Multi-Granularity Anchor-Contrastive Representation Learning for Semi-Supervised Skeleton-Based Action Recognition

Xiangbo Shu, *Senior Member, IEEE*, Binqian Xu, Liyan Zhang, and Jinhui Tang, *Senior Member, IEEE*

Abstract—In the semi-supervised skeleton-based action recognition task, obtaining more discriminative information from both labeled and unlabeled data is a challenging problem. As the current mainstream approach, contrastive learning can learn more representations of augmented data, which can be considered as the pretext task of action recognition. However, such a method still confronts three main limitations: 1) It usually learns global-granularity features that cannot well reflect the local motion information. 2) The positive/negative pairs are usually pre-defined, some of which are ambiguous. 3) It generally measures the distance between positive/negative pairs only within the same granularity, which neglects the contrasting between the cross-granularity positive and negative pairs. Toward these limitations, we propose a novel Multi-granularity Anchor-Contrastive representation Learning (dubbed as MAC-Learning) to learn multi-granularity representations by conducting inter- and intra-granularity contrastive pretext tasks on the learnable and structural-link skeletons among three types of granularities covering local, context, and global views. To avoid the disturbance of ambiguous pairs from noisy and outlier samples, we design a more reliable Multi-granularity Anchor-Contrastive Loss (dubbed as MAC-Loss) that measures the agreement/disagreement between high-confidence soft-positive/negative pairs based on the anchor graph instead of the hard-positive/negative pairs in the conventional contrastive loss. Extensive experiments on both NTU RGB+D and Northwestern-UCLA datasets show that the proposed MAC-Learning outperforms existing competitive methods in semi-supervised skeleton-based action recognition tasks.

Index Terms—Action recognition, skeleton, semi-supervised, contrastive learning, anchor graph

1 Introduction

Human action recognition is an essential problem in computer vision and pattern recognition fields, which is rapidly developing due to its wide applications in video retrieval, video surveillance, virtual reality, human-computer interaction, etc. [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. According to diverse types of input data, human action recognition tasks can usually be divided into RGB-based [2], [4], depth-based [11], [12], and skeleton-based action recognition tasks [13], [14]. Compared with RGB videos or depth data, skeleton sequences that consist of locations of key points are more robust to the human body scales, dynamic circumstances, camera viewpoints, and interferential background [15], [16], [17]. In addition, skeleton sequences can be regarded as a type of lightweight, compact, and high-level representation for human behaviors [18], [19]. Meanwhile, skeleton sequences can be easily obtained by depth sensors or pose estimation algorithms [20], [21]. Thus, skeleton-based action recognition has also attracted increasing attention in this community [13], [18], [22], [23], [24], [25].

To learn the discriminative representations of skeletons, some deep learning-based methods achieve remarkable performance by designing various Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) [22], [24], [26], [27]. Furthermore, considering the structural information of skeletons and the interdependence between joints, some researchers employ Graph Convolutional Networks (GCN) to learn the structural features of skeletons on a spatiotemporal graph [14], [28], [29], [30], [31]. Overall, most skeleton-based action recognition models are trained in a fully-supervised manner, which relies on a large amount of labeled data. As known, annotating skeleton sequences is always time and labor consuming. Thus, how to learn discriminative representations from both unlabeled and labeled skeleton sequences, which is called semi-supervised skeleton-based action recognition, is arising as a topic of concern.

Recently, in the semi-supervised skeleton-based action recognition scenario, some works leverage contrastive learning to learn more representations of augmented data as the pretext tasks [32], [33], [34]. Generally, such contrastive learning pretext tasks involved in these works measure the distance among the global features. However, some skeleton motions

- *Xiangbo Shu, Binqian Xu, and Jinhui Tang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: {shuxb, jinhuitang}@njust.edu.cn, xubinq11@gmail.com.*
- *Liyan Zhang is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China. E-mail: zhangliyan@nuaa.edu.cn.*

Manuscript received 14 January 2022; revised 3 August 2022; accepted 7 November 2022. Date of publication 17 November 2022; date of current version 5 May 2023.
This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102001, in part by the National Natural Science Foundation of China under Grants 61925204, 62072245, 62172212, and 61932020, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20211520.
(Corresponding author: Liyan Zhang.)
Recommended for acceptance by O. Russakovsky.
Digital Object Identifier no. 10.1109/TPAMI.2022.3222871

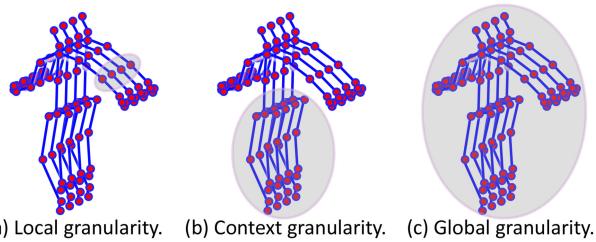


Fig. 1. The three types of granularities in human skeletons. (a) local granularity: covers one-joint skeleton sequence; (b) context granularity: covers partial-joint skeleton sequence; and (c) global granularity: covers all-joint skeleton sequence.

with different class labels are close from the global view, such as "reading" versus "writing", "drink water" versus "eat meal", which make the conventional contrastive learning confused. To better distinguish the details of similar actions, a natural way is to jointly consider the global and local feature distances among samples in contrastive learning process, as well as learning more abundant features for action recognition. More specifically, we define three granularities in skeleton sequences, namely local, context, and global granularity, as shown in Fig. 1. Among them, the local and context granularity refer to the situation in which the representations of skeletons can be learned from the one-joint and the partial-joint skeleton sequences, respectively, which supplement more discriminative information for recognizing human actions.

In summary, it is necessary to evolve contrastive learning to learn multi-granularity representations instead of single-granularity representations. In this work, we propose a novel Multi-granularity Anchor-Contrastive representation Learning (MAC-Learning) method that aims to learn the latent semantic links of human joints, and then obtain multi-granularity action representations. Specifically, MAC-Learning conducts inter- and intra-granularity contrastive pretext tasks on the learnable and structural-link skeletons among three types of granularities covering local, context, and global views. Here, the inter-granularity contrastive pretext task includes the local-context, local-global, and context-global granularity contrasts. And the intra-granularity contrastive pretext task includes the local-local, context-context, and global-global granularity contrasts. Correspondingly, we design a new Multi-granularity Anchor-Contrastive Loss (MAC-Loss) containing the inter- and intra-granularity contrastive losses on the learnable and structural-link skeletons among three types of granularities to encourage the agreement/disagreement between the soft-positive/negative pairs rather than the hard-negative/positive pairs in conventional contrastive loss. Here, to avoid the disturbance of ambiguous pairs from noise and outlier samples, we leverage Anchor Graph with anchor and sample adjacency to capture the high-confidence soft-positive/negative pairs for the first time.

The overall framework of the proposed MAC-Learning is shown in Fig. 2. It mainly consists of Graph Convolutional Network [35] (GCN), Context Graph Convolutional Network [35] (Context GCN), Global Average Pooling (GAP), Anchor Graph, Multi-granularity Anchor-Contrastive Loss (MAC-Loss), and Recognition Loss. First, given the skeleton sequences, their local features are obtained by feeding them into GCN on learnable-link graph or structural-link graph. Meanwhile, their global features and context features are obtained by feeding them into GCN and Context GCN on learnable-link graph or structural-link graph followed by GAP, respectively. Second, the multi-granularity features (including local, context, and global features) of one skeleton sequence (sample) are integrated into one fused feature, which can be seen as one node on an anchor graph. Then, the distribution of all samples on this anchor graph can be represented by a certain number of anchors [36]. Third, the sample adjacent matrix and anchor adjacent matrix on such anchor graph are leveraged to define the soft-positive/negative pairs, as well as weight the distance between soft-positive/negative pairs in the following MAC-Loss. Finally, MAC-Loss and Recognition Loss jointly train the whole model of MAC-Learning, as well as learn the semantic links among human joints. Among them, MAC-Loss containing inter- and intra-granularity contrastive losses closely pulls the distance between soft-positive pairs, while pushing away the distance between soft-negative pairs.

Overall, the main contributions in this work can be summarized as follows,

- To address the problem of semi-supervised skeleton-based action recognition, we propose a novel Multi-granularity Anchor-Contrastive Representation Learning (MAC-Learning) framework that learns the latent semantic links of human joints and obtains more multi-granularity action representations on both labeled and unlabeled data.
- To avoid the disturbance of ambiguous pairs from noise and outlier samples, we leverage the Anchor Graph with anchor and sample adjacency to capture the high-confidence soft-positive/negative pairs in contrastive representation learning for the first time.
- To obtain more multi-granularity representations, we design a new Multi-granularity Anchor-Contrastive Loss (MAC-Loss) that contains the inter- and intra-granularity contrastive losses on the learnable and structural-link skeletons among three types of granularities to jointly measure the agreement/disagreement between soft-positive/negative pairs.
- We conduct extensive experiments on two public benchmarks to illustrate the effectiveness of the proposed MAC-Learning method compared with state-of-the-art methods in a semi-supervised skeleton-based scenario.

2 Related Work

2.1 Supervised Skeleton-Based Action Recognition

For the skeleton-based action recognition task, traditional methods always design various handcrafted features to represent the human skeleton actions [37], [38]. However, their performance is limited, since the handcrafted features sometimes cannot fully adapt to the downstream tasks. Subsequently, various deep learning-based methods have been proposed to address the problem of skeleton-based action recognition by employing CNN or RNN to learn action representations, mainly including CNN-based methods [15], [27], [39], and RNN-based methods [22], [40], [41]. For example, Du et al. [27] regarded the joint coordinates of skeleton sequences as an image, and further proposed learning more discriminative

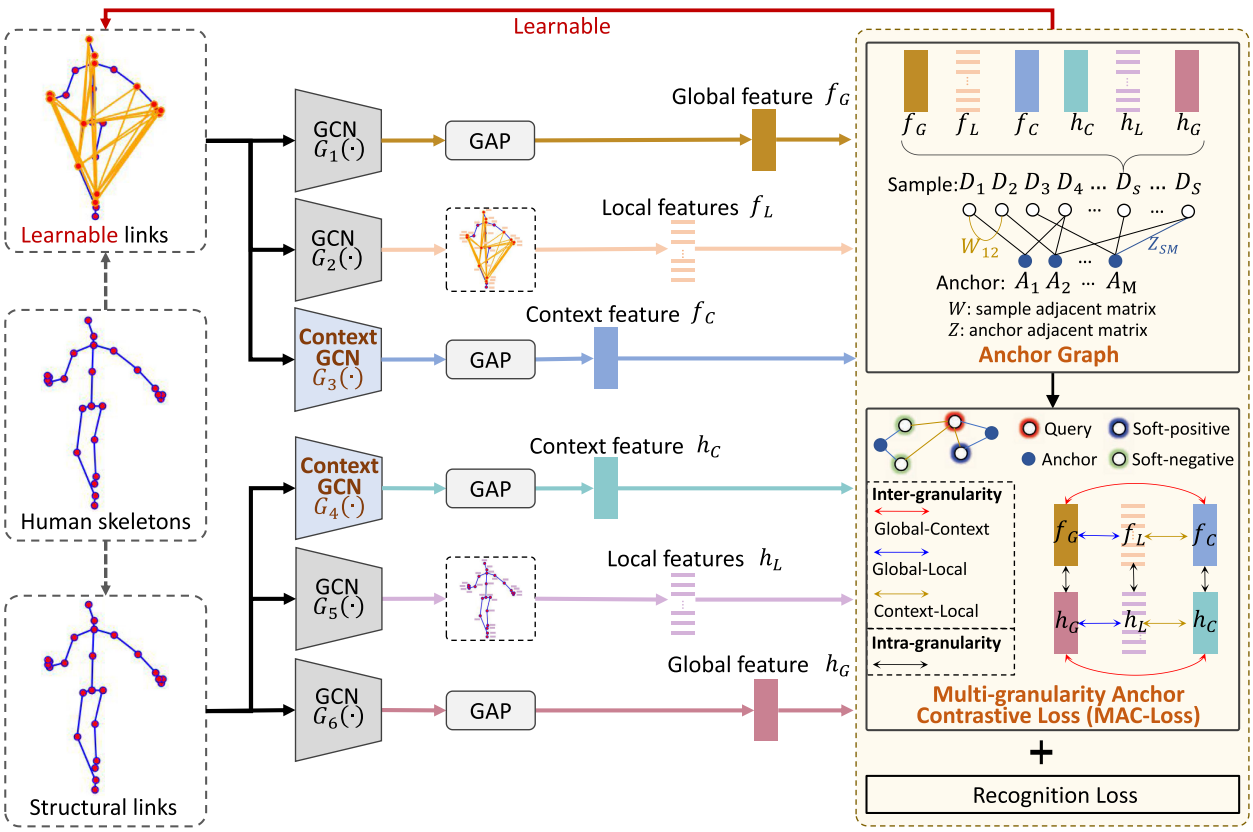


Fig. 2. The overall framework of MAC-Learning, mainly consisting of GCN, Context GCN, Global Average Pooling (GAP), Anchor Graph, Multi-granularity Anchor-Contrastive Loss (MAC-Loss), and Recognition Loss. MAC-Learning aims to learn the latent semantic links of human joints, and then obtain more multi-granularity action representations. The human skeleton data is constructed as two types of graphs with learnable links and structural links to obtain the global feature, local features, and context feature via GCNs, Context GCNs. Then, Anchor Graph is built to obtain the representative anchors with the corresponding anchor and sample adjacent matrices, which capture the high-confidence soft-positive/negative pairs in MAC-Loss. Finally, MAC-Loss containing inter- and intra-granularity losses measures the distance between soft-positive/negative pairs, and trains the whole model for action recognition allied with recognition loss.

joint information by CNN. Du et al. [22] utilized RNNs to model each body part and integrated the representation over time. However, CNN-based or RNN-based methods regard skeleton data as pseudo-images or frames over time, which cannot capture the structural information in skeleton motions.

