Pattern Recognition 167 (2025) 111710



Contents lists available at ScienceDirect

Pattern Recognition

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





# Frequency-Aware Self-Supervised Group Activity Recognition with skeleton sequences

Guoquan Wang, Mengyuan Liu *, Hong Liu, Jinyan Zhang, Peini Guo, Ruijia Fan, Siyu Chen

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

ARTICLE INFO

*Keywords:*
Self-supervised learning
Group activity recognition
Frequency-aware strategies
Amplitude-based loss

ABSTRACT

Self-supervised, skeleton-based techniques have recently demonstrated great potential for group activity recognition via contrastive learning. However, these methods have difficulty accommodating the dynamic and complex nature of spatio-temporal data, weakening the ability to conduct effective modeling and extract crucial features. To this end, we propose a novel **F**requency-**A**ware **G**roup **A**ctivity **R**ecognition (FAGAR) network, which offers a comprehensive solution by addressing three key subproblems. First, the challenge of extracting discriminative features is further exacerbated by pose estimation algorithms' limitations under random spatio-temporal data augmentation. To mitigate this, a frequency domain passing augmentation method that emphasizes individual collaborative changes is introduced, effectively filtering out noise interference. Second, the fixed connections in traditional relation modeling networks fail to adapt to dynamic scene changes. To address this, we design an adaptive frequency domain compression network, which dynamically adjusts to scene variations. Third, the temporal modeling process often leads to a loss of focus on key features, reducing the model's ability to assess individual contributions within a group. To resolve this, we propose an amplitude-aware loss function that guides the network in learning the relative importance of individuals, ensuring it maintains the correct learning direction. Our FAGAR achieves state-of-the-art performance on several datasets for self-supervised skeleton-based group activity recognition. Code is available at https://github.com/WGQ109/FAGAR.

## 1. Introduction

Group Activity Recognition (GAR) aims to analyze the behavior patterns and interrelationships of multiple individuals within a team [1,2]. It has crucial value in numerous fields such as public safety [3,4], intelligent transportation [5], and sports analysis [6].

In the early stage, GAR relied on manual rules to capture the relationships among actors. The researchers manually designed rules to describe behavior patterns [7,8]. For example, Liu et al. [9] judged the types of group behavior based on geometric features such as the crowd density threshold and the consistency of individual movement directions. However, the limitations of these methods are quite obvious. The manually designed features are too simple to cope with complex and changeable real-world scenarios.

Due to the progress of pose estimation algorithms and motion cameras [10,11], skeleton data is widely used in group behavior analysis. It emphasizes the position of human joint points, decreasing data redundancy. Nevertheless, the reliance on manual annotation of joint point positions for each frame presents a significant obstacle. This process demands a substantial investment of time and effort, which not only slows down the overall analysis but also increases the potential for human error.

Consequently, self-supervised learning methods are of particular importance for skeleton-based GAR [6,14,15]. The contrastive learning paradigm has been proven to be the most effective among these methods [16,17]. However, previous contrastive methods shown in Fig. 1(a) always have the following three drawbacks: (1) **Previous methods carried out randomized data augmentation on skeleton data.** This included randomly cropping and shuffling skeleton sequences in the temporal dimension [12], shearing skeleton sequences with random directional perturbations [18], and using the DMix strategy of randomly mixing skeleton sequences with different data augmentations [13]. However, in a volleyball spiking scenario, only the movements of the key points of the hands and legs of the spiker and the defender at the critical moment are crucial. Randomized data augmentation methods

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

https://doi.org/10.1016/j.patcog.2025.111710
Received 13 January 2025; Received in revised form 11 April 2025; Accepted 14 April 2025
Available online 21 April 2025
0031-3203/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

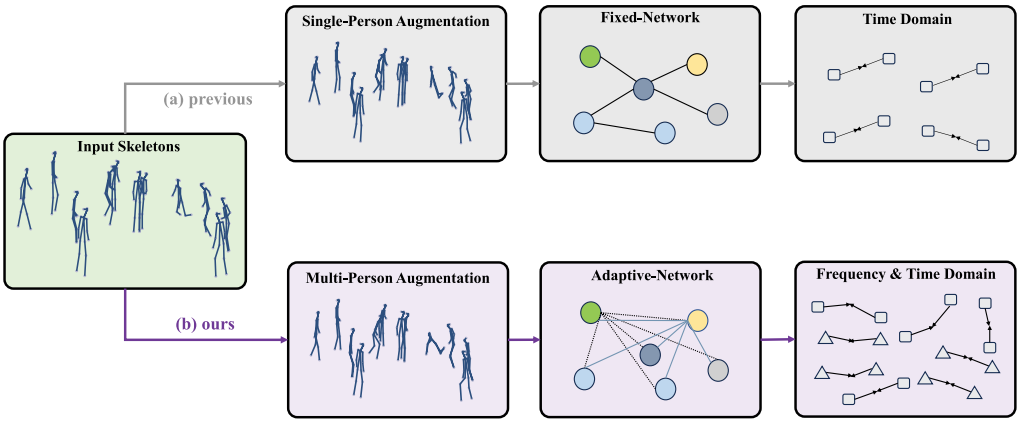
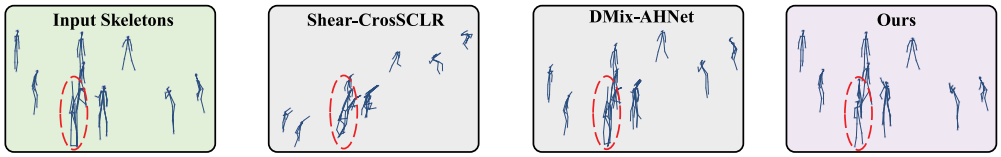



**Fig. 1.** (a) The general methods involve applying single-person augmentations, followed by a fixed network to generate representations in the time domain. (b) Our FAGAR utilizes multi-person augmentations and an adaptive network to produce representations in both the frequency and time domains, which is capable of effectively adapting to the dynamic and complex nature of spatio-temporal data.




(a) Origin sequence (b) Single-person augmented sequence (c) Multi-person augmented sequence (d) Frequency-aware augmented sequence

**Fig. 2.** (a) The original data with possible errors (red circle) in pose estimation algorithms. (b) A time-domain single-person data augmentation method *Shear* proposed in CrosSCLR [12] fails to deal with errors. (c) A time-domain multi-person data augmentation method $FMix$ proposed in AHNet [13] fails to deal with errors. (d) Our frequency-domain method calculates the motion intensity among actors and uses it as the weight basis for filtering, which effectively resolves the error estimation.

cannot locate those key features of group skeletons. In addition, the input group skeleton data is obtained through a pose estimation algorithm. When there are obvious estimation errors in the pose estimation algorithm as shown in Fig. 2, randomized data augmentation cannot solve this problem. This also poses challenges to the subsequent feature learning process. (2) **Previous approaches relied on manually constructed networks, employing static relational modeling that failed to capture evolving actor dynamics** [19]. AHNet [13] have tried to use a fixed hypergraph network constructed based on the positions of individuals and offensive strategies when modeling the relationships between actors. However, this fixed network structure makes it difficult to automatically learn new position and strategy patterns from the individuals, thus limiting its ability to extract features. Due to the rigidity of the network structure, the model performs poorly when facing new tactical changes and cannot flexibly adapt to different game scenarios. (3) **Previous methods modeled contrastive learning of skeleton data from the temporal domain perspective to learn data augmentation** [15,18,20]. However, in multi-person scenarios, overlapping actions on the time axis (e.g., players' limbs intertwining during a pick-and-roll) modeling make it challenging to effectively distinguish individual action patterns.

To this end, a **F**requency-**A**ware **G**roup **A**ctivity **R**ecognition (FAGAR) network shown in Fig. 1(b) is designed to better learn discriminative representations. Specifically, the Frequency Domain Passing (FDP) data augmentation method calculates the motion intensity between participants and uses it as the weight basis for frequency domain filtering. It enables the model to capture the collaborative changes between individuals, solving the problem that randomized data augmentation cannot understand the interaction relationships among individuals. Meanwhile, the Frequency Compression Network (FCN) designs a learnable cosine window function, enabling it to conduct adaptive learning in the complex domain. Finally, we extract the spectral amplitudes of key individuals to complement temporal domain contrastive learning for better capture of individual action characteristics.

Extensive experiments conducted on the Volleyball [7], Collective Activity [8], NTU-60 [21], NTU-120 [22], and PKU-MMD [23] datasets show that our FAGAR achieves state-of-the-art performance. The main contributions are summarized as follows:

- A network named FAGAR is proposed to perform self-supervised learning on group skeleton data from the frequency domain perspective, effectively recognizing activities in dynamic and complex scenarios.
- In terms of relation modeling, we introduce a frequency domain passing data augmentation method to capture the collaborative changes among individuals in the group.
- For key actor capture, we propose the amplitude-based loss to guide the frequency compression network to adaptively learn the importance of individuals to group behaviors in the frequency domain.
- Extensive experiments on the five datasets show that our FAGAR has a lower computational cost, and its performance is better than other state-of-the-art self-supervised methods.

## 2. Related work

### 2.1. Group activity recognition

Group Activity Recognition (GAR) refers to analyzing the behavior patterns exhibited by a group composed of multiple individuals. Early GAR methods were mainly based on RGB data [24,25]. For example, SBGAR [26] first generates video frame captions using LSTM and then predicts the activity category based on the captions. Subsequently, ARG [27] constructs a flexible and efficient actor relation graph by using two different graph convolutional networks, namely spatial local and temporal random ones. This method has improved the ability to model relationships among individuals within the group to a certain extent, but it is still limited by the insufficient expression of

individual detailed features in RGB data. With the development of technology, skeleton-based methods have gradually emerged [28–30]. Zoom Transformer [31] captures the details of interaction relations by simultaneously modeling the spatio-temporal graph and the attention relationship graph among multiple people. MP-GCN [32] takes the trajectories of interacting objects as an auxiliary modality, further enhancing the effect of skeleton-based GAR. However, all of the above methods are fully-supervised learning methods, which rely on a large amount of manually labeled data. In real-world applications, it is difficult and costly to obtain such data. Therefore, we propose to explore self-supervised methods which extract group behavior features from unlabeled data.

