Pattern Recognition 152 (2024) 110478



Contents lists available at ScienceDirect

Pattern Recognition

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





Augmented skeleton sequences with hypergraph network for self-supervised group activity recognition

Guoquan Wang, Mengyuan Liu *, Hong Liu, Peini Guo, Ti Wang, Jingwen Guo, Ruijia Fan

National Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, China

ARTICLE INFO

Keywords:
Self-supervised learning
Skeleton-based group activity recognition
Multi-person augmentation strategies
Hypergraph learning

ABSTRACT

Contrastive learning has been widely applied to self-supervised skeleton-based single-person action recognition. However, directly employing single-person contrastive learning techniques for multi-person skeleton-based Group Activity Recognition (GAR) suffers from some challenges. Firstly, single-person data augmentation strategies struggle to capture complex collaborations between actors in multi-person scenarios, resulting in poor generalization. Secondly, real-world uncertainties in the number of people make single-person methods fail to capture changing high-order actor relations. Finally, single-person methods treat each actor with equal importance for recognition, struggling to distinguish imbalanced contributions between individual and group activities. To this end, the coarse-to-fine **A**ugmented **H**ypergraph **Net**work (AHNet) is proposed for effective self-supervised GAR. Specifically, we introduce multi-person augmentation strategies to enhance the generalization of the model under complex actor collaboration scenarios. Moreover, a knowledge-masked hypergraph network is employed to enhance the adaptability of the model to capture varied high-order actor relations. Finally, coarse-to-fine contrast among key actors is conducted to mitigate the imbalanced contributions between individual and group levels. Extensive experiments on multiple datasets demonstrate that our AHNet achieves substantial improvements over state-of-the-art methods with various backbone architectures. Our code is available at https://github.com/WGQ109/AHNet.

1. Introduction

Group Activity Recognition (GAR) aims to identify and understand the collective behaviors of multiple individuals in a short video clip [1,2]. It can be applied to many fields, such as video surveillance, autonomous driving, and sports event analysis [3,4].

Previous approaches mainly relied on hand-crafted features or graphical models to capture spatial–temporal information [5,6]. However, these methods exhibited unsatisfactory performance. With the advent of deep neural networks, data-driven techniques have become the predominant approaches. Most of the early works first adopt Convolutional Neural Networks (CNNs) to extract the individual-level features and then aggregated these features with Recurrent Neural Networks (RNNs) [7,8]. More recent methods commonly adopt Graph Convolutional Networks (GCNs) or Transformers to model relations between individuals [9–11]. Given their promising results, subsequent works extended these two backbone architectures to the RGB-based GAR task [12–14].

Although the RGB-based methods have achieved remarkable performance, they are limited by the high computational expense. Consequently, the skeleton-based GAR task garners increasing attention due to its lightweight characteristic [15,16]. The majority of skeleton-based works employ a fully supervised manner [17–19]. However, GAR samples involve numerous individuals. Tracking and labeling individual skeletons frame by frame are time-consuming and labor-intensive. In this work, we focus on the self-supervised paradigm, aiming to avoid the laborious burden of manual annotation for the skeleton-based GAR task.

As shown in Fig. 1(a), self-supervised single-person action recognition approaches utilizing the siamese-like contrastive framework have exhibited remarkable efficacy [20–23]. However, directly applying these methods to self-supervised multi-person GAR suffers from the following limitations:

(1) Single-person data augmentation strategies typically rely on individual behavior patterns [21,24,25]. They are unsuitable for modeling complex spatial–temporal collaboration among multiple individuals. Directly applying them in multi-person scenes will lead to poor model generalization ability.

(2) Existing GCN-based methods are unable to capture high-order relations among varied actors [17,26]. As shown in Fig. 1(c), GCNs

* Corresponding author.
E-mail addresses: guoquanwang@stu.pku.edu.cn (G. Wang), nkliuyifang@gmail.com (M. Liu), hongliu@pku.edu.cn (H. Liu), guopeini@stu.pku.edu.cn (P. Guo), tiwang@stu.pku.edu.cn (T. Wang), jingwenguo@stu.pku.edu.cn (J. Guo), fanruijia@stu.pku.edu.cn (R. Fan).

https://doi.org/10.1016/j.patcog.2024.110478
Received 18 December 2023; Received in revised form 25 February 2024; Accepted 2 April 2024
Available online 4 April 2024
0031-3203/© 2024 Elsevier Ltd. All rights reserved.

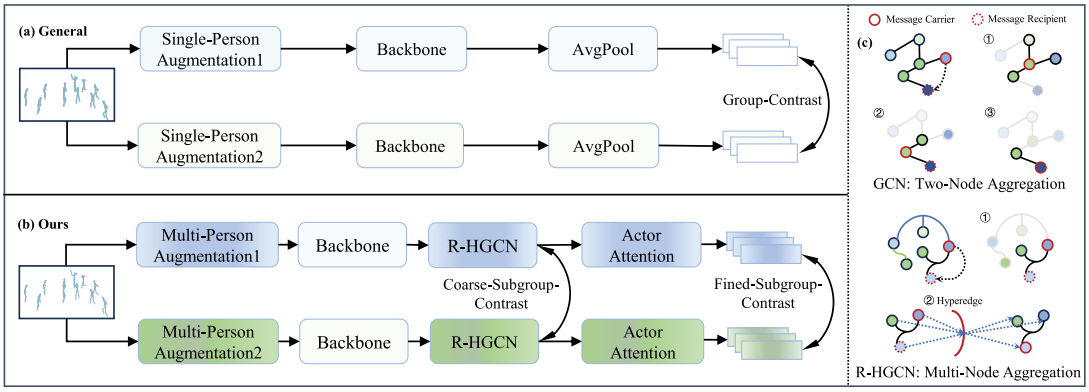


Fig. 1. Method comparison. (a) General methods use single-person augmentations for group-level contrast. (b) Our AHNet employs multi-person augmentations for coarse-to-fine contrast. (c) Message passing. GCN edges aggregate messages between two neighboring nodes, requiring multilayer transmission for distant nodes. R-HGCN edges simultaneously aggregate messages from multiple distant nodes for faster transmission.

necessitate multiple transmissions to establish relations between distant nodes. Additionally, the edges in GCNs can solely link two predefined nodes (one-order), while real-world actor relations extend beyond pairwise connections (high-order). Therefore, using GCNs makes it difficult to capture varied high-order actor relations in real scenarios.

(3) Existing methods are unmindful of the imbalanced contributions at individual and group levels. Self-supervised single-person action recognition methods treat each actor as equally important for inferring group activities [1,2]. Notably, actors for distinguishing group activities entail imbalanced spatial–temporal contributions. However, previous methods are constrained to coarse group-level features and fail to consider the fine-grained interactions among key actors.

To address these challenges, we propose an Augmented Hypergraph Network (AHNet) shown in Fig. 1(b) for the coarse-to-fine self-supervised skeleton-based GAR task. Our AHNet can effectively capture high-order actor correlations and selectively aggregate multi-scale features among key actors.

We address the **limitation (1)** related to spatial–temporal collaboration modeling by incorporating multi-person augmentation strategies. Namely, the intra-inter multi-person augmentation strategies are designed to enhance generalization of the model. It contains the intra-level Data-based cubecut Mixup strategy (DMix) and the inter-level Feature-based hyperbolic Mixup strategy (FMix). Specifically, DMix is employed to encourage the network to simultaneously learn spatially invariant features, temporally localizable features, and interactive features. Separately, FMix based on hyperbolic similarities in the embedding space captures complex multi-person collaborations. Then, the contrastive learning framework Barlow Twins [27] makes group skeleton embeddings invariant to these multi-person augmentations.

Furthermore, solutions are developed in response to the **limitation (2)** of being unable to capture high-order relations among varied actors. For one, the actor relations are represented as a Relational Hyper-Graph Convolutional Network (R-HGCN) shown in Fig. 1(c), aiming to leverage the high-order data correlations for an effective GAR process. For another, the knowledge-masked strategy and the graph-hypergraph contrast technique are utilized to enhance the adaptability of R-HGCN to varied actor scenarios.

Finally, a key actor matching mechanism with coarse-to-fine contrast is introduced to alleviate the **limitation (3)**, where actors in group activities valuable for the GAR task are ignored. Namely, key actors are selected to form subgroups whose interactions with each other are enhanced. More importantly, these identified key actors are then utilized for contrastive learning at both the coarse-grained subgroup and fine-grained interaction levels rather than coarse-grained group-level contrast. This approach effectively minimizes irrelevant relation information in the GAR task.

In summary, the contributions of this paper are as follows:

- We propose a novel framework, called AHNet, which pioneers the expansion of skeleton-based contrastive learning from single-person action recognition to multi-person group activity recognition.
- For relational inference, the intra-inter multi-person augmentation strategies are designed to model complex spatial–temporal collaboration. Then the knowledge-masked hypergraph convolutional network is utilized to capture high-order relations among varied actors.
- For key actor capture, the coarse-to-fine contrast technique is employed to learn the imbalanced contributions at individual and group levels.
- Our AHNet achieves state-of-the-art results across three different backbone models. Extensive evaluation protocols on the Volleyball and Collective Activity datasets demonstrate its superiority in the field of self-supervised group activity recognition.

2. Related work

2.1. Supervised group activity recognition

Supervised Group Activity Recognition (GAR) comprehensively analyzes activities by utilizing both individual and group labels, which can be categorized into two primary approaches: RGB-based GAR [9,28] and skeleton-based GAR [17,18]. Due to the imperative need for efficient data storage and fast inference time, recent researchers focus on analyzing group behaviors through lightweight skeleton-based GAR. ST-GCN [29] first takes joint coordinates as inputs and employs GCNs as the backbone for skeleton processing, but it overlooks inter-person relations. To this end, Zhai et al. [18] introduce the spatial Transformer to extract personal relations from individual features. However, it encountered difficulties in pinpointing local relations between actors. Therefore, ZOOM Transformer [17] combining GCNs and Transformers is proposed to focus on both global and local actor relation modeling. Furthermore, researchers divide the skeleton sequence into fine-grained entities to enhance GAR performance by consistent multi-scale representations [30,31]. However, these graph-based methods only consider one-order relations among actors. In this work, the hypergraph convolutional network is employed to capture high-order relations for skeleton-based GAR.