Recently, a great number of works have modeled skeleton data via graph convolutional network (GCN) to learn the structural representations of skeletons [13], [14], [29], [30], [31], [42]. Generally, these works regarded human skeletons as a graph by setting the joints and bones as nodes and edges, respectively. For example, Yan et al. [13] proposed a spatial-temporal graph convolution to aggregate joint features. Based on this, Shi et al. [29] proposed an adaptive graph convolutional network to learn flexible topology graphs in a data-driven manner instead of fixed graphs. Furthermore, Chen et al. [42] modeled the graph topology of skeletons to aggregate joint features by learning a shared topology as a generic prior and refining it with channel-specific correlations. In general, all the above deep-learning methods adopt a supervised training way that requires a large amount of labeled data.

2.2 Semi-Supervised Skeleton-Based Action Recognition

Semi-supervised learning learns from both labeled and unlabeled data [43]. To date, there have been semi-supervised skeleton-based action recognition methods [44], [45], [46], [47], [48], [49], [50]. In the early stage, some dimensionality reduction or clustering algorithms were utilized to alleviate the problem of insufficient labeled data [44], [45].

Due to its powerful representation ability, deep learning has become the dominant model to learn the features of unlabeled data [46], [47], [48], [49], [50]. These deep learning-based methods are mainly inspired by the ladder network [51], [52], which is one of the representative approaches in semi-supervised learning to simultaneously train a deep auto encoder on unlabeled data, and train a neural network on labeled data. Specifically, similar to the idea of ladder networks, researchers usually leverage an encoder model to learn the representations of unlabeled data being consistent with those of original augmented data or labeled data. For example, Liu et al. [47] added random augmentation or noise to unlabeled data, and learned that the representation of the original data being consistent with noisy data via an LSTM encoder. Si et al. [48] presented an adversarial encoder framework that aligned the feature distribution of labeled and unlabeled samples by exploring the data relations within a neighborhood in a self-supervised manner. Li et al. [50] employed an encoder-decoder RNN to learn the latent representations of unlabeled skeleton sequences based on the reconstructed consistency, and then performed active learning to select skeleton sequences to be labeled based on the cluster and classification uncertainty.

In this work, we employ contrastive learning to simultaneously learn more action representations of augmented data being consistent with those of the original data, and being inconsistent with those of other irrelevant data. Compared with previous semi-supervised skeleton-based action recognition methods, there are two main insights in this work: the augmentation is learnable; and the feature-level consistency and inconsistency are jointly considered.

2.3 Unsupervised Skeleton-Based Action Recognition

Unsupervised learning aims to learn action representations only from unlabeled data, and self-supervised learning is a type of unsupervised learning. For the past few years, self-supervised and unsupervised methods for skeleton-based action recognition have emerged [32], [33], [34], [53], [54], [55], [56], [57], [58], [59], [60], [61]. In the beginning, researches adopted the encoder-decoder scheme as the core to develop various frameworks for learning features of skeleton motions from unlabeled data. For example, Zheng et al. [53] presented a conditional skeleton inpainting framework with an encoder-decoder scheme to capture the long-term global motion dynamics in skeleton sequences guided by additional adversarial training strategies. Su et al. [56] proposed an encoder-decoder recurrent neural network to cluster similar motions by self-organizing the hidden states of sequences into a feature space. Kundu et al. [54] proposed a new hierarchical fusion of five different body parts, where body parts were first fused into the upper and lower body representations, and subsequently into a full-body representation. Such fusion strategy is heuristic that requires to pre-define the fine-grained parts. Different from [54], the proposed method adopts the attention mechanism to automatically learn the fusion of some key joints as the context features, besides local and global features. And then the local, context, and global features of multiple granularities are jointly learned and further integrated in the summing fusion way. For such multi-granularity scheme, Li et al. [62] proposed to divide the images into fine, medium, and coarse granularities according to different resolutions, which is very beneficial and meaningful for extracting local details. Our work divides the skeletons into local, context, and global granularities based on the biological structure of human body. Here, the local granularity and global granularity denote the single-joint information and global-skeleton information, and the context granularity denotes the aggregation of some joints that participate the key motions. Thus, the idea of our work is also meaningful for learning the abundant action features.

Recently, some contrastive representation learning methods have shown remarkable performance for either unsupervised or self-supervised skeleton-based action recognition tasks [32], [33], [34], [55], [57], [58], [60]. Among them, some methods are devoted to designing various augmentation strategies [32], [33], [60]. For example, Rao et al. [32] designed multiple augmentation strategies to learn the action representations in a contrastive learning framework. Gao et al. [33] presented an augmentation way with the combination of sample viewpoint and distance to explore invariant motion semantics in contrastive learning. Su et al. [60] presented a speed-changed and motion-broken based augmentation strategy to capture the dynamic motion consistency in contrastive learning. Moreover, some methods are also devoted to bringing in multiple pretext tasks to learn the action representations in terms of robustness or generalization [34], [55], [57], [58]. For example, Xu et al. [57] combined the prototypical contrast and the reversed prediction pretext tasks to jointly learn action representations and predicting the future skeleton motions. By integrating multiple different pretext tasks, including motion prediction, contrastive learning, and puzzle recognition, Lin et al. [55] proposed a multi-task self-supervised learning to learn more generalized action features that were adaptive for different tasks. Li et al. [34] proposed a cross-view contrastive learning framework to learn multi-view features by integrating cross-view contrastive and consistent knowledge mining tasks.

Overall, above methods based on contrastive learning conduct contrastive pretext tasks in the global granularity, which cannot well capture the local joint movements. In this work, MAC-Learning introduces inter- and intra-granularity contrastive pretext tasks to learn more multi-granularity action representations covering local, context, and global three granularities for capturing more discriminative information in skeleton sequences, especially for some local joint movements.

2.4 Contrastive Learning

In recent years, contrastive learning has attracted considerable attention [63], [64], [65], [66], [67]. For contrastive learning, researchers aim to build specific models for various tasks mainly by designing new augmentation strategies or evolving the contrastive formulation. On the one hand, many augmentation strategies have been presented, where some representative strategies have been introduced in section 2.3. On the other hand, various contrastive formulations have been designed. For example, He et al. [63] proposed a momentum contrast model that constructed a dynamic dictionary to store more negative samples, and introduced a momentum-based moving encoder to maintain the consistency among mini-batches. To model the view invariance, Tian et al. [64] proposed a multi-view contrastive learning model to maximize the mutual information in different views of the same scene. Furthermore, Chen et al. [65] explored compositions of different augmentations, and introduced a learnable nonlinear layer between representations and contrastive loss to improve the quality of the learned representations. Moreover, some works also explored the memory consumption in contrastive learning [66], [68]. For example, Caron et al. [66] conducted contrastive loss at the cluster level instead of the sample level for relieving the computational challenge to some extent.

Overall, the core of contrastive loss [69] is to enforce the feature consistency between query and positive samples, and the discrimination between query and negative samples by calculating the distance between sample representations. The proposed Multi-granularity Anchor-Contrastive Loss (MAC-Loss) improved from previous contrastive learning methods leverages the sample adjacent matrix and anchor adjacent matrix produced by the anchor graph to reinforce the agreement between soft-positive pairs and the disagreement between soft-negative pairs.

3 METHODOLOGY

3.1 Overview of MAC-Learning

The main framework of the proposed MAC-Learning is shown in Fig. 2. For human skeleton set $\mathcal{V} = \{v_s\}_{s=1}^{S}$, one skeleton data is denoted as $v_s \in \mathbb{R}^{C\times T\times Q\times P}$, C is the number of channels, T is the total number of frames, Q is the number of joint points of each person, and P is the number of people in each frame.

First, the skeleton data v_s is input into Graph Convolution Network (GCN) [35] $G_1(\cdot)$ on the learnable-link graph, followed by Global Average Pooling (GAP), to obtain the global feature f_G. And the local features f_L are obtained by inputting v_s into the GCN $G_2(\cdot)$ on the learnable-link graph. At the same time, v_s is input into Context Graph Convolution Network (Context GCN) [35] $G_3(\cdot)$ on the learnable-link graph and GAP in turn to obtain the context feature f_C. Similarly, v_s is input into Context GCN $G_4(\cdot)$ with GAP, GCN $G_5(\cdot)$, and GCN $G_6(\cdot)$ with GAP on the structural-link graph to obtain the context feature h_C, local features h_L, and global feature h_G, respectively. The multiple features $\{f_G, f_L, f_C, h_C, h_L, h_G\}$ corresponding to v_s are fused into a feature D_s by the summing operation. Thus, all fused features corresponding to $\{v_s\}_{s=1}^{S}$ can be denoted by $\{D_s\}_{s=1}^{S}$.

Second, we build an anchor graph by setting $\{D_s\}_{s=1}^{S}$ as nodes, and implementing the clustering algorithm to obtain M anchors. Then, the sample adjacent matrix W and anchor adjacent matrix Z based on this anchor graph are calculated [70], [71]. Here, W represents the relationship between samples. Z represents the relationship between the samples and anchors.

Third, we define the soft-positive/negative pairs instead of traditional positive/negative pairs based on the anchor adjacent matrix Z, and utilize the sample adjacent matrix W to weight the distance between soft-positive/negative pairs. As a result, Multi-granularity Anchor-Contrastive Loss (MAC-Loss) pulls the distance between soft-positive pairs closely while pushing away the distance between soft-negative pairs via the inter- and intra-granularity contrastive losses on the learnable and structural-link skeletons among the local, context, and global views. Finally, MAC-Loss and Recognition Loss jointly train the whole model of MAC-Learning, where the former aims to learn more effective action representations while the latter aims to learn the classifier.

3.2 Graph Convolution Network and Context Graph Convolution Network

We introduce Graph Convolution Network (GCN) and Context Graph Convolution Network (Context GCN) to extract the spatial-temporal features of skeleton sequences [35]. Specifically, GCN is stacked by multiple GCN blocks. Fig. 3 shows the architecture of each GCN block, which mainly consists of Spatial GCN (SGCN), Temporal GCN (TGCN), BatchNorm, and ReLU. Here, SGCN is defined as follows:

$$f_{out} = \sum_{k}^{K_v} W_k (f_{in} A_k), \tag{1}$$

where f_{in} and f_{out} are the input and output feature maps respectively, W_k is the parameter of network, K_v denotes the kernel size of the spatial dimension, $A_k = \Lambda_k^{-\frac{1}{2}} \bar{A}_k \Lambda_k^{-\frac{1}{2}}$, $\bar{A}_k$ is an adjacency matrix of human skeleton graph, and $\Lambda_k^{ii} = \sum_j (\bar{A}_k^{ij})$. TGCN is an ordinary $L \times 1$ convolutional layer to aggregate the contextual representations embedded in adjacent frames, where L denotes the length of temporal windows. Similar to GCN, Context GCN is stacked by multiple Context GCN blocks with attentions. As shown in Fig. 3, the difference between Context GCN and GCN blocks is that each Context GCN block additionally contains an attention module to capture the key joints as the context joints.

