Computer Vision and Image Understanding 229 (2023) 103655

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

# Global–local contrastive multiview representation learning for skeleton-based action recognition

Cunling Bian [a], Wei Feng [a,*], Fanbo Meng [b], Song Wang [c]

[a] School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[b] Institute of International Engineering, Tianjin University, Tianjin 300350, China
[c] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

ARTICLE INFO

Communicated by Nikos Paragios

*MSC:*
41A05
41A10
65D05
65D17

*Keywords:*
Skeleton-based action recognition
Contrastive representation learning
Multiview
Graph convolutional network

ABSTRACT

Skeleton-based human action recognition has been drawing more interest recently due to its low sensitivity to appearance changes and the accessibility of more skeleton data. However, the skeletons captured in practice are sensitive to the view of an actor, given the occlusion of different human-body joints and the errors in human joint localization. Each view is noisy and incomplete, but important factors, such as motion and semantics, should be shared between all views in action representation learning. We support the classic hypothesis that a powerful representation is one that models view-invariant factors, and so does unsupervised learning. Therefore, we study this hypothesis under the framework of contrastive multiview learning, where we learn a representation for action recognition that aims to maximize the mutual information between different views of the same action sequence. Apart from that, a global–local contrastive loss is proposed to model the multi-scale co-occurrence relationships in both spatial and temporal domains. Extensive experimental results show that the proposed method significantly boosts the performance of unsupervised skeleton-based human action methods on three challenging benchmarks of PKUMMD, NTU RGB+D 60, and NTU RGB+D 120.

## 1. Introduction

Human action recognition plays an important role in video surveillance, human-machine interaction, and sports video analysis (Herath et al., 2017). Different modality information, such as appearance, depth, optical flows, and body skeletons (Bhardwaj et al., 2019) has been used for human action recognition. Among them, the skeleton consists of compact positions of major body joints (Zhang et al., 2020) and can provide highly effective information on human motion underlying different actions (Johansson, 1973; Bian et al., 2021). Skeleton-based action recognition is robust to appearance inconsistencies, different environments, and varying illuminations and is getting more accessible with the rapid development of sensor technology for capturing the skeleton.

Most of the state-of-the-art methods for skeleton-based action recognition use supervised deep learning, which requires large-scale annotated data samples for training (Liu et al., 2020; Cheng et al., 2020). To address this problem, several recent studies attempt to leverage unsupervised learning for skeleton-based action recognition (Zheng et al., 2018; Lin et al., 2020; Su et al., 2020). In these studies, deep representations are learned for skeleton data sequences in terms of tasks like human motion prediction or regeneration, without using any action labels for supervision. For algorithm evaluation, a simple linear classifier is finally trained for action recognition based on both the learned representations and action labels of the training data. At present, there is still a relatively obvious performance gap between the supervised and unsupervised methods for skeleton-based action recognition. One point that we think can be further optimized is that the existing unsupervised skeleton representation learning methods tend to overlook the necessity of controlling irrelevant factors embedded in the inputs, such as view variation and pose deformation.

Skeletons simultaneously captured for the same person from different views are usually different (Zhang et al., 2019), as shown in Fig. 1, even if we try to transform them to the same coordinates. There are many reasons accounting for this phenomenon, such as altered reference coordinates, different occluded joints in different views, and inaccurate human pose estimation. In practice, the skeleton data used for action recognition may be captured from different and even time-varying views (Zhang et al., 2019; Nie et al., 2019). We revisit the classic hypothesis that good representations are the ones that are shared between multiple views. In other words, the viewpoint you view an action should not affect its semantics. For this reason, we propose a new approach to enhancing representation learning by tackling view variation in skeleton-based action recognition without using any manual action labels. Since the training data are unlabeled and have been

* Corresponding author.
*E-mail address:* wfeng@ieee.org (W. Feng).

https://doi.org/10.1016/j.cviu.2023.103655
Received 11 November 2022; Received in revised form 3 February 2023; Accepted 11 February 2023
Available online 14 February 2023

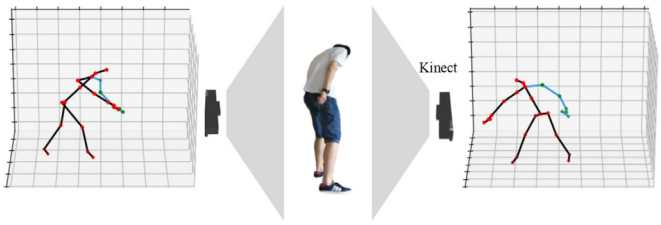



**Fig. 1.** An illustration of view variance of skeleton data: skeleton data simultaneously captured for a same person, but from different views, are usually different.

already formatted and related in data collection and prepossessing, we follow previous studies (Li et al., 2018b; Ji et al., 2021), where a surrogate task is designed to exploit the inherent structure of unlabeled multi-view data for representation learning, by calling our proposed method unsupervised here.

In this paper, we propose global–local contrastive multiview representation learning to enhance unsupervised skeleton-based action recognition. Our goal is therefore to learn representations that capture information shared between multiple views but that are otherwise compact (i.e. discard view-specific nuisance factors). More specifically, the proposed method maximizes the mutual information between features learned from skeletons that are simultaneously captured for the same person from different views. Such representations of different views but the same person are pulled closer to each other in the embedding space through the network training phase. Furthermore, the proposed training loss takes the form of a global–local contrastive one, which can also model the multi-scale co-occurrence relationships between the spatial and temporal domains. In the testing stage, just like in previous works we only take one skeleton data sequence captured from an unknown view as the input of the network for skeleton-based action recognition. We conduct comprehensive evaluation and analysis in our experiments to demonstrate that the proposed method can learn better representations for improving the performance of skeleton-based human action recognition. The proposed method significantly boosts the performance of unsupervised skeleton-based action recognition on three widely used multi-view benchmarks under the linear evaluation protocol.

The main contributions of this paper are as follows:

- A contrastive multiview learning framework for learning view-invariant representations for skeleton-based action recognition is proposed.
- We introduce a local–global spatial–temporal graph contrastive loss, combined with task uncertainty, to model the multi-scale co-occurrence relationship between spatial and temporal domains.
- Compared with existing methods that do not use ground-truth action labels in training, the proposed algorithm significantly boosts the performance on three widely used benchmarks of PKUMMD, NTU RGB+D 60, and NTU RGB+D 120.

The remainder of the paper is organized as follows. Section 2 gives a brief review of the related work on skeleton-based action recognition and contrastive learning. In Section 3, we describe our proposed global–local contrastive multiview representation learning approach. Section 4 describes the benchmark datasets and experimental setting, and reports the experiment results, followed by a brief conclusion in Section 5.

## 2. Related work

### 2.1. Skeleton-based action recognition

Skeleton-based action recognition is a very active and burgeoning area of research, due to its effective representation of motion dynamics. Much of the traditional skeleton-based action recognition work focuses on designing effective handcrafted features, especially the joint or body part based features (Vemulapalli et al., 2014; Yang and Tian, 2012; Xia et al., 2012; Hussein et al., 2013). New methods have recently emerged in the literature to address the skeleton-based action representation with deep learning, including Recurrent Neural Network (RNN), Convolutional Neural Networks (CNN), and Graph Convolutional Network (GCN). Most of them aim to find more effective ways to model temporal and spatial information of skeleton sequences. The structure of RNN is suitable for processing sequential data and prior works have shown that RNN is especially good for handling varying-length skeleton sequences (Wang and Wang, 2017). In order to extract discriminative spatial and temporal features of different actions, Song et al. (2017) propose a spatial and temporal attention module to assign different importance to each joint and frame within a sequence on top of RNN. CNN has the intrinsic ability to learn structural information from 2D or 3D grids, and it has also been used to encode skeleton sequences as pseudo-images for spatial–temporal representation learning (Ke et al., 2017). Liu et al. (2017b) firstly transform skeleton sequence into a series of color images and then enhance visual and motion local patterns through mathematical morphology, finally propose a multi-stream CNN-based model to extract and fuse deep features from the enhanced color images. GCN is the generalization of CNN to graphs and it can well represent the joint-based skeleton data. Therefore the use of GCN can automatically capture the patterns embedded in the spatial configuration of the joints as well as their temporal dynamics (Yan et al., 2018; Liu et al., 2020; Cheng et al., 2020). Cheng et al. (2020) take novel shift graph operations and lightweight point-wise convolutions to replace regular graph convolutions. This way it reduces computation cost and provides flexible receptive fields for both spatial graphs and temporal graphs.

