss weight, which is the loss weight ratio of *temporal shuffling* and *skeleton mixing*, more precisely. In the implementation, we set the weight of $\mathcal{L}_m$ to 1 and change the weight of the $\mathcal{L}_t$ term to analyze the effect. The results are presented in Fig. 6. When the weight increases, it is observed that the performance improves initially and then drops. The best performance is achieved when the weight is equal to 1.0.

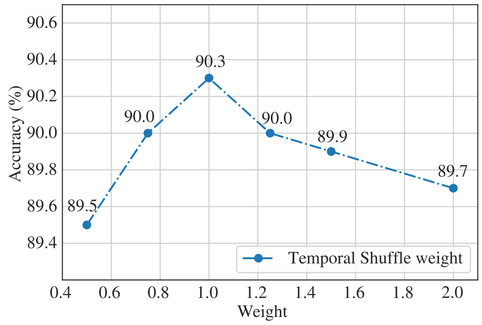


Fig. 6. Ablation study on the loss weight for *temporal shuffling*.

TABLE IX

ABLATION STUDY ON THE CLIP NUMBER C IN *Temporal Shuffle* TRANSFORMATION

Clip Number	Linear	KNN
$C = 3$	90.0%	88.1%
$C = 4$	**90.3%**	**89.3%**
$C = 5$	89.7%	88.2%
$C = 6$	75.7%	73.6%

TABLE X

ABLATION STUDY ON THE IMPLEMENTATION OF *skeleton Mixing* FOR EQUIVARIANT LEARNING

Equivariance Learning Implementation	Accuracy (%)
w/o Skeleton Mix	88.1
Predict the replaced joints	89.9
Maximize similarity with z'_m	**90.3**

- **Number of Clips**. The number of clips C controls the difficulty of predicting *the temporal shuffling* pseudolabel. Table IX shows the impact of the number of clips. We observe a significant performance degradation when C is 6. This is because when C is too large, the pretext task for *temporal shuffling* becomes too difficult, and the model is more susceptible to noise in the data, leading to the learning of useless feature representations. However, when C is too small, the prediction becomes relatively simple, which is not conducive to the model learning better feature representations. We set C to 4, which produces the best results in our experiment.

4) Equivariant Learning for Skeleton Mixing: We compare the two implementations for the equivariant learning of *skeleton mixing* in Table X. Intuitively, we simulate the implementation of *temporal shuffling* to predict the augmentation-related information, *e.g.* the replaced joints in the mixed skeleton sequence, by adding a predictor. In addition, inspired by the mixing label construction according to the mixing ratio λ, we also test the performance of constructing the target feature manually via Eq. 6. The second method performs slightly better than the first method, and both methods yield better results than not using this transformation.

5) FLOPS and Params Results: We estimate the space and computational complexities of the proposed model, as shown in Table XI. The reported results are targeted at the pretraining stage with a batch size of 128. Due to the introduction of different data transformations, our method relies on multiple

TABLE XI

FLOPS AND PARAMS RESULTS OF DIFFERENT MODELS. *Shuffle* AND *Mix* REPRESENT THE *temporal Shuffling* AND *skeleton Mixing* TRANSFORMATIONS, RESPECTIVELY

Models	Params ↓	FLOPs ↓	Accuracy
GL-Transformer [8]	214M	7596.8G	83.8%
ISC [6]	106M	1757.8G	85.2%
CMD [12]	99M	2217.3G	86.9%
Ours	103M	2217.8G	**90.3%**
Ours *w/o Shuffle*	99M	1847.7G	89.2%
Ours *w/o Mix, Shuffle*	99M	1478.2G	87.3%
Ours *w/o* Strong Trans	103M	1478.8G	89.1%