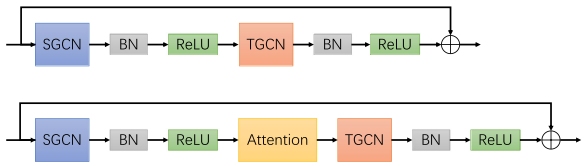


Fig. 3. Architecture of GCN block (above) and Context GCN block (below) in details. BN is short for BatchNorm, and ReLU is a nonlinear activation function.

Either for GCN or Context GCN, the skeleton graph is constructed by setting the joints as nodes, and the links between joints as edges. Similar to the construction of joint links in [35], there are two types of links, i.e., the learnable links[1] and structural links. The former can be learned with different structures for different action classes, while the latter is inherently based on the human body. In GCN $G_1(\cdot)$, $G_2(\cdot)$, and Context GCN $G_3(\cdot)$, the multi-granularity skeleton features are learned by graph convolution on the learnable-link graph for skeleton data. In Context GCN $G_4(\cdot)$, and GCN $G_5(\cdot)$, $G_6(\cdot)$, the multi-granularity skeleton features are learned by graph convolution on the structural-link graph for skeleton data. Finally, all features learned by $G_1(\cdot)$, $G_2(\cdot)$, $G_3(\cdot)$, $G_4(\cdot)$, $G_5(\cdot)$, and $G_6(\cdot)$ are complemented to obtain more powerful representations.

3.3 Anchor Graph

Anchor Graph [70], [71] has been successfully used in large-scale data mining and indexing. In this work, we leverage Anchor Graph to measure the relationship between skeleton data for boosting the performance of contrastive representation learning. As shown in Fig. 2, for one skeleton data v_s, we have the local features $f_L = G_2(v_s)$, $h_L = G_5(v_s)$, context features $f_C = \mathrm{GAP}(G_3(v_s))$, $h_C = GAP(G_4(v_s))$, and global features $f_G = GAP(G_1(v_s))$, $h_G = GAP(G_6(v_s))$. Then, all multi-granularity features $\{f_L, f_C, f_G, h_L, h_C, h_G\}$ are integrated to form one fused feature D_s by the summing operation. Furthermore, we set the fusion features $\{D_s\}_{s=1}^{S}$ corresponding to $\{v_s\}_{s=1}^{S}$ as nodes in an anchor graph. Next, $\{A_m\}_{m=1}^{M}$ anchors are obtained by the k-means clustering algorithm for representing the distribution of all samples, where $M < S$. Formally, the anchor adjacent matrix Z on anchor graph can be calculated as follows:

$$\begin{aligned} Z_{s,m} &= \frac{K_h(D_s, A_m)}{\sum_{m'\in\langle s\rangle} K_h(D_s, A_{m'})} \\ &= \frac{\exp(-\|D_s - A_m\|^2/2h^2)}{\sum_{m'\in\langle s\rangle} \exp(-\|D_s - A_{m'}\|^2/2h^2)}, \forall m \in \langle s\rangle, \end{aligned} \tag{2}$$

1. More details about learning links among joints can be found in [35].

where $Z_{s,m}$ denotes the distance between sample v_s and anchor A_m, $K_h(\cdot)$ adopts Gaussian kernel function, h is a hyperparameter, and $\langle s \rangle$ is the index set of the top-r closest anchors of sample v_s. Then, the sample adjacent matrix W denoting the relationship between samples can be calculated as follows:

$$W = Z\Lambda^{-1}Z^{\top}, \tag{3}$$

where the diagonal matrix $\Lambda \in \mathbb{R}^{M\times M}$ is defined as $\Lambda_{mm} = \sum_{s=1}^{S} Z_{sm}$. Thus far, the anchor adjacent matrix Z and the sample adjacent matrix W can be used to measure the relationship between skeleton data, which can help construct and confirm the soft-positive/negative pairs in the following contrastive learning process. To the best of our knowledge, this is the first work to leverage Anchor Graph to measure the relationship among skeleton data for boosting the performance of contrastive representation learning.

3.4 Multi-Granularity Anchor-Contrastive Loss (MAC-Loss)

Multi-granularity Anchor-Contrastive Loss (MAC-Loss) includes inter- and intra-granularity contrastive losses on the learnable and structural-link skeletons among three types of granularities, i.e., local, context, and global. Specifically, inter-granularity contrastive losses refer to contrasting the features of different granularity, and intra-granularity contrastive losses refer to contrasting the features of the same granularity, as shown in Fig. 2.

In Anchor Graph, the anchor graph with the anchor adjacent matrix Z and sample adjacent matrix W has been built. In MAC-Loss, for any two samples, we define them as the soft-positive pair if their closest anchors are the same, and the negative pair otherwise. Assuming that the batch size in the model training process is N, for skeleton data $\{v_n\}_{n=1}^{N}$, the corresponding multi-granularity features can be denoted as $\mathcal{F}_L = \{f_L^n\}_{n=1}^{N}$, $\mathcal{F}_C = \{f_C^n\}_{n=1}^{N}$, $\mathcal{F}_G = \{f_G^n\}_{n=1}^{N}$, $\mathcal{H}_L = \{h_L^n\}_{n=1}^{N}$, $\mathcal{H}_C = \{h_C^n\}_{n=1}^{N}$, and $\mathcal{H}_G = \{h_G^n\}_{n=1}^{N}$.

Formally, we denote $\mathcal{U}^1 = \mathcal{F}_G \cup \mathcal{F}_C$ as the global-context feature set within $2N$ features. The global-context contrastive loss $\mathcal{L}_{inter}^1$ between global features $\mathcal{F}_G$ and context features $\mathcal{F}_C$ is expressed as follows:

$$\mathcal{L}_{inter}^1 = -\sum_{i=1}^{2N} \log \frac{\sum_{j=1}^{2N} \mathbb{1}_{[j\neq i]} \cdot \mathbb{1}_{[Z_i'=Z_j']} \cdot W_{i,j} \cdot \exp(\langle g_i, g_j \rangle)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \cdot W_{i,k} \cdot \exp(\langle g_i, g_k \rangle)} \tag{4}$$

where $g_i, g_j \in \mathcal{U}^1$; Z' is changed from anchor adjacent matrix Z by setting all elements to zero except for the maximum value in each row, namely retaining only one closest anchor for each sample; $\langle u, v \rangle = (H_u u)^{\top} H_v v / (\tau \cdot \|H_u u\| \cdot \|H_v v\|)$, where τ is a hyperparameter, H_u and H_v are the projection matrices; and $\mathbb{1}_{[\cdot]}$ is an indicator function that is equal to 1 if the condition inside the square brackets is true, and 0 otherwise. Compared with the formulation of conventional contrastive loss, there are two main changes in Eq. (4): 1) The indicator function $\mathbb{1}_{[Z_i'=Z_j']}$ aims to find the soft-positive pairs. 2) The weight $W_{i,j}$ aims to measure the important degree of the distance between g_i and g_j in MAC-Loss.

Similar to the formulation of $\mathcal{L}_{inter}^1$ in Eq. (4), the other inter- and intra-granularity contrastive losses can also be formulated as $\mathcal{L}_{inter}^2$, $\mathcal{L}_{inter}^3$, $\mathcal{L}_{inter}^4$, $\mathcal{L}_{inter}^5$, $\mathcal{L}_{inter}^6$, $\mathcal{L}_{intra}^1$, $\mathcal{L}_{intra}^2$, and $\mathcal{L}_{intra}^3$, as shown in Table 1. It is noted that one local feature f_L in $\mathcal{F}_L$ can be seen as the combination of Q sub-features corresponding to all Q human joints. Thus, the contrastive loss of formula (4) is first calculated as the contrastive sub-loss between each local sub-feature in $\mathcal{F}_L$ and each global feature in $\mathcal{F}_G$, and then the Q sub-losses are averaged to obtain the loss $\mathcal{L}_{inter}^2$ (obtain $\mathcal{L}_{inter}^3$, $\mathcal{L}_{inter}^5$, $\mathcal{L}_{inter}^6$, $\mathcal{L}_{intra}^2$ in the same way). Finally, to integrate all inter- and intra-granularity contrastive losses, MAC-Loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{con} =& \mathcal{L}_{inter}^1 + \mathcal{L}_{inter}^2 + \mathcal{L}_{inter}^3 + \mathcal{L}_{inter}^4 + \mathcal{L}_{inter}^5 \\ &+ \mathcal{L}_{inter}^6 + \mathcal{L}_{intra}^1 + \mathcal{L}_{intra}^2 + \mathcal{L}_{intra}^3\end{aligned} \tag{5}$$

Meanwhile, the action recognition loss L_{reg} can be formulated as:

$$\mathcal{L}_{reg} = -y^{\top} \log(\hat{y}), \tag{6}$$

where $\hat{y} = \mathrm{softmax}(D_s)$, and y is the ground-truth label of the action. In this work, we utilize MAC-Loss and Recognition Loss to jointly train the whole model of MAC-Learning, and define the object function $\Psi(\boldsymbol{\theta})$ of MAC-Learning as follows:

$$\Psi(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\mathrm{minimize}}(\mathcal{L}_{con} + \mathcal{L}_{reg}) \tag{7}$$

where $\boldsymbol{\theta}$ is the parameter set of MAC-Learning. Algorithm 1 summarizes the main implementations of MAC-Learning.

TABLE 1
Definition of Inter- and Intra-Granularity Contrast Losses in MAC-Loss

Notations	Descriptions
$\mathcal{L}_{inter}^1$	Global-context contrastive loss among $\mathcal{U}^1 = \mathcal{F}_G \cup \mathcal{F}_C$
$\mathcal{L}_{inter}^2$	Global-local contrastive loss among $\mathcal{U}^2 = \mathcal{F}_G \cup \mathcal{F}_L$
$\mathcal{L}_{inter}^3$	Local-context contrastive loss among $\mathcal{U}^3 = \mathcal{F}_L \cup \mathcal{F}_C$
$\mathcal{L}_{inter}^4$	Global-context contrastive loss among $\mathcal{U}^4 = \mathcal{H}_G \cup \mathcal{H}_C$
$\mathcal{L}_{inter}^5$	Global-local contrastive loss among $\mathcal{U}^5 = \mathcal{H}_G \cup \mathcal{H}_L$
$\mathcal{L}_{inter}^6$	Local-context contrastive loss among $\mathcal{U}^6 = \mathcal{H}_L \cup \mathcal{H}_C$
$\mathcal{L}_{intra}^1$	Global-global contrastive loss among $\mathcal{O}^1 = \mathcal{F}_G \cup \mathcal{H}_G$
$\mathcal{L}_{intra}^2$	Local-local contrastive loss among $\mathcal{O}^2 = \mathcal{F}_L \cup \mathcal{H}_L$
$\mathcal{L}_{intra}^3$	Context-context contrastive loss among $\mathcal{O}^3 = \mathcal{F}_C \cup \mathcal{H}_C$

4 Experiments

4.1 Dataset

To evaluate the performance of the proposed method, we adopt two publicly accessible datasets as the benchmarks, including NTU RGB+D dataset [41], and Northwestern-UCLA dataset [72].

NTU RGB+D Dataset [41]. The NTU RGB+D dataset is a large-scale dataset including 56,578 skeleton action sequences from 60 different action classes, which are performed by 40 volunteers with 25 joints for each body, and collected by three Microsoft Kinect v2 cameras. We follow two standard evaluation protocols, namely Cross-Subject (CS) and Cross-View (CV) protocol. In the CS protocol, the training set includes 40,091 skeleton sequences from 20 volunteers, and

the testing set includes 16,487 skeleton sequences from the other 20 volunteers. In the CV protocol, the training set includes 37,646 skeleton sequences from Cameras 2 and 3, and the testing data contain 18,932 skeleton sequences from Camera 1. For the training setting in the semi-supervised scenario, we only use 5%, 10%, 20%, and 40% labeled data, and the corresponding remaining unlabeled data.