To avoid the laborious labeling of large-scale skeleton data, unsupervised skeleton-based action recognition has been studied by many researchers (Lin et al., 2020; Su et al., 2020). To make better use of the movement patterns introduced by extreme augmentations, Guo et al. (2022) propose a Contrastive Learning framework to utilize abundant information mining. Li et al. (2021) propose a cross-view contrastive learning framework by leveraging multiview complementary supervision signal. Nie and Liu (2021) propose a denoising autoencoder to learn intrinsic pose features through the task of recovering corrupted skeletons. Leveraging the colorized skeleton point cloud, Yang et al. (2021) design an auto-encoder framework that can learn spatial–temporal features. Thoker et al. (2021) propose to learn from multiple different input skeleton representations in a cross-contrastive manner. Kim et al. (2022) propose a transformer model for the task of unsupervised learning of skeleton motion sequences. Most existing methods perform the feature learning by an encoder–decoder structure, the input of which is a masked or original skeleton sequence, and the goal of training is to reconstruct the skeleton sequences from the encoded features. For the same reason, we focus on learning a powerful representation for skeleton-based action recognition that models view-invariant factors without any manual action labeling using a spatial–temporal graph network.

### 2.2. Contrastive learning

Contrastive learning aims to pull together an anchor and a "positive" sample in embedding space while pushing apart the anchor from many "negative" samples (Khosla et al., 2020). Therefore, contrastive losses are adopted to learn effective representations for pretext tasks in an unsupervised fashion. Closely related to contrastive learning is the family of losses based on metric distance learning or triplets that depend on the class label to supervise the choice of positive and negative pairs (Schroff et al., 2015). The key distinction between triplet losses and contrastive losses is that the former use exactly one positive and one negative pair per anchor and the positive pair of them is chosen from the same class and the negative pair is chosen from

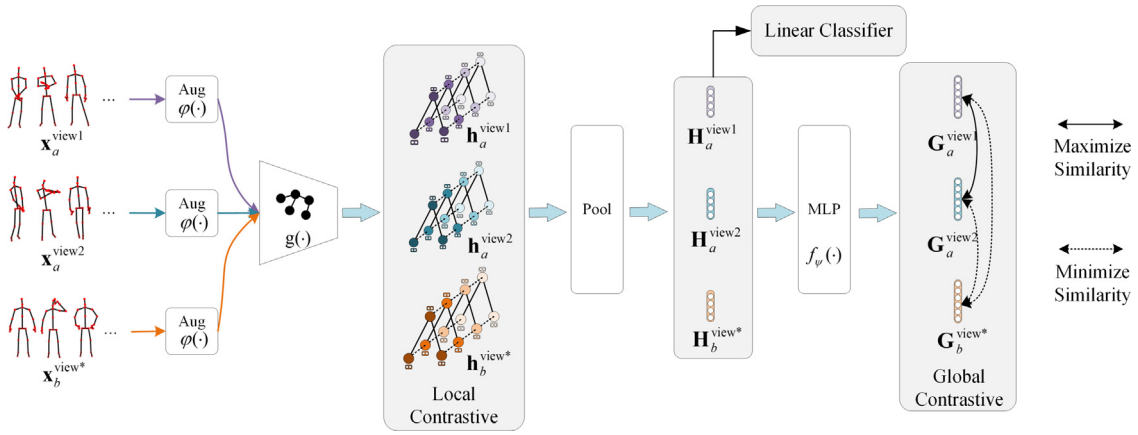



**Fig. 2.** The overall pipeline of the proposed global–local contrastive multiview representation learning for skeleton-based action recognition. $\mathbf{x}_a^{\text{view1}}$ and $\mathbf{x}_a^{\text{view2}}$ are from any two views of the multi-view skeleton sequence $\mathbf{X}_a$. $\mathbf{x}_b^{\text{view*}}$ is from any view of the multi-view skeleton sequence $\mathbf{X}_b$. This approach pulls together skeletons simultaneously captured for the same person from different views in embedding space, while pushing apart the others.

different classes. Contrastive learning generally uses just one positive pair for each anchor sample, selected using either co-occurrence (Hjelm et al., 2018; Henaff, 2020) or data augmentation (Chen et al., 2020). The introduction of contrastive learning leads to a surge of interest in unsupervised visual representation learning (Chen et al., 2020). Wu et al. (2018) maximize distinction between instances via a novel nonparametric softmax formulation and use a memory bank to store the instance class representation vector. For effective similarity measurement between samples in low-dimensional embedding space, other work explores the use of in-batch samples for negative sampling instead of a memory bank (Ye et al., 2019; Ji et al., 2019). Recently, researchers have attempted to relate the success of their methods to the maximization of mutual information between latent representations (Bachman et al., 2019; Henaff, 2020).

In probability theory and information theory, the mutual information of two random variables is a measure of their mutual dependence (Wikipedia, 2021). It has important applications to contrastive learning (Chen et al., 2020). By maximizing mutual information between node and graph representations, some works, focusing on general graphs, have achieved state-of-the-art results in unsupervised node and graph classification tasks (Veličković et al., 2018; Sun et al., 2019). Maximizing mutual information between features extracted from multiple views of a shared context is analogous to human learning to represent observations generated by a shared cause driven by a desire to predict other related observations (Bachman et al., 2019). Aiming at a specific spatial–temporal graph structure, we introduce a global–local contrastive multiview representation learning method for skeleton-based action recognition.

## 3. Global–local contrastive multiview representation learning

Inspired by recent contrastive learning algorithms, we propose an approach to learning a powerful representation that models view-invariant factors without any manual action labeling. It maximizes the mutual information between skeleton sequences that are simultaneously taken for the same person but from different views, via a global–local contrastive loss in the latent space. The overall pipeline of the proposed approach is illustrated in Fig. 2. Specifically, a stochastic data augmentation module $\varphi(\cdot)$ that transforms any given data example randomly to encourage learning a more robust representation for the downstream task. Then, a spatial–temporal graph convolution network (ST-GCN) structural encoder $g(\cdot)$ extracts representation vectors from augmented data examples. We maximize the representation agreement between samples simultaneously taken for the same person but from different views at a global level and a local level. At the global level, a small neural network projection head $f_\psi(\cdot)$ maps the representations to a latent space by applying a global contrastive loss. At the local level, as shown in Fig. 3, an ST-Graph partitioning function $\rho(\cdot)$ splits the graph structural representation of the whole skeleton sequence into multi-local subgraphs, and then a projection head $f_\phi(\cdot)$ maps the representations to a latent space by applying a local contrastive loss. Moreover, to effectively combine global and local contrastive losses, we adjust their relative weights based on task uncertainty.

Before getting into the details of the approach, we state the main notations. Similar to previous studies (Cheng et al., 2020; Zhang et al., 2020), we organize skeleton sequence of an action sample as an undirected spatial–temporal graph $\mathbf{x} = (\mathcal{J}, \mathcal{E})$, where $\mathcal{J} = \{\mathbf{j}_{ti} \mid t = 1,\dots,T; i = 1,\dots,M\}$ denotes a set of vertices, corresponding to $T$ frames and $M$ body joints per frame, and $\mathcal{E}$ is the set of edges, indicating the connections between nodes. Then, we represent a multi-view skeleton sample as $\mathbf{X} = \{\mathbf{x}^v\}_{v=1}^V$, where $V$ represents the number of views, which could be as many as needed, and $v$ indicates the specific $v$th view. For many multiview skeleton samples, we also use $\mathbf{x}_i^v$ to denote the $v$th view of the $i$th multiview skeleton sample $\mathbf{X}_i$.

### 3.1. Multiview skeletal data augmentation

Data augmentation aims to create novel and realistically rational data by applying a certain transformation to the original training data without affecting their semantic meanings. It has been demonstrated that contrastive learning usually needs stronger data augmentation than supervised learning (Chen et al., 2020). Meanwhile, for specific graphs, certain data augmentations might be more effective than the others (You et al., 2020). Let an augmented skeleton sequence be $\hat{\mathbf{x}}_i^v = \varphi(\mathbf{x}_i^v)$, where $\varphi(\cdot)$ is the augmentation function. In this paper, we apply temporal subgraph as the data augmentation, with definitions as follows: it samples a segment from $\mathbf{x}_i^v$ along the temporal dimension. As the length of a skeleton sequence is fixed to 100 frames, we randomly sample 95 consecutive frames and then extend it to 100 frames by linear interpolation. This data augmentation increases the robustness of action recognition when the starting and ending frames of the action cannot be accurately determined and the skeleton sequences captured from different views do not have perfect temporal alignment.