2.2. Hypergraph learning

GCNs primarily establish one-order relations between two predefined objects. In contrast, hyperedges can connect arbitrary objects to form a hypergraph that captures high-order dependencies. HGNN [32] first introduces a fixed hypergraph to capture high-order relations.

Subsequently, DHNN [33] constructs a dynamic hypergraph to address limitations of fixed hyperedge structure. Some studies have incorporated hypergraphs in human action analysis tasks. For example, AD-HNN [34] introduces an adaptive hypergraph to estimate human pose by considering high-order relations. HAM [35] assigns joints with different weights for high-order skeleton-based action recognition. The most relevant work is MNN [26], which employs spatial distance and appearance behavioral relations to construct a multi-hyperedge hypergraph for inference. However, MNN requires both individual and group labels to model a hypergraph and struggles with scenarios involving changing the number of actors. Distinctly, we creatively propose a self-supervised Relational HyperGraph Convolutional Network (R-HGCN) with the knowledge-masked strategy to capture high-order relations among varied actors.

2.3. Contrastive learning for group activity recognition

Self-supervised techniques design pretext tasks to learn powerful representations from unlabeled skeleton data. For example, AAGCM [36] employs the skeleton matching paradigm to identify which actors are from the same group. Recently, there has been a recent shift towards the contrastive learning paradigm due to its robust capabilities, as highlighted by works such as MoCo [37] and SimCLR [38]. Notably, STAHs [1] develops a group-level contrastive loss with single-person augmentation strategies to aggregate spatial–temporal features for each actor, while Con-RPM [2] equally aggregates the individual-level contrastive loss for GAR. Despite these advancements, the above methods directly apply self-supervised single-person techniques to group activity data and fail to effectively model imbalanced contributions at individual and group levels, resulting in sub-optimal learning outcomes. To this end, AHNet with multi-person augmentation strategies and coarse-to-fine subgroup contrast is proposed for self-supervised GAR.

3. Preliminaries

GCN & HGCN: The spatial–temporal graph of the multi-person skeleton is denoted by f(T, V, M), where T, V, and M denote time series, joints, and the number of actors, respectively [39–41]. In this work, the GCN architectures [29,42,43] are adopted as the backbone. The HyperGraph Convolutional Network (HGCN) for the multi-person relational reasoning is denoted by h(M, E), where M denotes the number of actors and E denotes their connections [32].

Barlow Twins. In this work, the Barlow Twins [27] framework is developed for the self-supervised GAR task. The cross-correlation matrix C of two augmented views is first employed as:

$$C_{ij} = \frac{\sum_b z^b_{1,i} z^b_{2,j}}{\sqrt{\sum_b \left(z^b_{1,i}\right)^2}\sqrt{\sum_b \left(z^b_{2,j}\right)^2}}, \tag{1}$$

where z^b denotes batch samples and i, j are the feature dimension indexes of the embeddings. The Barlow Twins loss is then calculated as:

$$\mathcal{L} = \sum_i \left(1 - C_{ii}\right)^2 + \lambda \sum_i \sum_{j \neq i} C^2_{ij}, \tag{2}$$

where the hyperparameter λ controls the trade-off between two terms.

4. AHNet

4.1. Overview

(a) The group skeleton sequences are separately fed into the *temporal scored-cutting* augmentation [22] module and the intra-level Data-based cubecut Mixup module (DMix). (b) Then the augmented sequences are fed into the GCN backbone and the relational inference networks R-HGCN. (c) Subsequently, the enhanced representations are sent into the actor attention module to obtain key actors. (d) The inter-level Feature-based hyperbolic Mixup (FMix) module then utilizes the augmented representations to yield the mixed features. (e) All the coarse-grained features and fine-grained features are matched with the key-actor index set to derive the person-level features. Their average features are regarded as subgroup-level features. (f) The graph-hypergraph contrast technique is introduced to make the model adapt to changes in the number of people. (g, h) The person-level feature and subgroup-level features are employed for coarse-to-fine subgroup-level contrastive learning. (i) Finally, the cross-correlation matrix is adjusted to be close to the identity matrix.

4.2. Multi-person augmentation strategies

Existing skeleton-based single-person augmentation strategies perform suboptimally for multi-person scenarios [36]. To enhance the generalization of the GAR model, we explore the intra-level Data-based cubecut Mixup augmentation strategy (see Section 4.2.1) and the inter-level Feature-based hyperbolic Mixup augmentation strategy (see Section 4.2.2).

4.2.1. Intra-level data-based cubecut mixup

Mixup-based data augmentation strategy has achieved considerable success as a regularization technique in the image domain [44,45]. However, its applicability to skeleton-based GAR, specifically in exploring the significance of collaboration between actors across time and joints, remains unexplored.

In this work, the intra-level Data-based cubecut Mixup (DMix) is proposed to incorporate spatial, temporal and relational information from skeleton sequence. It aims to effectively leverage both spatial–temporal dynamics and relational interactions for the self-supervised skeleton-based GAR task.

Fig. 3 shows the details of the DMix. Like mixup [44], DMix mixes sequences by a ratio γ from $Beta(\alpha, \alpha)$, where α is a hyperparameter. Unlike mixup, the binary Cubecut Mask ($\mathbf{CM} \in \{0,1\}^{L_S \times L_T \times L_P}$) enables mixing local areas across the spatial, temporal, and personal dimensions simultaneously. Here, L denotes the total length in each dimension, where S, T, and P represent spatial, temporal, and personal dimensions, respectively.

Generally, DMix has four steps: ① The mixing ratio γ first determines the beginning mixing coordinates ($B_\triangle$) and the cubecut length ($L^C_\triangle$) of the spatial, temporal, and personal dimensions, respectively:

$$B_\triangle \sim Random(0, L - L^C_\triangle), \quad L^C_\triangle = L_\triangle \sqrt[3]{1-\gamma}, \quad \triangle \in \{S, T, P\}. \tag{3}$$

② Next, the ending mixing coordinates ($E_\triangle$) are calculated as $E_\triangle = B_\triangle + L^C_\triangle$. ③ Then, as shown in Fig. 3, we intersect the range of the Spatial–Temporal–Personal Cube (STPC) from the start coordinates to the end coordinates and set the mixed value of **CM** at this intersection to 1, with the remaining values set to 0. ④ The final formula for DMix data is as follows:

$$X_{DMix} = \begin{cases} \mathbf{CM} \odot (\gamma X_{Aug1} + (1-\gamma) X_{Aug2}) + (\mathbf{1} - \mathbf{CM}) \odot X_{Aug1}, & \gamma < 0.5 \\ \mathbf{CM} \odot (\gamma X_{Aug1} + (1-\gamma) X_{Aug2}) + (\mathbf{1} - \mathbf{CM}) \odot X_{Aug2}, & \gamma \geq 0.5 \end{cases}, \tag{4}$$

where X_{DMix} represents cubecut mixed skeleton data, and $\odot$ denotes element-wise multiplication. γ is the mixing ratio.

4.2.2. Inter-level feature-based hyperbolic mixup

Although DMix augments intra-level samples effectively, it fails to consider the spatial distribution of dataset samples. To this end, a fully supervised solution interpolates hidden vectors among the same label samples to generate a mixed hidden vector at the feature level [46]. Inspired by this, the inter-level Feature-based Hyperbolic Mixup (FMix) is incorporated for the self-supervised augmentation. Specifically, FMix automatically identifies potential same-class samples based on their

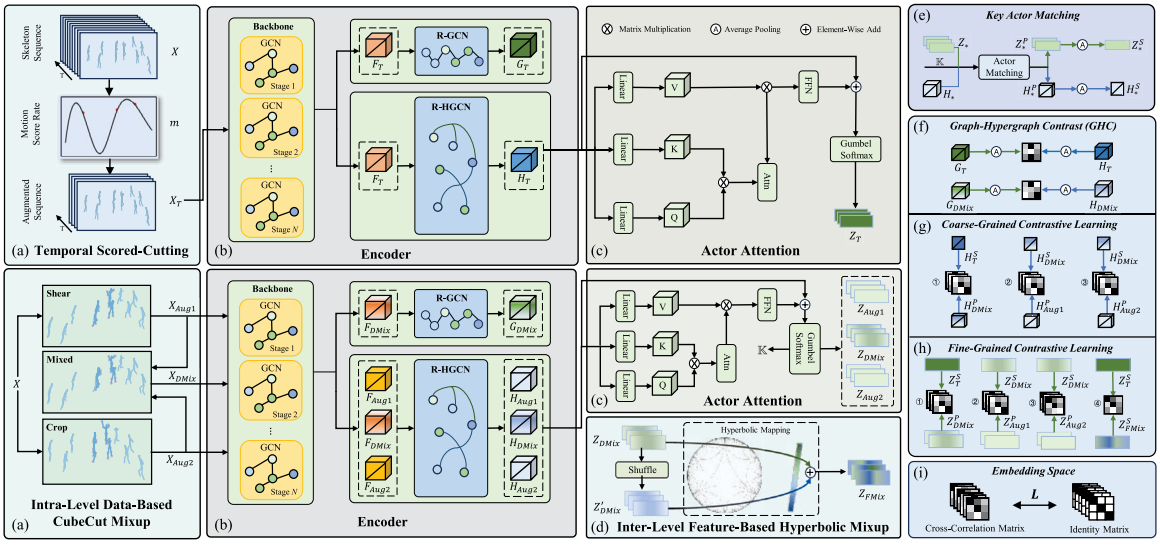


Fig. 2. Overview of the AHNet. Given augmented sequences X_*, the GCN backbone produces the skeleton features F_*. The Graph-Hypergraph Contrast (GHC) technique is then operated between G_* and H_*, which are the output of the R-GCN and the R-HGCN. Moreover, the enhanced features H_* are sent into the actor attention module to get the index set of K pivotal actors and constitute the subgroup-level representations Z^s_*. Specifically, $\mathbb{K}$ used for key actor matching in the temporal branch is obtained from the mixup branch. Finally, the coarse-to-fine contrast is used for GAR.