Algorithm 1. MAC-Learning

Input:
$\mathcal{V}$: skeleton sequences $\mathcal{V} = \{v_s\}_{s=1}^{S}$
K : total optimization steps
y : ground-truth label
h, τ : hyperparameter
for $k = 1$ **to** K **do**
// obtaining of multi-granularity features
$f_L,\ f_C,\ f_G = G_2(v_s),\ GAP(G_3(v_s)),\ GAP(G_1(v_s))$
$h_L,\ h_C,\ h_G = G_5(v_s),\ GAP(G_4(v_s)),\ GAP(G_6(v_s))$
// Anchor graph
$D_s = \text{fusion}(f_L, f_C, f_G, h_L, h_C, h_G)$
$\{A_m\}_{m=1}^{M} = \text{clustering}(\{D_s\}_{s=1}^{S})$

$$Z_{s,m} = \frac{K_h(D_s, A_m)}{\sum_{m' \in \langle s \rangle} K_h(D_s, A_{m'})}, \forall m \in \langle s \rangle$$
$$W = Z\Lambda^{-1}Z^{\top}$$

// Multi-granularity Anchor-Contrastive Loss (MAC-Loss)

$$\mathcal{L}_{inter}^{1} = -\sum_{i=1}^{2N} \log \frac{\sum_{j=1}^{2N} \mathbb{1}_{[j \neq i]} \cdot \mathbb{1}_{[Z_i' = Z_j']} \cdot W_{i,j} \cdot \exp(\langle g_i, g_j \rangle)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \cdot W_{i,k} \cdot \exp(\langle g_i, g_k \rangle)}$$
$$\mathcal{L}_{con} = \mathcal{L}_{inter}^{1} + \mathcal{L}_{inter}^{2} + \mathcal{L}_{inter}^{3} + \mathcal{L}_{inter}^{4} + \mathcal{L}_{inter}^{5} + \mathcal{L}_{inter}^{6} + \mathcal{L}_{intra}^{1} + \mathcal{L}_{intra}^{2} + \mathcal{L}_{intra}^{3}$$

// Object function
$\hat{y} = \text{softmax}(D_s)$
$\mathcal{L}_{reg} = -y^{\top} \log(\hat{y})$
$\Psi(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\text{minimize}}(\mathcal{L}_{con} + \mathcal{L}_{reg})$
Update parameter set $\boldsymbol{\theta}$ using SGD to optimize Ψ
end for

Northwestern-UCLA (NW-UCLA) Dataset [72]. The NW-UCLA dataset includes 1,494 samples from 10 different action classes, which were collected by three Kinect v1 cameras, and performed by 10 volunteers with 20 skeleton joints for each body. We follow the recommended evaluation protocol, where the training set contains 1,018 samples from the first two views, and the testing set includes 476 samples from the third view. For the training setting in the semi-supervised scenario, we only use 5%, 15%, 30%, and 40% labeled data, and the corresponding remaining unlabeled data.

4.2 Experimental Setting and Implementation

In the data preparation phase, all skeleton sequences are temporally resized to the fixed-length $T = 50$ frames by linear interpolation for both NTU RGB+D and NW-UCLA datasets, which is similar to [34]. In the semi-supervised training setting, we sample labeled data from NTU RGB+D and NW-UCLA datasets in a category-balanced strategy consistent with most methods, referring to [48]. Specifically, on NTU RGB+D, the training set contains approximately 33 (5%), 66 (10%), 132 (20%), and 264 (40%) labeled skeleton sequences per class in CS protocol, and contains approximately 31 (5%), 62 (10%), 124 (20%), and 248 (40%) labeled skeleton sequences per class in CV protocol. Likewise, on NW-UCLA, the training set contains approximately 5 (5%), 15 (15%), 30 (30%), and 40 (40%) skeletons per class. Here, the training set also contains the corresponding remaining unlabeled skeleton sequences. Finally, the testing data contain all labeled data on both NTU RGB+D and NW-UCLA datasets.

For the configuration of MAC-Learning, both GCN and Context GCN contain five blocks, and a fully connected layer is added at last. And the parameters of GCN $G_1(\cdot)$ and GCN $G_2(\cdot)$ are shared, as well as the parameters of GCN $G_5(\cdot)$ and GCN $G_6(\cdot)$ are shared. The summing fusion is used to fuse multiple features into a single feature. In Anchor Graph, h is set to the average of the maximum values of the distance between N sample nodes and top-r nearest anchors, M is set as the total number of classes (i.e., 60 and 10 in NTU RGB+D and NW-UCLA respectively), and r is set as 6 and 4 via diagnostic studies on NTU RGB+D and NW-UCLA respectively, where the diagnostic studies can be found in Section 4.6.2. Moreover, we adopt the k-means clustering algorithm [36] to calculate M cluster data as anchors because it is simple yet effective. In the process of k-means clustering, M samples are randomly selected from all training samples as the initial cluster centers. Then, the euclidean distance between each sample and each cluster center is calculated, and each sample is assigned to the closest cluster center to form a cluster. After all samples are assigned, the cluster centers are recalculated according to existing samples of clusters. These steps are repeated until no samples are reassigned to different clusters or the maximum number of iterations is reached. The final M cluster centers are the M anchors. To trade off computation and performance, we perform clustering and compute the anchor graph every 20 epochs. In the formulation of MAC-Loss, the value of hyperparameter τ is empirically set as 0.07. On NTU RGB+D, the batch size, momentum, initial learning rate, weight decay, and total epochs are set as 64, 0.9, 0.08, 10^{-4}, and 70, respectively. On NW-UCLA, the batch size, momentum, initial learning rate, weight decay, and total epochs are set as 16, 0.9, 0.1, 10^{-4}, and 100, respectively.

In the training process, Stochastic Gradient Descent (SGD) is employed to optimize the whole network, and the learning rate is reduced via cosine annealing. Moreover, a warmup strategy [78] is utilized in the first 10 and 20 epochs on NTU RGB+D and NW-UCLA, to make the training procedure more stable. On the NW-UCLA and NTU RGB+D datasets, the one-time computation time of anchor graph with clustering are approximately 38 seconds and 12 minutes, respectively, as well as the training time of one epoch are approximately 32 seconds and 14 minutes by using a Titan RTX GPU. All experiments are performed via the PyTorch deep learning framework on the Linux server equipped with a Titan RTX GPU. The source codes of MAC-Learning are publicly available at https://github.com/1xbq1/MAC-Learning.

Without loss of generality, Fig. 4 shows the changes of whole loss, recognition loss, and each contrastive loss during the whole training process on NTU RGB+D (CV) with 10% labeled data and NW-UCLA with 30% labeled data, respectively. Specifically, the left column represents the change of the whole loss along with the training, the middle column

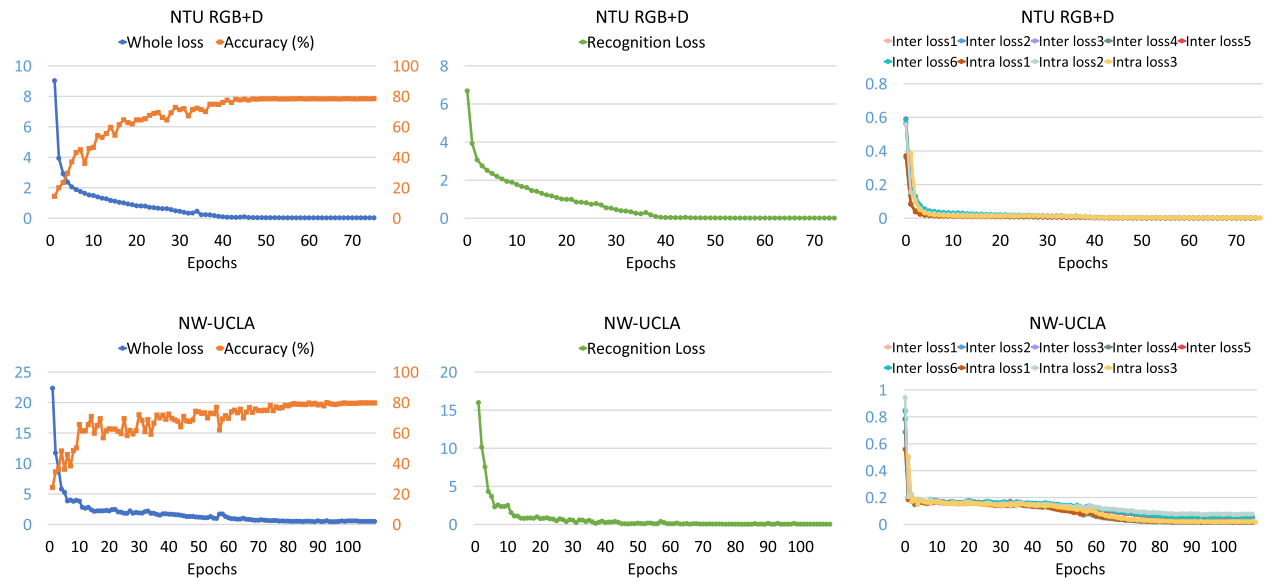


Fig. 4. The changes of the whole loss, recognition loss, and each contrastive loss on NTU RGB+D (CV) with 10% labeled data and NW-UCLA with 30% labeled data.

represents the change of the recognition loss along with the training, and the right column represents the changes of each contrastive loss along with the training. We can see that: 1) The change of the whole loss tends to be stable at approximately 70 and 100 epochs on NTU RGB+D and NW-UCLA, respectively; 2) Both recognition loss and MAC-Loss rapidly converge in early iterations; 3) MAC-Loss converges slightly faster than the recognition loss, and the convergence rates of all contrastive losses in MAC-Loss are similar.

4.3 Result and Analysis

4.3.1 Comparison on NTU RGB+D

We evaluate the performance of the proposed MAC-Learning on the NTU RGB+D dataset by comparing it with the currently representative methods, including semi-supervised methods (e.g., S^{4}L [73], Pseudolabels [74], VAT [75], VAT+EntMin [76], ASSL [48]), and unsupervised methods (e.g., AS-CAL [32], LongT GAN [53], Holden et al. [77], EnGAN-PoseRNN [54], MS^{2}L [55], Skeleton-Contrastive [58], and 3s-CrosSCLR [34]). The comparison of recognition accuracies obtained by different methods on the NTU RGB+D dataset is shown in Table 2. We can see that the proposed MAC-Learning achieves better accuracy than the alternatives. This illustrates that MAC-Learning is effective for addressing the problem of semi-supervised skeleton-based action recognition.

Specifically, compared with semi-supervised methods, MAC-Learning improves by 10.4% compared with the SOTA semi-supervised method (i.e, ASSL with an accuracy of 68.0%) on the CS protocol with 20% labeled data. Here, ASSL learns the single-granularity features by aligning the feature distribution of labeled and unlabeled data. Unlike ASSL, MAC-Learning learns the multi-granularity features of labeled and unlabeled data. Compared with unsupervised methods, MAC-Learning performs better than most of the other alternatives, and is comparable to 3s-CrosSCLR.