### 3.2. ST-GCN structural encoder

The ultimate goal of the proposed approach is to train a skeleton sequence encoder $g(\cdot)$ to get a powerful representation for skeleton-based action recognition without any manual action labeling. Specifically, to effectively model the co-occurrence relationships among joints in both spatial and temporal domains, we apply an ST-GCN structural encoder, which extracts representation $\mathbf{h}_i^v$ from the augmented skeleton sequence

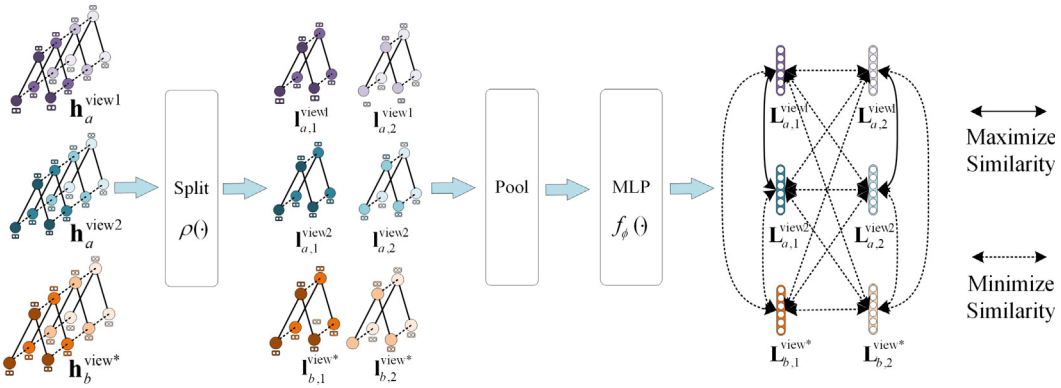



**Fig. 3.** An more detail illustration of local contrastive in Fig. 2. ST-Graph structural representations $\mathbf{h}_a^{\text{view1}}$ and $\mathbf{h}_a^{\text{view2}}$ are from any two views of the multi-view skeleton sequence $\mathbf{X}_a$ while $\mathbf{h}_b^{\text{view*}}$ is from any view of the multi-view skeleton sequence $\mathbf{X}_b$. This loss aims to pull together skeleton sequence regions simultaneously captured for the same person from different views in embedding space, while pushing apart the others.

$\hat{\mathbf{x}}_i^v$. Specifically, it contains two parts: spatial graph convolution and temporal graph convolution.

For spatial graph convolution, the neighbor set of joints is defined as an adjacent matrix $\mathbf{A} \in \{0,1\}^{M\times M}$ according to $\mathcal{E}$, which is typically partitioned into 3 partitions: the centripetal group containing neighboring nodes that are closer to the skeleton center, the node itself and otherwise the centrifugal group. For individual skeleton, let $\mathbf{F} \in \mathbb{R}^{M\times C}$ and $\mathbf{F}' \in \mathbb{R}^{M\times C'}$ denote the input and output feature during the processing respectively, where $C$ and $C'$ are the input and output feature dimensions. The graph convolution is computed as:

$$\mathbf{F}' = \sum_{p\in\mathcal{P}} \bar{\mathbf{A}}_p \mathbf{F} \mathbf{W}_p, \tag{1}$$

where $\mathcal{P} = \{\text{root}, \text{centripetal}, \text{centrifugal}\}$ denotes the spatial partitions, $\bar{\mathbf{A}}_p = \mathbf{\Lambda}_p^{-\frac{1}{2}} \mathbf{A}_p \mathbf{\Lambda}_p^{-\frac{1}{2}} \in \mathbb{R}^{M\times M}$ is the normalized adjacent matrix and $\mathbf{\Lambda}_p^{ij} = \sum_j (\mathbf{A}_p^{ij}) + \alpha$, $\alpha$ is set to 0.001 to avoid empty rows. $\mathbf{W}_p \in \mathbb{R}^{1\times 1\times C\times C'}$ is the weight of the $1 \times 1$ convolution for each partition group. For the temporal dimension, we construct a temporal graph by connecting identical joints in consecutive frames and use regular 1D convolution on the temporal dimension as the temporal graph convolution.

The ST-GCN structural encoder comprises a series of dynamic spatial–temporal graph convolution blocks stacked one above the other. In this form, there existed many specific models with subtle differences (Yan et al., 2018; Shi et al., 2019; Cheng et al., 2020). The proposed approach does not place any restriction on the ST-GCN structural encoder, as long as it maintains the feature of the spatial–temporal graph structure. In our implementation, we adopt the network recently proposed by Cheng et al. (2020) as the ST-GCN structural encoder.

### 3.3. ST-Graph partitioning function

As stated in Li et al. (2018a), the graph convolution operation can be considered Laplacian smoothing for node features over graph topology. The Laplacian smoothing computes the new node features as the weighted average of itself and its neighbors. It helps make nodes in the same cluster tend to learn similar representations. Nevertheless, it may also lead to the over-smoothing problem and make nodes indistinguishable as the number of network layers increases. Meanwhile, it may concentrate more on node features and make the learned embeddings lack structural information. In short, ST-GGN can handle most simple cases but may ignore local details on a complicated graph.

Given the above problems, we enhance the representation by giving more consideration to specific characteristics of local regions. Specifically, we include an ST-Graph partitioning function $\rho(\cdot)$ to split the feature of the whole skeleton sequence $\mathbf{h}_i^v$ into multi-local subgraphs $\mathbf{l}_{i,s}^v, s \in [1,\dots,S]$, where $S$ represents the number of generated subgraphs, $i$ and $s$ indicate sample index and subgraph index, respectively.

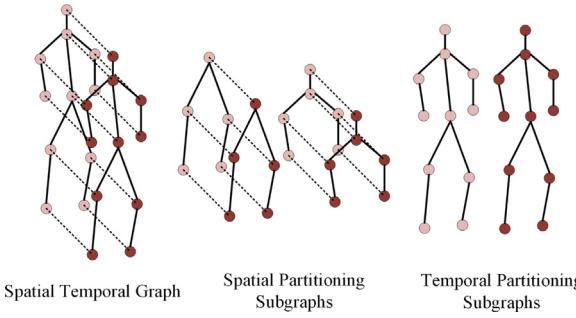



**Fig. 4.** ST-Graph spatial or temporal partitioning strategies. The spatial–temporal feature graph are evenly partitioned along different dimensions by cutting edges.

The choice of partitioning strategies has a strong impact on not only the performance of recognition networks but also the design of the networks (Fan et al., 2020). Several graph partitioning algorithms have already been developed and they are often either edge cut (Andreev and Racke, 2006), which evenly partitions vertices and cuts edges, or vertex cut (Bourse et al., 2014), which evenly partitions edges by replicating vertices. There have also been hybrid algorithms (Li et al., 2019), which cut both edges and vertices. In this paper, we adopt two simple rule-based edge cut style partitioning strategies to segment the skeleton spatial–temporal feature graph. Specifically, vertices of the ST-Graph are evenly partitioned into $S$ segments along the spatial dimension or the temporal dimension by cutting edges, as shown in Fig. 4.

### 3.4. Projection head

Recent work by Chen et al. (2020) found that mapping features to another latent space before contrastive loss calculation can be more effective. In this way, the features before a nonlinear projection are the learned representations, where information loss of raw data induced by the contrastive loss can be relieved. Therefore, in this paper, the representations $\mathbf{h}_i^v$ and $\mathbf{l}_{i,m}^v$ are mapped to another latent space through an MLP with one hidden layer, respectively. We name this module as projection head and add it to global and local contrastive learning subnetworks. Meanwhile, a global pooling is performed on $\mathbf{h}_i^v$ and $\mathbf{l}_{i,m}^v$ to get a fixed dimension feature vector for each ST-Graph to aggregate the node features before the projection head. Formally, the process is defined as:

$$\begin{aligned} \mathbf{G}_i^v &= f_\psi(\text{pool}(\mathbf{h}_i^v)) = \mathbf{W}^{(\psi,2)}\sigma(\mathbf{W}^{(\psi,1)}\text{pool}(\mathbf{h}_i^v)), \\ \mathbf{L}_{i,m}^v &= f_\phi(\text{pool}(\mathbf{l}_{i,m}^v)) = \mathbf{W}^{(\phi,2)}\sigma(\mathbf{W}^{(\phi,1)}\text{pool}(\mathbf{l}_{i,m}^v)), \end{aligned} \tag{2}$$

where $f_\psi(\cdot)$ and $f_\phi(\cdot)$ represent global and local projection heads. $\mathbf{G}_i^v$ and $\mathbf{L}_{i,m}^v$ are the global and local representations in another latent space.

pool(·) is a global pooling function. $\sigma$ is a ReLU nonlinearity and $\mathbf{W}$'s are learned weights of MLP. Note that the output of pool($\mathbf{h}_i^v$) is named as $\mathbf{H}_i^v$, which is the representation we learned that models view-invariant factors.