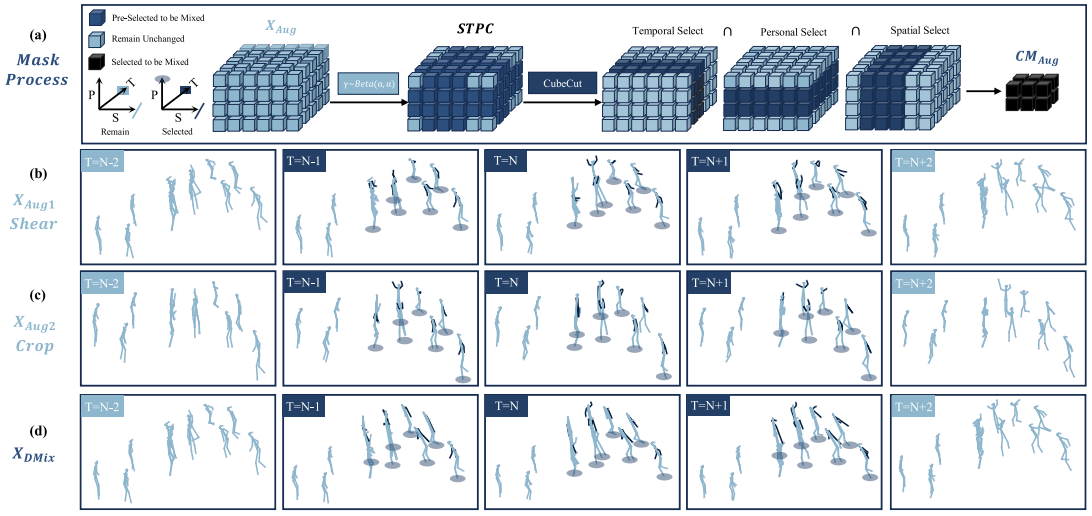


Fig. 3. Illustration of the DMix. (a) Generate the Spatial–Temporal–Personal Cube (STPC) based on the mixing ratio γ and then obtain the $\mathbf{CM}_{Aug}$ to be selected by taking the intersection of the three domains of the STPC matrix. (b) *shear* process. (c) *crop* process. (d) Take $\gamma < 0.5$ as an example to show the process of DMix. CM_{Aug} determines which features in each domain should be selected for local mixup. The axes in (a) distinguishes the difference between selected and unselected mixup samples in the three domains of spatial (S), temporal (T), and personal (P).

similarity without any labels. However, group skeleton features encompass hierarchical relations, and their similarities cannot be effectively captured in the Euclidean or Cosine spaces (See Experiments 5.3). Given the powerful performance of hyperbolic geometry in skeleton action analysis [47], the hyperbolic space is introduced to define the similarity between hidden vectors.

As depicted in Fig. 4, feature $Z_{DMix} = \{z_1, z_2, \dots, z_{b-1}, z_b\}$ is first shuffled along the batch dimension to obtain $Z'_{DMix} = \{z'_1, z'_2, \dots, z'_{b-1}, z'_b\}$, where b denotes mini-batch. Subsequently, we propose to adaptively interpolate the batch-wise representations by utilizing the hyperbolic distance of (z, z') pairs as mixing ratios. Mathematically, the Hyperbolic Distance Matrix (HDM) is defined as the hyperbolic distance between two feature pairs:

$$\mathbf{HDM}(z, z') = 2\tanh^{-1}(\|(-z) \bigcirc (z')\|), \quad (z, z') \in (Z_{DMix}, Z'_{DMix}), \tag{5}$$

where $\bigcirc$ is denoted as the Möbius summation on the Poincaré model [48] and then the Möbius summation operation is defined as:

$$z \bigcirc z' = \frac{(1 + 2\langle z, z'\rangle + \|z'\|^2)z + (1 - \|z\|^2)z'}{1 + 2\langle z, z'\rangle + \|z\|^2\|z'\|^2}, \tag{6}$$

where $\langle\cdot,\cdot\rangle$, $\|\cdot\|$ are Euclidean inner product and norm. HDM is employed as the mixing ratio to make shuffled z' with smaller similarity to original z get a smaller proportion for the reliable mixup. Finally, the inter-mixup process is executed at the feature level across all the batch-wise pairs:

$$z_{FMix} = (1 - \mathbf{HDM}(z, z'))z + \mathbf{HDM}(z, z')z', \quad (z, z') \in (Z_{DMix}, Z'_{DMix}). \tag{7}$$

where z_{FMix} denotes the batch-wise FMix feature and Z_{FMix} is obtained by stacking all the batch-wise mixed features.

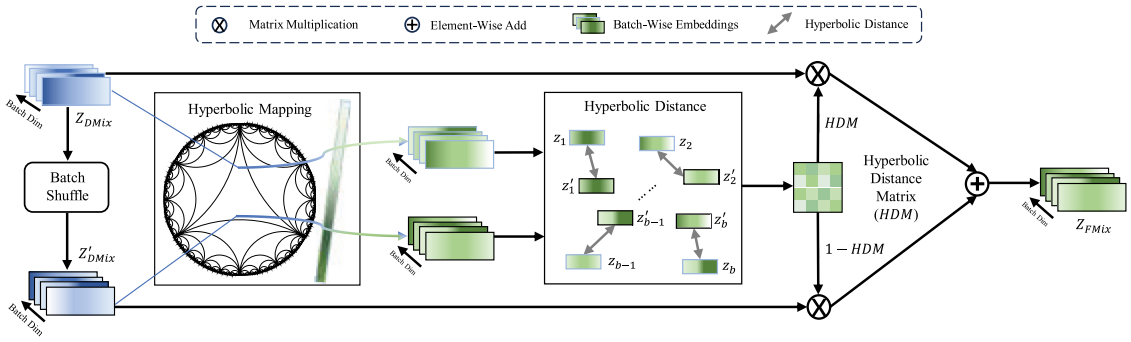


Fig. 4. Illustration of the FMix. FMix takes in two hidden features (Z_{DMix}, Z'_{DMix}) and adaptively mixes them with Hyperbolic Distance Matrix (HDM) into Z_{FMix}.

4.3. Relational hypergraph convolutional network

As illustrated in Fig. 5(a), previous methods utilized the connections of adjacent actors to depict pair-wise relations and construct the Relational Graph Convolutional Network (R-GCN) [17,26]. However, rather than simple one-to-one interactions, individual behaviors are typically triggered by more than one actor in group activities. To this end, the Relational HyperGraph Convolutional Network (R-HGCN) is employed with various hyperedge incidence matrices to explore high-order relations between actors.

① Position-connected hyperedge. As shown in Fig. 5(b), the hyperedges in the incidence matrix R_1 are tailored based on the current regions of the volleyball actors. Specifically, the front and back rows are divided according to the standing positions of the teams (resulting in a total of 4 regions for the two teams). The hyperedge is then employed to connect all the actors in each region, aiming to capture high-order relations among them.

② Strategy-connected hyperedge. This approach involves associating left and right actors who fulfill the same role (e.g. striker, setter). As shown in Fig. 5(b), the incidence matrix R_2 gains a comprehensive understanding of the action characteristics corresponding to each role. Notably, actor positions typically remain stable for coordination in volleyball games, with teams employing specific strategies. This stability ensures the validity and rationality of strategy-connected hyperedge.

③ Adaptive-connected hyperedge. To extend the applicability of hyperedges to diverse scenarios, the adaptive-connected incidence matrix R_{Ada} is introduced to automatically learn which nodes should be connected.

Although R-HGCN enhances the performance of our model, the changing number of actors poses a potential obstacle to efficient hypergraph relation modeling. To this end, a strategy involving the masking of hypergraph nodes is proposed, which simulates scenarios with changing actors in real-world situations. As discussed in Section 1, the contributions between individual and group behaviors is often unequal, with actors standing closer to the center exerting more influence on the emergence of group activities. In light of this, the hypergraph knowledge-masked strategy is proposed to assign higher mask probabilities for the actors situated on either side of the group.