### 2.2. Skeleton-based self-supervised learning

Skeleton-based self-supervised GAR learning aims to mine the intrinsic structure in unlabeled data [33]. It mainly includes generative methods and contrastive methods. In generative methods, AAGCM [14] cleverly makes the model focus on the spatio-temporal positions among different actors by designing a proxy task of discriminating the group occupancy order. However, contrastive methods can focus more precisely on key feature differences in the data in the skeleton field, which has gradually gained more favor [12,34]. GL-Transformer++ [1] combines multiple semantic granularities according to the multi-level characteristics of human actions, enabling the model to comprehensively learn the features of different degrees of actions. Con-RPM [6] conducts contrastive learning between individuals and groups, striving to analyze the unequal contribution relationships of key individuals to the overall behavior. AHNet [13] designs a knowledge-guided hypergraph masking network, effectively dealing with the situation of changes in the number of actors and enhancing the universality and flexibility of the model. Nevertheless, all the above methods start from the time-domain perspective, which are easily interfered by noise, making the trajectories of joint positions chaotic. Therefore, we consider exploring skeleton-based learning from the frequency-domain perspective to further improve the reliability of GAR.

### 2.3. Frequency domain learning

The Fourier transform decomposes a complex signal according to its frequency components. Currently, it is widely used in fields such as pose estimation [35], low-light action recognition [36], and object detection [37]. For example, PoseFormerV2 [35] utilizes the compact representation in the frequency domain to expand the receptive field. This enables it to achieve good results in the 3D pose estimation work. FAFA [37] trains a frequency-aware flow-based pose estimator on synthetic data for underwater object detection. FRAGRANT [36] prioritizes informative object regions to accurately characterize action and motion in low-light conditions, while effectively suppressing noise caused by poor illumination. However, there are few works that attempt to conduct self-supervised analysis of group skeletons within the frequency domain. Since the time-domain skeleton data is easily interfered by noise, resulting in measurement errors of joint positions [10]. Therefore, we design the FAGAR network to overcome the drawbacks of the time domain and improve the performance of self-supervised analysis of group skeletons.

## 3. Preliminaries

**Fourier Analysis.** Given the group skeleton sequence $x$, we could execute Discrete Fourier Transform (DFT) denoted as $\mathcal{F}(\cdot)$ to transform $x$ to the frequency domain:

$$X[k] = \mathcal{F}(x[i]) = \sum_{i=0}^{N-1} x[i] e^{-j(2\pi/N)ki} \tag{1}$$

where $X[k]$ is the spectrum of $x[i]$ at the frequency $2\pi k/N$. $\mathcal{F}(\cdot)$ is DFT, and $j$ is the imaginary unit. Due to the conjugate symmetric property in the frequency domain [35], only half of the spectrum is used. Thus, $\mathcal{F} : \mathbb{R}^N \to \mathbb{C}^L$ and $L = \lceil N/2 \rceil + 1$. In this paper, we apply the Fast Fourier Transform (FFT) [37] to efficiently compute the DFT, reducing time complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$. The Inverse Fast Fourier transform (IFFT) works similarly.

**Barlow Twins**. We propose to use the Barlow Twins [17] framework for self-supervised GAR, which includes the following steps: Firstly, two augmented views, namely $x_1$ and $x_2$, are separately generated respectively from the skeleton sequence $x$. Subsequently, these two views are fed into the network to obtain the corresponding representations $z_1$ and $z_2$. Finally, the cross-correlation matrix $C$ of two views is utilized in the following manner:

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

where the element $C_{ij}$ of the cross-correlation matrix $C$ is calculated as the ratio. This ratio is obtained by dividing the sum of the products of $z^b_{1,i}$ and $z^b_{2,i}$ over all batch samples $b$. In symbolic form, it can be written as:

$$\mathcal{L} = \sum_i \left(1 - C_{ii}\right)^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2, \tag{3}$$

where the first term aims to drive the diagonal elements of the cross-correlation matrix $C$ towards the value of 1. This has the effect of ensuring that the representations remain invariant with respect to the augmentations applied. The second term functions to minimize the redundancy between the embeddings. It achieves this by pushing the off-diagonal elements of the cross-correlation matrix $C$ towards 0. The hyperparameter $\lambda$ serves to regulate the balance or trade-off between these two terms of the loss function.

## 4. FAGAR

### 4.1. Overview

As illustrated in Fig. 3, we propose a frequency-aware network employed for skeleton-based self-supervised group behavior recognition. Firstly, the input sequence is passed through the Frequency Domain Passing (FDP) data augmentation method in (a) to obtain different views for the input model. Then, the generated views and the original input sequence are both sent to the backbone, where it deeply analyzes the original sequence and its views to mine valuable feature information. After feature extraction, the model further activates the Frequency Compression Network (FCN), with the aim of adaptively identifying the potential key persons in the group sequence. Once the key persons are determined, the network enters the contrastive learning stage. Besides performing contrastive learning between features in the temporal domain, the model also carries out the same operation in the amplitude domain of the frequency. Moreover, this contrastive learning covers the individual level and the group level, fully mining the behavior information of key persons from different aspects.

### 4.2. Frequency domain passing augmentation

Most of the self-supervised GAR methods have been limited to data augmentation in the time domain [13,14]. Differently, we propose a Frequency Domain Passing (FDP) data augmentation method, which can effectively break through the limitations of time-domain methods mentioned in Section 1.

As shown in Fig. 4, velocity attention scores are first calculated for input skeleton sequence data $x$. Concretely, the inter-frame velocity computation quantifies motion dynamics through skeletal coordinate differences between consecutive frames. For each actor $n$ at time step

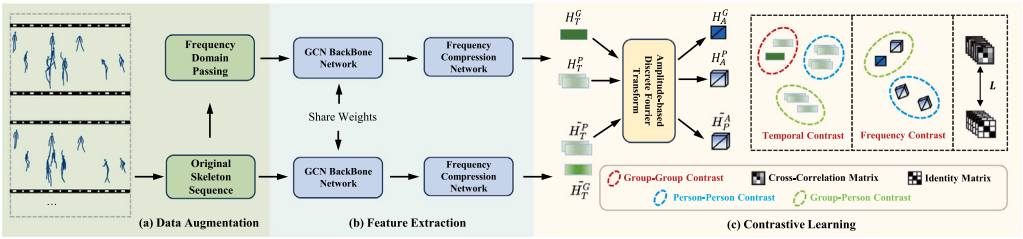



**Fig. 3.** Overview of the proposed FAGAR. (a) In the data augmentation stage, the original skeleton sequence and the sequence generated by the frequency conduction augmentation strategy are input into two different branches. (b) In the feature extraction stage, the two branches share the GCN skeleton network but use separate frequency domain compression networks to learn relationships between individuals. (c) In the contrastive learning stage, features from both branches are transformed from the time domain to the frequency domain, and multi-granularity contrastive learning is performed in both the time and frequency domain feature spaces.

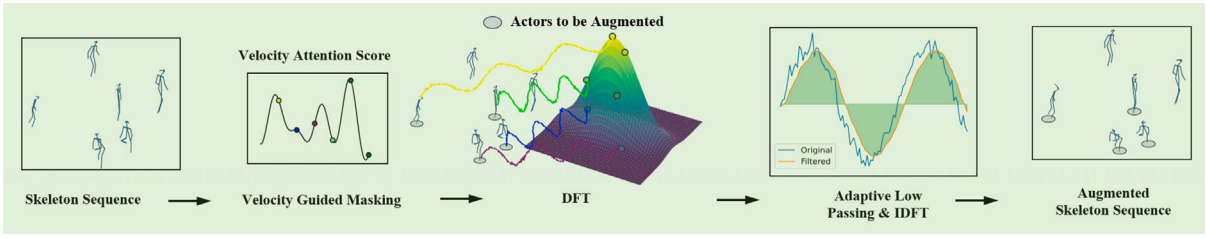



**Fig. 4.** Overview of the Frequency Domain Passing Augmentation. The frequency domain passing augmentation includes steps like velocity guided masking, DFT and adaptive low passing to produce the augmented skeleton sequence. It can effectively relieve the error in pose estimation.

$t$, the velocity is calculated as: $v_{t;n} = x_{t+1;n} - x_{t;n}, \quad t = 1, 2, \ldots, T-1$, where $x_{t;n}$ denotes the skeleton coordinates and $T$ represents the total frame count. This step captures the temporal changes in the skeleton sequence. Next, the motion energy is computed as the squared velocity to emphasize larger motion changes, which are often more significant in capturing meaningful motion patterns: $e_{t;n} = v_{t;n}^2$. Temporal attention scores are obtained by averaging the motion energy across actors, joints, and spatial dimensions. These scores highlight frames with significant motion changes, which are crucial for identifying key frames that represent the most informative parts of the sequence:

$$m_t = \frac{1}{N \cdot V \cdot M} \sum_{n=1}^{N} \sum_{v=1}^{V} \sum_{m=1}^{M} e_{t;n,v,m}, \tag{4}$$

where $N$ is the batch size, $V$ is the number of joints, and $M$ is the number of persons. The temporal attention scores are normalized using the sigmoid function to ensure that the attention scores are scaled to a range of [0, 1], making them suitable for further processing: $\hat{m}_t = \sigma(m_t)$, where $\sigma(\cdot)$ denotes the sigmoid function. The skeleton sequence is transformed into the frequency domain using the Fast Fourier Transform (FFT). Transforming the data into the frequency domain allows for the analysis and manipulation of its spectral properties, which is particularly useful for noise reduction and feature enhancement: $X = \mathcal{F}(x)$, where $\mathcal{F}$ represents the FFT operation. A frequency attention mask is applied based on the normalized temporal attention scores. This mask retains low-frequency components for frames with lower motion energy and allows higher frequencies for frames with higher motion energy, adapting the filtering process to the motion dynamics:

$$\text{mask}_{t;m} = \begin{cases} 1, & \text{if } m \le \hat{m}_t \cdot L \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $L$ is the number of frequency components. Finally, the masked frequency domain data is transformed back to the time domain using the Inverse Fast Fourier Transform (IFFT). This step reconstructs the filtered skeleton sequence in the time domain, enhancing meaningful motion patterns while reducing noise: $x_{\text{filtered}} = \mathcal{F}^{-1}(X \odot \text{mask})$, where $\mathcal{F}^{-1}$ represents the IFFT operation, and $\odot$ denotes element-wise multiplication. This pipeline adaptively modulates frequency responses according to motion saliency, preserving kinetically significant features while attenuating irrelevant components.