### 3.5. Global–local contrastive learning

A global representation can well capture the common knowledge of action patterns among all the regions in the skeleton sequence and hence possesses nice merit in terms of model generalization while a local representation targets the personalization of individual regions. As mentioned above, we propose several ST-Graph partitioning strategies to segment the graph into multiple local subgraphs. In this section, a global–local contrastive learning loss is proposed to effectively model the multi-scale co-occurrence relationship between spatial and temporal domains in the ST-Graph. For this, we define different positive pairs in global and local scenarios and maximize the consistency between the positive pairs compared with corresponding negative pairs using global and local contrastive loss functions. Meanwhile, the two contrastive loss functions are combined with task uncertainty in order to balance the trade-off between generalization and personalization of representation.

**Global contrastive loss.** Given two global representations $\mathbf{G}_a^{v1}$ and $\mathbf{G}_b^{v2}$, we specify that they form a positive pair if $a$ is equal to $b$, else they form a negative pair. It means multiple skeleton sequences, if simultaneously taken for the same person from different views, will be pulled together in embedding space, otherwise will be pulled apart, which is shown in Fig. 2. Therefore, not only skeleton representations can be effectively learned without any action label information, but also their view-invariant property of them can be enhanced during multi-view contrastive learning. To achieve this, we adopt the normalized temperature-scaled cross entropy loss (Chen et al., 2020). Specifically, we randomly sample a minibatch of $N$ examples and define the contrastive prediction task on pairs of skeleton sequences. Note that each example consists of $V$ skeleton sequences collected from $V$ different views, resulting in $VN$ data points. Given $V$ positive pairs in an example, we treat the other $V(N-1)$ data points within a minibatch as negative examples. Let $\mathbf{u}$ and $\mathbf{v}$ denote representations of two data points. To measure similarity, we define $\mathrm{sim}(\mathbf{u},\mathbf{v}) = \mathbf{u}^\top\mathbf{v}/\|\mathbf{u}\|\|\mathbf{v}\|$ that denotes the dot product between $\ell_2$ normalized $\mathbf{u}$ and $\mathbf{v}$. Then, the global loss function for positive pairs of example $i$ is defined as

$$\ell_i^{\text{global}} = -\log \frac{\sum_{v1,v2=1}^{V} \mathbb{1}[v1 \neq v2] \exp\left(\text{sim}\left(\mathbf{G}_i^{v1}, \mathbf{G}_i^{v2}\right)/\tau\right)}{\sum_{k=1}^{N}\sum_{v1,v2=1}^{V} \mathbb{1}\left[\begin{smallmatrix} k\neq i \\ v1\neq v2 \end{smallmatrix}\right] \exp\left(\text{sim}\left(\mathbf{G}_i^{v1}, \mathbf{G}_k^{v2}\right)/\tau\right)}, \tag{3}$$

where $\mathbb{1}[v1 \neq v2] \in \{0,1\}$ is an indicator function evaluating to 1 if $v1 \neq v2$, $\mathbb{1}\left[\begin{smallmatrix} k\neq i \\ v1\neq v2 \end{smallmatrix}\right]$ is also an indicator function evaluating to 1 if one of $k \neq i$ and $v1 \neq v2$ is satisfied, otherwise evaluating to 0. $\tau$ denotes a temperature parameter. For a minibatch, the global contrastive loss $\mathcal{L}_{\text{global}}$ is computed across all examples,

$$\mathcal{L}_{\text{global}} = \frac{1}{VN}\sum_{m=1}^{N} \ell_m^{\text{global}}, \tag{4}$$

where $N$ is the batchsize.

**Local contrastive loss.** Local contrastive loss is calculated among the local representations, as illustrated in Fig. 3. Given two local representations $\mathbf{L}_{a,s1}^{v1}$ and $\mathbf{L}_{b,s2}^{v2}$, we specify that they form a positive pair if both $a = b$ and $s1 = s2$ are satisfied, else they form a negative pair. From the composition of the positive and negative pairs, the contrastive loss achieves the same effect as the global one at the local scale when subgraph indices are consistent for all pairs. Besides, it can also handle the over-smoothing and the structural information lacking problems by contrasting among local regions in a sequence when sample indices are consistent for all pairs. The definition of local contrastive loss is basically the same as the global one. But because of the extra subgraph dimension, there are $V-1$ positive pairs and $V(SN-1)$ negative pairs in a sample. Formally, the local contrastive loss function for positive pairs of example $i$ is defined as

$$\ell_i^{\text{local}} = -\log \frac{\sum_{s=1}^{S}\sum_{v1,v2=1}^{V} \mathbb{1}[v1 \neq v2] \exp\left(\text{sim}\left(\mathbf{L}_{i,s}^{v1}, \mathbf{L}_{i,s}^{v2}\right)/\tau\right)}{\sum_{k=1}^{N}\sum_{s1,s2=1}^{S}\sum_{v1,v2=1}^{V} \mathbb{1}\left[\begin{smallmatrix} k\neq i \\ s1\neq s2 \\ v1\neq v2 \end{smallmatrix}\right] \exp\left(\text{sim}\left(\mathbf{L}_{i,s1}^{v1}, \mathbf{L}_{k,s2}^{v2}\right)/\tau\right)}, \tag{5}$$

where $\mathbb{1}\left[\begin{smallmatrix} k\neq i \\ s1\neq s2 \\ v1\neq v2 \end{smallmatrix}\right]$ is an indicator function that needs one of the three inequalities is true. $S$ is the number of split subgraphs. For a minibatch, the local contrastive loss $\mathcal{L}_{\text{local}}$ also needs to be computed across all examples,

$$\mathcal{L}_{\text{local}} = \frac{1}{SVN}\sum_{m=1}^{N} \ell_m^{\text{local}}. \tag{6}$$

Based on the above ST-Graph partitioning function, the ST-Graph can be evenly partitioned into multiply subgraphs along the spatial dimension or the temporal dimension by cutting edges. Corresponding to that, two different forms of local contrastive loss come up: $\mathcal{L}_{\text{spalocal}}$ and $\mathcal{L}_{\text{temlocal}}$, to learn various local representations.

**Global–Local contrastive loss.** Global–Local contrastive loss is concerned about jointly optimizing the related global and local contrastive loss functions. In this paper, the popular approach of using a linear combination of them as a total loss function is abandoned. Because manually tuning their weight hyper-parameters is expensive and intractable. Instead, following the work of Wang et al. (2020), we adjust each loss's relative weight in the total loss function by deriving a multi-task loss function based on maximizing the Gaussian likelihood with task-dependent uncertainty during model training. We define the global–local contrastive loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \frac{1}{\sigma_1^2}\mathcal{L}_{\text{global}} + \frac{1}{\sigma_2^2}\mathcal{L}_{\text{local}} + \log(\sigma_1^2) + \log(\sigma_2^2), \tag{7}$$

where $\sigma_1$ and $\sigma_1$ associate with the task uncertainty and can be interpreted as the relative weights of respective loss terms. $\log(\sigma_1^2)$ and $\log(\sigma_2^2)$ serve as regularizers to avoid over-fitting. All network parameters and the uncertainty task weights are trainable and optimized by gradient backpropagation.

The proposed global–local contrastive multiview representation learning is summarized as Algorithm 1.

## 4. Experiment

### 4.1. Dataset

We evaluate the proposed method on three public available multi-view action recognition benchmarks: NTU RGB+D 60 (Shahroudy et al., 2016), NTU RGB+D 120 (Liu et al., 2019), and PKUMMD (Liu et al., 2017a). We briefly describe them below.