TABLE 2
The Comparison Among Recognition Accuracies (%) Obtained by Different Methods on NTU RGB+D (CS, and CV) With 5%, 10%, 20%, and 40% Labeled Data of Training Set

Method	5%		10%		20%		40%	
	CS	CV	CS	CV	CS	CV	CS	CV
‡S^{4}L [73]	48.4	55.1	58.1	63.6	63.1	71.1	68.2	76.9
‡Pseudolabels [74]	50.9	56.3	58.4	65.8	63.9	71.2	69.5	77.7
‡VAT [75]	51.3	57.9	60.3	66.3	65.6	72.6	70.4	78.6
‡VAT+EntMin [76]	51.7	58.3	61.4	67.5	65.9	73.3	70.8	78.9
‡ASSL [48]	57.3	63.6	64.3	69.8	68.0	74.7	72.3	80.0
†AS-CAL [32]	-	-	52.2	57.3	-	-	-	-
†LongT GAN [53]	-	-	62.0	-	-	-	-	-
†Holden et al. [77]	-	-	-	-	-	-	72.9	81.1
†EnGAN-PoseRNN [54]	-	-	-	-	-	-	78.7	86.5
†MS^{2}L [55]	-	-	65.2	-	-	-	-	-
†Skeleton-Contrastive [58]	59.6	65.7	65.9	72.5	70.8	78.2	-	-
†3s-CrosSCLR [34]	-	-	**74.4**	77.8	-	-	-	-
MAC-Learning (Ours)	**63.3**	**70.4**	74.2	**78.5**	**78.4**	**84.6**	**81.1**	**89.6**

*The superscripts ‡ and † indicate the semi-supervised and unsupervised methods, respectively. The best and second-best values are highlighted in **bold** and underlined, respectively.*

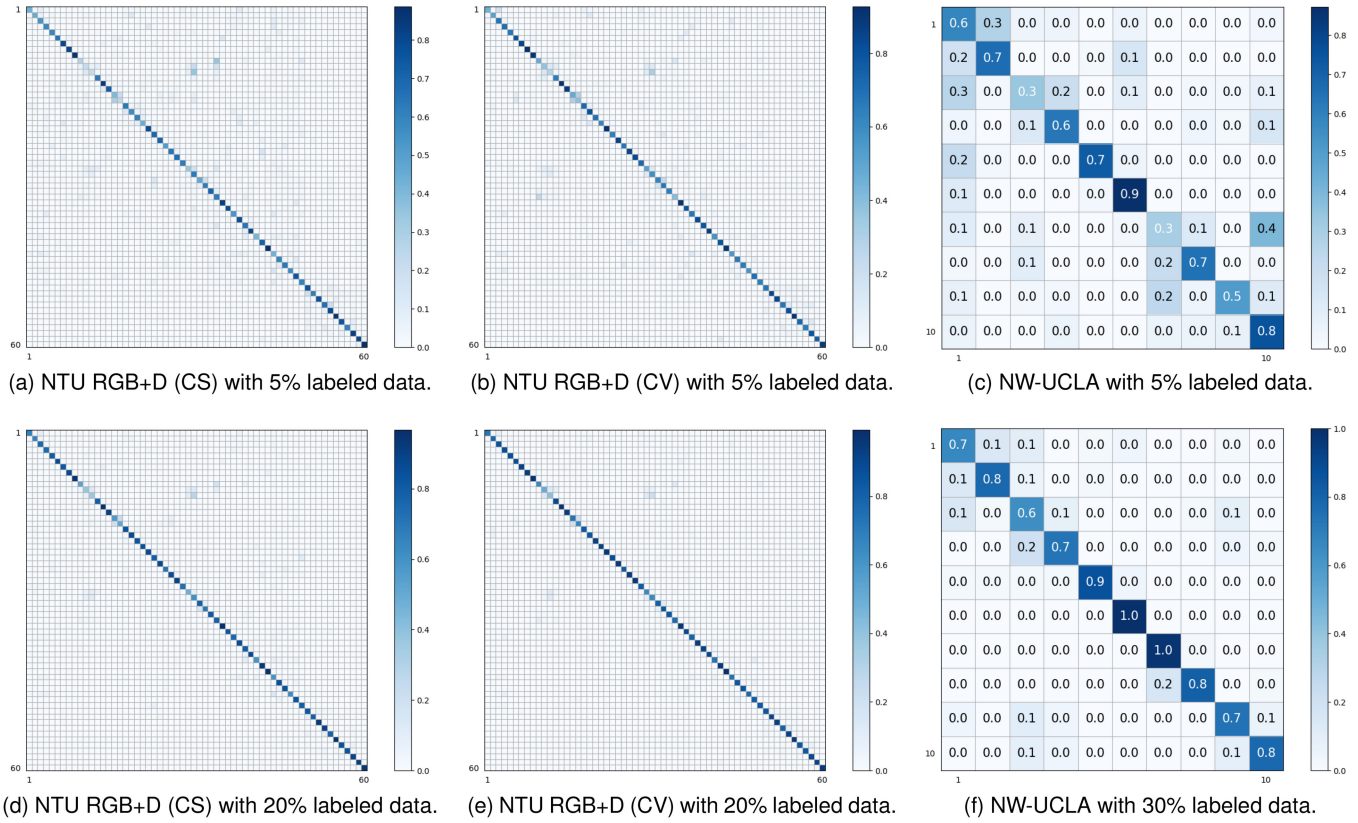


Fig. 5. Confusion matrices obtained by MAC-Learning on NTU RGB+D and NW-UCLA datasets. The accuracies (%) are reported on NTU RGB+D (CS) with 5% and 20% labeled data, NTU RGB+D (CV) with 5% and 20% labeled data, as well as NW-UCLA with 5% and 30% labeled data.

Here, 3s-CrosSCLR is also a contrastive learning method to pursue high-confidence positive/negative pairs by mining the multi-view knowledge of data. Similar to 3s-CrosSCLR, MAC-Learning pursues the high-confidence positive/negative pairs via and anchor graph. Thus, pursuing high-confidence positive/negative pairs is effective in contrastive learning. Compared with 3s-CrosSCLR, MAC-Learning provides another idea to pursue the high-confidence positive/negative pairs. Moreover, the proposed MAC-Learning improves by at least 7.6% on the CS protocol with 20% labeled data compared with some contrastive learning based unsupervised methods, i.e., AS-CAL [32], MS ^{2} L [55], and Skeleton-Contrastive [58], except for 3s-CrosSCLR. In addition, most methods perform better on Cross-View (CV) protocol than Cross-Subject (CS) protocol. For the same action on the CS protocol, the amplitude or some habitual local noise of actions performed by different people may be different. On the CV protocol, although the same action has different view difference, the input data are the location information of skeleton joints in 3D coordinates, and the relative distance between skeleton joints remains unchanged. Therefore, the performance on the CS protocol is not as good as the CV protocol.

The confusion matrices obtained by MAC-Learning on NTU RGB+D with 5% and 20% labeled data are shown in Fig. 5. From these confusion matrices (as shown in the first two-column figures), we can see that the main diagonal color of the 20% labeled data setting is darker than that of the 5% labeled data setting. This illustrates the overall improvement of all class accuracies when the number of labeled data increases.

Finally, Fig. 6 (the first two rows) shows some successful and failure recognition results obtained by the proposed method from the NTU RGB+D dataset. Since different actions have some similar local motions, it is easier to confuse the recognition of similar local actions. For example, the action "clapping" versus "rub two hands together", and the action "drink water" versus "brushing teeth" have similar local motions on the hands, so the failure recognition results happen.

4.3.2 *Comparison on NW-UCLA*

We also evaluate the performance of the proposed MAC-Learning on the NW-UCLA dataset by comparing it with the currently representative methods, including semi-supervised methods (e.g., S ^{4} L [73], Pseudolabels [74], VAT [75], VAT +EntMin [76], and ASSL [48]), and unsupervised method (e.g., MS ^{2} L [55]). The comparison among recognition accuracies obtained by different methods is shown in Table 3. The proposed MAC-Learning continually performs better than the alternatives, which further proves its effectiveness.

Specifically, compared with semi-supervised methods, MAC-Learning outperforms the SOTA semi-supervised method on the setting of 5% labeled data by a large margin, namely improves by 10.4% from 52.6% to 63.0%. Moreover, the accuracy obtained by MAC-Learning significantly outperforms the unsupervised method (i.e, MS ^{2} L with an accuracy of 60.5%), namely 18.3% is higher than the latter.

The confusion matrices obtained by MAC-Learning in the NW-UCLA dataset with 5% and 30% labeled data are shown in Fig. 5 (as shown in the last column of the figures).

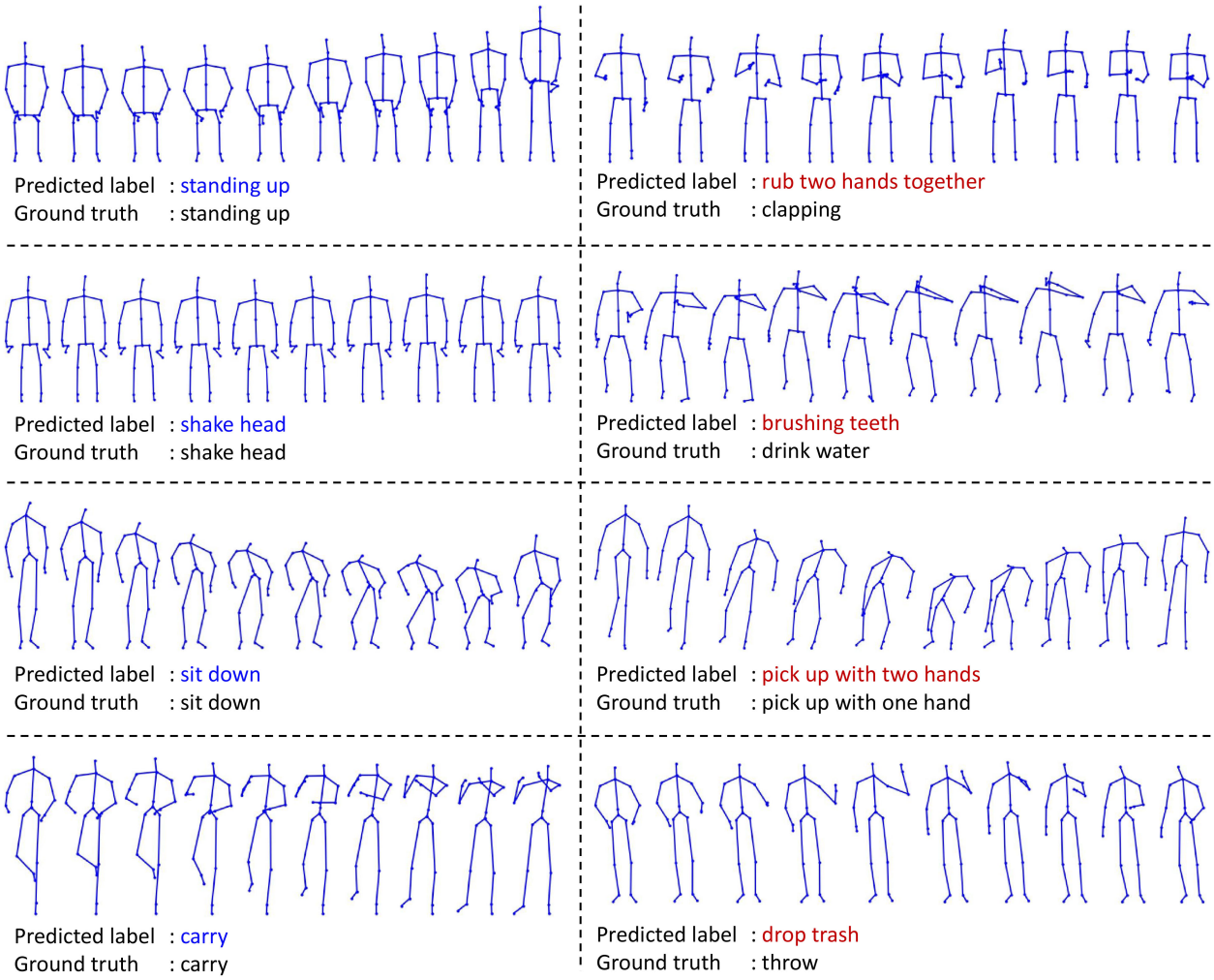


Fig. 6. Some recognition results obtained by the proposed method on the NTU RGB+D and NW-UCLA datasets. The first two rows are from NTU RGB+D, and the last two rows are from NW-UCLA.

We can see that the accuracies for most classes on the setting of the 30% labeled data are improved, compared with those on the setting of the 5% labeled data. In particular, for the third class "drop trash" and the seventh class "donning", the corresponding accuracies are obviously improved when the number of labeled data is changed from 5% to 30%.

Finally, Fig. 6 (the last two rows) also shows some correct and false recognition results obtained by the proposed method from the NW-UCLA dataset. It is noted that all samples of the NW-UCLA dataset come from different perspectives. For example, the action "pick up with one hand" and "pick up with two hand" are easily confused in the hand part, as well as the action "throw" and "drop trash" are easily confused due to their similar local motions.