Fig. 5 intuitively shows the effect of this method. The first row shows the original skeleton sequence, which reflects the initial state of the group skeleton. The second row presents the sequence after frequency passing in the low-frequency range. It can be observed that after the low-frequency passing, the main movement trend of the orange key person is highlighted, and the fit with the overall movement is higher. Meanwhile, for people like those in gray and blue, their movements are not obvious in the original frames. After the low-frequency passing, they also show relevant movement states following the overall movement trend, which helps to grasp the overall movement situation at the group level.

However, the effect of the skeleton sequence after high-frequency passing augmentation in the third row is not satisfactory. Due to the influence of the pose estimation algorithm, the green person does not truly reflect the characteristics of the essence of the movement. The high-frequency signal contains a large amount of chaotic information. This makes the model learn from the wrong perspective. In contrast, the low-pass smooths the differences caused by the pose estimation algorithm, enabling the model to focus more on the main trends and key features of the group movement.

### 4.3. Frequency compression network

Previous methods [13,15] have been relying on fixed network connection structures to learn the fixed relations among individuals. To this end, an intuitive solution is to reassign weights to each actor by calculating the correlations of different individual actions on the time dimension. However, as the number of network layers increases, this correlation-based reassignment leads to excessive mixing of information [38]. From the perspective of information theory, the self-attention operation in each layer reduces the information entropy, making the output information distribution approach a more uniform (but with lower information content) distribution. This means that the uniqueness of individual actions is gradually lost during the time-domain processing, ultimately resulting in the homogenization of action features.

To overcome the homogenization problem of the self-attention mechanism, learnable cosine filters shown in Fig. 6 are used to capture key actors. It can be adaptively adjusted according to the inherent

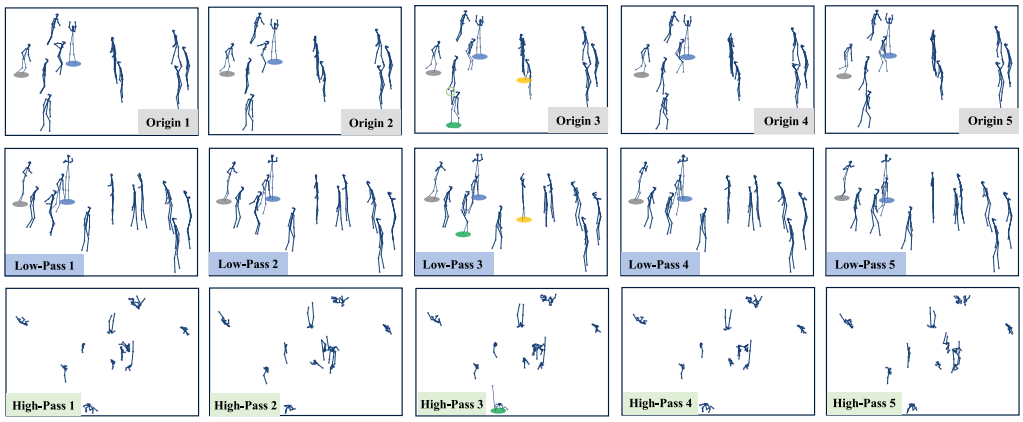



**Fig. 5.** Illustration of the frequency domain passing data augmentation. The first row shows the original skeleton sequence, the second row shows the low-pass filtered result, and the third row shows the high-pass filtered result. Orange highlights the key person's behavior, blue and gray enhance the features of individuals with less prominent motion, and green indicates that pose estimation errors are effectively corrected by low-pass filtering but amplified by high-pass filtering. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

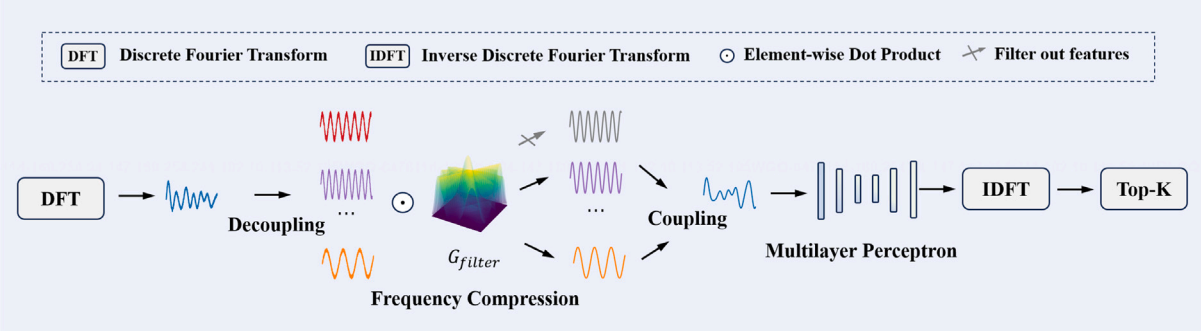



**Fig. 6.** The frequency compression network first applies a discrete Fourier transform to the input features, then decouples the resulting frequency domain features into different frequency bands. These decoupled features are filtered using a cosine filter $G_{filter}$, and the filtered features are then coupled. The MLP is used to learn features from the real and imaginary parts. Next, an inverse Fourier transform is applied to convert the frequency domain features back to the time domain. Finally, the Top-K key individuals with the highest scores are selected for contrastive learning.

frequency structure of group movements. Specifically, we introduce a specially designed filter bank $K = [k_1, k_2, \ldots, k_k]$, where $k$ represents the number of filters:

$$G_{filter} = \sum_{i=1}^{k} \frac{1}{L} |G|^2 \odot k_i cos(\frac{(2i-1)\pi}{2k}), \tag{6}$$

where $|G|^2$ denotes the power spectrum of $G$ and it reflects the energy intensity of a signal in each frequency component. $L$ denotes the length of $G$, and $\odot$ means the element-wise multiplication. As a modulation factor, $cos(\frac{(2i-1)\pi}{2k})$ performs frequency-selective modulation on different filters.

Each filter has specific frequency response characteristics, and they work together to perform multi-dimensional and multi-level screening and compression on the spectrum, so that the finally retained spectrum information can screen the important features of key actors to the greatest extent.

After the data is processed by the learnable cosine filters, $G_{filter}$ is divided into the real part and the imaginary part. They are input into the Multi-Layer Perceptron (MLP) respectively. The MLP captures the unique information in the imaginary part data, which complements the features reflected by the real part. They jointly construct a more complete description of group behavior features and get the hidden features $H$. However, not all individuals' features are of equal importance for the understanding and judgment of the overall behavior. Therefore, we retain the features corresponding to the Top-K individuals. These individuals can represent the behavior patterns of the entire group to a greater extent.

### 4.4. Amplitude-guided contrastive loss

To enable the model to focus on the invariance between data augmentations and the features corresponding to key persons, we first follow AHNet [13] and conduct contrastive learning shown in Fig. 7 at both the group–group level and the key person–person level:

$$\mathcal{L}_{GGT} = \mathcal{L}(H_T^G, \tilde{H_T^G}) = \sum_i \left(1 - C_{ii}^{gg}\right)^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^{gg\,2}, \tag{7}$$

$$\mathcal{L}_{PPT} = \mathcal{L}(H_T^P, \tilde{H_T^P}) = \sum_{p \in \mathbb{K}} (\sum_i \left(1 - C_{ii}^{pp}\right)^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^{pp\,2}), \tag{8}$$

where $C_*^{gg}$ and $C_*^{pp}$ are the cross-correlation matrices of the corresponding sample pairs across the time domain. $H_*$ and $\tilde{H}_*$ respectively represent the hidden layer features of the original sequence and the frequency-domain augmented sequence. Due to the minor time deviations during the data collection process, it can lead to indistinct differentiation of group skeleton features and the presence of superposition phenomena. To this end, the Fourier transform with time-shifting property is performed, which can obtain two components namely phase and amplitude simultaneously. The phase information reflects the starting position of the signal in time. The differences in the starting times will lead to the uncertainty of the phase information. The amplitude mainly depends on the scale and degree of change of group skeleton actions, rather than the specific starting time of the actions. This stability makes the amplitude information more suitable for contrastive learning:

$$\mathcal{L}_{PPA} = \mathcal{L}(H_A^P, \tilde{H_A^P}) = \sum_{p \in \mathbb{K}} (\sum_i \left(1 - \mathfrak{C}_{ii}^{pp}\right)^2 + \lambda \sum_i \sum_{j \neq i} \mathfrak{C}_{ij}^{pp\,2}), \tag{9}$$

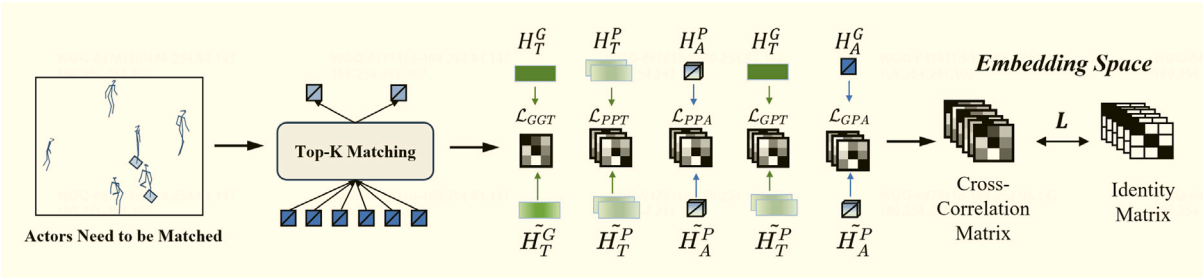



**Fig. 7.** The amplitude-guided contrast and temporal-aware contrast use Top-K matching to generate the group–person level pairs for the key actors capturing. It can help the model to focus on the key actors and their relations within the group.

where $\mathfrak{C}_{*}^{pp}$ are the cross-correlation matrices of the corresponding amplitude spectrum pairs across the frequency domain. In addition, through contrastive learning between key persons and their data augmentation variants as well as between key persons and group features, the model can enhance its perception of the diversity of key person features:

$$\mathcal{L}_{GPT} = \mathcal{L}(H_T^G, \tilde{H}_T^P) = \sum_{p\in\mathbb{K}}\Big(\sum_i \big(1 - C_{ii}^{gp}\big)^2 + \lambda \sum_i \sum_{j\neq i} C_{ij}^{gp2}\Big), \tag{10}$$

$$\mathcal{L}_{GPA} = \mathcal{L}(H_A^G, \tilde{H}_A^P) = \sum_{p\in\mathbb{K}}\Big(\sum_i \big(1 - \mathfrak{C}_{ii}^{gp}\big)^2 + \lambda \sum_i \sum_{j\neq i} \mathfrak{C}_{ij}^{gp2}\Big). \tag{11}$$

It can effectively distinguish key persons under different states and grasp the differences in the associations between key person features and group features. The final loss with the pre-defined $\beta$ coefficient is as follows:

$$\mathcal{L} = \beta_1\mathcal{L}_{GGT} + \beta_2\mathcal{L}_{PPT} + \beta_3\mathcal{L}_{PPA} + \beta_4\mathcal{L}_{GPT} + \beta_5\mathcal{L}_{GPA}. \tag{12}$$

Through multi-level dual-domain contrastive learning, the model can fully utilize the spatio-temporal frequency information in the skeleton sequence.