However, the sparse R-HGCN poses challenges to capture intrinsic relations when a significant number of actors are masked. To this end, the unmasked R-GCN shown in Fig. 5(a) is introduced as a self-supervised signal for masked R-HGCN pre-training. As illustrated in Fig. 6, we term this process as Graph-Hypergraph Contrast (GHC). Namely, in the R-GCN branch (green), the connection matrix is first partitioned into three submatrices, which are then concatenated into a single adjacency matrix $\mathbb{R}^{3\times M\times M}$ [29]. To facilitate the node aggregation of R-GCN, the Conv1D operation is employed to expand the input feature from $\mathbb{R}^{M\times D}$ to $\mathbb{R}^{3\times M\times D}$. Subsequently, we employ the matrix multiplication to obtain the R-GCN output feature. In the R-HGCN branching (blue), the hyperedge matrices R_1 and R_2 in Fig. 5(b) are concatenated to learn the high-order relations among actors. Similarly, the Conv1D operation is employed to expand the input feature of the R-HGCN branch into $\mathbb{R}^{2\times M\times D}$. Unlike the R-GCN branch, we first uses matrix multiplication to aggregate hyperedges and then aggregates hypernodes to get the R-HGCN output feature. Finally, the mutual information between the unmasked R-GCN and the masked R-HGCN is maximized through the contrastive learning process:

$$\begin{aligned}\mathcal{L}^{T}_{GHC} &= \mathcal{L}(G^{G}_{T}, H^{G}_{T}),\\ \mathcal{L}^{DMix}_{GHC} &= \mathcal{L}(G^{G}_{DMix}, H^{G}_{DMix}),\\ \mathcal{L}_{GHC} &= \mathcal{L}^{T}_{GHC} + \mathcal{L}^{DMix}_{GHC}.\end{aligned} \tag{8}$$

where $\mathcal{L}$ denotes the Barlow Twins loss shown in Eq. (2). As shown in Fig. 2(f), G^{G}_{T}, H^{G}_{T}, G^{G}_{DMix}, and H^{G}_{DMix} are the averaged features along the personal dimension of G_T, H_T, G_{DMix}, and H_{DMix}. In this way, the network can accommodate the changing number of actors while simultaneously considering both one-to-one and one-to-many connections.

4.4. Coarse-to-fine subgroup-level contrastive learning

Interactions at individual and group levels in skeleton sequence are unequal, and only a limited number of actors and their interactions are valuable for group activity recognition [30]. Traditional approaches [1,22] treat all actors as equally important and ignore the fine-grained differences among them, which is detrimental to group activity recognition.

For this reason, the actor attention mechanism is incorporated to identify key actors for contrastive learning. As shown in Fig. 2(c), the encoded features $H_\star \in \mathbb{R}^{M\times D}$ ($\star \in \{T, Aug1, Aug2, DMix\}$) calculate self-attention scores along personal dimension to obtain weighted feature $Z^{Att}_\star \in \mathbb{R}^{M\times D}$.

$$Z^{Att}_\star = H_\star + FFN(Softmax(\frac{Q_\star K^T_\star}{\sqrt{d_k}})V_\star). \tag{9}$$

We use Gumbel Softmax [49] to ensure the gradient in the following top-K sorting can be computed from dispersed samples of the original distribution:

$$Z_\star = Softmax((\log(Z^{Att}_\star) - \log(-\log(\mathbf{g})))/\tau), \tag{10}$$

where $\mathrm{g} \sim Gumbel(0, 1)$ and τ is the softmax temperature. Subsequently, the indexes corresponding to the K maxima of $Z_\star$ are selected to form the $\mathbb{K}$ set using the top-K function. The key actor matching mechanism shown in Fig. 2(e) then matches weighted feature $Z_\star \in \mathbb{R}^{M\times D}$ in the personal dimension with top-K indexes to obtain the person-level key actors $Z^{P}_\star \in \mathbb{R}^{K\times D}$. Finally, the average pooling operation is performed over the personal dimension to derive the subgroup-level features $Z^{S}_\star \in \mathbb{R}^{D}$.

Another problem is that existing methods predominantly engage in coarse-grained contrast between (F_{DMix}, F_T), which are the output of

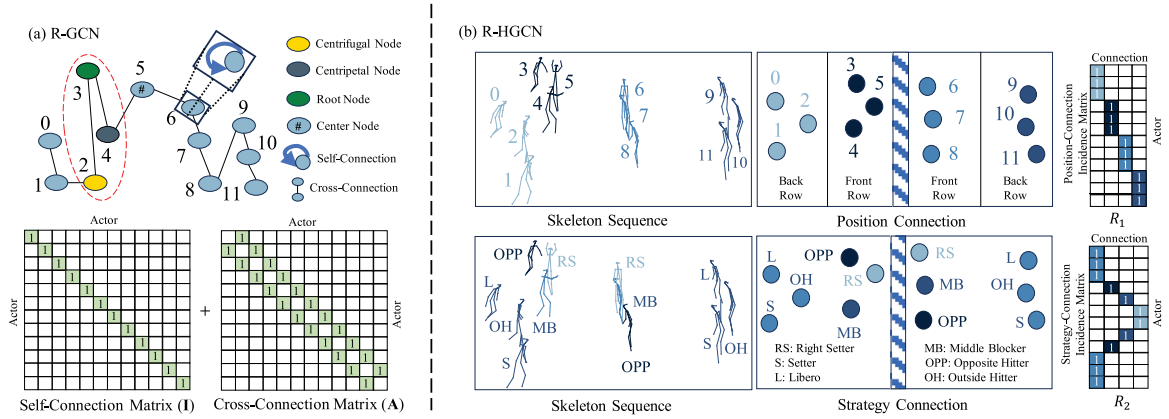


Fig. 5. The connection relations in R-GCN: circles denote actors, ellipses signify the self-connection (**I**) of each actor, and connecting lines between ellipses depict cross-connections (**A**) among actors. Additionally, R-GCN employs the spatial configuration partitioning strategy [29]. In contrast to R-GCN, which establishes node-to-node relations, R-HGCN establishes connections between nodes and edges. It derives incidence matrices from positional and strategic connections to facilitate high-order relation modeling.

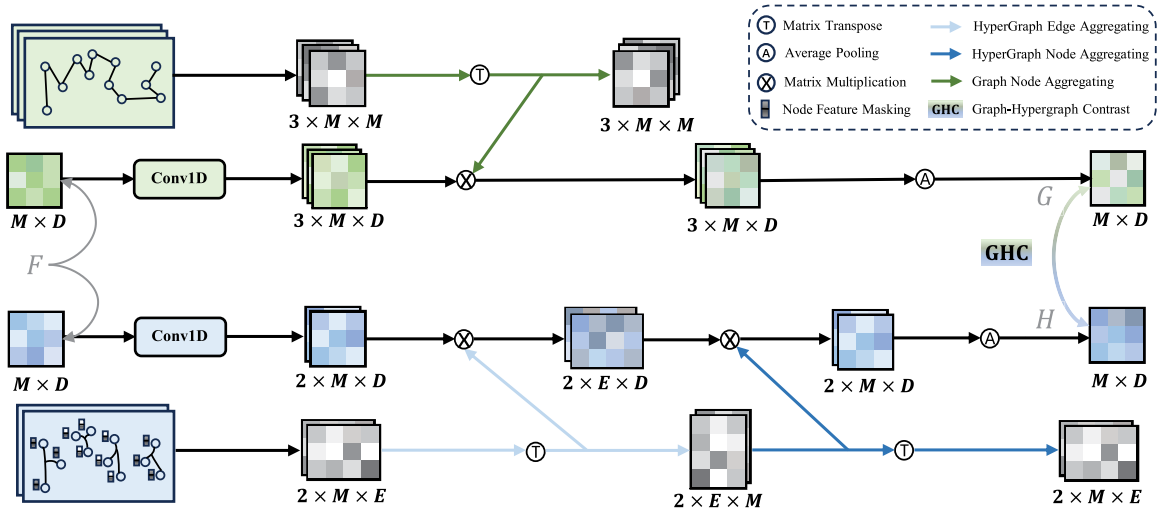


Fig. 6. R-GCN (green) is constituted by pair-wise connections of neighboring nodes, which are converted into three subgraphs with the spatial configuration partitioning strategy for aggregation. Simultaneously, R-HGCN (blue) employs the knowledge-masked strategy to adapt to variations of actor count, which first aggregates the features with the combination of position-connected and strategy-connected hyperedges. Finally, the aggregation result of unmasked R-GCN is used as a self-supervised signal for masked R-HGCN to process Graph-Hypergraph Contrast (GHC). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the GCN backbone at the mixup branch and the temporal branch [22, 23]:

$$\mathcal{L}(F^G_{DMix}, F^G_T) = \sum_i \left(1 - C^{gg}_{ii}\right)^2 + \lambda \sum_i \sum_{j \neq i} C^{gg\,2}_{ij}, \tag{11}$$

where C^{gg}_{ii} and C^{gg}_{ij} are the cross-correlation matrices of the group-level feature pair (F^G_{DMix}, F^G_T) that is obtained by the average pooling operation along the personal dimension of (F_{DMix}, F_T). However, the group-level contrast may yield imprecise representations by overstating the importance of irrelevant actors. To this end, we opt to replace group-level actors with subgroups comprised of key actors for coarse-grained contrastive learning:

$$\mathcal{L}(H^P_{DMix}, H^S_T) = \sum_{s \in \mathbb{K}} \Big(\sum_i \left(1 - \mathfrak{C}^{ps}_{ii}\right)^2 + \lambda \sum_i \sum_{j \neq i} \mathfrak{C}^{ps\,2}_{ij} \Big), \tag{12}$$

where $\mathfrak{C}^{ps}_{ii}$ and $\mathfrak{C}^{ps}_{ij}$ are the cross-correlation matrices of R-HGCN subgroup pair (H^P_{DMix}, H^S_T). As shown in Fig. 2(g)(①), H^P_{DMix}, H^P_T denotes the person-level DMix and temporal feature obtained by the key actor matching mechanism, while H^S_{DMix}, H^S_T denotes their averaged features.