TABLE 3
The Comparison Among Recognition Accuracies (%) Obtained by Different Methods on the NW-UCLA Dataset With 5%, 15%, 30%, and 40% Labeled Data of Training Set

Method	5%	15%	30%	40%
‡S ^{4} L [73]	35.3	46.6	54.5	60.6
‡Pseudolabels [74]	35.6	48.9	60.6	65.7
‡VAT [75]	44.8	63.8	73.7	73.9
‡VAT+EntMin [76]	46.8	66.2	75.4	75.6
‡ASSL [48]	52.6	74.8	78.0	78.4
†MS ^{2} L [55]	-	60.5	-	-
MAC-Learning (Ours)	**63.0**	**78.8**	**79.9**	**81.6**

*The superscripts ‡ and † indicate the semi-supervised and unsupervised methods, respectively. The best and second-best values are highlighted in **bold** and underlined, respectively.*

4.4 Qualitative Analysis

4.4.1 Visualization of Learned-Link Skeletons

For the learnable links among joints, MAC-Learning with MAC-Loss and Recognition Loss can learn different connections between joints for different actions to enhance the discriminative information for different action classes, which can be seen as the augmentation links of the original structural links among joints. It is noted that the ideal learned links are not the random augmented links, but the latent semantic links that reflect more discriminative information of each action class. Thus, we investigate the effectiveness of learnable links by visualizing some learned links on the NTU RGB+D and NW-UCLA datasets, as shown in Fig. 7. It can be seen that some body parts mainly performing the actions have more links. For example, for "brush hair", "cross hands in front", and "throw" actions, these actions are mainly performed by hands, so the learned links are mostly concentrated on the hand joints that connect to the

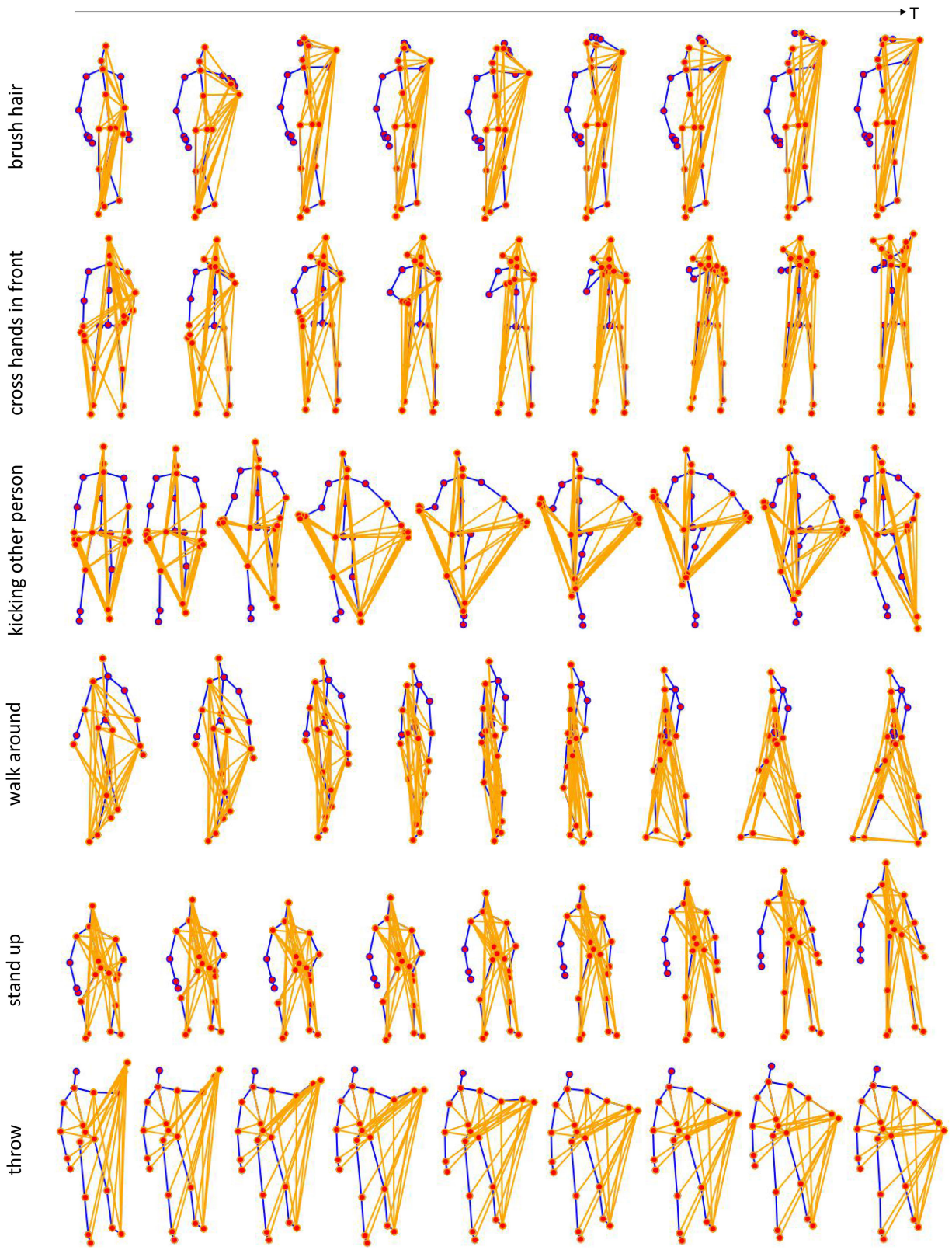
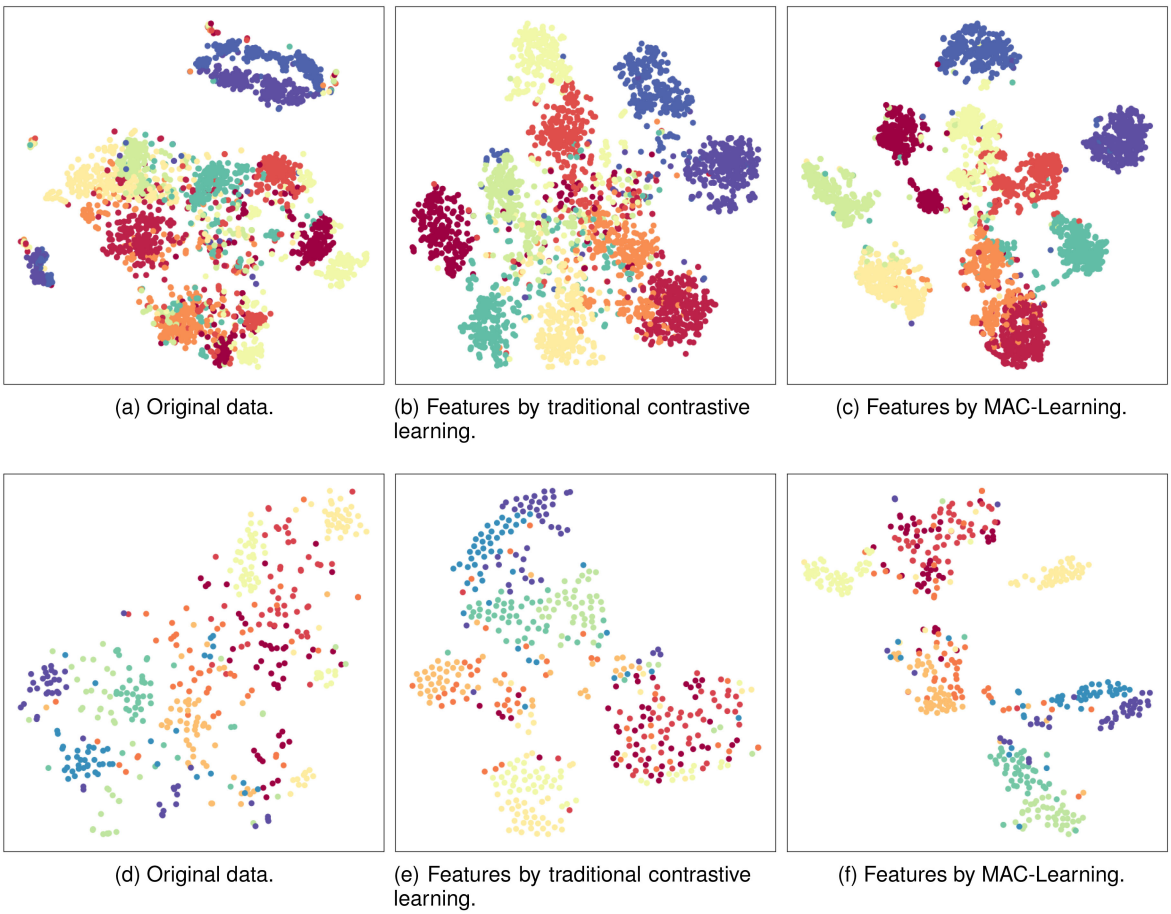


Fig. 7. Visualization of the learned links between joints by the proposed MAC-Learning on the NTU RGB+D and NW-UCLA datasets. The first three rows are from NTU RGB+D, and the last three rows are from NW-UCLA. The blue lines denote the structural links, and the orange lines denote learned links with top 45 and top 35 connections on NTU RGB+D and NW-UCLA dataset, respectively.

other joints. For "kicking other person" and "walk around" actions, they are more related to the movements of the feet, so the learned links are mainly concentrated on the foot joints that connect to the other joints. For "stand up" action, the learned links are mostly concentrated on the head, limbs and abdominal joints, which is also consistent with the fact. Overall, we find that when the human joints move more, the joint links are more concentrated. This proves that



(a) Original data. (b) Features by traditional contrastive learning. (c) Features by MAC-Learning.

(d) Original data. (e) Features by traditional contrastive learning. (f) Features by MAC-Learning.

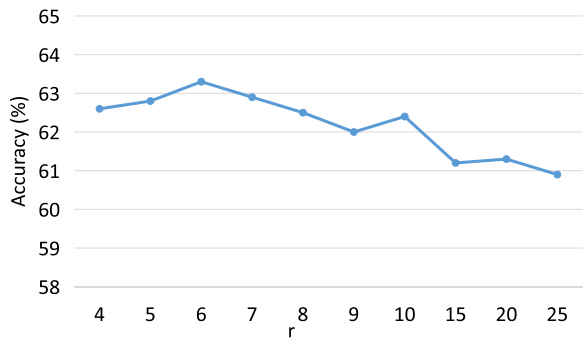
Fig. 8. The t-SNE visualization of action features learned by different methods on NTU RGB+D (1st row) and NW-UCLA (2nd row) with 5% labeled data. Following in [34], ten action classes on NTU RGB+D dataset are randomly selected and reported. Best view in color.

learning links among joints can effectively reflect more discriminative information of each action class, which is beneficial to learning the action representations.

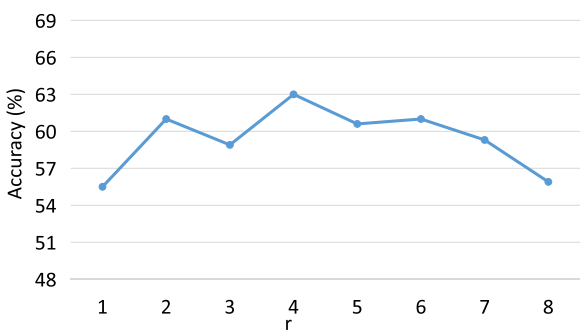
4.4.2 Visualization of learned features

Comparison of Features Learned by Different Methods. To illustrate the representation ability of the proposed MAC-Learning, we employ t-SNE to qualitatively visualize the distribution of the original data, the action features learned by traditional contrastive learning, and the action features learned by MAC-Learning. Here, for fair comparison, the traditional contrastive learning refers to the method only by adopting global-global contrastive loss in Table 1 with hard-positive/negative pairs in MAC-Learning. Fig. 8 visualize the distribution of the data and features on the NTU RGB+D and NW-UCLA datasets. We can see that the action features learned by contrastive learning become distinguishable. In particular, the action features learned by MAC-Learning are more distinguishable compared with those learned by the traditional contrastive learning method, especially on the NW-UCLA dataset. This illustrates more powerful ability of MAC-Learning in terms of representation learning.

Comparison of Features With Different-Granularities. To clearly compare the discriminative ability of different multi-granularity features, we also employ t-SNE to visualize the



(a) r on NTU RGB+D.



(b) r on NW-UCLA.

Fig. 9. Recognition accuracy (%) obtained by MAC-Learning with varying r on NTU RGB+D and NW-UCLA with 5% labeled data.