### 4.5. Model complexity analysis

Lightweight model design is crucial for efficiency [39]. A theoretical comparison of time complexity between our model and other methods. In the experiments, we select two typical methods for comparison: the single-person self-supervised method AimCLR [18] and the multi-person self-supervised method AHNet [13].

In AimCLR, the main time cost comes from feature extraction in the backbone network. The backbone consists of stacked GCN-TCN layers. The GCN has two steps: 2D convolution with time complexity $\mathcal{O}(\sum_{l=1}^{L} C_{l-1}C_l K_l^2 V M)$, and adjacency matrix aggregation with $\mathcal{O}(\sum_{l=1}^{L} C_l V M)$. Here, $C$ is the feature dimension, $D$ is the final layer's dimension ($C_L = D$), $V$ is the number of skeleton points, and $M$ is the number of actors. The TCN uses 1D convolution with $\mathcal{O}(\sum_{l=1}^{L} C_{l-1}C_l K_l V M)$. Our backbone uses the same architecture but only 3/10 of the layers (Our FAGAR $L$=3 vs AimCLR $L$=10), significantly reducing parameters.

For AHNet, the key difference is in the key person selection network. AHNet uses a hypergraph and Transformer Encoder. The hypergraph has convolution $\mathcal{O}(D^2KM)$ and hyperedge aggregation $\mathcal{O}(MD)$. The Transformer Encoder has $\mathcal{O}(M^2D) + \mathcal{O}(D^2M)$. The total complexity is $\mathcal{O}(D^2KM)+\mathcal{O}(MD)+\mathcal{O}(M^2D)+\mathcal{O}(D^2M)$. Our FAGAR uses a cosine filter for key person selection. It has three steps: FFT transform $\mathcal{O}(M\log M \cdot D)$, cosine filtering $\mathcal{O}(N\log M \cdot D)$ ($N$ is the number of filters), and fully connected weighting $\mathcal{O}(D^2)$. The total complexity for key person selection is $\mathcal{O}(M\log M \cdot D) + \mathcal{O}(D^2) + \mathcal{O}(N\log M \cdot D)$. Since $N$ and $M$ are much smaller than $D$, our method has much lower complexity than AHNet.

## 5. Experiments and discussions

### 5.1. Dataset

**Volleyball Dataset (VD)** [7] provides a multi-level annotation system including group activity labels, individual action annotations, poses, and position coordinates. By capturing both macro interaction patterns and micro individual behaviors in team sports, the annotation system covers semantic information from local to global levels, avoiding the one-sided understanding of behaviors that might arise from a single annotation level.

**Collective Activity Dataset (CAD)** [8] has 2481 activity segments, which is filmed using handheld devices in unconstrained environments. It contains over 40 short videos of complex real-world scenarios: dynamic lighting, random angles, moving backgrounds, and pedestrian interference. Compared to datasets collected in laboratory settings, its natural collection method fully preserves the randomness factors of the real world, effectively preventing model overfitting due to overly controlled environments and ensuring the robustness of algorithms in real-world applications.

**NTU RGB+D (NTU-60)** [21] holds 60 human activities along with the 3D coordinates of 25 joints. We carry out experiments in the cross-subject (xsub) and cross-view (xview) scenarios. It employs three cameras to capture actions from horizontal, overhead, and oblique angles, effectively supporting cross-view generalization research. The participation of 40 individuals aged between 10 and 35 enhances the diversity of action performances. By incorporating age differences and multi-angle data collection, the dataset effectively mitigates biases that could arise from a single demographic group.

**NTU RGB+D 120 (NTU-120)** [22] is an extension of NTU-60 and provides cross-subject (xsub) and cross-setup (xset) evaluation procedures. It encompasses 106 participants aged 10–57 from 15 countries, combined with 155 camera angles and 96 background environments, constructing complex scenes across cultures, ages, and lighting conditions. This dataset significantly reduces model bias caused by regional preferences or scene uniformity.

**PKU-MMD** [23] comprises 51 action classifications. It focuses on the temporal modeling of long continuous action sequences, ensuring the coherence and complexity of the collected actions. By continuously capturing natural action flows, it avoids the over-labeling of action boundaries typical of traditional segmented data, reducing temporal modeling bias due to manual segmentation.

### 5.2. Experimental settings

In the process of data pre-processing, we adopt the approach described in [24]. Specifically, the HRNet [11] that has been pre-trained on the COCO dataset [40] is utilized and adjusted as the pose estimation network to acquire group activity skeletons. The size of the mini-batch is configured to be 128.

**Self-supervised Pretext Training.** The default backbone model is built based on ST-GCN [28], and the quantity of stages is decreased

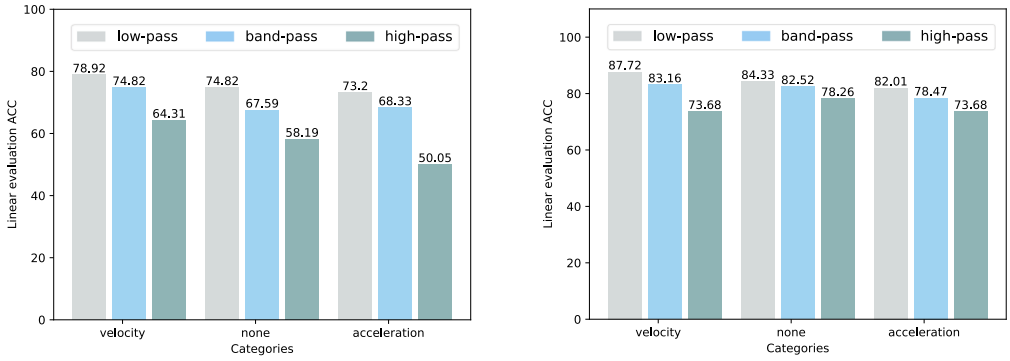



**Fig. 8.** Ablation experiments of frequency domain passing data augmentation on the VD dataset (left) and the CAD dataset (right).

**Table 1**
Ablation evaluation results through linear evaluation.

| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| FDP | ✗ | ✔ | ✗ | ✗ | ✔ | ✔ | ✗ | ✔ |
| FCN | ✗ | ✗ | ✔ | ✗ | ✔ | ✗ | ✔ | ✔ |
| AGCL | ✗ | ✗ | ✗ | ✔ | ✗ | ✔ | ✔ | ✔ |
| VD | 60.65 | 70.14 | 68.35 | 67.32 | 74.13 | 73.71 | 72.27 | **78.92** |
| CAD | 56.84 | 73.96 | 71.34 | 75.69 | 84.32 | 82.10 | 82.99 | **87.72** |

to three-tenths of the original setup. In terms of contrastive configurations, we comply with [17] to map features into 6144 dimensional embeddings. The parameter $\lambda$ within the Barlow Twins loss of each stream is assigned a value of 2e-4, and $\beta$ in the aggregate loss of each stream is set to 1. The hyperparameter $k$ utilized in FCN is set to 3. The weight decay is established as 1e-5. For the optimization process, we make use of the Adam optimizer and the CosineAnnealing scheduler for 300 epochs. The model undergoes training with a 10-epoch warmup to achieve stable contrastive learning.

**Linear Evaluation Protocol.** The models undergo verification via linear evaluation for the group activity recognition task, using classification accuracy (ACC) as the performance metric. Specifically, a linear classifier is trained while keeping the encoder fixed. The Adam optimizer, with an initial learning rate of 0.01 over 50 epochs, is utilized for the optimization process.

**Finetune Evaluation Protocol.** A linear classifier is attached to the pre-trained encoder and then train the entire model for the purpose of comparing it with fully-supervised strategies. For optimization, the Adam optimizer with an initial learning rate of 0.01 is employed.

### 5.3. Ablation study

Ablation experiments are carried out during multiple evaluations regarding the skeleton-based GAR task to confirm the efficacy of FAGAR.

**Effectiveness of Each Component.** Table 1 reveals the quantitative impact of including specific component. The Frequency Domain Passing (FDP) data augmentation method significantly enhances the performance of the model. When it is adopted, it is evidenced by a performance increase of 6.65% on VD (from 72.27% to 78.92%) and 4.73% on CAD. This is mainly because it can calculate motion intensity among actors in the frequency domain, facilitating the model to capture individual collaborative changes. Similarly, when the Frequency Compression Network (FCN) is incorporated, it results in a performance boost of 5.21% on VD and 5.62% on CAD. The key reason is its learnable cosine window function for adaptive complex domain learning. Meanwhile, Amplitude-Guided Contrastive Loss (AGCL) also proves to be essential. With its presence, it leads to an improvement of 4.79% on VD and 3.40% on CAD. This is due to its function of guiding the network to learn individual importance in group behaviors.