**NTU RGB+D 60 (NTU).** NTU is a large-scale multi-modal action recognition dataset. It is composed of 56,880 samples over 60 classes captured from 40 distinct subjects and three Kinect cameras. Each action in the samples involves one or two people. The dataset is very challenging due to the large intra-class and view variations. The original paper of the NTU recommends two benchmarks: (1) Cross-subject (CS): all samples from a selected group of subjects are used for training and the rest samples for testing. (2) Cross-view (CV): the training set contains samples that are captured by cameras 2 and 3, and the testing set contains videos that are captured by camera 1. We follow this convention and report performance on both benchmarks.

**NTU RGB+D 120 (NTU-120).** NTU-120 is an extended version of NTU. It is composed of 113,945 samples in 120 action categories. Two protocols are implemented here: (1) Cross-Subject (XSub): training

**Algorithm 1:** Global–local contrastive multiview representation learning algorithm

**Input**: Augmentation $\varphi(\cdot)$, global pooling pool$(\cdot)$, ST-Graph partitioning function $\rho(\cdot)$, ST-GCN structural encoder $g(\cdot)$, global and local projection heads $f_\psi(\cdot)$ and $f_\phi(\cdot)$, training multi-view skeleton sequences $\{\mathbf{X}_i = \{\mathbf{x}_i^v\}_{v=1}^V\}_{i=1}^N$, global contrastive loss $\mathcal{L}_{\text{global}}$, local contrastive loss $\mathcal{L}_{\text{local}}$, similarity measurement function $sim(\cdot)$.

**Parameters**: Learnable relative weight parameters for global and local contrastive loss: $\sigma_1$ and $\sigma_2$; number of views $V$; number of split subgraphs $S$; number of samples in one batch $K$; temperature parameter $\tau$.

**while** sampled batch $\{\{\mathbf{x}_i^v\}_{v=1}^V\}_{i=1}^K$ **do**
  **while** $i = 1$ to $K$ **do**
    **while** $v = 1$ to $V$ **do**
      $\mathbf{h}_i^v = g(\varphi(\mathbf{x}_i^v))$
      $\mathbf{G}_i^v = f_\psi(\text{pool}(\mathbf{h}_i^v))$
      $\{\mathbf{l}_{i,s}^v\}_{s=1}^S = \rho(\mathbf{h}_i^v)$
      **while** $s = 1$ to $S$ **do**
        $\mathbf{L}_{i,s}^v = f_\phi(\text{pool}(\mathbf{l}_{i,s}^v))$
      **end while**
    **end while**
  **end while**
  **while** $i = 1$ to $K$ **do**
    $$\ell_i^{\text{global}} = -\log \frac{\sum_{v1,v2=1}^V \mathbb{1}[v1\neq v2]\exp\left(\text{sim}\left(\mathbf{G}_i^{v1},\mathbf{G}_i^{v2}\right)/\tau\right)}{\sum_{k=1}^K \sum_{v1,v2=1}^V \mathbb{1}\left[\begin{smallmatrix} k\neq i \\ v1\neq v2\end{smallmatrix}\right]\exp\left(\text{sim}\left(\mathbf{G}_i^{v1},\mathbf{G}_k^{v2}\right)/\tau\right)}$$
    $$\ell_i^{\text{local}} = -\log \frac{\sum_{s=1}^S \sum_{v1,v2=1}^V \mathbb{1}[v1\neq v2]\exp\left(\text{sim}\left(\mathbf{L}_{i,s}^{v1},\mathbf{L}_{i,s}^{v2}\right)/\tau\right)}{\sum_{k=1}^K \sum_{s1,s2=1}^S \sum_{v1,v2=1}^V \mathbb{1}\left[\begin{smallmatrix} k\neq i \\ s1\neq s2 \\ v1\neq v2\end{smallmatrix}\right]\exp\left(\text{sim}\left(\mathbf{L}_{i,s1}^{v1},\mathbf{L}_{k,s2}^{v2}\right)/\tau\right)}$$
  **end while**
  $\mathcal{L}_{\text{global}} = \frac{1}{VK}\sum_{m=1}^K \ell_m^{\text{global}}$
  $\mathcal{L}_{\text{local}} = \frac{1}{SVK}\sum_{m=1}^K \ell_m^{\text{local}}$
  $\mathcal{L} = \frac{1}{\sigma_1^2}\mathcal{L}_{\text{global}} + \frac{1}{\sigma_2^2}\mathcal{L}_{\text{local}} + \log(\sigma_1^2) + \log(\sigma_2^2)$
  update networks $g(\cdot)$, $f_\psi(\cdot)$, $f_\phi(\cdot)$, $\sigma_1$ and $\sigma_2$ to minimize $\mathcal{L}$
**end while**
**return** encoder model $g(\cdot)$, and throw away projection heads $f_\psi(\cdot)$ and $f_\phi(\cdot)$

data and testing data are collected from different subjects. (2) Cross-Setup (XSet): training data and testing data are collected from different setups.

**PKUMMD.** PKUMMD is a new large-scale benchmark for continuous multi-modality 3D human action understanding and covers a wide range of complex human activities with well-annotated information. It contains almost 20,000 action instances in 51 action categories, performed by 66 subjects in three different view Kinect sensors. PKUMMD consists of two subsets: PKUMMD-I is an easier subset for action recognition, while PKUMMD-II is more challenging with more skeleton noise caused by large view variation. We conduct experiments under the cross-subject protocol on the two subsets.

### 4.2. Implementation details

#### 4.2.1. Pre-training without any action label information

In proposed approach, an ST-GCN structural encoder $g(\cdot)$, a global projection head $f_\phi(\cdot)$ and a local projection head $f_\psi(\cdot)$ are pre-trained using multi-view skeleton sequences without any action label information. We use SGD with Nesterov momentum 0.9 to pre-train them for 40 epochs. The learning rate is set to 0.1 and divided by 10 at epochs 20, 30, and 35. The batch size is set to 16 for all experiments. The sequence length $T$ is set to 100. The temperature parameter for global–local contrastive loss is set to 0.07. The number of subgraph $S$ is set to 5. $V$ is set to 2, which means each sample includes two skeleton sequences, simultaneously taken from different views.

#### 4.2.2. Evaluation protocol

To validate the effectiveness of the proposed representation learning method, we follow the linear evaluation protocol (Wang et al., 2020; Chen et al., 2020), which is commonly used to evaluate unsupervised learning methods. In this way, a linear classifier attached to the frozen encoder model $g(\cdot)$ is trained with the annotated dataset. We report Top-1 accuracy on the testing set as a quantitative evaluation indicator. The classifier is trained for 45 epochs, with the learning rate divided by 10 at epochs 25, 35, and 40. The other settings remain the same as the pre-training.

### 4.3. Comparison experiments

To quantitatively evaluate the performance, Tables 1 and 3 list the linear evaluation results of our approach and other state-of-the-art unsupervised methods on PKUMMD and NTU benchmarks. The model which only trains the linear classifier and freezes the randomly initialized encoder is denoted as ST-Graph Rand. We regard this model as one of our baselines. The models implementing ST-Graph contrastive learning in single-view and multi-view scenarios are denoted as ST-Graph CSRL and ST-Graph CMRL, respectively. In the single view version, we maximize the mutual information between skeleton ST-Graph representations of one augmented instance and another augmented instance of an identical skeleton sequence, to learn inherent action patterns of different skeleton transformations. For the evaluation of P&C FW on the action recognition task, we reproduce the coder of P&C FW with a linear evaluation protocol. The temporal subgraph is the default data augmentation method we adopt in these experiments.