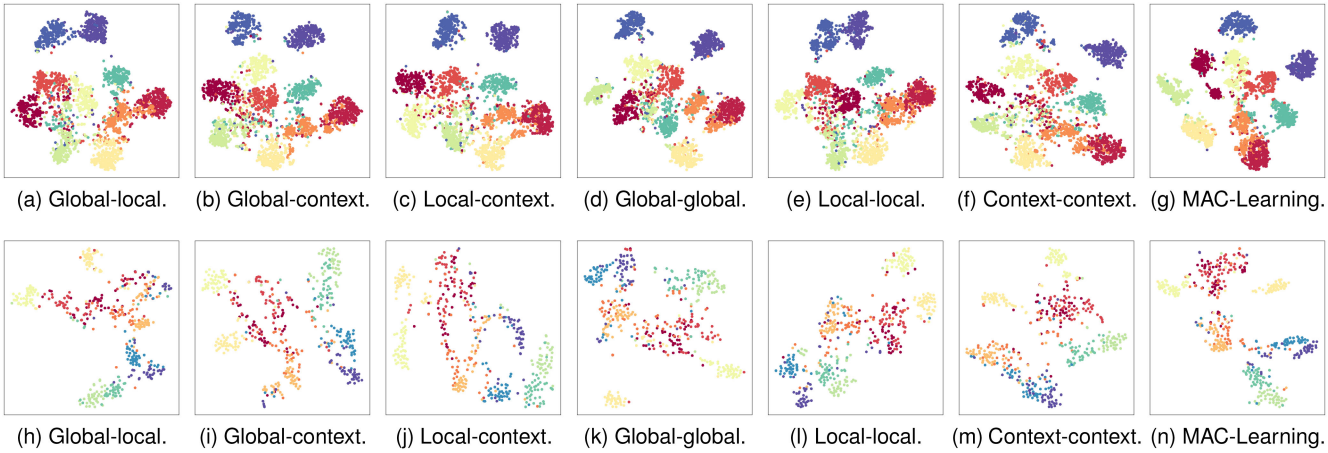


Fig. 10. The t-SNE visualization of the action features with different granularities learned by MAC-Learning on NTU RGB+D (1st row) and NW-UCLA (2nd row) with 5% labeled data. Best view in color.

distribution of the action features with different granularities learned by MAC-Learning. We first learn two-granularity features (i.e., global-local granularity features, global-context granularity features, local-context granularity features), single-granularity features (i.e., global granularity features, local granularity features, and context granularity features), and three-granularity features (i.e., global-context-local granularity features) by MAC-Learning on NTU RGB+D and NW-UCLA with 5% labeled data. For example, the global-local granularity features are learned by MAC-Learning only with global-local contrastive loss. All of the above features can be learned in a similar way. The t-SNE visualization of these features is shown in Fig. 10. We can see that the distribution of the three-granularity features is more distinguishable than those of either two-granularity or single-granularity features. Here, the distribution of two-granularity features and single-granularity features are comparable since both are learned by MAC-Learning with only one type of contrastive loss. This well demonstrates the advantage of three-granularity features learned by MAC-Learning in terms of discriminative ability.

4.5 Ablation Studies

To illustrate the superior idea of the multi-granularity strategy and MAC-Loss in MAC-Learning, we conduct the ablation studies on the NTU RGB+D (CS) dataset with 5% labeled data. In this work, we first set seven baselines as follows,

- *B1* *w/ Single-Granularity w/o Contrastive Learning*. It only uses the global-granularity features of the labeled data, which are fed into recognition loss for predicting the action classes. It can be seen as the single-granularity supervised baseline.
- *B2* *w/ Multi-Granularity w/o Contrastive Learning*. It uses the multi-granularity features of the labeled data, which are fed into recognition loss for predicting the action classes. It can be seen as the multi-granularity supervised baseline, which aims to test the superiority of multi-granularity features compared with single-granularity features in B1.
- *B3* *w/ Single-granularity w/ Traditional contrastive learning*. It uses the global-granularity features of the labeled and unlabeled data, which are fed into recognition loss and traditional contrastive loss [65]. It can be seen as the single-granularity semi-supervised learning baseline.
- *B4* *w/ Multi-granularity w/ Traditional contrastive learning*. It uses the multi-granularity features of the labeled and unlabeled data, which are fed into recognition loss and traditional contrastive loss [65]. It can be seen as the multi-granularity semi-supervised learning baseline.
- *B5* *w/ Single-granularity w/ MAC-Loss*. It uses the global-granularity features of the labeled and unlabeled data, which are fed into recognition loss and MAC-Loss. It aims to test the superiority of MAC-Loss compared with B3.
- *B6* *w/ Multi-granularity w/ MAC-Loss w/o Anchor graph*. It uses the multi-granularity features of the labeled and unlabeled data, which are fed into recognition loss and MAC-Loss without Anchor Graph. It performs contrastive learning on hard positive/negative pairs obtained by clustering.
- *B7* *MAC-Learning (w/ Multi-granularity w/ MAC-Loss)*.

Table 4 shows the accuracies obtained by different baselines on NTU RGB+D (CS) with 5% labeled data. B2 using multi-granularity features improves the recognition accuracy compared with B1 using single-granularity features. This indicates that multi-granularity strategy is beneficial to learning richer features compared with the single-granularity strategy. B4(B3) with contrastive learning performs better than B2(B1) without contrastive learning, which illustrates that learning representations on unlabeled data via contrastive learning can provide more discriminative features for the training model. In addition, B5 (with an accuracy of 59.2%) improves by 2.6% over B3 (with an accuracy of 56.6%), and even then it is comparable to B4, though B4 uses richer multi-granularity features. This indicates that MAC-Loss is more effective than traditional contrastive loss by measuring the distance between the high-confidence soft-positive/negative pairs. B6 performs contrastive learning on hard positive/negative pairs obtained by clustering. Compared with B4 with traditional contrastive learning, B6 improves by 2.1% indicating the contribution of the MAC-Loss without the anchor graph. Finally, B7 (namely MAC-

TABLE 4
Accuracies (%) Obtained by Different Baselines on NTU RGB+D (CS) With 5% Labeled Data

Baseline	Accuracy (%)
B1 (w/ Single-granularity w/o Contrastive learning)	55.6
B2 (w/ Multi-granularity w/o Contrastive learning)	57.9
B3 (w/ Single-granularity w/ Traditional contrastive learning)	56.6
B4 (w/ Multi-granularity w/ Traditional contrastive learning)	58.4
B5 (w/ Single-granularity w/ MAC-Loss)	59.2
B6 (w/ Multi-granularity w/ MAC-Loss w/o Anchor graph)	60.5
B7 (Ours)	**63.3**

Learning) with accuracy of 63.3% outperforms all baselines, and significantly improves by 2.8%, 4.9%, and 7.7% over B6, B4, and B1, respectively. This illustrates that MAC-Learning with the multi-granularity strategy, MAC-Loss, and Anchor Graph is superior to the supervised baselines, the traditional contrastive learning baselines, and the MAC-Loss without anchor graph baseline.

4.6 Diagnostic Studies

4.6.1 Effect of Different-Granularity Contrastive Losses

To investigate the effect of different-granularity losses, we conduct the diagnostic studies to test the effect of the inter- and intra-granularity contrastive representation learning. Specifically, we evaluate the recognition performance of MAC-Learning with local-context, global-local, context-global, global-global, local-local, and context-context contrastive losses on NTU RGB+D with 5% and 20% labeled data, as shown in Table 5. Here, we set A1 (only uses global-global contrastive loss), A2 (only uses local-local contrastive loss), A3 (only uses context-context contrastive loss) baselines in intra-granularity contrastive representation learning, A4 (only uses global-local contrastive loss), A5 (only uses global-context contrastive loss), and A6 (uses local-context granularity representations) baselines in inter-granularity contrastive learning. MAC-Loss (uses both inter- and intra-granularity contrastive losses) significantly improves the recognition accuracy compared with all baselines. The recognition accuracy achieved by either inter-granularity contrastive representation learning or intra-granularity contrastive representation learning is insignificant. This proves that MAC-Loss with both inter- and intra-granularity contrastive losses is useful for learning the multi-granularity representations.

4.6.2 Effect of the Top-r Closest Anchors

In Anchor Graph, parameter r denotes the top-r closest anchors influences the calculation of the anchor adjacent matrix and sample adjacent matrix, which affects the recognition performance of MAC-Learning to some extent. Thus, we conduct the diagnostic studies to investigate how r affects the final recognition performance. Specifically, based on the number of anchors, we empirically set $r \in \{4, 5, 6, 7, 8, 9, 10, 15, 20, 25\}$ and $r \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ to implement MAC-Learning and tune this parameter on NTU RGB+D and NW-UCLA with 5% labeled data. Fig. 9 shows the recognition accuracy obtained by MAC-Learning with varying r on NTU RGB+D and NW-UCLA, respectively. We can see that: 1) The best performance is achieved when $r = 6$ and $r = 4$ on NTU RGB+D and NW-UCLA respectively. 2) The larger or smaller values affect the performance to some extent. Thus, we set $r = 6$ and $r = 4$ in default on NTU RGB+D and NW-UCLA, respectively.

4.6.3 Effect of Different Fusion Mechanisms

In this work, we adopt the summing operation to fuse multi-granularity features into a single feature. Specifically, local, context, and global features are jointly learned with the same dimension, and then fused into a single feature by the summing operation. To explore the effect of different fusion mechanisms, we also conduct experiments to compare the summing fusion and concatenating fusion. Here, in

TABLE 5
Accuracies (%) Obtained by MAC-Learning With Different-Granularity Contrastive Losses on NTU RGB+D with 5%, 20% Labeled Data of Training Set

Baseline	Inter-granularity contrastive loss			Intra-granularity contrastive loss			5%		20%	
	G-L	G-C	L-C	G-G	L-L	C-C	CS	CV	CS	CV
A1				✓			59.2	66.1	74.5	80.4
A2					✓		58.7	64.3	74.2	80.3
A3						✓	59.5	66.3	75.0	80.4
A4	✓						59.1	66.0	74.3	80.1
A5		✓					58.9	65.7	73.8	80.7
A6			✓				59.1	65.6	73.9	79.8
MAC-Loss (Ours)	✓	✓	✓	✓	✓	✓	**63.3**	**70.4**	**78.4**	**84.6**

G-L, G-C, L-C, G-G, L-L, and C-C indicate the global-local, global-context, local-context, global-global, local-local, and context-context contrastive losses, respectively.

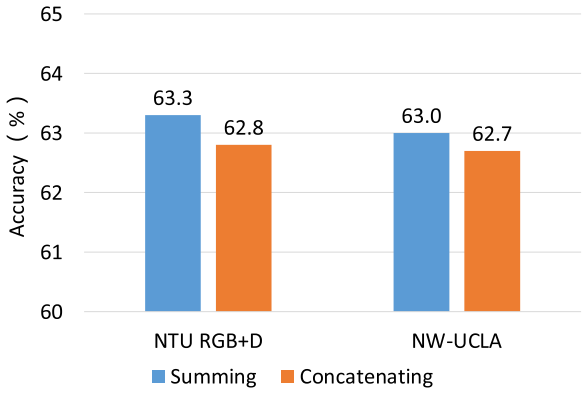


Fig. 11. The comparison of different fusion mechanisms (i.e., summing, and concatenating) on NTU RGB+D (CS) and NW-UCLA with 5% labeled data.

the concatenating fusion, the local, context, and global features are concatenated into a single feature.

The comparison results of summing and concatenating fusion on NTU RGB+D (CS) and NW-UCLA with 5% labeled data are shown in Fig. 11. Either on NTU RGB+D or NW-UCLA, the recognition performance of MAC-Learning via summing fusion and concatenating fusion is comparable, which illustrates that MAC-Learning is relatively robust with different fusion mechanisms.

4.6.4 Effect of the View-Invariant Augmentation

Several previous works [23], [64], [79] have validated that view-invariant structures can help improving the recognition performance. To verify whether adding a view-invariant structure into the proposed MAC-Learning can further improve the recognition performance, and also show the flexible extension of the proposed method, we refer to [33], [79] to apply the view-invariant augmentation to MAC-Learning. Specifically, we adopt a rotation transformation of the random angle along the x, y, and z axes to realize the view-invariant augmentation, which is implemented to rotate the input skeleton data before feeding into the GCNs and Context GCNs.

We compare the recognition accuracies of MAC-Learning with/without view-invariant augmentation on NTU RGB+D (CS) and NW-UCLA with 5% labeled data, as shown in Fig. 12. We can see that, while using view-invariant augmentation, the recognition accuracies of MAC-Learning are improved by 3.2% and 3.6% on NTU RGB+D and NW-UCLA, respectively. Therefore, it is flexible to equip the view-invariant augmentation into MAC-Learning for further improving the recognition performance.