**Effectiveness of the Frequency Domain Passing Data Augmentation Method.** As shown in Fig. 8, we conduct ablation experiments on the adaptive selection basis of FDP and different filtering methods. “velocity" represents our method. It uses the velocity of each actor per frame to calculate scores for generating adaptive weights for low-pass, band-pass, and high-pass filters. “none" represents manually designed filter ranges. Specifically, the low-pass filter range is 0-1/3$L$, the band-pass filter range is 1/3$L$-2/3$L$, and the high-pass filter range is 2/3$L$-$L$, where $L$ is the length of the filter sequence. “acceleration" represents using the acceleration of each actor to calculate scores for generating adaptive weights for different filters. It is worth noting that no matter which adaptive strategy is chosen, the low-pass filtering performs the best on each dataset. This is because it retains the basic form of actions, making it easier for the model to learn stable feature patterns. Besides, when using low-pass filtering for data augmentation, it can be found that the mask strategy based on acceleration is not stable. Its performance on the VD and CAD datasets is respectively 1.62% and 2.32% lower than that without using any adaptive strategy (“none"). The reason is that the pose estimation algorithm can lead to sudden changes between the previous and the following frames, which causes large deviations in the calculation results of acceleration and wrongly identify the state changes of key persons.

In contrast, the velocity-based strategy achieves optimal performance by accurately tracking smooth motion transitions while naturally filtering out estimation noise. To demonstrate that incorporating frequency-domain enhancement strategy coefficients improves the robustness of our proposed method, we added zero-mean Gaussian noise to the pose estimation results on VD, making the skeleton-based group activity recognition task more challenging. Experimental results in Fig. 9 show that as the standard deviation (Sigma) of Gaussian noise increases, FAGAR exhibits a smaller performance drop while maintaining higher efficiency. Specifically, when Sigma increases from 0.5 to 1.0, SkeAttnCLR [20] performance decreases sharply, widening the gap with FAGAR from 10.92% to 17.09%. Additionally, compared to other methods, FAGAR demonstrates superior robustness, maintaining stable performance even under strong noise conditions.

**Effectiveness of the Frequency Compression Network.** In Fig. 10, the curves in different colors represent the filter selection of the Frequency Compression Network (FCN). The green curve describes a low-pass filter serving as the primary component of the FCN, while the blue and orange curves represent cosine filters that function as key components of the FCN. From the green curve, it can be seen that the cut-off frequency of the low-pass filter significantly affects performance. When using a low-pass filter with a low cut-off frequency, it removes the frequency bands that represent fast interactive actions. For example, on the VD dataset, actions requiring instant completion like “spike" and “win" are significantly affected, as seen in the confusion matrix of Fig. 11(a). Compared to the cosine filter (w/o c), the low-pass filter (c=2) performs noticeably worse in recognizing these actions. On the other hand, when the low-pass filter’s cut-off frequency is too

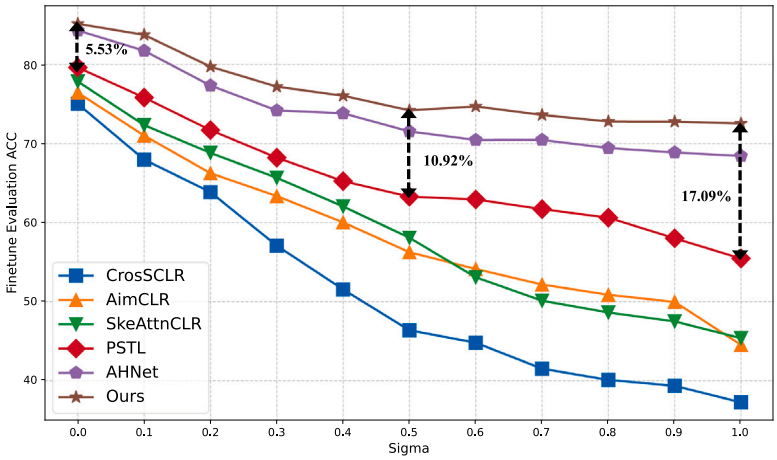
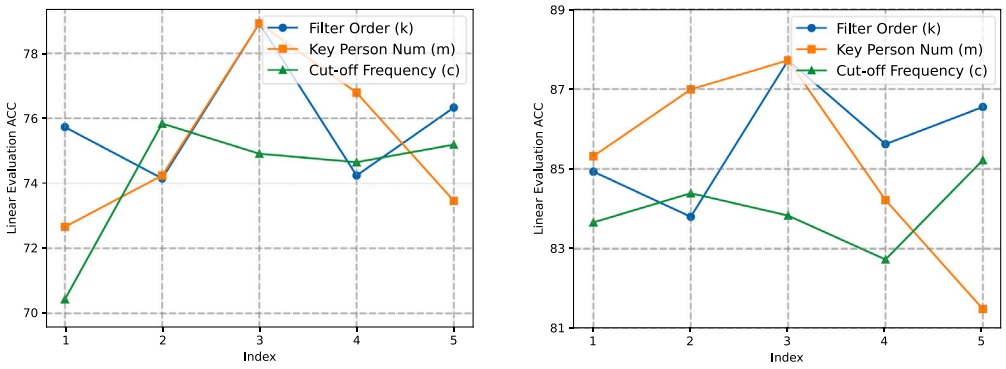



**Fig. 9.** Finetune evaluation for robustness test on the VD datasets.




(a) VD (b) CAD

**Fig. 10.** Hyper-parameter selection experiments of frequency compression network on VD and CAD datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

high, it keeps high-frequency details but also introduces noise from pose estimation into the feature space. On the CAD dataset, high-frequency details are crucial for performance since the behaviors to be distinguished are fine-grained. However, Fig. 10 shows that with a cut-off frequency of 5, the low-pass filter works best but still performs 2.50% worse than the cosine filter (k=3) (87.72% vs. 85.22%).

The low-pass filter has two main drawbacks when used as the main component of FCN. First, its hyper-parameters are highly sensitive. The optimal cut-off frequency varies greatly across datasets without a clear pattern. Second, it consistently underperforms compared to the cosine filter on multiple datasets. Given these limitations, experiments show that using an adaptive cosine filter as the main component of FCN is more suitable, delivering more stable and superior performance across different datasets. From the blue curve, it can be seen that the filter order determines the precision of frequency-domain feature extraction. When the filter order is too low (from k=3 to k=1), the feature spectrum is overly smoothed. On the CAD dataset, the confusion matrix in Fig. 11(b) shows that the misjudgment rate for similar actions, such as “waiting" and “queuing”, increases by 13%. When the filter order is too high, it creates a sharper cutoff. However, this also leads to the Gibbs phenomenon, which appears as oscillation artifacts in the joint coordinates over time. For example, a filter with an order of 5 generates high-frequency jitter in fast continuous motion trajectories, leading to a significant increase of 12% in the misjudgment rate for the “moving" and “waiting" actions compared with the order of 3. Therefore, we use a 3rd-order cosine filter as the main component of the frequency compression network, which effectively balances the retention of key motion frequency bands while suppressing over-smoothing and oscillation artifacts. Additionally, we conduct ablation experiments on the selection of the number of key individuals with the orange curve. On the VD dataset, the number of key individuals has a significant impact on model performance. When the number is too small, the model lacks confidence and tends to perform contrastive learning between non-key individuals and the group. When the number is too large, the contribution of key individuals is diluted. Through ablation experiments, we find that setting the number of key individuals m to 3 yields the best results, effectively avoiding learning bias caused by an inappropriate number.

**Effectiveness of the Amplitude-Guided Contrastive Loss.** Table 2 presents the impact of different contrastive learning methods and contrastive levels on the model. Firstly, an ablation study is conducted on the two different contrastive learning methods of amplitude and phase in the frequency domain. Owing to the fact that the distribution of phase data is more complex and irregular, the model encounters more significant difficulties when conducting similarity measurements. Therefore, the phase-based contrastive learning method fails to bring positive gains. In addition, when considering the contrastive learning of both amplitude and phase simultaneously, the model’s performance is also inferior to that of only enhancing the amplitude (ID 3 vs 9, 78.92% vs 71.78%). From the perspective of contrastive learning levels, the effect at the person–person level is often weaker than that at the group–person level. It shows that when contrast is only limited to different data augmentations of the same person makes the network unable to capture the dynamic connections of key individuals for group behavior.

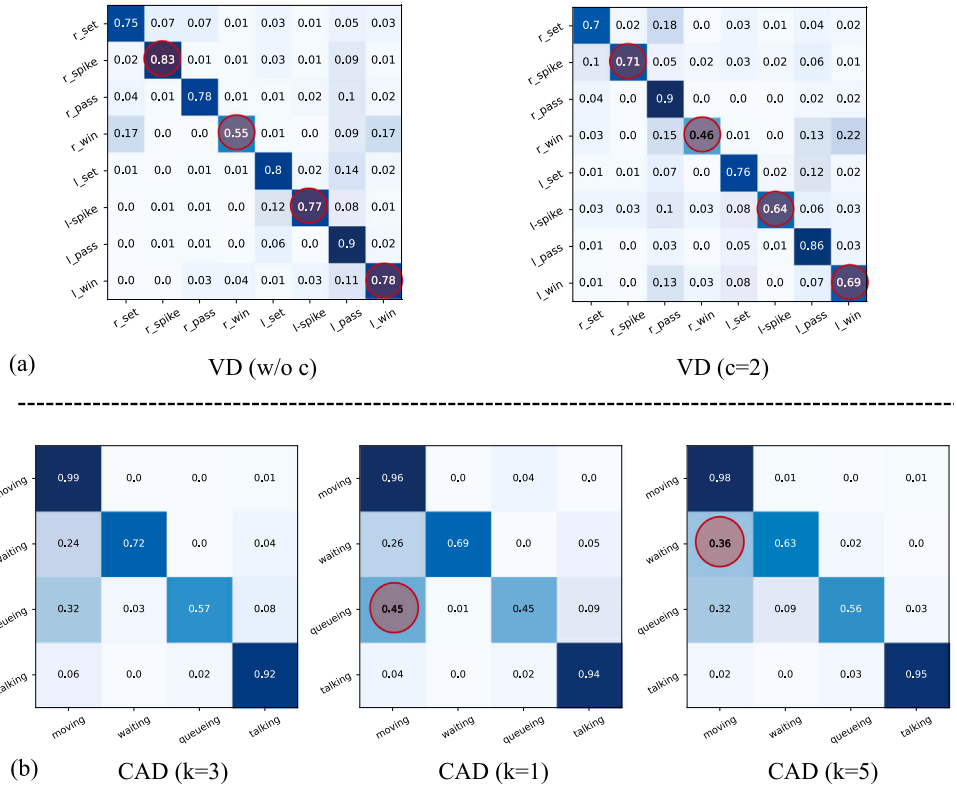



**Fig. 11.** Visualization experiments of hyper-parameter confusion matrices on VD and CAD datasets.

**Table 2**
Exploration of different type of contrast through linear evaluation.

| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Amplitude Guided | ✗ | ✔ | ✔ | ✔ | ✗ | ✗ | ✗ | ✔ | ✔ | ✔ |
| Phase Guided | ✗ | ✗ | ✗ | ✗ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| Person–Person Contrast | ✗ | ✔ | ✗ | ✔ | ✔ | ✗ | ✔ | ✔ | ✗ | ✔ |
| Group–Person Contrast | ✗ | ✗ | ✔ | ✔ | ✗ | ✔ | ✔ | ✗ | ✔ | ✔ |
| VD | 60.65 | 71.03 | 73.71 | **78.92** | 67.88 | 64.31 | 71.15 | 70.64 | 71.18 | 75.47 |
| CAD | 56.84 | 81.80 | 82.41 | **87.72** | 80.17 | 78.84 | 82.56 | 79.56 | 80.81 | 83.08 |