Moreover, to enhance the DMix process, we follow the Eq. (1) and calculate the coarse-grained Barlow Twins loss to maximize the mutual information between the mixed samples and the augmented samples: $\mathcal{L}^{Aug1}_{Coarse} = \mathcal{L}(H^P_{Aug1}, H^S_{DMix})$, $\mathcal{L}^{Aug2}_{Coarse} = \mathcal{L}(H^P_{Aug2}, H^S_{DMix})$ shown in Fig. 2(g) (②, ③). Subsequently, $\mathcal{L}^{DMix}_{Coarse}$ is designed to aggregate $\mathcal{L}^{Aug1}_{Coarse}$ and $\mathcal{L}^{Aug2}_{Coarse}$ according to the ratio of mixing and the proportion of local mixup (correspond Eq. (4)):

$$\mathcal{L}^{DMix}_{Coarse} = \begin{cases} m(\gamma \mathcal{L}^{Aug1}_{Coarse} + (1-\gamma)\mathcal{L}^{Aug2}_{Coarse}) + (1-m)\mathcal{L}^{Aug1}_{Coarse}, & \gamma < 0.5 \\ m(\gamma \mathcal{L}^{Aug1}_{Coarse} + (1-\gamma)\mathcal{L}^{Aug2}_{Coarse}) + (1-m)\mathcal{L}^{Aug2}_{Coarse}, & \gamma \geq 0.5 \end{cases}, \tag{13}$$

where γ denotes the mixing ratio and m denotes the proportion of the original data in the mixup region (correspond Eq. (3)):

$$m = \frac{L^C_S \times L^C_T \times L^C_P}{L_S \times L_T \times L_P} = 1 - \gamma. \tag{14}$$

Generally, the coarse-grained contrastive learning shown in Fig. 2(g) is expressed as:

$$\mathcal{L}_{Coarse} = \mathcal{L}(H^P_{DMix}, H^S_T) + \mathcal{L}^{DMix}_{Coarse}. \tag{15}$$

Meanwhile, the fine-grained contrasts depicted in Fig. 2(h) are conducted to ensure the model to focus on interactions among key actors.

Notably, the fine-grained computation in Fig. 2(h)(①, ②, ③) follows the same way as Fig. 2(g)(①, ②, ③), except that the input for the loss changes from the coarse-grained features of the R-HGCN to the fine-grained features of the actor attention module. Different from coarse-grained process, we introduce extra $\mathcal{L}_{Fine}^{FMix} = \mathcal{L}(Z_T^S, Z_{FMix}^S)$ shown in Fig. 2(g)(④), aiming for the feature-based mixed features to capture intrinsic relations when interacting with the temporal-augmented features. As shown in Fig. 2(h), the fine-grained contrastive learning is expressed as:

$$\mathcal{L}_{Fine} = \mathcal{L}(Z_{DMix}^P, Z_T^S) + \mathcal{L}_{Fine}^{DMix} + \mathcal{L}_{Fine}^{FMix}. \tag{16}$$

In conclusion, we follow the Barlow Twins framework [27] to make the cross-correlation matrices in the graph-hypergraph contrastive learning process (Fig. 2(f)), coarse-grained contrastive learning process (Fig. 2(g)), and fine-grained contrastive learning process (Fig. 2(h)) both close to the identity matrix (Fig. 2(i)) with the predefined coefficient β:

$$\mathcal{L}_{Total} = \beta_1 \mathcal{L}_{GHC} + \beta_2 \mathcal{L}_{Coarse} + \beta_3 \mathcal{L}_{Fine}. \tag{17}$$

5. Experiments and discussions

5.1. Dataset

Volleyball Dataset (VD) [4] comprises 4,830 video clips, with 3,493 designated for training and 1,337 for testing. These clips are sourced from 55 volleyball matches.

Collective Activity Dataset (CAD) [8] contains 2481 activity clips captured in dynamic view by a low-resolution handheld camera. We follow [2] to merge the label crossing with walking to the label moving.

NTU RGB+D (NTU-60) [50] contains 60 human activities with 3D coordinates of 25 joints. We perform experiments on the cross-subject (xsub) and cross-view (xview) settings. It contains 11 classes of mutual actions.

NTU RGB+D 120 (NTU-120) [51] is an extension of the NTU-60, which provides cross-subject (xsub) and cross-setup (xset) evaluation protocols. In total there are 24,794 videos from 26 interaction only classes.

PKU-MMD [52] are developed for human action recognition and contain 51 action categories, of which 10 actions are mutual-action classes.

5.2. Experimental settings

For data pre-processing, the HRNet [41] pre-trained on the COCO dataset [53] is adapted as the pose estimation network to obtain group activity skeletons. The mini-batch size is set to 32. Our method is designed for multi-person scenarios, but most of the actions in NTU and PKU are completed in single-person scenarios. Direct use of these two datasets cannot reflect the advantages of our method. It is worth noting that some actions in NTU and PKU occur in two-person interaction scenarios. We choose the types of interactive motion involved in NTU and PKU (NTU-60-Mutual, NTU-120-Mutual, and PKU-Mutual) for self-supervised pre-training and evaluation.

Self-supervised Pretext Training. The default backbone model is constructed with ST-GCN [29], the number of stages is reduced to 2/5 of the original setting. For contrastive settings, we follow [27] to project features into 6144-dim embeddings and choose *shear* [24] and *crop* [21] and *temporal score-cutting* [22] as the normal augmentation strategy. λ in the Barlow Twins loss of each stream is set to 2e-4, and β in the total loss of each stream is set to 1. Hyperparameter α used in DMix process is set to 1.2. The weight decay is set to 1e-5. For optimization, we adopt the Adam optimizer and the CosineAnnealing scheduler with 300 epochs. The model is trained with a 10-epoch warmup for the stable contrastive learning.

Table 1
Ablation evaluation results through linear evaluation.

MA	R-HGCN	CFSCL	VD	CAD
✗	✗	✗	60.65	56.84
✗	✔	✔	67.83	80.22
✔	✗	✔	66.91	81.93
✔	✔	✗	64.92	76.52
✔	✔	✔	**72.58**	**84.31**

Table 2
Exploration of different relation networks through linear evaluation.

None	R-GCN	R-HGCN	GHC	VD	CAD
✔	✗	✗	✗	66.91	81.93
✗	✔	✗	✗	68.56	80.69
✗	✗	✔	✗	70.64	82.37
✗	✔	✗	✔	69.42	83.76
✗	✗	✔	✔	**72.58**	**84.31**

Linear Evaluation Protocol. The models are verified by linear evaluation for the group activity recognition task, which utilized classification accuracy (ACC) as a performance metric. Specifically, a linear classifier is trained with a encoder fixed. The Adam optimizer with an initial learning rate of 0.01 in 50 epochs is used for optimization.

Finetune Evaluation Protocol. We append a linear classifier to the trained encoder and train the whole model to compare it with fully supervised strategies. For optimization, the Adam optimizer and the CosineAnnealing scheduler with an initial learning rate of 0.01 in 50 epochs is applied.

5.3. Ablation study

To validate the effectiveness of AHNet, ablation studies are performed under multiple evaluations with the task of skeleton-based GAR.

Effectiveness of Each Component. Table 1 reveals the quantitative impact of removing specific component. Multi-person Augmentations (MA) significantly improves the generalization of the model, as evidenced by a performance decrease of 4.75% on VD (from 72.58% to 67.83%) and 4.09% on CAD when it is omitted. Similarly, the exclusion of the Relation HyperGCN (R-HGCN) results in a performance drop of 5.67% on VD and 2.38% on CAD, which emphasizes its crucial role in capturing high-order relations among varied actors. Meanwhile, Coarse-to-Fine Subgroup Contrastive Learning (CFSCL) also proves to be pivotal, with its absence leading to a decline of 7.66% on VD and 7.79% on CAD, showing its capability in modeling imbalanced contributions at individual and group levels.

Effectiveness of the Combinations of Relation Inference. Table 2 illustrates the impact of various relation modeling approaches. Initially, None yields a baseline performance of 66.91% on VD and 81.93% on CAD. Introducing R-GCN results in a 1.24% decrease on CAD. Conversely, R-HGCN gets 0.44% improvements on CAD, showcasing its ability in reasoning relations in multi-person scenarios. Furthermore, the combination of R-HGCN and GHC achieves the highest performance on both VD and CAD, indicating its effectiveness in capturing high-order relations among actors.

Effectiveness of the Combinations of Multi-Person Augmentation Strategies. Table 3 presents the evaluation of various data augmentation techniques for intra-level Data-based Mixup (DMix) and inter-level Feature-based Mixup (FMix). Notably, Spatial Data-based Mixup (DMix-S), Temporal Data-based Mixup (DMix-T) and Personal Data-based Mixup (DMix-P) exhibit a signature trend, with a 2.38%, 0.28%, and 2.60% boost on CAD, respectively. In terms of feature-based mixup techniques, Hyperbolic Distance Feature Mixup (FMix-Hyb) shows a significant improvement of 1.79% and 4.54% on VD with respect

Table 3
Exploring diverse multi-person augmentation strategies through linear evaluation.

NO	DMix-S	DMix-T	DMix-P	FMix-Cos	FMix-Euc	FMix-Hyb	VD	CAD
1	✗	✔	✔	✗	✗	✔	70.20	81.93
2	✔	✗	✔	✗	✗	✔	72.30	83.12
3	✔	✔	✗	✗	✗	✔	69.98	79.68
4	✔	✔	✔	✔	✗	✗	70.79	83.41
5	✔	✔	✔	✗	✔	✗	68.04	82.49
6	✔	✔	✔	✗	✗	✔	**72.58**	**84.31**

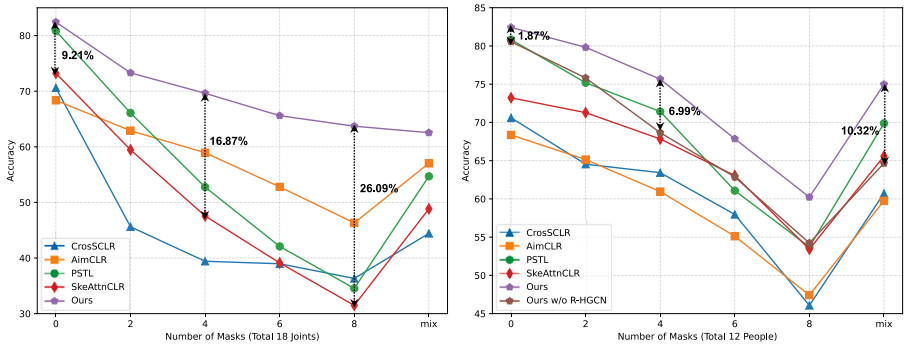


Fig. 7. Finetune evaluation for robustness test on VD. 'mix' means mask is not fixed (range 0–8). AHNet shows powerful robustness in a variety of complex scenarios.