#### 4.3.1. Comparison with state-of-the-art

In existing studies (Su et al., 2020; Lin et al., 2020), the pre-training and evaluation are usually conducted on the same dataset. An overall summary of the results is given in Tables 1 and 2, where the proposed method has returned significantly improved performance in the unsupervised methods that do not use action labels for training. As we can see, ST-Graph CMRL is far beyond the performance of random baseline and other state-of-the-art unsupervised methods and greatly reduces the gap to the models trained with action annotation. NTU (CV) is a suitable benchmark to evaluate the model's robustness to the

**Table 1**
Comparison of action recognition performance of the proposed approaches and other state-of-the-art methods.

| Supervised | Models | PKUMMD-I | PKUMMD-II | NTU (CS) | NTU (CV) |
|---|---|---|---|---|---|
| Yes | ST-Graph | 94.5 | 56.8 | 87.8 | 95.1 |
| No | ST-Graph Rand | 30.1 | 10.6 | 19.6 | 23.2 |
| No | LongT GAN (Zheng et al., 2018) | 67.7 | 26.0 | 52.1 | – |
| No | P&C FW (Su et al., 2020) | 67.6 | 35.9 | 32.5 | 35.7 |
| No | M$^2$SL (Lin et al., 2020) | 64.9 | 27.6 | 52.6 | – |
| No | CAE+ (Rao et al., 2021) | – | – | 58.5 | 64.8 |
| No | SkeletonCRL (Li et al., 2021) | 80.9 | – | 68.3 | 76.4 |
| No | CrosSCRL (Li et al., 2021) | – | – | 72.9 | 79.9 |
| No | AimCRL (Guo et al., 2022) | 83.4 | – | 74.3 | 79.7 |
| No | SeBiReNet (Nie and Liu, 2021) | – | – | – | 79.7 |
| No | Colorization (Yang et al., 2021) | – | – | 71.6 | 79.9 |
| No | ST-Graph CSRL (**Ours**) | 68.4 | 31.8 | 60.2 | 59.8 |
| No | ST-Graph CMRL (**Ours**) | **83.6** | **39.9** | **74.7** | **82.6** |

**Table 2**
Unsupervised results on NTU-120.

| Models | NTU-120(XSub) | NTU-120(XSet) |
|---|---|---|
| LongT GAN (Zheng et al., 2018) | 39.7 | 35.6 |
| P&C FW (Su et al., 2020) | 44.1 | 41.1 |
| CAE+ (Rao et al., 2021) | 48.6 | 49.2 |
| SkeletonCLR (Li et al., 2021) | 56.8 | 55.9 |
| AimCLR (Guo et al., 2022) | 63.4 | 63.4 |
| Skeleton Contrastive (Thoker et al., 2021) | 67.9 | 67.1 |
| ST-Graph CMRL (**Ours**) | **69.2** | **68.7** |

**Table 3**
Performance of transfer learning setting in linear evaluation.

| Supervised | Models | PKUMMD-I | PKUMMD-II |
|---|---|---|---|
| Yes | ST-Graph | 90.6 | 55.0 |
| No | P&C FW (Su et al., 2020) | 63.3 | 23.6 |
| No | M$^2$SL (Lin et al., 2020) | – | 44.8 |
| No | LongT GAN (Zheng et al., 2018) | – | 45.8 |
| No | Skeleton Contrastive (Thoker et al., 2021) | – | 45.9 |
| No | ST-Graph CSRL (**Ours**) | 76.3 | 39.8 |
| No | ST-Graph CMRL (**Ours**) | **82.2** | **47.0** |

viewpoint difference. Here, we can see that model's Top-1 accuracy of ST-Graph CMRL in NTU (CV) is 82.6%, while ST-Graph Rand and P&C FW are only 23.2% and 35.7%, respectively. Therefore, the multi-view contrastive learning significantly improved the view-invariant property of skeleton representation. As most recent unsupervised results are reported on the NTU-120 dataset, we also compare the proposed method with unsupervised methods. As shown in Table 2, our method defeats the other unsupervised method on both XSub and XSet protocols. Even in a single view scenario, under the truly unsupervised setting, the performances of ST-Graph CSRL are quite outstanding, which performs better than almost all the baselines. It achieves high recognition accuracies of 60.2% and 59.8% on NTU (CS) and NTU (CV), respectively, which proves that our global–local contrastive learning of augmented skeletons of the same sample also works well. From the comparison of ST-Graph CSRL and CMRL, we can see that significant improvements are made in each benchmark. It proves that CRL between the multi-view skeletons brings in a giant performance leap for unsupervised skeleton-based action recognition.

#### 4.3.2. Transfer learning performance

To further evaluate whether the proposed approach can gain knowledge to related tasks, we investigate the transfer learning performance of our model (Lin et al., 2020). As the representations learned from large-scale data are more generalizable, we regard the NTU as the source dataset and PKUMMD-I and PKUMMD-II as the target datasets. We conduct the pre-training on source datasets and the evaluation on target datasets. Under this setting, the samples used for pre-training and linear evaluation are completely different in terms of viewpoints, action patterns, and so on, which is more following the practical scenarios. The results are summarized in Table 3, from which our approach gets better results of 82.2% for PKUMMD-I and 47.0% for PKUMMD-II, when models are pre-trained without action annotations. Apart from that, together with Table 1, we can see that the accuracies of P&C FW are reduced from 67.6% and 35.9% to 63.3% and 23.6%, respectively, while our approaches boost the accuracies from 68.4%, 31.8%, 83.6%, and 39.9% to 76.3%, 39.8%, 82.2%, and 47.0%, respectively, when the training and testing datasets are from consistent to inconsistent. Meanwhile, the performances of models pre-trained with action annotations also decrease in this transfer learning setting from 94.5% and 56.8% to 90.6% and 55.0%, respectively. One possible reason is that our approach can take advantage of a larger training set more effectively with less influence from the data distribution difference between different datasets. It can be concluded that the representations learned in the proposed approach have a good generalization ability.

**Table 4**
Analysis of global–local contrastive multiview learning loss function.

| Loss Function | NTU (CS) | NTU (CV) |
|---|---|---|
| $\mathcal{L}_{\text{global}}$ | 69.7 | 73.7 |
| $\mathcal{L}_{\text{temlocal}}$ | 67.4 | 74.9 |
| $\mathcal{L}_{\text{spalocal}}$ | 61.0 | 65.9 |
| $\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{temlocal}}$ | **74.7** | **82.6** |
| $\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{spalocal}}$ | 73.6 | 79.0 |
| $\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{temlocal}} + \mathcal{L}_{\text{spalocal}}$ | 74.2 | 81.5 |

### 4.4. Ablation experiments

For a specific ST-Graph structural encoder, the performance of the proposed approach is mainly determined by the following four components: multi-view skeleton contrastive mechanism, data augmentation, projection head, and global–local contrastive loss. From the results of ST-Graph CSRL and ST-Graph CMRL in Tables 1 and 3, the performance of the multi-view skeleton contrastive mechanism is shown to be impressive in all cases. To further assess the other factors, we conduct several ablation experiments on NTU with a linear evaluation protocol.

#### 4.4.1. The effect of global–local contrastive loss

In this experiment, we evaluate different forms of the contrastive loss function. Experimental results are summarized in Table 4. Based on the results, we make the following observations. As the accuracy of $\mathcal{L}_{\text{temlocal}}$ is higher than $\mathcal{L}_{\text{spalocal}}$ by 6.4%(CS) and 9.0%(CV), the temporal splitting method is superior to the spatial splitting in this experiment for local contrastive loss. The impacts of the global and the local losses are different but complementary. Compared with using only one of them, the combined global–local loss function $\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{temlocal}}$ leads to substantially better performance in two benchmarks, i.e., 74.7%(CS) and 82.6%(CV). In Table 4, it can be found that $\mathcal{L}_{\text{spalocal}}$ only produces poor performance. We think the reason might be related to the multi-view action datasets. Specifically, as most multiview action datasets

**Table 5**
Performance by using different methods to combine the global and local contrastive learning losses.

| Viewpoint | Loss combination | NTU (CS) | NTU (CV) |
|---|---|---|---|
| Single-view | Linear | 54.1 | 56.2 |
| | Task uncertainty | 60.2 | 59.8 |
| Multi-view | Linear | 71.1 | 78.7 |
| | Task uncertainty | **74.7** | **82.6** |

**Table 6**
Linear evaluation of representations with different projection heads and various dimensions of output. The representation, before projection, is 256-dimensional here.

| Projection head | Identity mapping | Nonlinear projection | | | | |
|---|---|---|---|---|---|---|
| Output dimension | 256 | 32 | 64 | 128 | 256 | 512 |
| Accuracy | 66.2 | 74.3 | 74.3 | 74.4 | 74.4 | **74.7** |

**Table 7**
Performance of ST-Graph CMRL using different augmentation strategies.

| Augmentation | NTU (CS) | NTU (CV) |
|---|---|---|
| Original | 69.9 | 78.4 |
| Node dropping | 69.1 | 77.4 |
| Node perturbation | 66.3 | 75.0 |
| View rotation | 70.4 | 78.0 |
| Shear | 69.6 | 77.6 |
| Temporal subgraph | **74.7** | **82.6** |

do not provide the corresponding relation between persons in different views, the spatial partitioning strategy is likely to lead to a phenomenon that the positive pairs of local parts are from different persons when an action is performed by two people. In this case, the effect of $\mathcal{L}_{\text{spalocal}}$ is inconsistent with our expectation and one of its impacts, learning a fine-grained view-irrelevant representation in the spatial dimension, will fail, while others still work. However, when combined with $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{temlocal}}$, the main role of $\mathcal{L}_{\text{spalocal}}$ is reflected in learning a fine-grained view irrelevant representation in the spatial dimension. This is why its accuracy goes down. Therefore, the default form of global–local contrastive loss consists of $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{temlocal}}$, combined with task uncertainty in this paper. We also compare linear and task uncertainty-based methods to combine the global and local contrastive losses. Note that all the weight parameters are uniformly set to 1 in the linear combination method. Results are shown in Table 5, from which we can see that the task uncertainty-based combination method outperforms the linear combination methods in both single-view and multi-view scenarios.