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

**Linear Evaluation.** To demonstrate the generality of the proposed self-supervised learning paradigm, a series of experiments are carried out with diverse backbones. Table 3 presents the linear evaluation results of comparing FAGAR with other self-supervised learning methods. Notably, self-supervised learning methods achieve a significant improvement compared with those with randomly initialized parameters. For example, our FAGAR achieves a 45.03% improvement on VD when using 2S-AGCN [29] as the backbone network compared with the method with randomly initialized parameters. It fully demonstrates the effectiveness of self-supervised learning. Besides, when using ST-GCN [28] as the backbone on VD, our FAGAR outperforms the state-of-the-art method AHNet [13] by 6.34%, demonstrating the necessity of frequency-domain learning for skeleton-based GAR.

**Finetune Evaluation.** Table 4 shows the performance comparison between FAGAR and other self-supervised learning methods under different backbones. AimCLR [18] improves the representation ability of the model by designing eight kinds of time-domain data augmentation strategies based on skeletons. On the CAD dataset, even though our method only uses the FDP data augmentation method in the frequency domain, in the experimental setting where Shift-GCN [30] is used as the backbone network, its performance is still 21.47% higher than that of AimCLR. This strongly indicates that performing data augmentation in the frequency domain helps to capture the potential spatio-temporal relationships among actors. AHNet [13] conducts contrastive learning in a hypergraph network with fixed edges, making the network sensitive to the positions of key people. FAGAR locates key people by designing an adaptive cosine filter. Moreover, when 2s-AGCN [29] is used as the backbone network on the CAD dataset, the performance of our FAGAR is 4.29% higher than that of AHNet [13]. It shows that the cosine filter has more advantages in accurately locating key people compared with the preset network structure. It can be observed that under the linear and finetune evaluation protocols, our FAGAR achieves the best results. Especially on the CAD dataset, using MS-G3D [41] as the backbone improves performance by 3.88% compared to AHNet [13] (from 85.28% to 89.16%) under finetune evaluation test.

Table 5 shows the performance comparison results between our method and other more general methods. FAGAR-Tiny represents the feature dimension reduced to 1/8 of the original based on FAGAR. Apparently, our method remains competitive when compared to the fully-supervised methods that depend on label training. For example, compared with the ISTIN [42] method that uses RGB data, our method performs better on the CAD dataset. However, on VD, there is still a certain gap compared with the current fully-supervised methods. This is because the goal of fully-supervised learning focuses precisely on classification tasks. In contrast, the goal of self-supervised learning is relatively broad. Thus, its performance in fine-grained classification

**Table 3**
Linear results compared with other self-supervised methods on VD and CAD.

| Method | VD | | | | CAD | | | |
|---|---|---|---|---|---|---|---|---|
| BackBone | ST-GCN | 2S-AGCN | Shift-GCN | MS-G3D | ST-GCN | 2S-AGCN | Shift-GCN | MS-G3D |
| Random | 25.08 | 30.29 | 28.88 | 29.55 | 40.13 | 42.78 | 39.82 | 45.65 |
| CrosSCLR (CVPR' 21) [12] | 41.97 | 53.25 | 47.76 | 54.82 | 42.61 | 55.54 | 46.64 | 69.67 |
| AimCLR (AAAI' 22) [18] | 54.64 | 56.47 | 61.26 | 40.00 | 44.18 | 52.78 | 45.61 | 62.36 |
| PSTL (AAAI' 23) [34] | 58.02 | 62.45 | 56.43 | 42.80 | 48.83 | 52.00 | 47.52 | 67.32 |
| SkeAttnCLR (IJCAI' 23) [20] | 58.49 | 60.23 | 57.89 | 54.75 | 57.39 | 60.65 | 56.88 | 68.10 |
| AHNet (PR' 24) [13] | 72.58 | 73.83 | 70.91 | 62.15 | 84.11 | 81.90 | 77.10 | 80.20 |
| **FAGAR (Ours)** | **78.92** | **75.32** | **75.83** | **67.69** | **87.72** | **88.77** | **84.30** | **83.62** |

**Table 4**
Finetune results compared with other self-supervised methods on VD and CAD.

| Method | VD | | | | CAD | | | |
|---|---|---|---|---|---|---|---|---|
| BackBone | ST-GCN | 2S-AGCN | Shift-GCN | MS-G3D | ST-GCN | 2S-AGCN | Shift-GCN | MS-G3D |
| Fully-Supervised | 64.38 | 72.17 | 73.04 | 71.58 | 67.71 | 79.15 | 72.64 | 71.44 |
| CrosSCLR (CVPR' 21) [12] | 75.02 | 70.83 | 70.38 | 69.67 | 67.32 | 65.88 | 76.73 | 79.61 |
| AimCLR (AAAI' 22) [18] | 76.44 | 74.50 | 76.29 | 67.19 | 83.01 | 83.92 | 71.37 | 70.55 |
| PSTL (AAAI' 23) [34] | 77.85 | 73.22 | 76.30 | 79.37 | 78.56 | 81.63 | 76.47 | 72.32 |
| SkeAttnCLR (IJCAI' 23) [20] | 79.62 | 79.90 | 81.06 | 70.32 | 69.92 | 82.55 | 79.62 | 74.76 |
| AHNet (PR' 24) [13] | 84.31 | 82.68 | 83.01 | 81.85 | 89.32 | 87.74 | 80.71 | 85.28 |
| **FAGAR (Ours)** | **85.15** | **83.82** | **84.39** | **83.23** | **92.14** | **92.03** | **92.84** | **89.16** |

**Table 5**
Finetune results compared with other GAR methods on VD and CAD.

| Method | Type | Params | FLOPs | VD | CAD |
|---|---|---|---|---|---|
| *Supervised:* | | | | | |
| ZOOM Transformer (TCSVT' 22) [31] | Skeleton | 2.93M | 5.6G | 71.20 | – |
| Zhai et al. (ACCV' 22) [24] | Skeleton | – | – | 83.90 | **95.00** |
| ISTIN (MTA' 23) [42] | RGB | 165.46M | 5.3G | 88.80 | 86.50 |
| ARG (CVPR' 19) [27] | RGB | 61.19M | 161.6G | 92.50 | 91.00 |
| Actor-Transformer (CVPR' 20) [43] | Skeleton | – | – | 92.30 | – |
| Actor-Transformer (CVPR' 20) [43] | RGB+Skeleton | – | – | **93.50** | 91.00 |
| *Self-Supervised:* | | | | | |
| CrosSCLR (CVPR' 21) [12] | Skeleton | 0.84M | 0.57G | 75.02 | 67.32 |
| AimCLR (AAAI' 22) [18] | Skeleton | 0.84M | 0.57G | 76.44 | 83.01 |
| PSTL (AAAI' 23) [34] | Skeleton | 0.84M | 0.57G | 79.62 | 78.56 |
| SkeAttnCLR (IJCAI' 23) [20] | Skeleton | 3.06M | 2.8G | 77.85 | 69.92 |
| Con-RPM (TMM' 23) [6] | RGB+Skeleton | – | – | 81.50 | 75.40 |
| Skeleton2vec (ARXIV' 24) [44] | Skeleton | 6.34M | 3.79G | 78.41 | 85.44 |
| AHNet-Small (PR' 24) [13] | Skeleton | 0.32M | 0.17G | 77.05 | 89.45 |
| AHNet-Middle (PR' 24) [13] | Skeleton | 0.72M | 0.47G | 80.69 | 88.76 |
| AHNet-Large (PR' 24) [13] | Skeleton | 3.83M | 2.7G | 84.31 | 89.32 |
| **FAGAR-Tiny (Ours)** | Skeleton | **0.11M** | **0.13G** | 82.10 | 90.19 |
| **FAGAR (Ours)** | Skeleton | 0.71M | 0.79G | **85.15** | **92.14** |

**Table 6**
Comparison results with SOTA methods on NTU-60, NTU-120 and PKU-MMD.

| Method | Params | FLOPs | NTU-60 | | | | NTU-120 | | | | PKU-MMD | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | xsub | | xview | | xsub | | xset | | part I | part II |
| | | | Linear | Finetune | Linear | Finetune | Linear | Finetune | Linear | Finetune | Linear | Linear |
| AimCLR (AAAI' 22) [18] | 0.84M | 0.57G | 74.30 | 83.00 | 79.70 | 89.20 | 63.40 | 76.40 | 63.40 | 76.70 | 83.40 | 38.50 |
| PSTL (AAAI' 23) [34] | 0.84M | 0.57G | 77.30 | 84.50 | 81.80 | 92.00 | 66.20 | 78.60 | 67.70 | 78.90 | 88.40 | 49.30 |
| AHNet (PR' 24) [13] | 3.83M | 2.7G | 74.56 | 84.56 | 79.49 | 90.71 | 63.08 | 76.47 | 64.11 | 77.89 | 87.75 | 37.44 |
| AimCLR++ (PR' 24) [15] | 0.84M | 0.57G | 77.20 | 84.50 | 81.50 | 91.00 | 65.50 | **78.10** | 67.80 | 79.10 | 89.60 | 40.80 |
| **FAGAR* (Ours)** | **0.84M** | **0.57G** | **80.84** | **85.48** | **85.77** | **92.43** | **71.34** | 78.03 | **72.13** | **79.44** | **90.38** | **50.42** |

tasks is not as good as that of fully-supervised methods. This is also reflected in single-person self-supervised behavior analysis [12,18,20,34]. When compared with the time-domain contrast method Con-RPM [6] that uses both RGB and skeleton data at the same time, our method has a 3.65% higher performance on VD. It shows that the performance gain brought by the amplitude-guided dual-domain contrastive learning method can outperform that resulting from introducing additional types of data.