Table 4
Exploration of different hyperedges through linear evaluation.

PC	SC	AC	VD	CAD
✗	✗	✗	66.91	81.93
✔	✗	✗	70.58	82.38
✗	✔	✗	71.22	82.53
✗	✗	✔	63.31	82.31
✔	✔	✗	**72.58**	**84.31**

to Cosine Similarity Feature Mixup (FMix-Cos) and Euclidean Distance Feature Mixup (FMix-Euc), proving to be an effective measure of feature-level skeleton-based interactions.

Exploration of the Different Hyperedge Connection. Table 4 shows the selection of hyperedges for relation inference. Positional Connection (PC) and Strategic Connection (SC) both result in a moderate performance on two datasets. Adaptive Connection (AC) provides a 0.38% boost on CAD accuracy while yielding a slight decrease on VD (63.31% vs baseline 66.91%), showing its sensitivity to the strategic volleyball game analysis.

5.4. Comparison with state-of-the-art

Linear Evaluation Protocol. Table 5 presents the results of comparing AHNet with other self-supervised learning methods. To showcase the generality of the proposed self-supervised learning paradigm, experiments are conducted on different backbones. Both self-supervised learning paradigms achieve a significant performance improvement compared with randomly initialized parameters, demonstrating the advantages of self-supervised learning. CrosSCLR [20] introduces additional motion information for auxiliary learning and treats all the actors as consistently important for skeleton-based self-supervised learning. Notably, our AHNet solely uses joint coordinates and outperforms CrosSCLR by 23.15% on VD with 2S-AGCN backbone [43], which is attributed to the efficient capture of key actors to deal with the unequal contributions between actors and group behaviors in GAR.

In addition, our AHNet designs applicable data augmentation strategies for multi-person scenarios and outperforms AimCLR [21], which introduces a variety of extreme single-person skeleton data augmentation strategies, by 17.94% on VD with ST-GCN backbone [29]. While SkeAttnCLR [23] focuses on obtaining fine-grained representations by introducing contrastive learning at local skeleton joints. However, the abundance of noise in group skeleton data leads to sub-optimal results. In contrast, our approach emphasizes fine-grained modeling at the global skeleton joint domain (i.e. the actor level) and employs a relational hypergraph network (R-HGCN) to capture high-order relations, achieving a 20.22% gain on the CAD dataset with Shift-GCN backbone [42] compared to SkeAttnCLR [23].

Finetune Evaluation Protocol. As shown in Table 6, our AHNet and other SOTA methods undergo finetune evaluation under different backbones. We first present fully supervised training results of randomly initialized networks as a reference to assess the improvement of self-supervised learning methods. AHNet increases the accuracy by 21.61%, 8.59%, and 8.07% compared with random initialized models on three backbone architectures on CAD, respectively. To enhance the practicability of our model in different scenarios, we provide different feature dimension versions of AHNet. The feature dimensions of AHNet-Large are consistent with the feature dimensions of the backbone network. The feature dimensions of AHNet-Middle are half of the backbone network. The feature dimensions of AHNet-Small are 1/4 of the backbone network. Unless otherwise specified, the default model in our manuscript is AHNet-Large. As shown in Table 7, AHNet-Middle and AHNet-Small are superior to existing self-supervised learning methods [20–23] in terms of model complexity. The AHNet-Middle model can still maintain optimal performance compared with these methods [20–23]. It can be found that our AHNet is still competitive. For example, AHNet-Small, AHNet-Middle, and AHNet-Large models outperform the fully supervised model ZOOM Transformer [17] by 5.85%, 9.7%, and 13.7% on VD, respectively. Notably, AHNet-Small exceeds the fully supervised model ISTIN [11] on CAD by 2.95%, but the number of model parameters is only 1/517 of ISTIN.

It is worth noting that there is also a certain gap between our method and Zappardino et al. [54] in the field of self-supervised learning. We analyze the following three reasons: (1) Group label. Although Zappardino et al. does not use the label of each individual, it still uses the label of the group activity in the pre-training process. For a fair comparison with previous methods [20–23], our AHNet does not use any label in the pre-training process. (2) Pre-training method. Zappardino et al. first trains on the extra dataset with individual labels to get the knowledge. Then the knowledge is transferred to the VD dataset without individual labels. Other self-supervised learning methods [20–23] in Table 7 employ contrastive learning methods for

Table 5
Linear results compared with other self-supervised methods on VD and CAD.

Method	VD						CAD					
	ST-GCN		2S-AGCN		Shift-GCN		ST-GCN		2S-AGCN		Shift-GCN	
	ACC	Δ	ACC	Δ	ACC	Δ	ACC	Δ	ACC	Δ	ACC	Δ
Random	25.08	0.00	30.29	0.00	28.88	0.00	40.13	0.00	42.78	0.00	39.82	0.00
CrosSCLR (CVPR' 21) [20]	41.97	+16.89	53.25	+22.96	47.76	+18.88	42.61	+2.48	55.54	+12.76	46.64	+6.82
AimCLR (AAAI' 22) [21]	54.64	+29.56	56.47	+26.18	61.26	+32.38	44.18	+4.05	52.78	+10.00	45.61	+5.79
PSTL (AAAI' 23) [22]	58.02	+32.94	62.45	+32.16	56.43	+27.55	48.83	+8.70	52.00	+9.22	47.52	+7.70
SkeAttnCLR (IJCAI' 23) [23]	58.49	+33.41	60.23	+29.94	57.89	+29.01	57.39	+17.26	60.65	+17.87	56.88	+17.06
AHNet (Ours)	**72.58**	**+47.50**	**73.83**	**+43.54**	**70.91**	**+41.63**	**84.11**	**+43.98**	**81.90**	**+39.12**	**77.10**	**+37.28**

Table 6
Finetune results compared with other self-supervised methods on VD and CAD.

Method	VD						CAD					
	ST-GCN		2S-AGCN		Shift-GCN		ST-GCN		2S-AGCN		Shift-GCN	
	ACC	Δ	ACC	Δ	ACC	Δ	ACC	Δ	ACC	Δ	ACC	Δ
Fully-Supervised	64.83	0.00	72.17	0.00	73.04	0.00	67.71	0.00	79.15	0.00	72.64	0.00
CrosSCLR (CVPR' 21) [20]	75.02	+10.19	70.83	−1.34	70.38	−2.66	67.32	−0.39	65.88	−13.27	76.73	+4.09
AimCLR (AAAI' 22) [21]	76.44	+11.61	74.50	+2.33	76.29	+3.25	83.01	+15.30	83.92	+4.77	71.37	−1.27
SkeAttnCLR (IJCAI' 23) [23]	77.85	+13.02	73.22	+1.05	76.30	+3.26	78.56	+10.85	81.63	+2.48	76.47	+3.83
PSTL (AAAI' 23) [22]	79.62	+14.79	79.90	+7.73	81.06	+8.02	69.92	+2.21	82.55	+3.40	79.62	+6.98
AHNet (Ours)	**84.31**	**+19.48**	**82.68**	**+10.51**	**83.01**	**+9.97**	**89.32**	**+21.61**	**87.74**	**+8.59**	**80.71**	**+8.07**

Table 7
Finetune results compared with other GAR methods on VD and CAD.

Method	Extra Supervision	Type	Params	FLOPs	VD	CAD
Supervised:						
ZOOM Transformer (TCSVT' 22) [17]	✗	Skeleton	2.93M	5.6G	71.20	–
Zhai et al. (ACCV' 22) [18]	✗	Skeleton	–	–	83.90	**95.00**
ISTIN (MTA' 23) [11]	✗	RGB	165.46M	5.3G	88.80	86.50
ARG (CVPR' 19) [9]	✗	RGB	61.19M	161.6G	92.50	91.00
Actor-Transformer (CVPR' 20) [10]	✗	Skeleton	–	–	92.30	–
Actor-Transformer (CVPR' 20) [10]	✗	RGB+Skeleton	–	–	**93.50**	91.00
Self-Supervised:						
CrosSCLR (CVPR' 21) [20]	✗	Skeleton	0.84M	0.57G	75.02	67.32
AimCLR (AAAI' 22) [21]	✗	Skeleton	0.84M	0.57G	76.44	83.01
PSTL (AAAI' 23) [22]	✗	Skeleton	0.84M	0.57G	79.62	78.56
SkeAttnCLR (IJCAI' 23) [23]	✗	Skeleton	3.06M	2.8G	77.85	69.92
Con-RPM (TMM' 23) [2]	✗	RGB+Skeleton	–	–	81.50	75.40
AHNet-Small (Ours)	✗	Skeleton	**0.32M**	**0.17G**	77.05	89.45
AHNet-Middle (Ours)	✗	Skeleton	0.72M	0.47G	80.69	88.76
AHNet-Large (Ours)	✗	Skeleton	3.83M	2.7G	84.31	89.32
3s-Zappardino et al. (ICPR' 20) [54]	✗	Skeleton	–	–	80.90	–
3s-Zappardino et al. [a] (ICPR' 20) [54]	✗	Skeleton	–	–	84.90	–
3s-Zappardino et al. [a] (ICPR' 20) [54]	✔	Skeleton	–	–	**91.00**	–
3s-AHNet-Small (Ours)	✗	Skeleton	0.97M	0.52G	78.83	**93.33**
3s-AHNet-Middle (Ours)	✗	Skeleton	2.16M	1.4G	84.14	91.76
3s-AHNet-Large (Ours)	✗	Skeleton	11.4M	8.2G	85.34	92.02

[a] denotes the method using complex feature fusion.
"Extra Supervision" denotes using extra pre-trained model.

pre-training. They do not use any labels or additional data during the pre-training process. Therefore, it is unfair to directly compare with Zappardino et al. which utilizes the source domain dataset as an additional supervision signal. When removing the additional supervision signal, our 3s-AHNet-Large outperforms Zappardino et al. by 0.44% (85.34% vs 84.90%) on VD. (3) Additional streams are introduced. The skeleton sequence contains three streams: joint, bone, and motion [20–23]. Zappardino et al. employs a complex feature fusion method for three streams through CNNs. For a fair comparison, we follow AimCLR [21] and PSTL [22] and supplement the 3s experimental results with a simple decision fusion method. To be specific, we added the classification scores of the three streams for group activity recognition. Our 3s-AHNet-Large outperforms Zappardino et al. by 0.44% (85.34% vs 84.90%) on VD even though Zappardino et al. uses a stronger fusion method. When Zappardino et al. also uses decision fusion, our model even outperforms it by 4.44% (85.34% vs 80.90%) on VD.