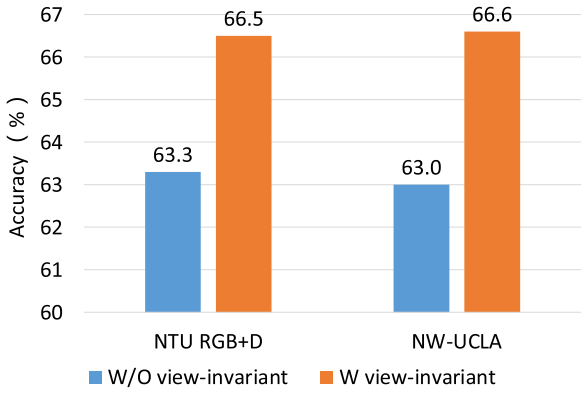


Fig. 12. Recognition accuracies (%) of MAC-Learning with/without view-invariant augmentation on NTU RGB+D (CS) and NW-UCLA with 5% labeled data.

TABLE 6
The Comparison of Different Methods in Terms of #Parameters (M) and FLOPs (G)

Methods	#Params (M)	FLOPs (G)
SGN [82]	0.7	15.4
Shift-GCN [83]	2.8	19.2
ST-GCN [13]	3.1	16.7
MS-G3D [31]	6.4	98.0
2s-AGCN [29]	6.9	37.4
AS-GCN [14]	7.2	35.5
2s-AAGCN [35]	7.6	39.1
DGNN [80]	8.1	71.1
DeCoup-GCN [81]	13.7	102.3
MAC-Learning (Ours)	8.1	40.7

The abbreviations "M" and "G" denote Mega and Giga, respectively.

4.6.5 Complexity Analysis: Parameters and FLOPs

To provide more details of the complexity analysis of the proposed MAC-Learning, we have calculated the total number of parameters (#Params) and the FLOPs, and provided the complexity comparison among different GCN-based methods in Table 6. Although there are six GCNs in our framework, the parameters of GCN $G_1(\cdot)$ and GCN $G_2(\cdot)$ are shared, as well as the parameters of GCN $G_5(\cdot)$ and GCN $G_6(\cdot)$ are shared in the implementation process. The main parameters of MAC-Learning exist in these six GCNs. Each (Context) GCN contains five blocks, and there are 20 blocks in MAC-Learning framework, which are equal to 20 blocks in 2s-AAGCN [35]. Thus, the number of parameters and FLOPs in MAC-Learning and 2s-AAGCN are comparable. In addition, compared with different methods (using GCN/GNN as the main backbone) in terms of the number of parameters and FLOPs, we can see that MAC-Learning costs less FLOPs than some popular models, e.g., DGNN [80] and DeCoup-GCN [81]. This illustrates that MAC-Learning is acceptable in terms of computational complexity.

4.7 Extensive Experiment

To further demonstrate the effectiveness of MAC-Learning on the larger-scale dataset, we conduct the comparative experiment on a more challenging Kinetics-skeleton dataset. Specifically, Kinetics dataset [84] includes 300,000 raw video clips of 400 classes without skeletal data, collected from YouTube videos. Following [13], we obtain the kinetics-skeleton data from the kinetics dataset by using the publicly available OpenPose toolbox [21] to estimate 18 human skeleton joints of each person. And then, we adopt the top-1 and top-5 evaluation criteria, where the training and testing sets contain 240,000, and 20,000 samples, respectively.

The comparison of recognition accuracies obtained by different methods on the Kinetics-skeleton dataset is shown in Table 7. It is noted that all comparative methods adopt

TABLE 7
The Comparison Among Recognition Accuracies (%) Obtained by Different Methods on the Kinetics-Skeleton Dataset in the Fully-Supervised Manner

Method	Top-1 (%)	Top-5 (%)
Deep LSTM [41]	16.4	35.3
TCN [15]	20.3	40.0
ST-GCN [13]	30.7	52.8
AS-GCN [14]	34.8	56.5
2s-AGCN [29]	36.1	58.7
DGNN [80]	36.9	59.6
GCN-NAS [85]	37.1	60.1
MAC-Learning (Ours)	37.9	60.6

the fully-supervised learning manner. For fair comparison, the proposed MAC-Learning also adopts the fully-supervised learning manner by using the 100% labeled data of training set. In Table 7, we can see that MAC-Learning achieves the competitive performance compared with the other methods. Specifically, the recognition performance achieved by MAC-Learning and GCN-NAS [85] are comparable.

5 Conclusion

In this work, we proposed a novel Multi-granularity Anchor-Contrastive Representation Learning (MAC-Learning) framework to address the problem of semi-supervised skeleton-based action recognition by learning multi-granularity action features. Specifically, MAC-Learning conducts inter- and intra-granularity contrastive pretext tasks on the learnable and structural-link skeletons among local, context, and global granularities. Overall, there are two main insights in the proposed MAC-Learning. First, MAC-Learning creatively captures the high-confidence soft-positive/negative pairs in contrastive learning to avoid the disturbance of ambiguous pairs from noise and outlier samples. Second, MAC-Learning leverages Multi-granularity Anchor-Contrastive Loss (MAC-Loss) containing the inter- and intra-granularity contrastive losses to measure the agreement/disagreement between the soft-positive/negative pairs on the learnable and structural-link skeletons among three types of granularities. Extensive experimental results on NTU RGB+D and Northwestern-UCLA datasets show the promising performance of MAC-Learning in terms of semi-supervised skeleton-based action recognition task. In the future, due to the flexibility of MAC-Loss, it can be regarded as a plug-and-play module, and then be pushed into other semi-supervised/unsupervised representation learning frameworks.

References

[1] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.
[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.
[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.
[5] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.
[6] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.
[7] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with enhanced motion vector CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2718–2726.
[8] X. Shu, J. Tang, G. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 1110–1118, Mar. 2021.
[9] X. Shu, L. Zhang, Y. Sun, and J. Tang, "Host–parasite: Graph LSTM-in-LSTM for group activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 663–674, Feb. 2021.
[10] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph LSTM for group activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 636–647, Feb. 2022.
[11] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 716–723.
[12] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: Histogram of oriented principal components of 3d pointclouds for action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 742–757.
[13] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.
[14] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3595–3603.
[15] T. S. Kim and A. Reiter, "Interpretable 3 D human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1623–1631.
[16] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5457–5466.
[17] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3300–3315, Jun. 2022.
[18] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," 2018, *arXiv:1804.06055*.
[19] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, "Skeleton-aided articulated motion generation," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 199–207.
[20] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.
[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.
[22] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1110–1118.
[23] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2117–2126.
[24] 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.*, 2017, pp. 3288–3297.
[25] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5323–5332.
[26] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2017.
[27] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit.*, 2015, pp. 579–583.

[28] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
[29] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12 026–12 035.
[30] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1625–1633.
[31] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 143–152.
[32] 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, 2021.
[33] X. Gao, Y. Yang, and S. Du, "Contrastive self-supervised learning for skeleton action recognition," in *Proc. Int. Conf. Neural Inf. Process. Syst. Workshops*, 2021, pp. 51–61.
[34] 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 Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4741–4750.
[35] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020.
[36] J. MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Berkeley Symp. Math. Statist. Probability*, 1967, pp. 281–297.
[37] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2466–2472.
[38] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 588–595.
[39] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, 2017.
[40] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 816–833.
[41] 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.*, 2016, pp. 1010–1019.
[42] 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 Int. Conf. Comput. Vis.*, 2021, pp. 13 359–13 368.
[43] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, O. et al. eds.; 2006)[book reviews]," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, pp. 542–542, Mar. 2009.
[44] X. Zhao, X. Li, C. Pang, and S. Wang, "Human action recognition based on semi-supervised discriminant analysis with global constraint," *Neurocomputing*, vol. 105, pp. 45–50, 2013.
[45] H. Yuan, "A semi-supervised human action recognition algorithm based on skeleton feature," *J. Inf. Hiding Multimedia Signal Process.*, vol. 6, no. 1, pp. 175–182, 2015.
[46] G. Pikramenos et al., "An adversarial semi-supervised approach for action recognition from pose information," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17181–17195, 2020.
[47] H. Liu, C. Liu, and R. Ding, "Semi-supervised long short-term memory for human action recognition," *J. Eng.*, vol. 2020, no. 13, pp. 373–378, 2020.
[48] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, and J. Feng, "Adversarial self-supervised learning for semi-supervised 3D action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 35–51.
[49] J. Li and E. Shlizerman, "Iterate & cluster: Iterative semi-supervised action recognition," 2020, *arXiv:2006.06911*.
[50] J. Li and E. Shlizerman, "Sparse semi-supervised action recognition with active learning," 2020, *arXiv:2012.01740*.
[51] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 3546–3554.
[52] F. Cricri, X. Ni, M. Honkala, E. Aksu, and M. Gabbouj, "Video ladder networks," 2016, *arXiv:1612.01756*.
[53] 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.*, 2018, pp. 1–6.
[54] J. N. Kundu, M. Gor, P. K. Uppala, and V. B. Radhakrishnan, "Unsupervised feature learning of human actions as trajectories in pose embedding manifold," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2019, pp. 1459–1467.
[55] 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*, 2020, pp. 2490–2498.
[56] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9631–9640.
[57] S. Xu, H. Rao, X. Hu, and B. Hu, "Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition," 2020, *arXiv:2011.07236*.
[58] F. M. Thoker, H. Doughty, and C. G. Snoek, "Skeleton-contrastive 3D action representation learning," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 1655–1663.
[59] Y.-B. Cheng, X. Chen, D. Zhang, and L. Lin, "Motion-transformer: Self-supervised pre-training for skeleton-based action recognition," in *Proc. ACM Int. Conf. Multimedia Asia*, 2021, pp. 1–6.
[60] Y. Su, G. Lin, and Q. Wu, "Self-supervised 3D skeleton action representation learning with motion consistency and continuity," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 13328–13338.
[61] 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 Int. Conf. Comput. Vis.*, 2021, pp. 13423–13433.
[62] L. Li, X. Li, K. Wu, K. Lin, and S. Wu, "Multi-granularity feature interaction and relation reasoning for 3D dense alignment and face reconstruction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 4265–4269.
[63] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[64] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.
[65] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[66] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 9912–9924.
[67] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.
[68] C. Feichtenhofer, H. Fan, B. Xiong, R. Girshick, and K. He, "A large-scale study on unsupervised spatiotemporal representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3299–3309.
[69] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5171–5180.
[70] W. Liu, J. He, and S.-F. Chang, "Large graph construction for scalable semi-supervised learning," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 679–686.
[71] J. Tang, X. Shu, Z. Li, Y.-G. Jiang, and Q. Tian, "Social anchor-unit graph regularized tensor completion for large-scale image retagging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 2027–2034, 2019.
[72] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2649–2656.
[73] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1476–1485.
[74] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. Workshops*, 2013, Art. no. 896.
[75] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[76] Y. Grandvalet and Y. Bengio et al., "Semi-supervised learning by entropy minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 529–536.
[77] D. Holden, J. Saito, T. Komura, and T. Joyce, "Learning motion manifolds with convolutional autoencoders," in *Proc. SIGGRAPH Asia Tech. Briefs*, 2015, pp. 1–4.
[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[79] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Unsupervised learning of view-invariant action representations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1254–1264.
[80] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7912–7921.
[81] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling GCN with dropgraph module for skeleton-based action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 536–553.
[82] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1112–1121.
[83] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 183–192.
[84] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.
[85] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 2669–2676.

Xiangbo Shu (Senior Member, IEEE) received the PhD degree from the Nanjing University of Science and Technology, in 2016. He is currently a professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. From 2014 to 2015, he worked as an visiting scholar with the National University of Singapore, Singapore. His current research interests include computer vision, and multimedia. He has received the Best Student Paper Award in MMM 2016, and the Best Paper Runner-up in ACM MM 2015. He is also the Member of ACM, the Senior Member of CCF.

Binqian Xu is currently working toward the master's degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Her current research interest is computer vision, and deep learning.

Liyan Zhang received the PhD degree in computer science from the University of California, Irvine, CA, USA, in 2014. She is currently a professor with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her research interests include multimedia analysis, computer vision, and deep learning. She was a recipient of the best paper awards in ACM ICMR 2013 and ACM MM Asia 2021, the best student paper awards in MMM 2016 and ICIMCS 2017.

Jinhui Tang (Senior Member, IEEE) received the BE and PhD degrees from the University of Science and Technology of China, in 2003 and 2008, respectively. He is currently a professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He has authored more than 200 papers in top-tier journals and conferences. His research interests include multimedia analysis and computer vision. He has been serving or served as an associate editor of the *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Neural Networks and Learning Systems* and *IEEE Transactions on Circuits and Systems for Video Technology*. He is a fellow of IAPR.