We specifically carried out a detailed analysis on datasets with a smaller number of people and presented the relevant results in Table 6. The core of the proposed amplitude-guided contrastive learning method is to learn by comparing key individual features with group features, making it unsuitable for single-person scenarios. To ensure a fair comparison, we remove the amplitude-guided contrastive module when testing on the full NTU-60, NTU-120, and PKU-MMD datasets. This version is labeled as FAGAR* and used the same backbone network and training settings as PSTL [34] to ensure fairness when comparing with self-supervised single-person action recognition methods. As shown in Table 6, the FAGAR* model performs exceptionally well in multiple benchmarks. In the NTU-60 dataset's xsub linear evaluation, it outperforms PSTL [34] by 3.54% (80.84% vs. 77.30%). In the xview finetune test, it surpasses AimCLR [18] by 6.07% (85.77% vs. 79.70%). For NTU-120's xsub linear evaluation, it outperforms AimCLR++ [15] by 5.84% (71.34% vs. 65.50%), and in the xset linear evaluation, it is

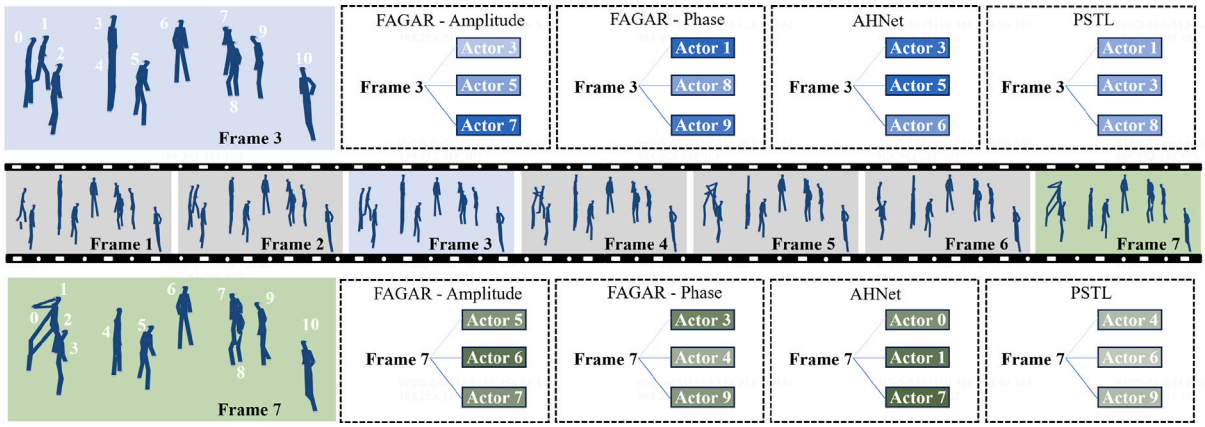
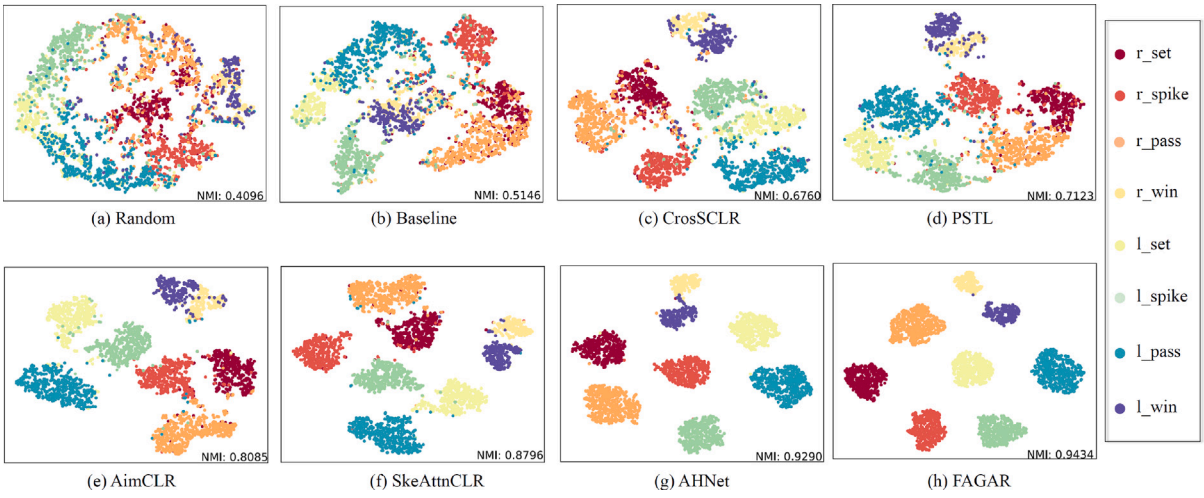



**Fig. 12.** Key actor visualization for "l-win" activity with connected lines. Brighter color indicates the more crucial the actor. (a) When the pose estimation algorithm is accurate, our FAGAR and AHNet [13] based on multi-person relation modeling can capture the key actors. However, the single-person method PSTL [34] and our FAGAR which models the phase in the frequency domain fail to capture the key actors. It proves the necessity of relation modeling and the effectiveness of amplitude-guided contrastive learning. (b) When the pose estimation algorithm makes mistakes, AHNet [13] focuses on the person where the abnormally estimated points are located instead of the key actors. Owing to the passing data augmentation algorithm, our method can still achieve a robust effect.




(a) Random (b) Baseline (c) CrosSCLR (d) PSTL

(e) AimCLR (f) SkeAttnCLR (g) AHNet (h) FAGAR

**Fig. 13.** The T-SNE visualization of embeddings on VD through finetune evaluation. Larger Normalized Mutual Information (NMI) means more discriminative distribution. Our proposed FAGAR achieves compactness with the highest NMI.

ahead by 4.33% (72.13% vs. 67.80%). On PKU-MMD, it improves over AHNet [13] by 2.63% and 12.98% in part I and part II, respectively. The experimental results demonstrate that our FAGAR can achieve significant effectiveness in single-person action recognition through appropriate modifications.

### 5.5. Visualization results

Fig. 12 presents a key actor visualization for the "l-win" activity. Notably, the brightness of the color assigned to each actor plays a significant role in this visualization. (a) In the context of the "l-win" activity, actors 3, 4, 5, 6, and 7 have been identified as the key actors. These actors are consistently involved in the most pivotal moments and decision-making processes throughout the "l-win" activity. In frequency domain learning, amplitude learning excels in capturing key actors as it better detects significant fluctuations. In contrast, phase learning is sensitive to temporal shifts and masked by noise or irrelevant activities, thus less reliable for identifying crucial individuals. Moreover, multi-person method AHNet [13] outperform single-person method PSTL [34] in capturing key actors, and this superiority can be attributed to the design of the relation modeling network. (b) However, when the pose estimation algorithm encounters errors, AHNet [13] focuses on the actors associated with the abnormally estimated points. In contrast, our FAGAR, benefiting from a frequency domain passing data augmentation algorithm, manages to maintain a robust performance.

Fig. 13 presents the visualization results of T-SNE [45], which is designed to visually compare the impacts of different methods on data classification. To conduct a more convincing analysis, quantitative Normalized Mutual Information (NMI) metrics are also supplied. In the eight classes shown in the figure, samples belonging to the same class are depicted in the same color, while those from different classes are in different colors. Compared with the other SOTA methods, our FAGAR exhibits a higher classification efficiency, as evidenced by its NMI value reaching an impressive 94.34%, which is the highest among them. This outstanding classification performance of FAGAR can be attributed to its utilization of frequency-domain modeling for actor relationships and the support of an adaptive network.

As shown in Fig. 14, we visualize the attention scores for both correct and incorrect pose-estimated group skeleton sequences. For temporal features, we follow AHNet's method [13] using Transformer Encoder Layers to highlight key individuals. For frequency features, we use Frequency Compression Network (FCN) to weight key individuals' features. In Fig. 14(a), when pose estimation contains obvious errors (red skeleton points), the frequency features remain more stable. Individuals 5, 6, 8, and 9 receive higher attention scores, excluding the severely misestimated individual 4. In contrast, temporal features