TABLE XII

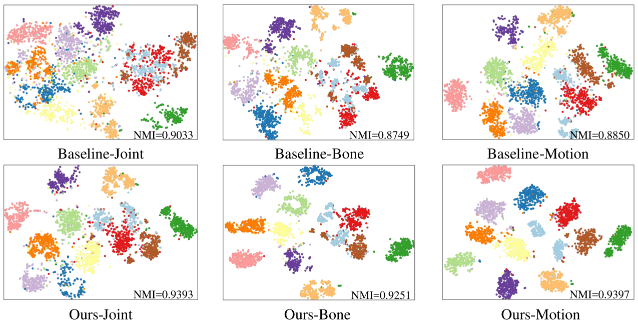
FLOPS AND PARAMS RESULTS OF DIFFERENT ARCHITECTURES

Models	Params ↓	FLOPs ↓	Accuracy	Δ
Transformer	18.54M	3101.6G	84.2%	3.4 ↑
GCN	0.83M	439.8G	83.4%	1.4 ↑
GRU	103M	2217.3G	**90.3%**	3.4 ↑

forward encoding processes for invariant or equivariant learning. This increases the computational cost of our method. For the parameters, we adopt Bi-GRU of the same hidden size as in recent works [6], [12]. The extra space overhead comes mainly from the classification head for the prediction of *temporal shuffling*. First, we compare our method with state-of-the-art methods, *i.e.*, GL-Transformer, ISC, and CMD. Although not optimal in terms of complexity, our approach achieves a significant performance improvement at an acceptable computational cost.

6) Architectures Results: To enhance the generality and versatility of our proposed approach, we applied it to various architectures. Specifically, we test our approach on both GCN, Transformer and GRU architectures, yielding consistent performance improvements across the board in Table XII. GCN exhibits a balance between efficiency and performance, outperforming GRU in terms of model size and resource demand. However, GRU achieves superior performance compared to GCN. The performance of the Transformer is comparable to that of GCN, but it requires more computational resources. Notably, our approach significantly enhances performance, especially when paired with the GRU architecture. This comprehensive experiment culminates in our selection of GRU as the optimal backbone for our method.

7) Visualization Results: To qualitatively demonstrate the effectiveness of our approach, we present a visualization of the results of our method. We compare our method with the baseline method using t-SNE [67] with the same settings. As mentioned before, the baseline method optimizes only $\mathcal{L}_{Info}^{b}$ and employs only invariant contrastive learning. All other training settings and strategies remain the same as those of our method. We randomly select the data of 12 action classes from the validation set for visualization. Meanwhile, we calculate the normalized mutual information (NMI) for the objective. A higher NMI indicates a representation space of higher quality for downstream tasks. As shown in Fig. 7, our method can improve the representation space consistently for all joint, bone, and motion streams compared with the baseline method. Additionally, the NMI results support our claim that our method can generate a more meaningful representation space for the downstream tasks.



Baseline-Joint Baseline-Bone Baseline-Motion

Ours-Joint Ours-Bone Ours-Motion

Fig. 7. t-SNE visualization results of features on the NTU-60 xview benchmark. Our method can improve the representation quality significantly compared with the baseline method.

V. CONCLUSION

In this paper, we propose to combine invariant contrastive learning and equivariant contrastive learning for 3D action representation learning, and achieve significant performance improvement on the downstream action recognition task. For invariant contrastive learning, a self-distillation loss is introduced to optimize the feature consistency, which utilizes the knowledge learned from basic transformations to guide invariant learning under a strong transformation set. For equivariant contrastive learning, we design two data transformations, *skeleton mixing* and *temporal shuffling*, to generate more novel motion patterns. In addition, these equivariant transformations make the model to preserve important data properties for the downstream task. Our experiments show that invariant transformations prioritize consistency, while equivariant transformations prioritize diversity. We demonstrate that equivariant transformations can improve representation quality by alleviating dimensional collapse.

REFERENCES

[1] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12018–12027.

[2] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, “Channel-wise topology refinement graph convolution for skeleton-based action recognition,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13339–13348.

[3] 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 *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 2644–2651.