Table 8 shows the comparison of our model with other self-supervised learning methods on NTU-60-Mutual, NTU-120-Mutual, and PKU-Mutual. AHNet-Middle can outperform other self-supervised learning methods with a smaller number of parameters.

Fig. 7 simulates the situation of occlusion and people changes by carrying out robustness tests with different scale masks of skeleton joints and actors. As shown in Fig. 7(a), the gap between our AHNet and SkeAttnCLR [23] becomes more obvious when the proportion of masked skeleton joints increases (from 9.24% to 26.09%, with the gap reaching 1.82 times), which proves the robustness of our AHNet. Moreover, this ablation process is also highlighted in Fig. 7(b) to argue the adaptability of the proposed knowledge-masked hypergraph for changing actors. The advantage of R-HGCN becomes increasingly evident as the proportion of actor masking intensifies.

Visualization Results. T-SNE [55] is applied to show the embedding distribution in Fig. 8 through finetune evaluation on CAD. All

Table 8
Comparison results with other SOTA methods on NTU-Mutual and PKU-Mutual.

Method	Params	FLOPs	NTU60-Mutual				NTU120-Mutual				PKU-Mutual	
			xsub		xview		xsub		xset			
			Linear	Finetune	Linear	Finetune	Linear	Finetune	Linear	Finetune	Linear	Finetune
AimCLR (AAAI' 22) [21]	0.84M	0.57G	57.33	84.10	76.12	89.32	75.56	82.62	72.40	82.45	81.36	96.12
PSTL (AAAI' 23) [22]	0.84M	0.57G	56.79	86.05	56.39	91.74	42.21	83.88	51.24	82.65	80.16	97.03
AHNet-Small (Ours)	**0.32M**	**0.17G**	75.81	87.13	73.04	91.01	70.06	81.01	71.94	82.09	82.81	96.88
AHNet-Middle (Ours)	0.72M	0.47G	80.29	88.05	**81.87**	92.27	74.49	84.89	75.69	83.81	87.23	97.27
AHNet-Large (Ours)	3.83M	2.7G	**80.33**	**90.85**	81.64	**93.38**	**79.09**	**86.43**	**79.57**	**86.64**	**87.66**	**98.13**

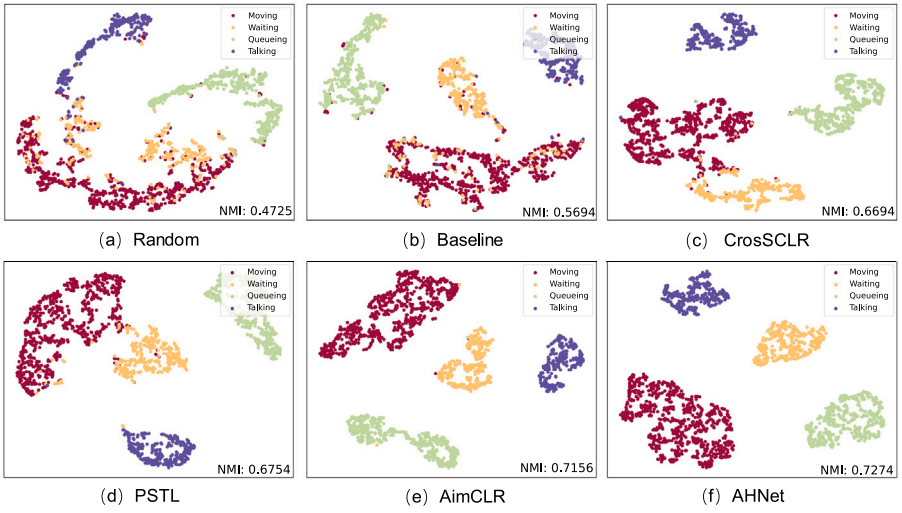


Fig. 8. The T-SNE visualization of embeddings on CAD through finetune evaluation. Larger Normalized Mutual Information (NMI) means more discriminative distribution.

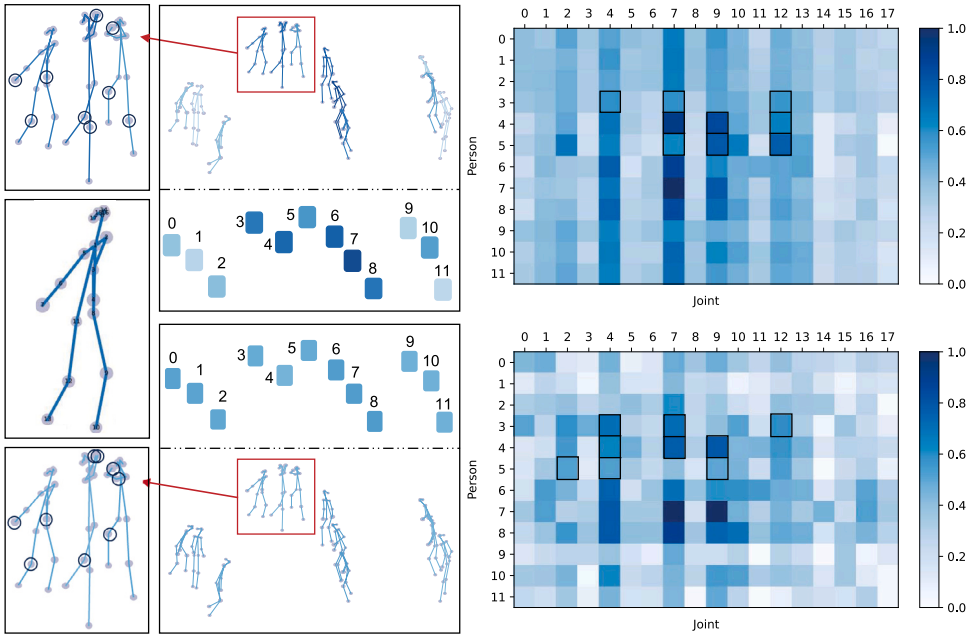


Fig. 9. Qualitative results of AHNet w or w/wo R-HGCN on VD.

self-supervised features cluster the embeddings of the same class closer than random initialization, with compact intra-class distributions and more distinguishable inter-class distributions.

Fig. 9 visualizes the skeleton maps for different layers of the same frame of a sample (from bottom to top: the GCN backbone output and the R-HGCN output, as shown in Fig. 2). The skeleton is drawn based on the physical connections of the human body. Each circle represents a joint, and its size indicates the joint feature values. Rectangles represent actors, and the depth of their color reflects the size of their values. The figure captures the volleyball "spiking" game, where the middle blocker (actors 4, 7) and right setter (actors 5, 6) are more critical for discrimination. It is evident that their values are higher in R-HGCN

compared to other actors, but the output of the GCN backbone treats all actors with nearly the same contribution. Notably, both output features can highlight the joints (i.e. the knees and hands of the actors) crucial for GAR. The above results indicate that the skeleton map of the GCN backbone emphasizes the connections between joints but neglects the importance of each actor for GAR. In contrast, R-HGCN not only explores spatial–temporal dependencies between skeleton joints, but also effectively models high-order relations of actors.

6. Conclusion

We employ the coarse-to-fine Augmented Hypergraph Network (AHNet) for self-supervised skeleton-based group activity recognition. Specifically, multi-person augmentation strategies are employed to effectively enrich the spatial–temporal dynamics and personal interactions for the encoded representations. Additionally, the knowledge-masked hypergraph enables the adaptability of the model to the situations of occlusion and people changes. Finally, the key actor matching mechanism with coarse-to-fine contrast is introduced to select the actors critical for the recognition. Extensive evaluations validate the efficacy of our approach.

Limitations. There is still room of improvement to adaptively construct hyperedges instead of manually set.

CRediT authorship contribution statement

Guoquan Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Mengyuan Liu:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources. **Hong Liu:** Writing – review & editing, Writing – original draft, Validation, Funding acquisition, Resources, Software, Supervision. **Peini Guo:** Writing – review & editing, Software, Resources, Investigation, Funding acquisition, Data curation, Conceptualization. **Ti Wang:** Writing – review & editing, Writing – original draft, Software, Project administration, Investigation, Conceptualization. **Jingwen Guo:** Software, Investigation, Writing – original draft, Writing – review & editing. **Ruijia Fan:** Conceptualization, Resources, Validation, Writing – review & editing.

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

Data will be made available on request.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62203476), National Natural Science Foundation of China (No. 62073004), and Natural Science Foundation of Shenzhen (No. JCYJ20230807120801002).