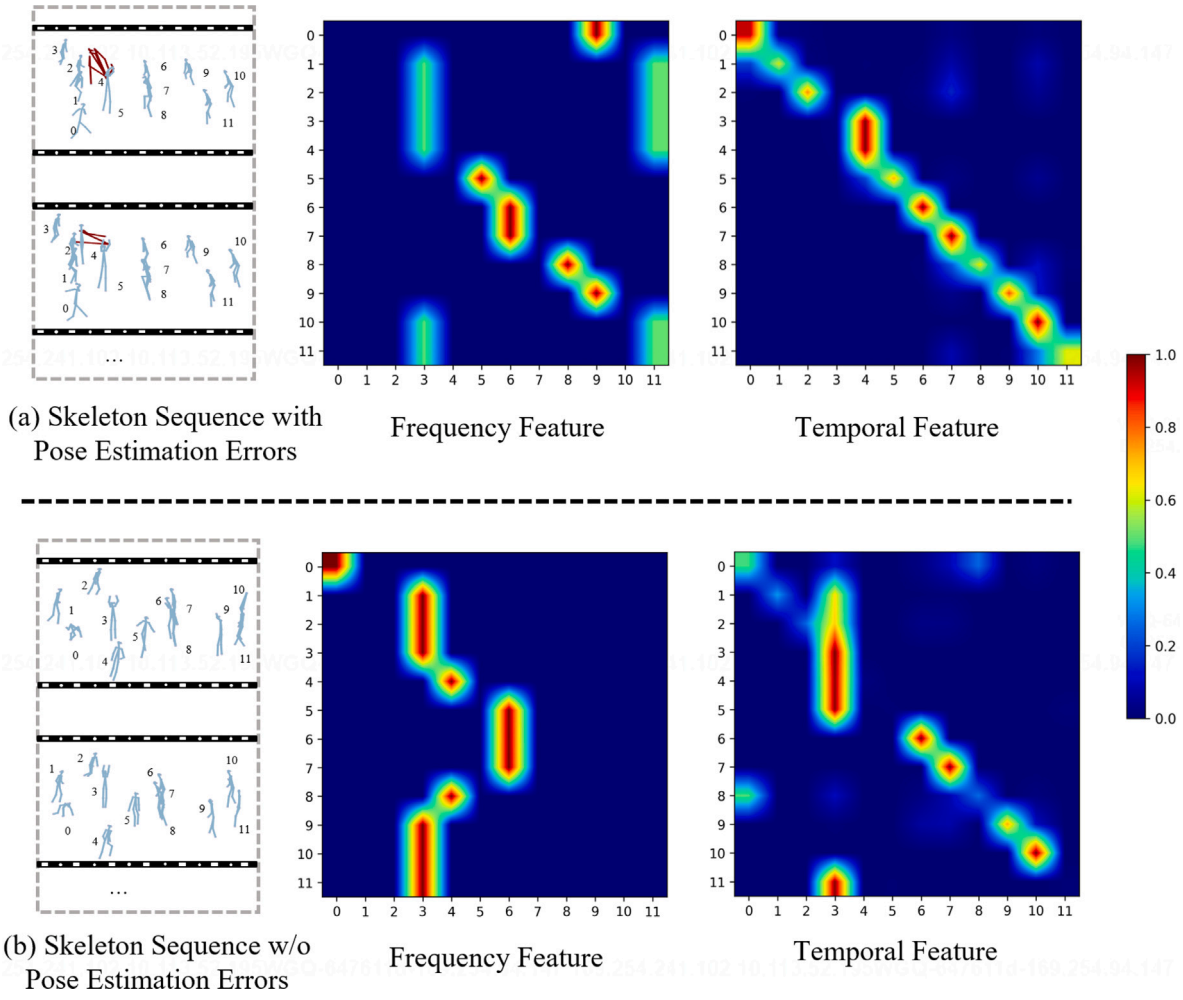



**Fig. 14.** Visualization of attention scores for temporal and frequency features under correct/incorrect pose estimation conditions.

incorrectly assign high attention scores to individual 4. When pose estimation is accurate, frequency features in Fig. 14(b) show more concentrated distributions, while temporal features appear more scattered. This suggests frequency features provide better discriminative power for key individual identification tasks.

## 6. Conclusion

We propose the **F**requency-**A**ware **G**roup **A**ctivity **R**ecognition (FAGAR) network to adapt to the dynamic and complex nature of spatio-temporal skeleton sequneces. Specifically, the frequency domain passing augmentation method focuses on individual collaborative changes and effectively filters out noise point interference. Additionally, an adaptive frequency domain compression network is designed to address the issue of fixed-connection relation modeling networks that cannot handle scene dynamic changes. Moreover, the amplitude-aware loss is introduced to guide the network to learn the importance of individuals to the group, preventing it from getting lost during the temporal domain modeling process. Extensive evaluations validate that our FAGAR achieves state-of-the-art performance on several datasets for self-supervised skeleton-based group activity recognition.

## 7. Limitations and future work

The frequency domain passing augmentation and adaptive network are designed with a focus on group activity recognition, which limits their direct transferability and generalization to other related tasks such as multi-person pose estimation and action detection.

For future work, we target at enhancing model versatility. One way is to design a multi-task framework for FAGAR to cover not only group activity recognition but also multi-person related tasks. Meanwhile, more in-depth research on data augmentation strategies will be carried out to boost their adaptability and efficacy across different tasks and datasets.

## CRediT authorship contribution statement

**Guoquan Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mengyuan Liu:** Supervision, Investigation, Funding acquisition, Conceptualization. **Hong Liu:** Validation, Supervision, Resources, Funding acquisition. **Jinyan Zhang:** Writing – original draft, Visualization, Validation, Investigation. **Peini Guo:** Writing – review & editing, Visualization, Validation. **Ruijia Fan:** Writing – review & editing, Methodology, Investigation. **Siyu Chen:** Writing – review & editing, Investigation.

## 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.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62203476) and Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064).

## Data availability

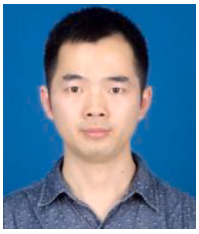
Data will be made available on request.

## References

[1] B. Kim, J. Kim, H.J. Chang, T.-H. Oh, A unified framework for unsupervised action learning via global-to-local motion transformer, PR (2024) 111118.
[2] J. Huang, L. Li, L. Qing, W. Tang, P. Wang, L. Guo, Y. Peng, Spatio-temporal interactive reasoning model for multi-group activity recognition, PR (2024) 111104.
[3] M. Dai, Z. Sun, T. Wang, J. Feng, K. Jia, Global spatio-temporal synergistic topology learning for skeleton-based action recognition, PR (2023) 109540.
[4] G.B. Mensah, M.M. Mijwil, M. Abotaleb, G. Ali, P.K. Dutta, T. Mzili, M.M. Eid, Explainable AI for healthcare: Training healthcare workers to use artificial intelligence techniques to reduce medical negligence in ghana's public health act, 2012 (act 851), EDRAAK (2025) 1–6.
[5] 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.
[6] W. Zhou, L. Kong, Y. Han, J. Qin, Z. Sun, Contextualized relation predictive model for self-supervised group activity representation learning, TMM (2023) 353–366.
[7] 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.
[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] Y. Liu, L. Nie, L. Liu, D.S. Rosenblum, From action to activity: sensor-based activity recognition, Neurocomputing (2016) 108–115.
[10] J. Zhang, M. Liu, H. Liu, G. Wang, W. Li, APP: Adaptive pose pooling for 3D human pose estimation from videos, in: ACM MM, 2024, pp. 1672–1681.
[11] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: CVPR, 2019, pp. 5693–5703.
[12] 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.
[13] G. Wang, M. Liu, H. Liu, P. Guo, T. Wang, J. Guo, R. Fan, Augmented skeleton sequences with hypergraph network for self-supervised group activity recognition, PR (2024) 110478.
[14] C. Bian, W. Feng, S. Wang, Self-supervised representation learning for skeleton-based group activity recognition, in: ACM MM, 2022, pp. 5990–5998.
[15] T. Guo, M. Liu, H. Liu, G. Wang, W. Li, Improving self-supervised action recognition from extremely augmented skeleton sequences, PR (2024) 110333.
[16] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: CVPR, 2020, pp. 9729–9738.
[17] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: ICML, 2021, pp. 12310–12320.
[18] 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.
[19] Y. Kong, Y. Fu, Human action recognition and prediction: A survey, IJCV (2022) 1366–1401.
[20] 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.
[21] 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.
[22] 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.
[23] J. Liu, S. Song, C. Liu, Y. Li, Y. Hu, A benchmark dataset and comparison study for multi-modal human action analytics, ACM T Multimed. Comput. (2020) 1–24.
[24] 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.
[25] 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.
[26] X. Li, M. Choo Chuah, Sbgar: Semantics based group activity recognition, in: ICCV, 2017, pp. 2876–2885.
[27] J. Wu, L. Wang, L. Wang, J. Guo, G. Wu, Learning actor relation graphs for group activity recognition, in: CVPR, 2019, pp. 9964–9974.
[28] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI, 2018, pp. 7444–7452.
[29] 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.
[30] 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.
[31] J. Zhang, Y. Jia, W. Xie, Z. Tu, Zoom transformer for skeleton-based group activity recognition, TCSVT (2022) 8646–8659.
[32] Z. Li, X. Chang, Y. Li, J. Su, Skeleton-based group activity recognition via spatial-temporal panoramic graph, in: ECCV, 2024, pp. 252–269.
[33] 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.
[34] 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.
[35] Q. Zhao, C. Zheng, M. Liu, P. Wang, C. Chen, Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation, in: CVPR, 2023, pp. 8877–8886.
[36] W. Liu, X. Jia, Y. Ju, Y. Ju, K. Jiang, S. Wu, L. Zhong, X. Zhong, Fragrant: frequency-auxiliary guided relational attention network for low-light action recognition, TVC (2025) 1379–1394.
[37] J. Tang, G. Wang, Z. Chen, S. Li, X. Li, X. Ji, FAFA: Frequency-aware flow-aided self-supervision for underwater object pose estimation, in: ECCV, 2024, pp. 352–369.
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, NeurIPS (2017) 5998–6008.
[39] K. Balasubramani, U.M. Natarajan, Improving bus passenger flow prediction using bi-LSTM fusion model and SMO algorithm, BJAI (2024) 73–82.
[40] 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.
[41] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: CVPR, 2020, pp. 143–152.
[42] 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.
[43] K. Gavrilyuk, R. Sanford, M. Javan, C.G. Snoek, Actor-transformers for group activity recognition, in: CVPR, 2020, pp. 839–848.
[44] R. Xu, L. Huang, M. Wang, J. Hu, W. Deng, Skeleton2vec: A self-supervised learning framework with contextualized target representations for skeleton sequence, 2024, arXiv:2401.00921.
[45] 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 his Ph.D. under the supervision of Prof. Hong Liu at the School of Electrical Engineering and Computer Science (EE&CS), Peking University (PKU), China. He subsequently worked as a research fellow with Prof. Junsong Yuan and Prof. KaiKuang Ma at the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore. Currently, he is an Assistant Professor at Peking University, Shenzhen Graduate School. His research primarily focuses on human-centric perception and human–robot interaction. His work has been published in leading conferences and journals, including

NeurIPS, CVPR, ICRA, T-IP, T-CSVT, T-MM, and PR. In addition, he actively serves as a reviewer for several top-tier international conferences and journals, such as NeurIPS, CVPR, T-IP, T-RO, and T-PAMI.



**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.



**Jinyan Zhang** received the B.S. degree from China University of Geosciences (CUG), China. He is a research graduate student studying at Peking University (PKU), China, under the supervision of Prof. Mengyuan Liu. His current research interest lies in 3D human pose and shape estimation.



**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.



**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.



**Siyu Chen** received her bachelor's degree from the University of Science and Technology Beijing (USTB). She is currently a research graduate student studying at Peking University (PKU), China, under the supervision of Prof. Hong Liu. Her current research interest include computer vision and depth estimation.