[4] L. Lin, S. Song, W. Yang, and J. Liu, “MS2L: Multi-task self-supervised learning for skeleton based action recognition,” in *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2020, pp. 2490–2498.

[5] K. Su, X. Liu, and E. Shlizerman, “PREDICT & CLUSTER: Unsupervised skeleton based action recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9628–9637.

[6] F. M. Thoker, H. Doughty, and C. G. Snoek, “Skeleton-contrastive 3D action representation learning,” in *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2021, pp. 1655–1663.

[7] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Skeleton cloud colorization for unsupervised 3D action representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13403–13413.

[8] B. Kim, H. J. Chang, J. Kim, and J. Y. Choi, "Global-local motion transformer for unsupervised skeleton-based action learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 209–225.

[9] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition," *Inf. Sci.*, vol. 569, pp. 90–109, Aug. 2021.

[10] 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 *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 762–770.

[11] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3D human action representation learning via cross-view consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4739–4748.

[12] Y. Mao, W. Zhou, Z. Lu, J. Deng, and H. Li, "CMD: Self-supervised 3D action representation learning with cross-modal mutual distillation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 734–752.

[13] J. Zhang, L. Lin, and J. Liu, "Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations," 2022, *arXiv:2211.13466*.

[14] R. Vemulapalli and R. Chellappa, "Rolling rotations for recognizing human actions from 3D skeletal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4471–4479.

[15] Y. Goutsu, W. Takano, and Y. Nakamura, "Motion recognition employing multiple kernel learning of Fisher vectors using local skeleton features," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 321–328.

[16] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[17] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4263–4270.

[18] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based LSTM networks for 3D action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3459–3471, Jul. 2018.

[19] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. Int. Conf. Signal Process. Mach. Learn.*, 2019, pp. 79–84.

[20] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4570–4579.

[21] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[22] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 7444–7452.

[23] H.-G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "InfoGCN: Representation learning for human skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20154–20164.

[24] H. Duan, J. Wang, K. Chen, and D. Lin, "PYSKL: Towards good practices for skeleton action recognition," in *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2022, pp. 7351–7354.

[25] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.

[26] D. Ahn, S. Kim, H. Hong, and B. Chul Ko, "STAR-transformer: A spatio-temporal cross attention transformer for human action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3319–3328.

[27] Y. Zhou et al., "Hypergraph transformer for skeleton-based action recognition," 2022, *arXiv:2211.09590*.

[28] Y. Yang, G. Liu, and X. Gao, "Motion guided attention learning for self-supervised 3D human action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8623–8634, Dec. 2022.

[29] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2006, pp. 1735–1742.

[30] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," 2019, *arXiv:1906.05849*.

[31] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[32] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 15535–15545.

[33] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6203–6212.

[34] P. Isola, D. Zoran, D. Krishnan, and E. H. Adelson, "Learning visual groups from co-occurrences in space and time," 2015, *arXiv:1511.06811*.

[35] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.

[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[37] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.

[38] X. Chen and K. He, "Exploring simple Siamese representation learning," 2020, *arXiv:2011.10566*.

[39] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020, *arXiv:2006.09882*.

[40] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 6827–6839.

[41] X. Wang and G.-J. Qi, "Contrastive learning with stronger augmentations," 2021, *arXiv:2104.07713*.

[42] X. Huang et al., "Graph contrastive learning for skeleton-based action recognition," 2023, *arXiv:2301.10900*.

[43] Y. Zhou, H. Duan, A. Rao, B. Su, and J. Wang, "Self-supervised action representation learning from partial spatio-temporal skeleton sequences," 2023, *arXiv:2302.09018*.

[44] R. Dangovski et al., "Equivariant contrastive learning," 2021, *arXiv:2111.00899*.

[45] A. Devillers and M. Lefort, "EquiMod: An equivariance module to improve self-supervised learning," 2022, *arXiv:2211.01244*.