References

[1] Z. Du, X. Wang, Q. Wang, Self-supervised global spatio-temporal interaction pre-training for group activity recognition, TCSVT (2023) 5076–5088.
[2] W. Zhou, L. Kong, Y. Han, J. Qin, Z. Sun, Contextualized relation predictive model for self-supervised group activity representation learning, TMM (2023) 1–14.
[3] K. Gedamu, Y. Ji, L. Gao, Y. Yang, H.T. Shen, Relation-mining self-attention network for skeleton-based human action recognition, PR (2023) 109455.
[4] M.S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, G. Mori, A hierarchical deep temporal model for group activity recognition, in: CVPR, 2016, pp. 1971–1980.
[5] T. Lan, L. Sigal, G. Mori, Social roles in hierarchical models for human activity recognition, in: CVPR, 2012, pp. 1354–1361.
[6] W. Choi, S. Savarese, Understanding collective activities of people from videos, TPAMI (2013) 1242–1257.
[7] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, L. Van Gool, StagNet: An attentive semantic RNN for group activity and individual action recognition, TCSVT (2019) 549–565.
[8] W. Choi, K. Shahid, S. Savarese, What are they doing?: Collective activity classification using spatio-temporal relationship among people, in: ICCVW, 2009, pp. 1282–1289.
[9] J. Wu, L. Wang, L. Wang, J. Guo, G. Wu, Learning actor relation graphs for group activity recognition, in: CVPR, 2019, pp. 9964–9974.
[10] K. Gavrilyuk, R. Sanford, M. Javan, C.G. Snoek, Actor-transformers for group activity recognition, in: CVPR, 2020, pp. 839–848.
[11] Y. Li, Y. Liu, R. Yu, H. Zong, W. Xie, Dual attention based spatial-temporal inference network for volleyball group activity recognition, MTA (2023) 15515–15533.
[12] R. Yan, L. Xie, J. Tang, X. Shu, Q. Tian, HiGCIN: Hierarchical graph-based cross inference network for group activity recognition, TPAMI (2020) 6955–6968.
[13] H. Yuan, D. Ni, Learning visual context for group activity recognition, in: AAAI, 2021, pp. 3261–3269.
[14] X. Shu, J. Yang, R. Yan, Y. Song, Expansion-squeeze-excitation fusion network for elderly activity recognition, TCSVT (2022) 5281–5292.
[15] W. Li, H. Liu, H. Tang, P. Wang, Multi-hypothesis representation learning for transformer-based 3D human pose estimation, PR (2023) 109631.
[16] G. Wang, H. Liu, T. Guo, J. Guo, T. Wang, Y. Li, Self-supervised 3D skeleton representation learning with active sampling and adaptive relabeling for action recognition, in: ICIP, 2023, pp. 56–60.
[17] J. Zhang, Y. Jia, W. Xie, Z. Tu, Zoom transformer for skeleton-based group activity recognition, TCSVT (2022) 8646–8659.
[18] X. Zhai, Z. Hu, D. Yang, L. Zhou, J. Liu, Spatial temporal network for image and skeleton based group activity recognition, in: ACCV, 2022, pp. 20–38.
[19] H. Zhou, A. Kadav, A. Shamsian, S. Geng, F. Lai, L. Zhao, T. Liu, M. Kapadia, H.P. Graf, Composer: Compositional reasoning of group activity in videos with keypoint-only modality, in: ECCV, 2022, pp. 249–266.
[20] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, W. Zhang, 3D human action representation learning via cross-view consistency pursuit, in: CVPR, 2021, pp. 4741–4750.
[21] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, R. Ding, Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition, in: AAAI, 2022, pp. 762–770.
[22] Y. Zhou, H. Duan, A. Rao, B. Su, J. Wang, Self-supervised action representation learning from partial spatio-temporal skeleton sequences, in: AAAI, 2023, pp. 3825–3833.
[23] Y. Hua, W. Wu, C. Zheng, A. Lu, M. Liu, C. Chen, S. Wu, Part aware contrastive learning for self-supervised action recognition, in: IJCAI, 2023, pp. 855–863.
[24] H. Rao, S. Xu, X. Hu, J. Cheng, B. Hu, Augmented skeleton-based contrastive action learning with momentum LSTM for unsupervised action recognition, Inform. Sci. (2021) 90–109.
[25] L. Wu, C. Zhang, Y. Zou, SpatioTemporal focus for skeleton-based action recognition, PR (2023) 109231.
[26] W. Li, W. Xie, Z. Tu, W. Wang, L. Jin, Multi-hyperedge hypergraph for group activity recognition, in: IJCNN, 2022, pp. 1–7.
[27] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: ICML, 2021, pp. 12310–12320.
[28] S. Li, Q. Cao, L. Liu, K. Yang, S. Liu, J. Hou, S. Yi, Groupformer: Group activity recognition with clustered spatial-temporal transformer, in: CVPR, 2021, pp. 13668–13677.
[29] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI, 2018, pp. 7444–7452.
[30] H. Zhou, A. Kadav, A. Shamsian, S. Geng, F. Lai, L. Zhao, T. Liu, M. Kapadia, H.P. Graf, Composer: Compositional reasoning of group activity in videos with keypoint-only modality, in: ECCV, 2022, pp. 249–266.
[31] M. Dai, Z. Sun, T. Wang, J. Feng, K. Jia, Global spatio-temporal synergistic topology learning for skeleton-based action recognition, PR (2023) 109540.
[32] Y. Feng, H. You, Z. Zhang, R. Ji, Y. Gao, Hypergraph neural networks, in: AAAI, 2019, pp. 3558–3565.
[33] J. Jiang, Y. Wei, Y. Feng, J. Cao, Y. Gao, Dynamic hypergraph neural networks, in: IJCAI, 2019, pp. 2635–2641.
[34] X. Xu, Q. Zou, X. Lin, Adaptive hypergraph neural network for multi-person pose estimation, in: AAAI, 2022, pp. 2955–2963.
[35] X. Hao, J. Li, Y. Guo, T. Jiang, M. Yu, Hypergraph neural network for skeleton-based action recognition, TIP (2021) 2263–2275.
[36] C. Bian, W. Feng, S. Wang, Self-supervised representation learning for skeleton-based group activity recognition, in: ACM MM, 2022, pp. 5990–5998.
[37] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: CVPR, 2020, pp. 9729–9738.
[38] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: ICML, 2020, pp. 1597–1607.
[39] H. Qiu, B. Hou, Multi-grained clip focus for skeleton-based action recognition, PR (2023) 110188.

[40] M. Perez, J. Liu, A.C. Kot, Skeleton-based relational reasoning for group activity analysis, PR (2022) 108360.
[41] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: CVPR, 2019, pp. 5693–5703.
[42] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in: CVPR, 2020, pp. 183–192.
[43] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: CVPR, 2019, pp. 12026–12035.
[44] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: ICLR, 2018.
[45] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: CVPR, 2019, pp. 6023–6032.
[46] J. Chen, Z. Yang, D. Yang, MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification, in: ACL, 2020, pp. 2147–2157.
[47] J. Chen, Z. Jin, Q. Wang, H. Meng, Self-supervised 3D behavior representation learning based on homotopic hyperbolic embedding, TIP (2023) 6061–6074.
[48] O.-E. Ganea, G. Bécigneul, T. Hofmann, Hyperbolic neural networks, NeurIPS (2019) 5345–5355.
[49] E. Jang, S. Gu, B. Poole, Categorical reparameterization with gumbel-softmax, in: ICLR, 2016.
[50] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, in: ICCV, 2016, pp. 1010–1019.
[51] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding, TPAMI (2019) 2684–2701.
[52] J. Liu, S. Song, C. Liu, Y. Li, Y. Hu, A benchmark dataset and comparison study for multi-modal human action analytics, ACM T Multim. Comput. (2020) 1–24.
[53] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: ECCV, 2014, pp. 740–755.
[54] F. Zappardino, T. Uricchio, L. Seidenari, A. Del Bimbo, Learning group activities from skeletons without individual action labels, in: ICPR, 2021, pp. 10412–10417.
[55] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. (2008) 2579–2605.



Guoquan Wang received the B.S. degree from Hefei University of Technology (HFUT), China. He is a research graduate student studying at Peking University (PKU), China. His research interest lies in human action recognition and contrastive learning.



Prof. Mengyuan Liu received the Ph.D. degree from the School of EECS, Peking University (PKU), Shenzhen, China, in 2017, under the supervision of Prof. Hong Liu. He currently serves as an Assistant Professor at the School of EECS, Peking University, Shenzhen, China. He has authored or coauthored in T-IP, T-CSVT, T-MM, PR, CVPR, ECCV, AAAI, and IJCAI.



Prof. Hong Liu received the Ph.D. degree in Mechanical Electronics and Automation in 1996 and serves as a Full Professor in the School of EE&CS, Peking University (PKU), China. His research interests include computer vision, pattern recognition, and robot motion planning. He is also a reviewer for many international journals such as PR, T-PAMI, T-IP.



Peini Guo received her bachelor's degree from Shanghai University and is currently a graduate student at the School of EECS, Peking University, Shenzhen, China, under the supervision of Prof. Hong Liu. Her research interests include computer vision, human analysis, and person re-identification.



Ti Wang is a master student in the School of Electronic and Computer Engineering, Peking University (PKU), China, under the supervision of Prof. H. Liu. He received the B.E. degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology in 2021. His current research mainly focuses on 3D human pose and shape estimation.



Jingwen Guo received the B.S. degree in Computer Science and Technology from Soochow University in 2021. She is currently a graduate student in the School of Electronics and Computer Engineering, Peking University (PKU), China, under the supervision of Prof. H. Liu. Her research interests include computer vision and machine learning.



Ruijia Fan received her bachelor's degree from Hunan University and is currently a graduate student at the School of EECS, Peking University, Shenzhen, China, under the supervision of Prof. Hong Liu. Her research interests include human body analysis and multimodal learning.