#### 4.4.2. Analysis of the projection head

We study the importance of including a projection head, i.e. $f_\psi(\cdot)$ and $f_\phi(\cdot)$. Table 6 shows the linear evaluation results using two different architectures for the head: identity mapping and the nonlinear projection with one additional hidden layer. We can observe that a nonlinear projection head, regardless of its output representation dimension, performs better than identity mapping in terms of recognition accuracy. Therefore, it can be concluded that the hidden layer before the projection head is a better representation than the layer after.

#### 4.4.3. The effects of data augmentation

Apart from the temporal subgraph, we also explore other four popular skeleton data augmentations in experiments including node dropping, node perturbation, view rotation, and skeleton shearing, with definitions as follows:

Node dropping. It randomly discards body joints in the input skeleton sequence $\mathbf{x}_i^v$. Specifically, with a 50% chance, we randomly drop 10% of nodes, where the corresponding joint coordinates are set to zero. It is a common phenomenon that a subset of joints, e.g., those occluded ones, cannot be detected. The augmentation of node dropping enables the crucial action patterns can still be learned from a subset of joints.

Node perturbation. The coordinates of joints are perturbed using a normal Gaussian distribution. The mean of the distribution is set to 0 while the standard deviation is set to 0.05. The detected joint locations, even for those without occlusion, always contain errors due to sensor and estimation accuracies in practice. The augmentation of node perturbation enables the action recognition to be robust to such errors.

View rotation. It randomly rotates the joint coordinates in a skeleton sequence along three axes in terms of a rotation matrix. Specifically, we randomly select three degrees $\alpha, \beta, \gamma$, all uniformly in the range of $[-17°, 17°]$ for each sequence. Three basic rotation matrices with rotation angles about X, Y, and Z axis are given as follows:

$$\begin{aligned}
\mathbf{R}_{\text{X}}(\alpha) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix}, \\
\mathbf{R}_{\text{Y}}(\beta) &= \begin{bmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{bmatrix}, \\
\mathbf{R}_{\text{Z}}(\gamma) &= \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}.
\end{aligned} \tag{8}$$

Based on these three basic rotation matrices, the final rotation matrix is

$$\mathbf{R} = \mathbf{R}_{\text{X}}(\alpha)\mathbf{R}_{\text{Y}}(\beta)\mathbf{R}_{\text{Z}}(\gamma). \tag{9}$$

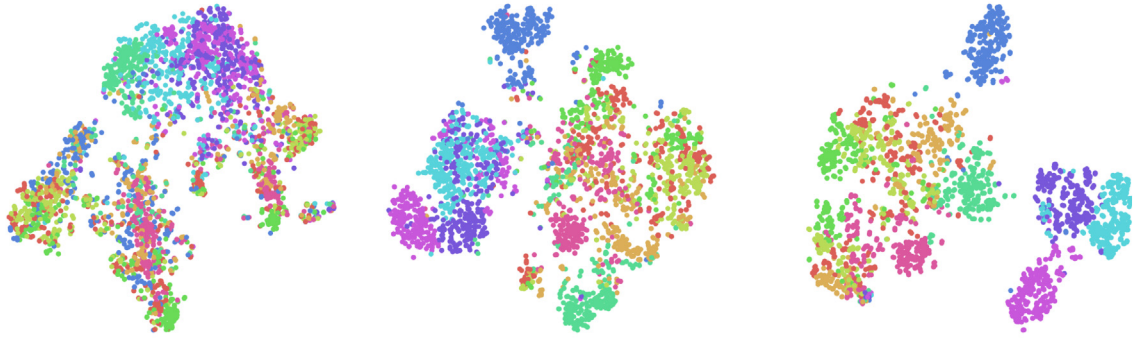
We apply the rotation matrix $\mathbf{R}$ to the original coordinates of the skeleton sequence and get the transformed coordinates. It simulates the view changes of the camera. This augmentation enables the action recognition to be robust to the camera view changes.

Skeleton shearing. It slants the shape of the skeleton at a random angle. The shearing factors are drawn from a uniform distribution in $[0.01, 0.1]$. The transformation matrix can be written as

$$\mathbf{S} = \begin{bmatrix} 1 & s_{\text{X}}^{\text{Y}} & s_{\text{X}}^{\text{Z}} \\ s_{\text{Y}}^{\text{X}} & 1 & s_{\text{Y}}^{\text{Z}} \\ s_{\text{Z}}^{\text{X}} & s_{\text{Z}}^{\text{Y}} & 1 \end{bmatrix}, \tag{10}$$

where $s_{\text{X}}^{\text{Y}}$, $s_{\text{X}}^{\text{Z}}$, $s_{\text{Y}}^{\text{X}}$, $s_{\text{Y}}^{\text{Z}}$, $s_{\text{Z}}^{\text{X}}$, $s_{\text{Z}}^{\text{Y}}$ are shearing factors. All joint coordinates of the original skeleton sequence are transformed with the shearing matrix $\mathbf{S}$. The augmentation of skeleton shearing further increases the robustness of action recognition to more nonrigid transformations of the skeleton sequence.

We denote the model without any data augmentation as the original. The results are shown in Table 7. Much to our surprise, compared with directly using the original sequence, only the temporal subgraph strategy to ST-Graph CMRL can significantly improve the accuracy by 4.8%(CS) and 4.2%(CV). Two possible reasons are: (1) defining precise frame-level starting and ending time for action is almost impossible, and (2) it is hard to achieve a strict temporal alignment of skeleton sequences captured by multiple cameras. Applying inconsistent temporal subgraphs for different views can improve the robustness of these unavoidable problems without breaking their original relationships. The counterproductive of other data augmentations may be due to the fact that the original correspondences in spatial structure among skeletons, which are simultaneously taken from different views, are destroyed after these random transformations. For example, compared with other augmentations, node perturbation, which changes the values of joints with a normal Gaussian distribution, is most damaging to spatial structure correspondences and leads to the sharpest performance drop of 3.6%(CS) and 3.4%(CV). Thus, the temporal subgraph is the default data augmentation we adopt in this work.



(a) P&C FW (Su et al., 2020) (b) ST-Graph CSRL (c) ST-Graph CMRL

**Fig. 5.** TSNE-embedding visualizations of the learned representations from 10 classes randomly selected in NTU (CS) testing set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 8**
Contrastive learning between multiple skeleton generation methods results on NTU.

| Skeleton generation method | NTU (CS) | NTU (CV) |
|---|---|---|
| OpenPose | 58.4 | 60.9 |
| OpenPose, OpenPifPaf | 69.4 | 78.7 |
| OpenPose, OpenPifPaf, Detectron | 71.5 | 80.3 |

### *4.5. Application without multi-view data*

The core of our motivation is that important features for skeleton-based action recognition, such as motion and semantics, should be irrelevant to some variable factors in action representation learning. In this paper, we focus on viewpoint variation. However, multi-view data would become infeasible in some scenarios. In this scenario, the proposed approach can also maximize the mutual information between other irrelevant factors of the same action sequence to learn better action representation. Here we take the skeleton generation method as an example. Specifically, we first pre-train model on NTU60 with multiple skeleton generation methods, including OpenPose (Cao et al., 2017), OpenPifPaf (Kreiss et al., 2021), and Detectron (Wu et al., 2019), and then transfer it to the dataset with skeleton data generated by OpenPose for linear evaluation. In Table 8, the proposed approach also can bring significant improvement for unsupervised skeleton-based action recognition.