[46] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[47] Y. Bai, Y. Yang, W. Zhang, and T. Mei, "Directional self-supervised learning for heavy image augmentations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16671–16680.

[48] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3962–3971.

[49] S. Kim, G. Lee, S. Bae, and S.-Y. Yun, "MixCo: Mix-up contrastive learning for visual representation," 2020, *arXiv:2010.06300*.

[50] K. Lee, Y. Zhu, K. Sohn, C.-L. Li, J. Shin, and H. Lee, "I-mix: A domain-agnostic strategy for contrastive representation learning," 2020, *arXiv:2010.08887*.

[51] Z. Shen, Z. Liu, Z. Liu, M. Savvides, T. Darrell, and E. Xing, "Un-mix: Rethinking image mixtures for unsupervised visual representation learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 2216–2224.

[52] Z. Chen, H. Liu, T. Guo, Z. Chen, P. Song, and H. Tang, "Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition," 2022, *arXiv:2207.03065*.

[53] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[54] S. Ren et al., "A simple data mixing prior for improving self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14575–14584.

[55] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[56] S. Xu, H. Rao, X. Hu, J. Cheng, and B. Hu, “Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition,” *IEEE Trans. Multimedia*, vol. 25, pp. 624–634, 2023.

[57] P. Wang, J. Wen, C. Si, Y. Qian, and L. Wang, “Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition,” *IEEE Trans. Image Process.*, vol. 31, pp. 6224–6238, 2022.

[58] H. Zhang, Y. Hou, W. Zhang, and W. Li, “Contrastive positive mining for unsupervised 3D action representation learning,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 36–51.

[59] 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 *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[60] J. Dong, S. Sun, Z. Liu, S. Chen, B. Liu, and X. Wang, “Hierarchical contrast for unsupervised skeleton-based action representation learning,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2023, pp. 525–533.

[61] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[62] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[63] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, “A benchmark dataset and comparison study for multi-modal human action analytics,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–24, May 2020.

[64] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, and J. Feng, “Adversarial self-supervised learning for semi-supervised 3D action recognition,” 2020, *arXiv:2007.05934*.

[65] B. Xu, X. Shu, and Y. Song, “X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition,” *IEEE Trans. Image Process.*, vol. 31, pp. 3852–3867, 2022.

[66] X. Shu, B. Xu, L. Zhang, and J. Tang, “Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7559–7576, Jun. 2023.

[67] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

Lilang Lin (Graduate Student Member, IEEE) received the B.S. degree in data science from Peking University, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology. His current research interests include action recognition, self-supervised learning, and unsupervised learning.

Jiahang Zhang received the B.S. degree in computer science from Peking University, Beijing, China, in 2023, where he is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology. His current research interests include action recognition and self-supervised learning.

Jiaying Liu (Senior Member, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing, China, in 2010. She is currently an Associate Professor and a Boya Young Fellow with the Wangxuan Institute of Computer Technology, Peking University. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 70 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She is a Senior Member of CSIG and a Distinguished Member of CCF. She was a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA, from 2007 to 2008. She was a Visiting Researcher with Microsoft Research Asia in 2015, supported by the Star Track Young Faculties Award. She has served as a member of the Multimedia Systems and Applications Technical Committee (MSA TC) and the Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society. She received the IEEE ICME 2020 Best Paper Award and the IEEE MMSP 2015 Top10% Paper Award. She has also served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS SYSTEMS FOR VIDEO TECHNOLOGY, and *Journal of Visual Communication and Image Representation*, the Technical Program Chair for ACM MM Asia-2023/IEEE ICME-2021/ACM ICMR-2021/IEEE VCIP-2019, the Area Chair for CVPR-2021/ECCV-2020/ICCV-2019, an ACM ICMR Steering Committee Member, and a CAS Representative for the ICME Steering Committee. She was an APSIPA Distinguished Lecturer from 2016 to 2017.