### *4.6. Visualization of skeleton representation*

Superior performance of ST-Graph CRL over the existing methods is largely due to the use of the multi-view skeleton contrastive mechanism. Hence apart from the quantitative evaluation, we also visualize the feature changes by using this mechanism. We randomly select ten classes in the NTU testing set and visualize the TSNE-embeddings of the features obtained from P&C FW (Su et al., 2020), ST-Graph CSRL, and ST-Graph CMRL for the same skeleton sequences in Fig. 5. Here we observe that even in this 2D embedding it is clearly evident that the features for different classes are better separated by ours than P&C FW. Points of different colors are mixed up in (a) while they are more separated in (b) and (c). Meanwhile, points of the same color in (c) are more concentrated than those in (b). For example, there is a clear line among points with a different color at the bottom right of (c) while they are mixed up at the bottom left of (b). This supports the conclusion that CRL between multi-view skeletons makes the learned representation more discriminative.

## 5. Conclusion

In this paper, we studied the problem of learning powerful representations for skeleton-based action recognition without any manual action labeling. Our proposal is to make the representations good at modeling view-invariant factors. To this end, a global–local contrastive multiview representation learning approach was developed to maximize the mutual information between the representations extracted from multiple skeleton data simultaneously taken from different views. Specifically, we explored five popular skeleton data augmentation methods and found only the temporal subgraph can make a positive role in the CRL framework. Then, in order to support our global–local CRL, partitioning functions were designed to segment ST-Graph into multiple subgraphs along spatial or temporal dimensions and projection heads were added to map the learned representations to another latent space. Besides, we proposed a local–global spatial–temporal graph contrastive loss, combined with task uncertainty, to model the multi-scale co-occurrence relationship between spatial and temporal domains. Experiments on three multi-view action datasets showed that our proposed approach, no matter in single-view or multi-view scenarios, got competitive performance compared with the random baseline and other state-of-the-art unsupervised skeleton-based action recognition methods. In the future, we will explore new approaches to effectively handle multi-view multi-person scenarios.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Acknowledgments

This work was supported, in part, by the National Key Research and Development Program of China under Grants 2020YFC1522701 and National Natural Science Foundation of China under Grants 62072334, U1803264.

## References

Andreev, K., Racke, H., 2006. Balanced graph partitioning. Theory Comput. Syst. 39 (6), 929–939.

Bachman, P., Hjelm, R.D., Buchwalter, W., 2019. Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910.

Bhardwaj, S., Srinivasan, M., Khapra, M.M., 2019. Efficient video classification using fewer frames. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Bian, C., Feng, W., Wan, L., Wang, S., 2021. Structural knowledge distillation for efficient skeleton-based action recognition. IEEE Trans. Image Process. 30, 2963–2976.

Bourse, F., Lelarge, M., Vojnovic, M., 2014. Balanced graph edge partition. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y., 2017. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709.

Cheng, K., Zhang, Y., He, X., Chen, W., Lu, H., 2020. Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Fan, W., Jin, R., Liu, M., Lu, P., Luo, X., Xu, R., Yin, Q., Yu, W., Zhou, J., 2020. Application driven graph partitioning. In: Proceedings of the ACM SIGMOD International Conference on Management of Data.

Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., Ding, R., 2022. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, (1), pp. 762–770.

Henaff, O., 2020. Data-efficient image recognition with contrastive predictive coding. In: Proceedings of the International Conference on Machine Learning.

Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition: A survey. Image Vis. Comput. 60 (1), 4–21.

Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y., 2018. Learning deep representations by mutual information estimation and maximization. In: Proceedings of the International Conference on Learning Representations.

Hussein, M.E., Torki, M., Gowayyed, M.A., El-Saban, M., 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In: Proceedings of the International Joint Conference on Artificial Intelligence.

Ji, X., Henriques, J.F., Vedaldi, A., 2019. Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE International Conference on Computer Vision.

Ji, Y., Yang, Y., Shen, H.T., Harada, T., 2021. View-invariant action recognition via unsupervised attention transfer (UANT). Pattern Recognit. 113, 107807.

Johansson, G., 1973. Visual perception of biological motion and a model for its analysis. Percept. Psychophys. 14 (2), 201–211.

Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F., 2017. A new representation of skeleton sequences for 3D action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. arXiv preprint arXiv:2004.11362.

Kim, B., Chang, H.J., Kim, J., Choi, J.Y., 2022. Global-local motion transformer for unsupervised skeleton-based action learning. In: Proceedings of the European Conference on Computer Vision. Springer.

Kreiss, S., Bertoni, L., Alahi, A., 2021. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. IEEE Trans. Intell. Transp. Syst. 23 (8), 13498–13511.

Li, Q., Han, Z., Wu, X.-M., 2018a. Deeper insights into graph convolutional networks for semi-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, (1).

Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W., 2021. 3D human action representation learning via cross-view consistency pursuit. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4741–4750.

Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S., 2018b. Unsupervised learning of view-invariant action representations. In: Proceedings of the International Conference on Neural Information Processing Systems.

Li, D., Zhang, Y., Wang, J., Tan, K.-L., 2019. TopoX: Topology refactorization for efficient graph partitioning and processing. Proc. VLDB Endow. 12 (8), 891–905.

Lin, L., Song, S., Yang, W., Liu, J., 2020. MS2L: Multi-task self-supervised learning for skeleton based action recognition. In: Proceedings of the ACM International Conference on Multimedia.

Liu, C., Hu, Y., Li, Y., Song, S., Liu, J., 2017a. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. arXiv preprint arXiv:1703.07475.

Liu, M., Liu, H., Chen, C., 2017b. Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognit. 68, 346–362.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., Kot, A.C., 2019. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. 42 (10), 2684–2701.

Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W., 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Nie, Q., Liu, Y., 2021. View transfer on human skeleton pose: Automatically disentangle the view-variant and view-invariant information for pose representation learning. Int. J. Comput. Vis. 129 (1), 1–22.

Nie, Q., Wang, J., Wang, X., Liu, Y., 2019. View-invariant human action recognition based on a 3D bio-constrained skeleton model. IEEE Trans. Image Process. 28 (8), 3959–3972.

Rao, H., Xu, S., Hu, X., Cheng, J., Hu, B., 2021. Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition. Inform. Sci. 569, 90–109.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Shahroudy, A., Liu, J., Ng, T.-T., Wang, G., 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Shi, L., Zhang, Y., Cheng, J., Lu, H., 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Song, S., Lan, C., Xing, J., Zeng, W., Liu, J., 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Su, K., Liu, X., Shlizerman, E., 2020. Predict & cluster: Unsupervised skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Sun, F.-Y., Hoffman, J., Verma, V., Tang, J., 2019. InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In: Proceedings of the International Conference on Learning Representations.

Thoker, F.M., Doughty, H., Snoek, C.G.M., 2021. Skeleton-contrastive 3D action representation learning. In: Proceedings of the ACM International Conference on Multimedia.

Veličković, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., Hjelm, R.D., 2018. Deep graph infomax. In: Proceedings of the International Conference on Learning Representations.

Vemulapalli, R., Arrate, F., Chellappa, R., 2014. Human action recognition by representing 3D skeletons as points in a lie group. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Wang, M., Lin, Y., Lin, G., Yang, K., Wu, X.-m., 2020. M2GRL: A multi-task multi-view graph representation learning framework for web-scale recommender systems. arXiv preprint arXiv:2005.10110.

Wang, H., Wang, L., 2017. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Wikipedia, 2021. Mutual information. https://en.wikipedia.ahut.cf/wiki/Mutual_information.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Xia, L., Chen, C.-C., Aggarwal, J.K., 2012. View invariant human action recognition using histograms of 3D joints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Yan, S., Xiong, Y., Lin, D., 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Yang, S., Liu, J., Lu, S., Er, M.H., Kot, A.C., 2021. Skeleton cloud colorization for unsupervised 3D action representation learning. In: Proceedings of the IEEE International Conference on Computer Vision.

Yang, X., Tian, Y.L., 2012. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Ye, M., Zhang, X., Yuen, P.C., Chang, S.-F., 2019. Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y., 2020. Graph contrastive learning with augmentations. In: Proceedings of the Advances in Neural Information Processing Systems.

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N., 2019. View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 41 (8), 1963–1978.

Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N., 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z., 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence.