



**Please cite the Published Version**

Huang, Jianglan, Li, Lindong, Qing, Linbo , Tang, Wang, Wang, Pingyu, Guo, Li  and Peng, Yonghong (2024) Spatio-temporal interactive reasoning model for multi-group activity recognition. Pattern Recognition, 159. 111104 ISSN 0031-3203

**DOI:** <https://doi.org/10.1016/j.patcog.2024.111104>

**Publisher:** Elsevier

**Version:** Accepted Version

**Downloaded from:** <https://e-space.mmu.ac.uk/637189/>

**Usage rights:**  [Creative Commons: Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

**Additional Information:** This is an author accepted manuscript of an article published in Pattern Recognition, by Elsevier.

**Data Access Statement:** The authors do not have permission to share data.

**Enquiries:**

If you have questions about this document, contact [openresearch@mmu.ac.uk](mailto:openresearch@mmu.ac.uk). Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

# Spatio-Temporal Interactive Reasoning Model for Multi-Group Activity Recognition

Jianglan Huang<sup>1a</sup>, Lindong Li<sup>1a</sup>, Linbo Qing<sup>a,\*</sup>, Wang Tang<sup>a</sup>, Pingyu Wang<sup>a</sup>, Li Guo<sup>b</sup>, Yonghong Peng<sup>b</sup>

<sup>a</sup>College of Electronics and Information Engineering, Sichuan University, Chengdu, Sichuan, 610064, China

<sup>b</sup>Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M1 5GD, UK

---

## Abstract

Multi-group activity recognition aims to recognize sub-group activities in multi-person scenes. Existing works explore group-level features by simply using graph neural networks for reasoning about the individual interactions and directly aggregating individual features, which cannot fully mine the interactions between people and between sub-groups, resulting in the loss of useful information for group activity recognition. To address this problem, this paper proposes a Spatio-Temporal Interactive Reasoning Model (STIRM) to better exploit potential spatio-temporal interactions for multi-group activity recognition. In particular, we present an interactive feature extraction strategy to explore correlation features between individuals by analyzing the features of their nearest neighbor. We design a new clustering module that combines the action similarity feature and spatio-temporal trajectory feature to divide people into small groups. In addition, to obtain rich and accurate group-level features, a group interaction reasoning module is constructed to explore the interactions between different small groups and among people in the same group and exclude people who have less impact on group activities according to their importance. Extensive experiments on the Social-CAD, PLPS and JRDB-PAR datasets indicate the superiority of the proposed method over state-of-the-art methods.

*Keywords:* Action Recognition, Multi-Group Activity Recognition, Group Clustering, Interaction Reasoning, Spatio-Temporal Trajectory

---

<sup>1</sup>These authors contributed equally to this work.

\*corresponding author

Email address: qing\_lb@scu.edu.cn (Linbo Qing)

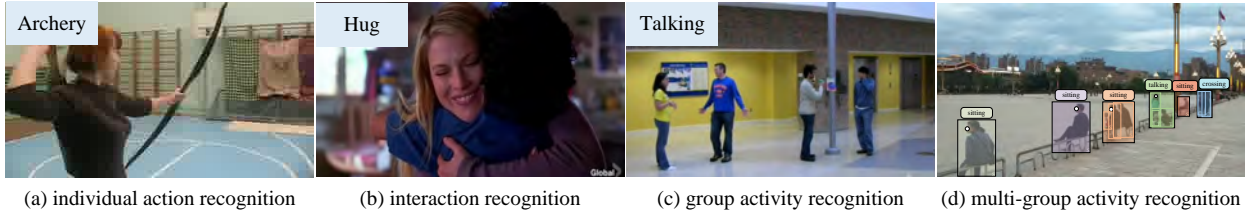


Figure 1: Examples of different types of human activity recognition.

## 1. Introduction

Human activity understanding based on video is the core research field of computer vision, which has various practical applications in the real world, such as surveillance systems and social scene understanding. Research on human activity understanding has focused mainly on recognizing a single person’s action (Fig. 1(a)) and on the identification of interactions between individuals (Fig. 1(b)) in a simple scenario with single background and explicit action [1].

Recently, the task of understanding human activity has been expanded to more demanding and real scenes. Most existing works regard all individuals as one group and recognize the activity of a single group (Fig. 1(c)). However, in real scenes, people are likely to form different social groups according to the social interactions between them (Fig. 1(d)). Therefore, studies on multi-group activity recognition (MGAR) are more meaningful and essentially aim to understand “Who is with whom and what they are doing together”. Typically, MGAR includes three subtasks: individual action recognition for each person from social groups or skeletons, group clustering, and group activity recognition for each social group and singleton. Among these, individual action recognition aims to identify the action of each individual, group clustering aims to divide individuals into small groups based on their social interactions (“who is with whom”), and group activity recognition of each group aims to recognize the activity of each small group based on group clustering results (“what they are doing together”).

Studies of MGAR have recently been extended from static views [2, 3] to panoramic and dynamic views [4]. In such studies, mining and inferring potential interactions between individuals or sub-groups to form comprehensive group-level features is crucial. Graph neural networks (GNNs) are now widely used to infer

the potential relationships between nodes. Therefore, individual features are often regarded as nodes, and the edges are the relationships between nodes. Once nodes and edges are initialized, GNN can automatically explore the correlation between persons to build rich group-level features for further reasoning. For static view, Ehsanpour et al. [2] utilized graph attention networks (GATs) [5] to exploit the potential interactions among people to divide them into small groups and recognize multi-group activities. Qing et al. [3] designed individual and group graph reasoning modules to infer the relationships between them, respectively. From a panoramic and dynamic view, Han et al. [4] developed a hierarchical graph neural network to gradually represent and model multi-granular human activities and mutual social relationships among populations. Cao et al. [6] proposed a multi-granularity unified perception framework to mine the intra-relevant and cross-relevant semantics. Gan et al. [7] considered both the unique characteristics of each task and the synergies between different tasks simultaneously. However, most existing MGAR approaches extract and reason for only individual features, adopt graph spectral clustering technique or use the appearance similarity of behaviors to cluster people, and then aggregate individual features in sub-groups directly without further exploiting deep spatio-temporal interactions between persons, which fails to obtain a rich representation of individual/group-level features.

To better exploit potential spatio-temporal cues and form a rich representation of final features, we propose a new Spatio-Temporal Interactive Reasoning Model (STIRM) for MGAR that focuses on exploiting spatio-temporal interaction between people and sub-groups. Specifically, we design a group interaction reasoning module to explore interactions between people in the same group and between different sub-groups from global and local perspectives. Moreover, people who have less impact on group activity recognition results are ignored in this module. In addition, we obtain spatial interactive features between people through reasoning about their nearest neighbor to enhance the original individual’s features. We combine this approach and spatio-temporal trajectory features [8] to divide individuals into small parts. Our key contributions are summarized as follows:

1. A nearest neighbor interaction extraction module is designed to explore spatial interactive features between individuals. This module not only reasons about the interactive appearance features of people

but also exploits spatial group topology and the relative positions between them.

2. A spatio-temporal clustering module is constructed to better understand the relationships between people by incorporating action similarity features and spatio-temporal trajectory features. This module divides people into multiple clusters for better learning activity-aware semantic representations.
3. A group interaction reasoning module is built to infer representative group-level features, which contains two sub-modules: Inter-Trajectory Relation Graph Reasoning Module and Graph Aggregation and Pooling Module. This module can effectively explore spatio-temporal interactions between different sub-groups and between people and ignore unimportant people according to their impact on group activity recognition.
4. Extensive experiments were conducted on three challenging benchmarks to illustrate the effectiveness of our approach. The experiments demonstrate that the Spatio-Temporal Interactive Reasoning Model (STIRM) can effectively explore the spatio-temporal interaction between people and between different sub-groups. The results demonstrate the excellent performance of our method.

The rest of the paper is organized as follows. Section 2 present related work on action recognition, group clustering methods, and group activity recognition methods. Section 3 describes the proposed approach in detail. The experimental results and discussion are analyzed in Section 4. Finally, Section 5 presents the conclusions of this paper.

## 2. Related Work

MGAR aims to identify sub-group activities in multi-person scenes; this involves simultaneously performing the individual action recognition, group clustering, and group activity recognition tasks. Therefore, the related works on action recognition, group clustering and group activity recognition are briefly introduced in this section.

### 2.1. Action Recognition

Action recognition is a fundamental task in computer vision that takes a video as input and outputs an action category [1, 9]. The existing action recognition approaches can be divided into two categories:

appearance-based methods and the skeleton-based methods. For appearance-based methods, Ma and Wang [10] utilized Transformers to encode the spatial and temporal dynamics of visual features for action recognition. Wang et al. [11] leveraged the rich semantic priors in CLIP to obtain reliable prototypes and achieve accurate few-shot classification. For skeleton-based approaches, Qin and Hou [12] designed different attention mechanisms to focus on joint-level and part-level information and help distinguish similar actions. Wu et al. [13] divided human joints into the trunk, hands and legs and built spatial-temporal hypergraphs to extract their high-order features for multi-view data lightweight action recognition. Temporal dependency is also vital to action recognition via videos. Therefore, CNNs with 3D kernels have been proposed to better exploit spatial and temporal cues [14, 15]. Note that the video data commonly employed for this task are typically gathered from laboratory settings or sourced from movies/website videos. The videos usually contain only one or very few humans with people occupying the central focus of each frame. However, multiple small people are captured in the real world, making it difficult to extract and analyze the appearance features and skeletons of people, which poses a significant challenge in this task.

## *2.2. Group Clustering*

Group clustering endeavors to divide individuals into small segments for subsequent applications. Accurate clustering methods can more correctly divide closely related individuals into the same group, thereby promoting the inference of multi-group activity. The existing works for group clustering can be categorized into two groups: the first approach clusters people on the basis of diverse motion patterns exhibited by crowds [16] and is effective in medium- or high-density crowd scenarios such as crowd flow monitoring. The methods in the second category focus on clustering people to form small or social groups by observing social interactions among individuals and is primarily applied in low or medium density scenes. This category serves as the foundation for studying group-based activity understanding, and we mainly focus on this field in the following paragraphs. In these studies, various measures have been employed to assess inter-group proximity for group detection, including pairwise proximity [17], velocity information [17], trajectory data [18, 19] and interaction characteristics [18, 19]. For deep learning-based approaches, researchers have proposed GAN-based methods [20], DNN-based methods [21], GCN-based methods [22] and graph attention

network (GAT)-based methods [2] to process extracted features. Specifically, Akbari et al. [21] utilized deep neural networks (DNNs) to integrate the Euclidean distance, proximity distance, motion causality and trajectory shape between pairs of pedestrians for social group detection. Sun et al. [22] introduced a group-based social interaction model to explore the connections among individuals by constructing a Recursive Social Behavior Graph (RSBG). Ehsanpour et al. [2] applied graph attention networks (GATs) to capture potential interaction relationships among individuals, enabling the clustering of social groups and recognition of multi-group activities.

In the MGAR task, accurate clustering results benefit group activity reasoning. However, most existing approaches consider only the Euclidean distance in a static view or use only shallow features. Therefore, in our spatio-temporal clustering module, we employ both spatio-temporal trajectory cues and the similarity of individual activities to reflect their social connections further.

### *2.3. Group Activity Recognition*

Group activity recognition (GAR) is another task in the realm of understanding human activity. Its objective is to identify the activity being performed by a group of individuals. Recently, many existing works have considered all people as a single group, which is called (single) group activity recognition. In such studies, it is highly important to obtain better representations of group-level features. Some studies max-pool individual features directly to form group-level features [3], but this leads to a loss of useful information for activity recognition. Therefore, attention mechanisms have been widely used. Some researchers have adopted attention mechanisms to find key actors that have the greatest influence on group activity [23] or produce coefficients to aggregate actor-level features to group-level features [24]. In addition, exploring interactions between people can benefit group activity recognition results and mainstream methods often utilize Graph Neural Networks (GNNs) [25, 26], Transformers [27, 28], and Recurrent Neural Networks (RNNs) [29, 30] to reason about the relationships between humans. To obtain more accurate individual features for group reasoning, some works have focused on reasoning the latent spatio-temporal dependencies among body regions [31]. In addition, some studies have either used self-supervised inference of spatio-temporal correlations to identify group activity [32, 33] or cluster people into small parts, where the correlation is

tight in the same cluster but distant in different clusters [34]. The aforementioned work takes videos as input, which requires a significant amount of computational resources. To address this issue, Wang et al. [35] and Perez et al. [36] explored skeleton-based group activity recognition.

However, it is not reasonable to regard all people as one group because not all people have social connections. Therefore, multi-group activity recognition (MGAR) scheme have been proposed to determine group activities in every small group. Ehsanpour et al. [2] employed graph attention networks (GATs) [5] to capture interactions among people to divide social groups and recognize multi-group activities. Qing et al. [3] developed individual and group graph reasoning modules to infer their relationships. Han et al. [4] presented a hierarchical graph neural network to progressively model multi-granular human activities and social relations for crowds. Cao et al. [6] proposed a multi-granularity unified perception to mine intra-granularity and inter-granularity semantics. Gan et al. [7] designed a mix-parameters Transformer to consider the complementary effect between tasks of different granularities.

Most existing MGAR approaches extract and reason about individual features and aggregate them directly without further exploiting spatio-temporal interactions, leading to poor effectiveness for better representing group-level features. In this work, we focus on exploring spatio-temporal interactive cues between people and between sub-groups via the interaction graph reasoning module and build rich group-level features after key actors are automatically selected.

### **3. Method**

#### *3.1. Overview*

We proposed a new Spatio-Temporal Interactive Reasoning Model (STIRM) for MGAR, which focuses on exploiting spatio-temporal interaction between people and sub-groups and forming representative group-level features. As illustrated in Fig. 2, this framework includes five parts: Individual Feature Extraction Module, Inter-Actor Relation Graph Reasoning Module, Nearest Neighbor Interaction Extraction Module, Spatio-Temporal Clustering Module and Group Interaction Reasoning Module. Firstly, we employ the backbone of 3D-ResNet50-NonLocal [37], attention mechanism of Squeeze-and-excitation Networks (SENet)



[38] and RoI-Align to extract the actor features  $f_{act}$ , which are regarded as nodes and then fed into the Graph Attention Network (GAT) [5] and Dynamic Inference Network (DIN) [26] to reason about pair-wise interactions and achieve spatio-temporal person-specific inferences. The two features are added as  $f_{actor}$  and sent into individual action classification.

Then, according to the individuals' positions and their pair-wise distances, we explore the features of the individuals' nearest neighbors  $f_{ai}$  as individual interactive features to capture relative context features and location features  $f_p$  are cascaded to individual features, which are considered effective supplements for determining group topology. Next, we combine the output of the nearest neighbor interaction extraction module  $f_{ap}$  and the preprocessed spatio-temporal trajectory feature  $f_{ST}$  to group people. To better reason about the interactions between people and groups, first, we utilize  $f_{ap}$  as nodes and the similarity of the global pair-wise trajectory as edges and feed them into the GCN. Then, incorporating  $f_{act}$ ,  $f_{apt}$  and  $f_{ap}$  as  $f_g$ , graph aggregation and the self-attention graph pooling method are introduced to explore relations within and between groups and discard unimportant people according to the importance of individuals. Finally, the multi-group activity recognition results are obtained according to the group clustering result and group final feature  $f_{group}$ . We describe the method in details in the following sections.

### 3.2. Individual Feature Extraction

We first design an individual feature extraction module to extract spatial and temporal information about individuals. In particular, we adopt 3D-ResNet50-NonLocal [37] pre-trained on ImageNet [39] as the backbone to extract the global features from RGB images of dimension  $H \times W \times 3$ , where we take 32 frames as input to capture temporal dynamic information. In the last layer of the backbone, we insert SENet [38] to better obtain the correlation of different channels and restrain the less impactful channels. Then, the global features are fed into RoI-Align, whose crop size is  $5 \times 5$ , to extract individual features. After that, a fully connected layer is applied to the aligned features to obtain a  $d$  dimensional appearance feature  $f_{act}$  for each actor. The total number of bounding boxes in all input frames is denoted as  $N$ ; thus a  $N \times d$  matrix  $f_{act}$  is utilized to represent the appearance features of the actors.

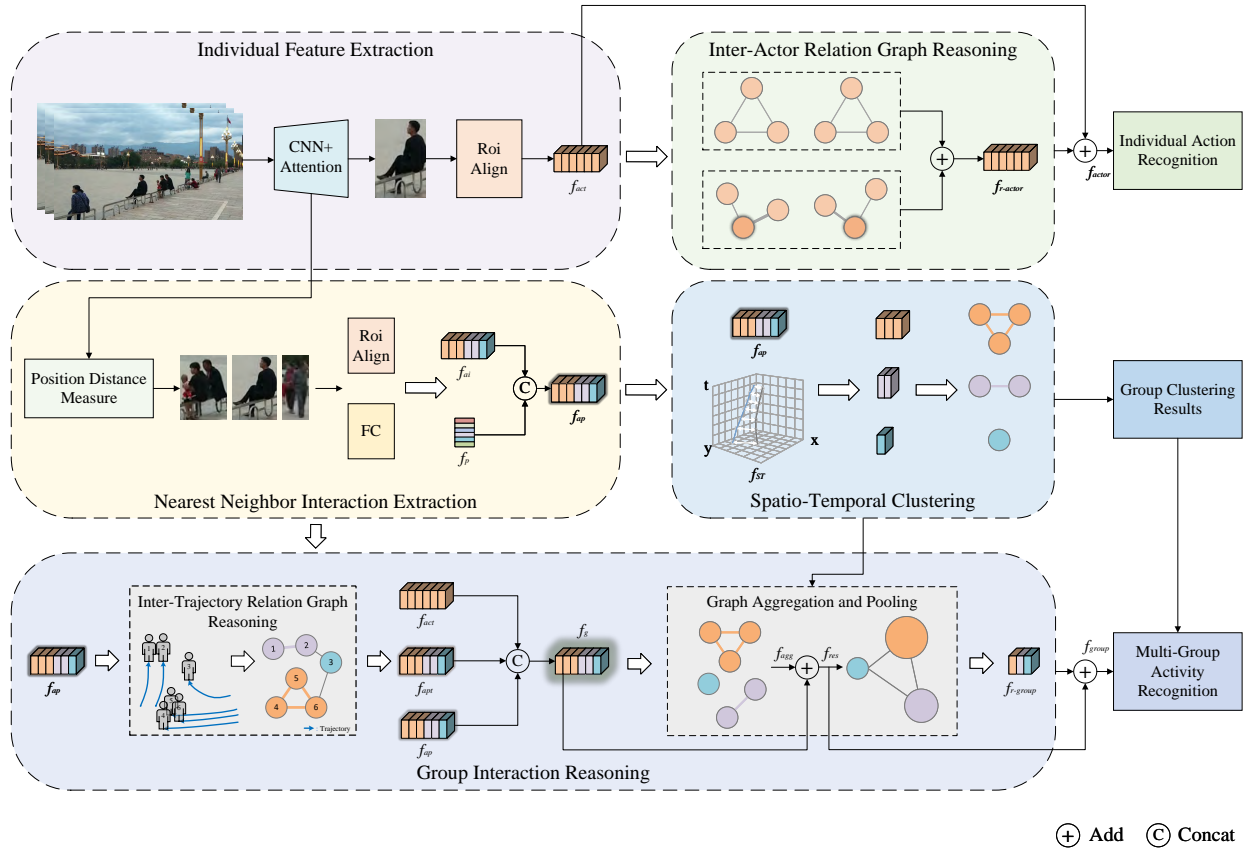


Figure 2: Overview framework of the proposed method. We first extract individual features and explore their relationships via the graph reasoning module. Then, we explore features of the individual’s nearest neighbor and combine them with position features to gain rich interactive representations. Next, we incorporate the above appearance-position interactive features and preprocessed spatio-temporal features to cluster people. Then, we perform reasoning on people according to the similarity of their global trajectories. Finally, the graph aggregation and pooling module is introduced to exploit relations within and between groups, ignore unimportant people, and obtain group-level features.

### 3.3. Inter-Actor Relation Graph Reasoning

Individual action is unique and easily affected by others in a multi-person scenario. Therefore, this task can be modeled by a graph, where the nodes represent the individuals' features and the edges denote the interactions among people. We apply GAT [5] and DIN [26] to learn the underlying interactions between individuals and achieve person-specific inference simultaneously. The whole process is shown in Fig. 3.

First, on the basis of the individual appearance features  $f_{act}$ , we regard them as nodes and connect all pairs in the GAT. The GAT enables the flexible learning of attention weights between nodes through parameterized operations, utilizing a self-attention strategy, which is a good choice for uncovering subtle interactions among individuals.

Moreover, individual appearance features  $f_{act}$  are fed into the DIN to form a person-specific graph. In particular, DIN contains two main modules: DR and DW [26]. DR predicts the relation matrix for a given person, whereas DW predicts the dynamic walk offsets to introduce a global interaction field to the interaction graph.

Then, the output features from GAT and DIN are fused through element-wise addition, expressed as follows:

$$f_{GAT} + f_{DIN} = f_{r-actor}. \quad (1)$$

Then, we add  $f_{act}$  and  $f_{r-actor}$  create the final output of individual features  $f_{actor}$ , and individual action results are obtained.

### 3.4. Nearest Neighbor Interaction Extraction

Individuals close in space may have social relationships and may also engage in the same activities and interact. Therefore, we propose the nearest neighbor interaction extraction module for mining spatial interactive information.

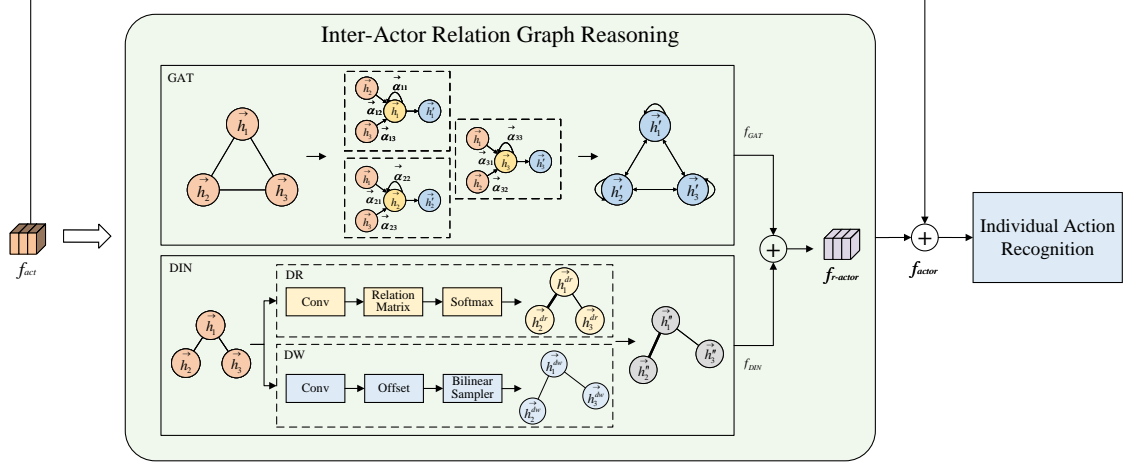


Figure 3: Example diagram of inter-actor relation graph reasoning.

#### 3.4.1. Reconstruction of the joint region boundary box

Reconstruction of the joint region boundary box of the nearest neighbor relationship is the basis of this section. Suppose that there are  $N$  people in the scene, the original rectangular box of individual  $i$  is  $box_i = (lx_i, ly_i, rx_i, ry_i)$ , and the center position coordinates of the  $i$ -th and  $j$ -th target individuals are calculated as  $P_i = (x_i, y_i)$  and  $P_j = (x_j, y_j)$ , respectively according to Eq. 2 as follows:

$$(x_i, y_i) = \left( \frac{lx_i + rx_i}{2}, \frac{ly_i + ry_i}{2} \right). \quad (2)$$

The Euclidean distance between two people is then calculated as follows,

$$Dis_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad (3)$$

and the element of the distance matrix  $D_m$  is the Euclidean distance for  $N$  people in pairs, represented as follows,



Figure 4: Example diagram of nearest neighbor interaction. Take A and B for example, we find their nearest neighbor C according to their distances among individuals. Then, we crop individuals and their nearest neighbors based on the threshold  $\mu$ . Finally, we extract features of the Rebox regions as every individual's feature.

$$D_m = \begin{bmatrix} Dis_{11} & \cdots & Dis_{1i} & Dis_{1N} \\ \cdots & \cdots & \cdots & \cdots \\ Dis_{i1} & \cdots & Dis_{ii} & \cdots \\ Dis_{N1} & \cdots & \cdots & Dis_{NN} \end{bmatrix}, \quad (4)$$

where the diagonal distance value is 0 because  $Dis_{ii}$  represents the distance between an individual and themselves.

According to the pair-wise distance  $Dis_{ij}$ , we can find the nearest neighbor individual by sorting the distance vector corresponding to the individual  $i$ . The indices of nearest neighbor and distance between them are denoted by  $newId_i$  and  $newDis_i$ , respectively and are given by:

$$newId_i, newDis_i = \text{Rank}(\{\text{Sort}(Dis_{ij}), j \neq i\}, 1). \quad (5)$$

If the nearest distance is lower than the threshold  $\mu$ , which means that they are closely related to each other, the bounding box of the  $i$  person is reconstructed as the minimum circumscribed rectangle of the individual bounding box with indices  $i$  and  $newId_i$ , expressed as follows:

$$\begin{cases} newId_i = newId_i, & newDis_i \leq \mu, \\ newId_i = i, & newDis_i > \mu. \end{cases} \quad (6)$$

If the coordinate vector of the original rectangular box of individual  $i$  is  $box_i = (lx_i, ly_i, rx_i, ry_i)$ , then the reconstructed boundary box  $Rebox_i$  is obtained as follows:

$$Rebox_i = \{\min(lx_i, lx_{newId_i}), \min(ly_i, ly_{newId_i}), \\ \max(rx_i, rx_{newId_i}), \max(ry_i, ry_{newId_i})\}. \quad (7)$$

According to the reconstructed boundary box  $Rebox$ , we can crop the persons and their interactive neighbors, as shown in Fig. 4.

#### 3.4.2. Appearance and position interactive feature extraction

To some extent, human position and activity may be highly correlated. For example, in team activities, each member's position and action depend on the behaviors of other members, particularly the nearest member. Additionally, individuals close to each other in the scene have a similar spatial locations and relative distance distributions, their corresponding individual behaviors are strongly correlated, and they tend to have specific social relationships. Therefore, we add the appearance interactive feature of the nearest neighbor as a whole and explore the position feature to supplement the appearance feature to exploit the interactions between them. The whole process is shown in Fig. 5.

According to the  $Rebox$  from Section 3.4.1, nearest neighbor areas are fed into RoI-Align to extract appearance interactive features  $f_{ai}$ . Then, we concatenate the coordinate of each individual and the distance between every two individuals and send them to fully connected layer to obtain the position interactive feature  $f_p$ . Finally, appearance interactive features  $f_{ai}$  and position interactive feature  $f_p$  are fused as refined interactive features  $f_{ap}$ , which are applied in clustering and sub-group activity recognition.

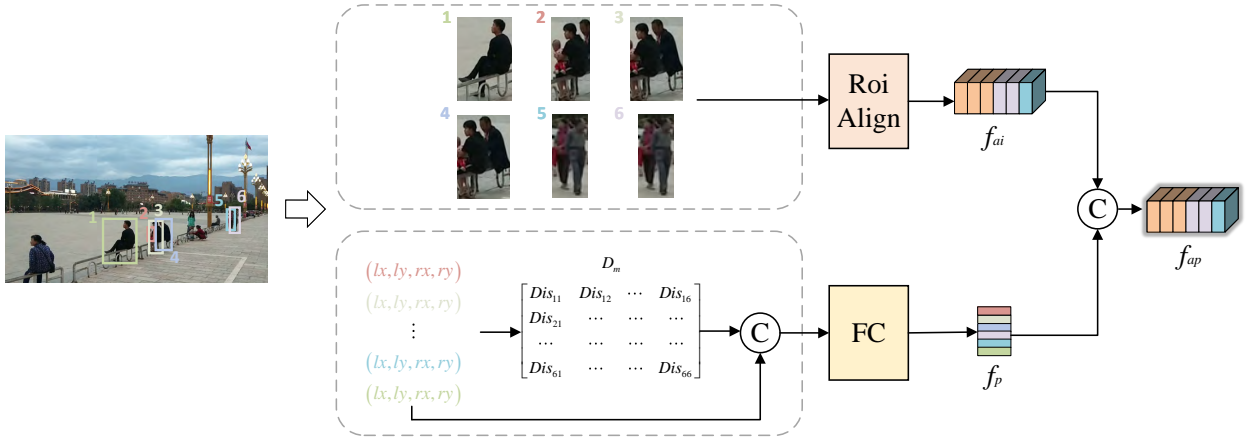


Figure 5: The detailed process of appearance and position interactive feature extraction.

### 3.5. Spatio-Temporal Clustering

Group clustering aims to partition individuals into distinct sub-groups. In MGAR research, a group is considered a collection of individuals who exhibit common consistent behavior and are relatively close in distance. Therefore, we consider action similarity characteristics and distance information between pair-wise trajectories in this clustering module.

The entire process is illustrated in Fig. 6, and includes action similarity measurements and spatio-temporal trajectory distance measurements. Based on these measurements, we can obtain the action similarity result matrix  $M_{ap}$  and spatio-temporal distance(ST-distance) result matrix  $M_{ST}$ . The final group clustering result matrix  $M$  is subsequently calculated via AND operations on  $M_{ap}$  and  $M_{ST}$ , where 0 represents a social interaction between a person pair, whereas 1 means that the two people are in the same group. We describe this procedure in detail below.

#### 3.5.1. Action similarity measurement

To measure action similarity, we utilize the refined interactive feature  $f_{ap}$  described in Section 3.4.2 as one of the inputs of this module. Then, we calculate the similarity between two persons using the cosine similarity measure as follows:

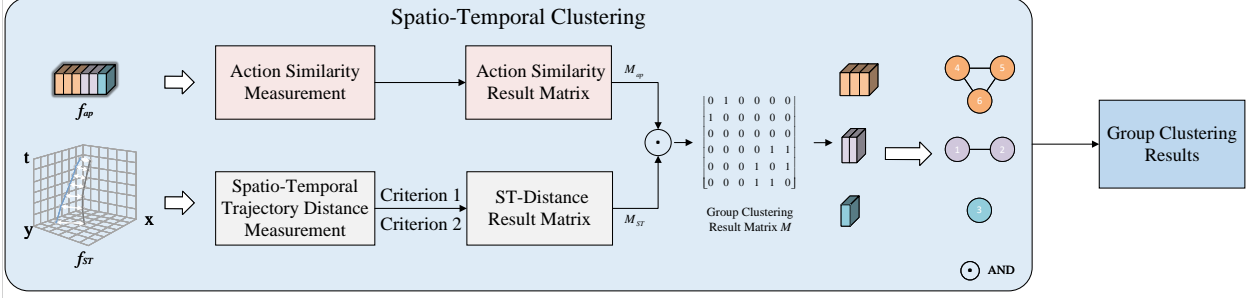


Figure 6: Example diagram of spatio-temporal clustering.

$$F_a(f_{ap}^i, f_{ap}^j) = \frac{\theta(f_{ap}^i)^T \phi(f_{ap}^j)}{\|\theta(f_{ap}^i)\| \times \|\phi(f_{ap}^j)\|}, \quad (8)$$

where  $f_{ap}^i \in \mathbb{R}^d$  and  $f_{ap}^j \in \mathbb{R}^d$  represents the features of the  $i$ -th and  $j$ -th person, respectively.  $\theta(f_{ap}^i) = W_\theta f_{ap}^i + b_\theta$  and  $\phi(f_{ap}^j) = W_\varphi f_{ap}^j + b_\varphi$  are learnable linear transformations.  $W_\theta \in \mathbb{R}^{d_k \times d}$  and  $W_\varphi \in \mathbb{R}^{d_k \times d}$  are weight matrices, and  $b_\theta \in \mathbb{R}^{d_k}$  and  $b_\varphi \in \mathbb{R}^{d_k}$  are weight vectors.

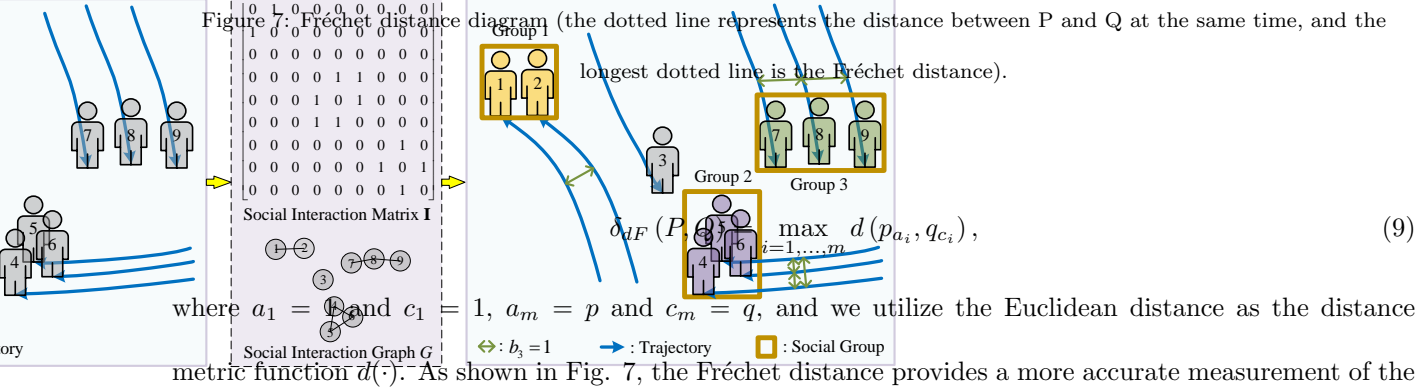
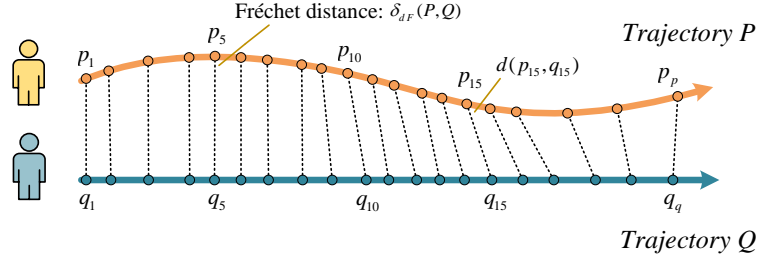
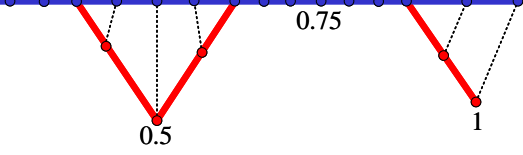
Setting a threshold  $\mu$ , we compare every similarity value with the threshold individually. If the similarity value between  $i$  and  $j$  is larger than the threshold, the element value of the result matrix  $M_{ap}^{ij}$  is 1, suggesting that person  $i$  and person  $j$  are in the same group; otherwise, the value is 0, which means that they do not belong to the same group.

### 3.5.2. Spatio-temporal trajectory distance measurement

For the distance information between pair-wise trajectories, we adopt spatio-temporal interpersonal distance measurement [8] as another piece of information for determining whether individuals belong to the same group. Unlike traditional distance measurements, which consider only static distances, we utilize spatio-temporal distances instead.

In particular, the two trajectory curves  $P$  and  $Q$  are discretely represented by a set of sampling points ( $p$  and  $q$ ), also written as  $P = (p_1, \dots, p_p)$  and  $Q = (q_1, \dots, q_q)$ , respectively. The discrete Fréchet distance is calculated by taking the maximum value among the distances between corresponding sampling points, represented by:





According to the spatio-temporal interpersonal distance measurement method mentioned above, we can determine the existence of a social interaction relationship between pairs, which includes two judgment conditions [8].

The outputs of two judgments are written as  $b_1$  and  $b_2$ , and the final judgment condition for recognizing the presence of a social interaction between individuals can be determined via a logical expression as follows:

$$b_3 = b_1 \otimes b_2, \quad (10)$$

where  $\otimes$  represents a logical OR. If the binary variable  $b_3$  is 1, the outcomes of the social interaction recognition for the individual pairs are positive. Notably, the element of the adjacent matrix  $M_{ST}$  is the value of  $b_3$ , indicating whether the individuals are in the same group.

Through the action similarity cues  $M_{ap}$  and distance cues  $M_{ST}$ , we can obtain group clustering results via the AND operation, expressed as follows:

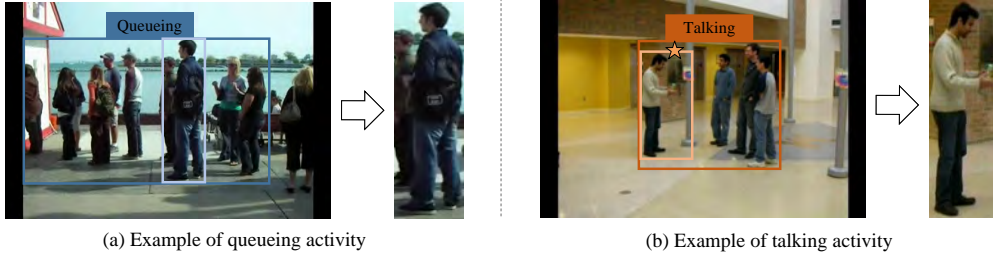
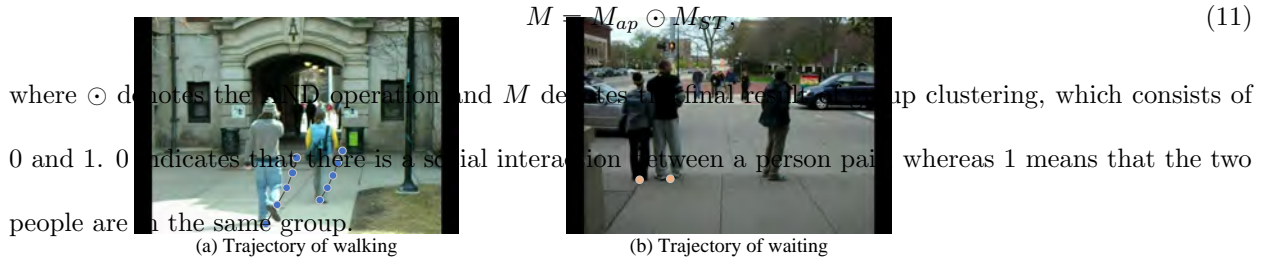


Figure 8: Examples of different group activities.



### 3.6. Group Interaction Reasoning

Group activity is related not only to individual actions but also to interactions between people. For example, when several people are queuing, if we only see one of them and ignore the influence of others, we may regard the activity as standing (as shown in Fig. 8(a)). If we notice the interactions of the people surrounding him, we can easily identify the queuing activity. In addition, the key person in a group also determines the group activity. As shown in Fig. 8(b), we can observe that the key person in the talking activity is the person who is speaking while others are listening. The activity can be identify more accurately if we can identify the key person.

Therefore, to explore interactions within and across the sub-groups, we propose Group Interaction Reasoning Module, which contains two main parts: the inter-trajectory relation graph reasoning module and the graph aggregation and pooling module (as shown in Fig. 9). We describe these modules in detail below.

#### 3.6.1. Inter-trajectory relation graph reasoning

Individuals with very similar trajectories in structure and shape tend to act similarly. Therefore, we design an inter-trajectory relation graph reasoning module to measure pair-wise interactions.

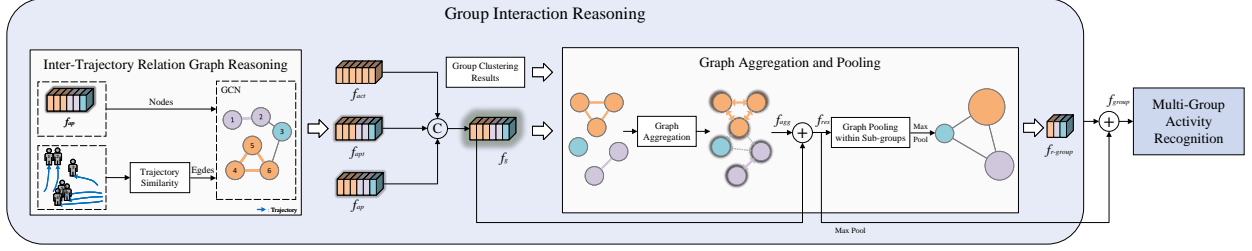


Figure 9: Example diagram of group interaction reasoning.

In this module, we take the refined interactive features  $f_{ap}$  mentioned in Section 3.4.2 as nodes and the global trajectory similarity between individuals as edges and input them into the GCN to reason about their interactions. The edge weights increase for greater similarity between the trajectories and for more closely connected individuals. In the following paragraphs, we introduce how the method for computing the trajectory similarity.

The trajectory is composed of the vectors of adjacent sampling points. As shown in Fig. 10, the trajectory  $P$  is represented by a local vector  $\{p_i\}_{i=1}^n$  and a global vector  $p_{global}$ , whereas the trajectory  $Q$  is represented by a local vector  $\{q_i\}_{i=1}^n$  and a global vector  $q_{global}$ , where  $p_i$  and  $q_i$  represent the  $i$ -th trajectory vectors of the corresponding trajectory. Since the angle and length information of the trajectory provide an important information about the trajectory shape and structure, the cosine similarity distance and trajectory length similarity distance at the local and global scales are selected to calculate the similarity distance between the trajectory  $P$  and the trajectory  $Q$ . Here, a smaller the distance corresponds to higher similarity. The distance is calculated according to Eq. 12-14:

$$\text{Dis}(P, Q) = \theta_{\text{global}} l_{\text{global}} \cdot \sum_{i=1}^n \theta_i l_i, \quad (12)$$

$$\theta_i(p_i, q_i) = \arccos\left(\frac{p_i \cdot q_i}{\|p_i\|_2 \|q_i\|_2}\right), \quad \theta_i(p_i, q_i) \in (0, \pi), \quad (13)$$

$$l_i(p_i, q_i) = \frac{\max(\|p_i\|_2, \|q_i\|_2)}{\min(\|p_i\|_2, \|q_i\|_2)}, \quad (14)$$

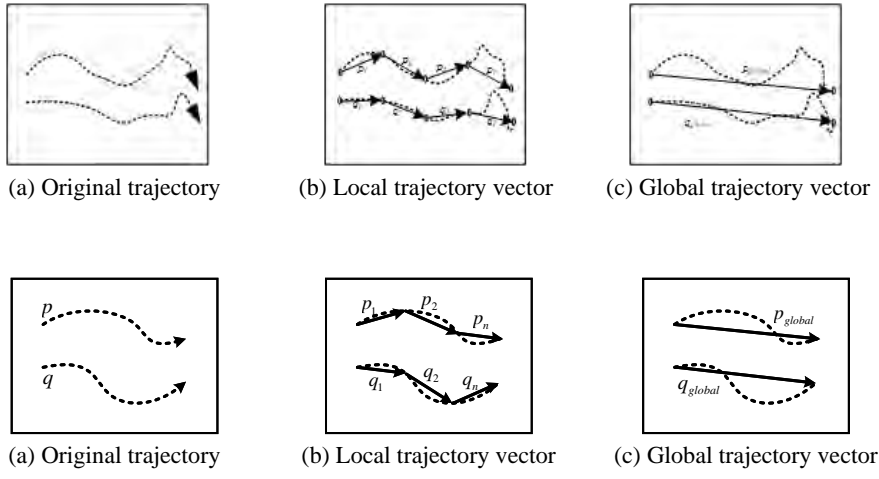


Figure 10: Vector graph representation on the basis of trajectory structure similarity.

where  $\theta_i$  represents the similarity distance between trajectory vectors  $p_i$  and  $q_i$  and  $l_i$  denotes the corresponding length similarity distance.  $\theta_{global}$  and  $l_{global}$  represent the angle similarity distance and length similarity distance of the global trajectory vectors  $p_{global}$  and  $q_{global}$ , respectively.

According to the similarity distance between any two trajectories, we can obtain a similarity coefficient between them that is defined as follows:

$$\text{Sim}(P, Q) = \text{Softmax}(-\ln \text{Dis}(P, Q)). \quad (15)$$

Then, we feed refined interactive features  $f_{ap}$  as nodes and similar coefficients as edges into the GCN to explore the interaction between individuals, where the output feature can be written as  $f_{apt}$ . After that,  $f_{act}$ ,  $f_{apt}$  and  $f_{ap}$  are concatenated as  $f_g$ .

### 3.6.2. Graph aggregation and pooling

The graph aggregation and self-attention graph pooling methods are introduced to explore relationships within and between groups, and unimportant people are discarded according to their importance.

As shown in Fig. 9, all individuals in the same group have strong social interaction, so that full connections unite all nodes in the same group, and a relationship path is established between groups to explore the influence of different groups. To learn the overall representation within a group, Master2Token (M2T) [40] is used to capture the correlation of interactive fusion features between individuals and adjacent nodes. This method considers the self-attention of total aggregation within a group and local aggregation between groups.

We take  $f_g = \{v_i \mid i = 0, \dots, N\}$  and the group clustering results mentioned in Section 3.5 as the inputs

of M2T. After that, we add  $f_g$  and the output of M2T  $f_{agg}$  as  $f_{res}$ . Next,  $f_{res}$  are input into the self-attention graph pooling (SAGPool) module [41] to focus on key actors to generate group-level features. In the SAGPool module, self-attention scores  $S$  are computed to select important actors. Finally, group-level features  $f_{group} \in \mathbb{R}^d$  are obtained via the max-pooling operation, which are fed into the group activity classifier to recognize the multi-group activity. The whole process is shown in Eq. 16:

$$R = f_{g\text{-classifier}}(f_{group}) = f_{g\text{-classifier}}(m(f_{res} \circ S) + m(f_{res})), \quad (16)$$

where  $m(\cdot)$  is the max-pooling operation and where  $\circ$  denotes the dot product.

### 3.7. Training Loss

The whole model is trained with gradient descent via back propagation. By incorporating the standard cross-entropy loss, we establish the final loss function via Eq. 17:

$$L = \lambda_1 \sum_s L_{sgp} \left( O_s^{SG}, O_s^{\hat{SG}} \right) + \lambda_2 \sum_n L_{ind} \left( O_n^I, \hat{O}_n^I \right) + \lambda_3 L_c \left( O^\alpha, \hat{O}^\alpha \right), \quad (17)$$

where  $L_c$  is the binary cross-entropy loss used to reduce the discrepancy between the relationship matrix  $\hat{O}^\alpha$  constructed in Section 3.5 and the ground-truth social group labels  $O^\alpha$ .  $L_{sgp}$  and  $L_{ind}$  denote cross-entropy loss for the group activity of each small group and individual action classification in the Social-CAD and PLPS experiments, respectively, and represent binary cross-entropy with logits in the JRDB-PAR experiment because this dataset is annotated by multile labels.  $O_n^I$  and  $O_s^{SG}$  represent the ground-truth individual action and sub-group activity labels, respectively, and  $\hat{O}_n^I$  and  $\hat{O}_s^{SG}$  denote the corresponding predicted labels.  $\lambda_1, \lambda_2, \lambda_3$  represents the coefficient used to balance the loss function.

## 4. Experiments

In this section, we first introduced three dataset and the evaluation metrics that we use in this work. Then, we present the implementation details. After that, we compared our method with state-of-the-art methods on the three datasets. Finally, ablation experiment results are presented to illustrate each module’s effectiveness.

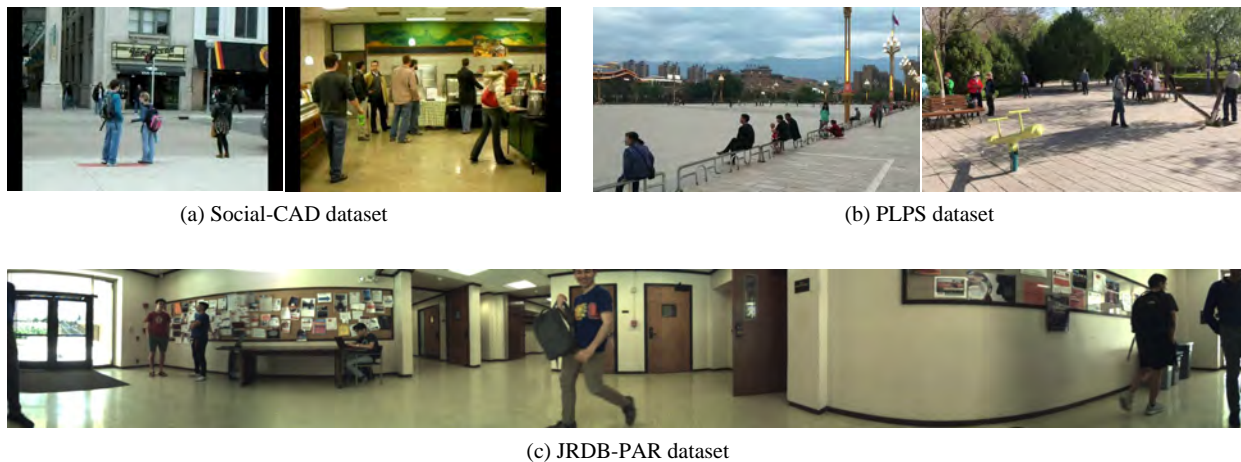


Figure 11: Examples of three datasets.

#### 4.1. Datasets and Metrics

We conducted experiments on three publicly available multi-group activity recognition datasets, namely, the Social Collective Activity Dataset (Social-CAD) [2], the Public Life in Public Space (PLPS) Dataset [3] and the JRDB-Panoramic Human Activity Recognition (JRDB-PAR) Dataset [4]. Sample illustrations of the three datasets are presented in Fig. 11. We also note that Tamura et al. [42] extended the original single-group volleyball dataset to a multi-group setting with group annotations. In contrast to Social-CAD, PLPS, and JRBD-PAR, their work focuses on game activities rather than social contexts, resulting in a different group partitioning rule. Additionally, Tamura et al. did not release their data annotations. Our work, on the other hand, emphasizes interactions in social contexts, and the trajectories can reflect social interactions in social contexts rather than games. Therefore, we conducted experiments for three sub-tasks on the first three datasets: 1) group clustering, 2) group activity recognition for each group and 3) individuals' action recognition. We introduce these datasets and evaluation metrics below.

##### 4.1.1. Collective Activity Dataset (Social-CAD)

Social-CAD is an extension of the Collective Activity Dataset (CAD), originally used for Single-Group Activity Recognition. As the first dataset for MGAR, it serves as a foundation for this research area. It includes 44 short video sequences from 6 group activities and 6 individual actions (NA, Crossing, Waiting,

Queueing, Walking and Talking). In addition, it divides different sub-groups according to their social interactions and assigns group activity labels to all sub-groups. We followed the same evaluation scheme as [2] and allocated 1/3 of the video sequences for testing and used the remaining portion for training [2].

For the Social-CAD experiments, we adopted the same evaluation metric as [2]. For (1), we calculate the membership accuracy by measuring the accuracy of predicting the assignment of each individual to a specific social group, formulated as follows:

$$GC - Acc = \max_m \frac{\sum_{i=1}^N 1[l_i = m(c_i)]}{N}, \quad (18)$$

where  $N$  is the total number of persons,  $l_i$  and  $c_i$  denote the ground truth and prediction results, respectively, and  $m(\cdot)$  represents all possible assignments. For (2), we evaluated whether a person’s membership and social activity label were jointly right. If so, we regarded this instance as a true positive; otherwise, it was false. Therefore, the social accuracy was written as the ratio between the number of true positives and the number of predictions. For (3), a correctly predicted individual’s action was regarded as a true positive, expressed as:

$$Acc_{\text{action}} = \frac{\sum_{i=1}^s TP_i}{\sum_{i=1}^s (TP_i + FN_i)}, \quad (19)$$

where  $s$  is the number of action categories  $TP_i$ , and  $FN_i$  denote the number of correct and wrong predictions for the  $i$ -th type of action.

#### 4.1.2. Public Life in Public Space Dataset (PLPS)

PLPS dataset focuses on more complex scenes in urban public spaces, offering a diverse range of scenarios for MGAR. It includes 71 clips from public spaces, with eleven individual action labels, group activity labels (NA, Queueing, Waiting, Crossing, Talking, Dancing, Doing Sport, Sitting, Jogging, Playing, Riding), and group clustering labels. We follow the same evaluation scheme as [3] and the train-test split in [3].

For the PLPS experiments, we utilized the same evaluation metric as [4]. For (1), assuming that  $N$  persons are present, we use an  $N \times N$  matrix  $M$  to show individual relations, where  $M(i, j) = 1$  denotes

that the  $i$ -th and  $j$ -th individuals are in the same group and  $M(i, j) = 0$  by default. Accordingly,  $M_g$  and  $M_p$  represent the ground truth and prediction, respectively, and the accuracy of group clustering can be calculated using Eq. 20-21:

$$Acc_{\text{cluster}} = \frac{\sum_{i=1}^N \sum_{j=1}^N \delta \{M_g(i, j) = M_p(i, j)\}}{N \cdot N}, \quad (20)$$

$$\delta(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases}. \quad (21)$$

For (2), since every individual was tagged with an activity label, its calculation was the same as that of individual action recognition. For (3), the calculation was the same as that for the individuals' action recognition in the Social-CAD experiments.

#### 4.1.3. JRDB-Panoramic Human Activity Recognition (JRDB-PAR)

JRDB-PAR was collected by a social robot, simulating a human's point of view, providing a unique perspective for activity recognition. It contains 27 videos, which are split into 20 for training and 7 for testing, following the training/validation splitting in the JRDB dataset [4]. It includes 27 categories of individual actions, e.g., walking, and talking, which is the same as JRDB-Act [43], 11 categories of social group activities, 7 categories of global activities, and group clustering labels. Note that individual action and activity annotations are multi-class labels; i.e., each individual/group is assigned one or more labels.

In the JRDB-PAR experiments, we utilized the same evaluation metric as [4]. For (1), we used the classical Half metric used in the group detection task, which was denoted as IOU@0.5. Then, we calculated the AUC (Area Under the Curve) score as the metric, namely, the IOU@AUC. In addition, we computed the matrix IOU score as Mat.IOU as Eq. 22:

$$\text{Mat.IOU} = \frac{\sum AND(R, \tilde{R})}{\sum OR(R, \tilde{R})}, \quad (22)$$

where  $R$  and  $\tilde{R}$  are the ground truth and the predicted binary human group relation matrix, respectively,



with 1 denoting two subjects in the same group. *AND* and *OR* represent the functions of element-wise logical and/or operations.

For (2), we regarded the group member  $\text{IoU} > 0.5$  in the predicted group and the ground-truth group as the true detected group. It is important to note that we only accounted for groups that included more than one individual. The true detected groups, where the activity category prediction was correct, were considered as accurate predictions for social group activities. Then, we calculate the precision, recall and  $F_1$  score as the social activity recognition metrics. For (3), we adopted the same metrics as those used for task (2), namely, precision, recall and the  $F_1$  score as evaluation metrics.

#### 4.2. Implementation Details

We run our proposed framework using Pytorch on Ubuntu 20.04 with a single NVIDIA GeForce RTX 3090. We keep the original frame size in the Social-CAD dataset, and resize each frame to  $240 \times 1880$  for the JRDB-PAR dataset and  $540 \times 660$  for the PLPS dataset. The channels of the feature map output by the backbone corresponded to 1024 dimensions. In addition, we extracted 256-dimensional position features as a supplement to the original features.

We trained the network in two stages. First, we trained the network without reasoning modules, which meant that we only extracted the original features as individual features and max-pooled the features in each sub-group as small-group activity features. Specifically, we used ground-truth group detection results in this stage. We then fine-tuned the whole network in the second stage. Adam was adopted to learn the parameters. The learning rate for all the datasets was 0.0005, and the training batch size was 8 with 200 epochs. For Social-CAD, we set  $\lambda_1 = 1.0$ ,  $\lambda_2 = 8.0$ ,  $\lambda_3 = 2.0$ . For the PLPS dataset, we set  $\lambda_1 = 1.0$ ,  $\lambda_2 = 5.0$ ,  $\lambda_3 = 1.0$ . For the JRDB-PAR dataset, we set  $\lambda_1 = 1.0$ ,  $\lambda_2 = 2.0$ ,  $\lambda_3 = 1.0$ .

#### 4.3. Comparisons with State-of-the-Art Methods

To verify the effectiveness of our proposed method, we compared our experimental results with the results obtained by other reported mainstream methods on three widely used datasets.

Table 1: Experimental results on Social-CAD.

Methods	Group Clustering Accuracy(%)	Group Activity Accuracy(%)	Individual Action Accuracy(%)
JLSG [2]	83.0	69.0	<b>83.3</b>
Ours	<b>83.8</b>	<b>69.6</b>	83.1

#### 4.3.1. Performance on the Social-CAD Dataset

The performance of the proposed method on Social-CAD is described in Table 1. Since only JLSG [2] has conducted experiments on multi-group activity recognition on Social-CAD, we only compared our results with those obtained by JLSG in Table 1. The proposed method performed better for the group clustering task and the subgroup activity recognition task (83.8% and 69.6%, respectively). JLSG only employs GAT to capture the interactions among individuals to obtain individual action and group activity recognition results and uses the GAT coefficient for spectral clustering to obtain group detection results. However, we utilized graph structure to reason about the inter-group and intra relations and introduced local-global information (i.e., neighbors’ information and spatio-temporal trajectory) to obtain better results. However, the individual action recognition result was 0.2% lower than that of JLSG. This is because our method may provide more global information and ignore the analysis of refined individual features. Fig. 12 shows the confusion matrix of our method. We can see that “Walking” and “Crossing” are easily confused, as both activities involve walking and are divided into two activities only because of the different scenes. Besides, “NA” is confused with other activities because of the small number of samples used for this activity. However, overall, our method achieves better results.

Fig. 12 illustrates the results of our method. The people in the same group are placed in purple boxes, with sub-group activity results at the top.

#### 4.3.2. Performance on the PLPS Dataset

The performance of the proposed method on the PLPS dataset is described in Table 2. Since only PLPS [3] has conducted experiments on multi-group activity recognition on the PLPS dataset, we only compared results with PLPS in Table 2. The group clustering accuracy, Group Activity Accuracy and Individual

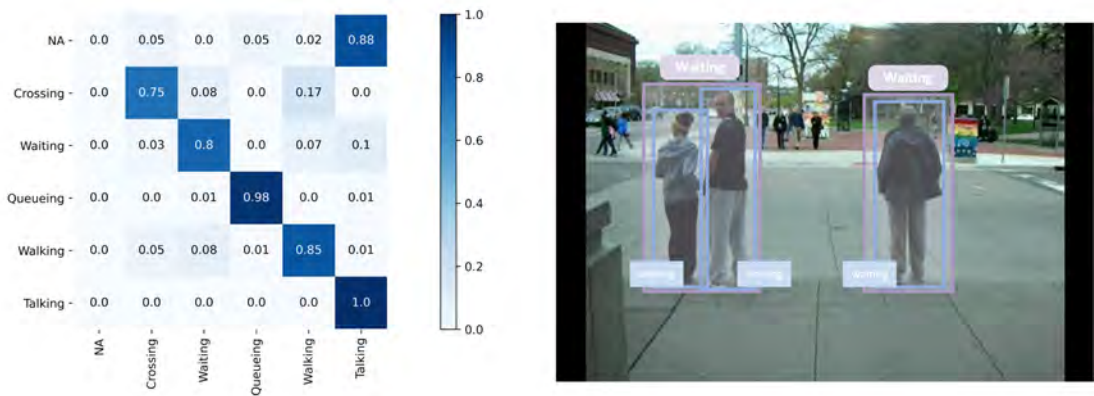


Figure 12: Confusion matrix (left) and predict results (right) on the Social-CAD datasets.

Table 2: Experimental results on the PLPS dataset.

Methods	Group Clustering Accuracy(%)	Group Activity Accuracy(%)	Individual Action Accuracy(%)
PLPS [3]	64.17	55.04	51.07
Ours	<b>77.72</b>	<b>76.70</b>	<b>73.00</b>

Action Accuracy of the proposed method were 77.72%, 76.70% and 73.00%, respectively, which were better than those of compared methods. PLPS uses 2D-CNN to extract features and focuses on inferring the relationships between actors; thus it ignores temporal information and cannot deeply mine the relationships between individuals and groups. In our method, we employed 3D-CNN to better capture temporal features and designed reasoning modules to explore interactions between individuals and groups to infer group-level features. In the group clustering module, PLPS uses only action similarity between actors for group detection without considering global information, whereas we use both types of information in our method. Fig. 13 presents the confusion matrix and the predict results of our method. We can observe that two riding persons belong to the same group with the activity of “riding” and that the person who is isolated from others is crossing.

#### 4.3.3. Performance on the JRDB-PAR dataset

The performance of the proposed method on the JRDB-PAR dataset is described in Table 3. Our results outperform those of the SOTA methods on the JRDB-PAR dataset. As mentioned in the introduction

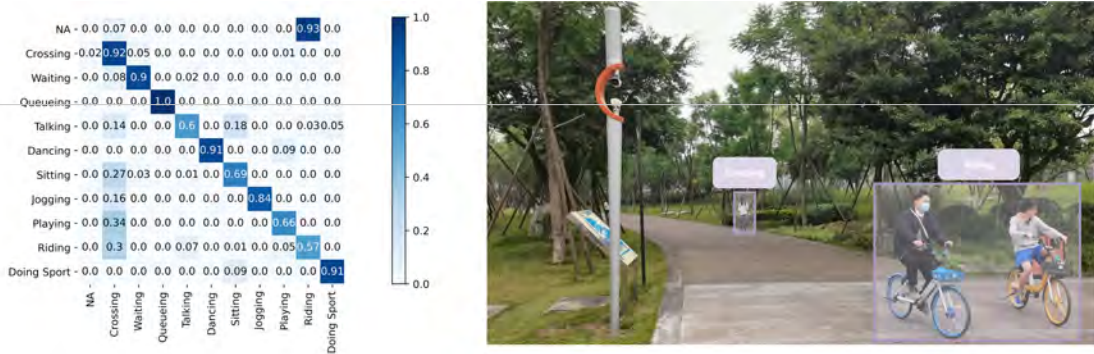


Figure 13: Confusion matrix (left) and predict results (right) on the PLPS datasets.

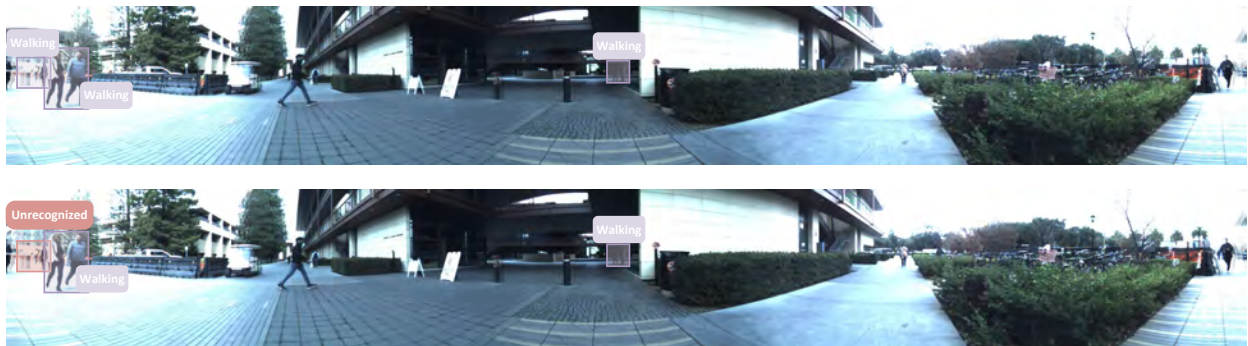


Figure 14: Experimental results for the JRDB-PAR datasets: predicted results (bottom) and ground truths (top).

of this dataset, JRDB-PAR is a panoramic video dataset with multiple labels for individual actions and group activities. Our method not only fits standard video datasets but also performs well on panoramic datasets. The visualization of the experimental results is shown in Fig. 14. The challenges of this dataset are the movement of the camera, multiple labels, and the small number of people in panoramic views. Camera movement and the small size of people hinder feature extraction, and multiple labels with actions and activities result in the subtle reasoning and exploration between people and groups. However, our proposed method focuses on the static view of individual action and group activity recognition with single labels. Although our method outperforms other models because of its effective spatio-temporal interactive reasoning, visualization shows that some additional effort is needed to refine our approach and solve the challenges mentioned above.

Table 3: Experimental results on JRDB-PAR dataset.

Methods	Group Clustering Accuracy(%)			Group Activity Accuracy(%)			Individual Action Accuracy(%)		
	IOU@0.5	IOU@AUC	Mat.IOU	P	R	F	P	R	F
ARG [25]	35.2	21.6	19.3	8.7	8.0	8.2	39.9	30.7	33.2
JLSG [2]	29.1	20.4	16.6	8.8	8.9	8.8	44.8	40.4	40.3
JRDB-Base[43]	38.4	26.3	20.6	14.3	12.2	12.8	19.1	34.4	23.6
JRDB-PAR [4]	53.9	38.1	30.6	24.7	26.0	24.8	51.0	40.5	43.4
MUP [6]	-	-	-	25.4	26.6	25.1	55.4	44.8	47.7
MPT [7]	-	-	-	25.5	27.3	25.4	<b>59.2</b>	<b>58.6</b>	<b>56.0</b>
Ours	<b>65.1</b>	<b>39.6</b>	<b>51.8</b>	<b>29.0</b>	<b>28.6</b>	<b>28.7</b>	55.1	43.1	46.5

Notably, our method outperforms both MUP [6] and MPT [7] in terms of group activity accuracy, but falls short in individual action accuracy. The slight decrease in accuracy relative to MUP is attributed to our smaller input frames ( $240 \times 1880$ ) compared to the original size ( $480 \times 3760$ ), which is due to computational resource constraints. The superior performance of MPT in individual action recognition is due primarily to its design, which focuses on modeling the temporal dynamics of each individual using a temporal encoder. By contrast, our method prioritizes capturing the spatio-temporal interactions among individuals and social groups, particularly with the augmentation of trajectory data. Consequently, our method achieves a significant improvement in group activity recognition and group clustering, which is the primary focus of the multi-group activity recognition task.

#### 4.4. Ablation Study

We also designed the following ablation experiments to illustrate each module’s effectiveness.

**B1 Baseline:** This baseline employs 3D-ResNet50-NonLocal [37] for feature extraction without any further reasoning modules. In the individual action recognition part, we add only the SENet [38] attention mechanism to suppress the less influential channels and then send the features to the classifier. In the clustering module, we utilized only the similarity of appearance features between individuals. For the multi-group activity recognition module, we used the max-pooling strategy on individual features and obtained

the final results while considering the group detection results.

**B2 Baseline+IGR:** Inter-Actor Relation Graph Reasoning Module is added to the baseline to explore individual interactions further.

**B3 Baseline+IGR+STC:** This baseline adds the Spatio-Temporal Clustering Module, which introduces the similarity of action features between individuals and spatio-temporal interpersonal distances to hunt for deep cues for group detection.

**B4 Baseline+IGR+STC+NNIE:** This baseline adds Nearest Neighbor Interaction Extraction Module to reason about individual interactive appearance, location and distance features.

**B5 Baseline+IGR+STC+NNIE+GTGR:** This baseline introduces Inter-Trajectory Relation Graph Reasoning Module to obtain global information for group interaction reasoning.

**Ours Baseline+IGR+STC+NNIE+GTGR+GAPM:** Our full model is equipped with Graph Aggregation and Pooling Module to exploit relationships within and between groups and focus on key actors in the same group to generate group-level features.

As shown in Table 4 and Table 6, we take the Social-CAD dataset as an example for analysis. B1 integrated only individual appearance features as multi-group activity features. It only uses the similarity of appearance for the clustering process, thus only reaching 70.2%, 57.9% and 81.4% for group clustering accuracy (GCA), group activity accuracy (GAA) and individual action accuracy (IAA), respectively. Compared with B1, B2 adds the Individual Graph Reasoning Module to explore their relationships, increasing IAA by 1.3%. B3 added spatio-temporal trajectory features for clustering, leading to a large improvement in GCA from 70.5% to 83.5%. Because a person’s activity is closely related to his/her neighbors, B4 is equipped with Nearest Neighbor Interaction Extraction Module to reason about individual interactive appearance, location and distance features, increasing GAA by 1.5%. B5 adds the global trajectory information to exploit the relationships among individuals, increasing GAA from 68.8% to 68.9%. Finally, the whole model introduces the Graph Aggregation and Pooling Module based on B5, which not only reasons relationships within and between groups but also focuses on key actors in the same group to infer group-level features. The final results of GCA, GAA and IAA reached 83.8%, 69.6% and 83.1%, respectively, validating our method’s

Table 4: Results of ablation experiments on the Social-CAD (column 2-4) and the PLPS datasets (column 5-7).

Methods	GCA(%)	GAA(%)	IAA(%)	GCA(%)	GAA(%)	IAA(%)
B1: Baseline	70.2	57.9	81.4	68.00	67.23	70.56
B2: Baseline+IGR	70.5	58.1	82.7	68.32	68.74	71.53
B3: Baseline+IGR+STC	83.5	67.3	82.7	77.65	68.52	71.18
B4: Baseline+IGR+STC+NNIE	83.5	68.8	82.3	77.60	71.97	72.19
B5: Baseline+IGR+STC+NNIE+GTGR	83.5	68.9	82.1	77.72	73.69	72.97
Ours: Baseline+IGR+STC+NNIE+GTGR+GAPM	<b>83.8</b>	<b>69.6</b>	<b>83.1</b>	<b>77.72</b>	<b>76.70</b>	<b>73.00</b>

Table 5: Results of ablation experiments on the JRDB-PAR dataset.

Methods	GCA(%)			GAA(%)			IAA(%)		
	IOU@0.5	IOU@AUC	Mat.IOU	P	R	F	P	R	F
B1: Baseline	23.2	5.0	23.8	4.0	3.8	3.8	24.4	17.9	19.8
B2: Baseline+IGR	23.8	5.4	24.3	7.1	7.0	7.0	26.7	19.5	21.6
B3: Baseline+IGR+STC	64.7	37.3	51.5	16.5	18.1	16.8	31.1	23.7	25.9
B4: Baseline+IGR+STC+NNIE	64.9	39.5	51.7	24.3	24.0	24.1	37.1	27.7	30.4
B5: Baseline+IGR+STC+NNIE+GTGR	64.9	39.2	51.7	27.2	26.9	27.0	47.5	36.1	39.4
Ours: Baseline+IGR+STC+NNIE+GTGR+GAPM	<b>65.1</b>	<b>39.6</b>	<b>51.8</b>	<b>29.0</b>	<b>28.6</b>	<b>28.7</b>	<b>55.1</b>	<b>43.1</b>	<b>46.5</b>

effectiveness.

Additionally, it is observed that for the PLPS dataset GAA decreases after the addition of the STC module. This decrease is attributed to the imbalanced data of the PLPS dataset, i.e., the long-tail distribution. Notably, the addition of the STC module (B3) yields a significant improvement of 9.33% in GCA, demonstrating its effectiveness. While most groups are correctly clustered with this module, the classifier still struggles to recognize group activities because of the limitations imposed by data imbalance. However, when the subsequent group activity reasoning modules (NNIE, GTGR, and GAPM) are incorporated, this issue is alleviated, and GAA is increased by 8.18%.

To analyze the differences of group activity recognition between social contexts and games, we conducted the ablation experiments on the Volleyball dataset under the single-group setting. As shown in Table 6, the IGR and NNIE modules significantly improve group activity recognition accuracy by enhancing interaction reasoning among individuals. However, the addition of ITRGR results in a decrease in group activity

Table 6: Ablation experiments on the Volleyball dataset under the single-group setting.

Method	Group Activity Accuracy (%)
Baseline	90.052
Baseline+IGR	90.127
Baseline+IGR+NNIE	90.651
Baseline+IGR+NNIE+ITRGR	90.277

accuracy, demonstrating that interactions among individual trajectories differ from those in social contexts and are not effective in games. Intuitively, individuals’ trajectories within a social group are similar but it is not the case in games. Moreover, our designed STC Module and GAPM for multi-group interactions are not suitable for single-group activity recognition in games.

## 5. Conclusion

This paper proposes a new Spatio-Temporal Interactive Reasoning Model (STIRM) for Multi-Group Activity Recognition, which focuses on exploiting spatio-temporal interaction within the same group and across sub-groups and forming rich group-level features. In particular, interactive feature extraction aims to explore spatial interactions between nearest neighbors and individuals themselves. To better exploit spatio-temporal relations between individuals, we combine the interactive action feature and spatio-temporal trajectory feature to divide them into small groups. In addition, a group interaction reasoning module is constructed to exploit rich and accurate group-level representations by reasoning correlations within and between groups and discarding people who have less impact on group activity. Extensive experiments on three datasets demonstrated that our method shows outstanding performance in exploring spatio-temporal interactions and improving accuracy compared with state-of-the-art methods.

In summary, the two primary strengths of this method lie in its ability to reason about interactions among different social groups and its augmentation of visual features with non-vision trajectory data. In the context of human activity understanding, particularly in multi-person scenarios, various interactions



Table 7: Results of comparative experiments on different backbones.

Baseline	Social-CAD			PLPS			JRDB-PAR		
	GCA	GAA	IAA	GCA	GAA	IAA	GCA	GAA	IAA
ResNet18	69.7	51.1	69.5	58.2	18.7	35.5	24.0	3.0	3.6
InceptionV3	69.8	53.9	74.5	59.4	27.9	45.7	23.7	3.7	6.6
Ours: 3D-ResNet50-NonLocal	70.2	58.0	81.4	68.0	67.2	70.6	23.8	3.8	19.8

(e.g., individual and spatio-temporal interactions) are crucial aspects of research. However, previous studies have overlooked the connections between different social groups, which is essential for more accurate activity understanding. By incorporating trajectory data, our work effectively enhances the interactions among social groups, thereby addressing this limitation.

Despite these strengths, our work is not without its weaknesses. A notable limitation is the neglect of background interactions. Individuals interact not only with others but also with objects in their surroundings, such as chairs, tables, and smartphones. These interactions can significantly enhance individual representations and further enrich the interactions among individuals and social groups. As shown in Table 7, although our baseline shows superior performance compared to the others, it exhibits weaknesses in individual action feature extraction, especially when applied to the JRDB-PAR dataset. Consequently, incorporating background information is essential to ameliorate this deficiency.

In our future work, we plan to design a scene-aware multi-level interaction reasoning model for MGAR that is based on complex scene understanding. Additionally, we will consider the correlations between multiple tasks (i.e., individual action recognition, group clustering, and group activity recognition) and the inter-activity dependencies, given the characteristics of multiple activity labels.

## Acknowledgment

The research in our paper is sponsored by the National Natural Science Foundation of China (No. 62301346), the Natural Science Foundation of Sichuan Province, China (No. 24NSFSC3408), the Sichuan

Science and Technology Program, China (No. 2023YFS0195), the Key Research and Development Program of Chengdu, China (No. 2023-YF09-00019-SN) and the Fundamental Research Funds for the Central Universities, China (No. YJ202326).

## References

- [1] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (3) (2023) 3200–3225.
- [2] M. Ehsanpour, A. Abedin, F. Saleh, J. Shi, I. Reid, H. Rezaatofghi, Joint learning of social groups, individuals action and sub-group activities in videos, in: *Computer Vision–ECCV 2020*, 2020, pp. 177–195.
- [3] L. Qing, L. Li, S. Xu, Y. Huang, M. Liu, R. Jin, B. Liu, T. Niu, H. Wen, Y. Wang, et al., Public life in public space (plps): A multi-task, multi-group video dataset for public life research, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3618–3627.
- [4] R. Han, H. Yan, J. Li, S. Wang, W. Feng, S. Wang, Panoramic human activity recognition, in: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 2022, pp. 244–261.
- [5] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, in: *Proceedings of the International Conference on Learning Representations*, 2018.
- [6] M. Cao, R. Yan, X. Shu, J. Zhang, J. Wang, G. Xie, MUP: multi-granularity unified perception for panoramic activity recognition, in: *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7666–7675.
- [7] W. Gan, Y. Sun, F. Liu, X. Luo, Mpt-par: Mix-parameters transformer for panoramic activity recognition (2024). [arXiv: 2408.00420](https://arxiv.org/abs/2408.00420).
- [8] J. Su, J. Huang, L. Qing, X. He, H. Chen, A new approach for social group detection based on spatio-temporal interpersonal distance measurement, *Heliyon* 8 (10) (2022) e11038.
- [9] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, C. Tang, Rgb-d-based action recognition datasets: A survey, *Pattern Recognition* 60 (2016) 86–105.
- [10] Y. Ma, R. Wang, Relative-position embedding based spatially and temporally decoupled transformer for action recognition, *Pattern Recognition* 145 (2024) 109905.
- [11] X. Wang, S. Zhang, J. Cen, C. Gao, Y. Zhang, D. Zhao, N. Sang, Clip-guided prototype modulating for few-shot action recognition, *International Journal of Computer Vision* 134 (2024) 1899–1912.
- [12] H. Qiu, B. Hou, Multi-grained clip focus for skeleton-based action recognition, *Pattern Recognition* 148 (2024) 110188.
- [13] Z. Wu, N. Ma, C. Wang, C. Xu, G. Xu, M. Li, Spatial-temporal hypergraph based on dual-stage attention network for multi-view data lightweight action recognition, *Pattern Recognition* 151 (2024) 110427.

- [14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
- [15] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
- [16] X. Zhang, D. Ma, H. Yu, Y. Huang, P. Howell, B. Stevens, Scene perception guided crowd anomaly detection, *Neurocomputing* 414 (2020) 291–302.
- [17] W. Ge, R. T. Collins, R. B. Ruback, Vision-based analysis of small groups in pedestrian crowds, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (5) (2012) 1003–1016.
- [18] L. Li, L. Qing, L. Guo, Y. Peng, Relationship existence recognition-based social group detection in urban public spaces, *Neurocomputing* 516 (2023) 92–105.
- [19] X. Wang, X. Zhang, Y. Zhu, Y. Guo, X. Yuan, L. Xiang, Z. Wang, G. Ding, D. Brady, Q. Dai, et al., Panda: A gigapixel-level human-centric video dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3268–3278.
- [20] T. Fernando, S. Denman, S. Sridharan, C. Fookes, Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds, in: *Computer Vision—ACCV*, 2019, pp. 314–330.
- [21] A. Akbari, H. Farsi, S. Mohamadzadeh, Deep neural network with extracted features for social group detection, *Journal of Electrical and Computer Engineering Innovations* 9 (1) (2021) 47–56.
- [22] J. Sun, Q. Jiang, C. Lu, Recursive social behavior graph for trajectory prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 660–669.
- [23] Y. Duan, J. Wang, Learning key actors and their interactions for group activity recognition, in: *Pattern Recognition and Computer Vision: 4th Chinese Conference*, 2021, pp. 53–65.
- [24] R. Yan, X. Shu, C. Yuan, Q. Tian, J. Tang, Position-aware participation-contributed temporal dynamic model for group activity recognition, *IEEE Transactions on Neural Networks and Learning Systems* 33 (12) (2021) 7574–7588.
- [25] J. Wu, L. Wang, L. Wang, J. Guo, G. Wu, Learning actor relation graphs for group activity recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9964–9974.
- [26] H. Yuan, D. Ni, M. Wang, Spatio-temporal dynamic inference network for group activity recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7476–7485.
- [27] K. Gavriluyk, R. Sanford, M. Javan, C. G. M. Snoek, Actor-transformers for group activity recognition, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 836–845.
- [28] X. Zhu, Y. Zhou, D. Wang, W. Ouyang, R. Su, Mlst-former: Multi-level spatial-temporal transformer for group activity recognition, *IEEE Transactions on Circuits and Systems for Video Technology* 33 (7) (2023) 3383–3397.
- [29] 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, *IEEE Transactions on Circuits and Systems for Video Technology* 30 (2) (2020) 549–565.

- [30] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, G. Mori, A hierarchical deep temporal model for group activity recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1971–1980.
- [31] R. Yan, L. Xie, J. Tang, X. Shu, Q. Tian, Hgcin: Hierarchical graph-based cross inference network for group activity recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
- [32] Z. Du, X. Wang, Q. Wang, Self-supervised global spatio-temporal interaction pre-training for group activity recognition, *IEEE Transactions on Circuits and Systems for Video Technology* (2023).
- [33] N. V. Chappa, P. Nguyen, A. H. Nelson, H.-S. Seo, X. Li, P. D. Dobbs, K. Luu, Spartan: Self-supervised spatiotemporal transformers approach to group activity recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5157–5167.
- [34] K. Mao, P. Jin, Y. Ping, B. Tang, Modeling multi-scale sub-group context for group activity recognition, *Applied Intelligence* 53 (1) (2023) 1149–1161.
- [35] 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, *Pattern Recognition* (2024) 110478.
- [36] M. Perez, J. Liu, A. C. Kot, Skeleton-based relational reasoning for group activity analysis, *Pattern Recognition* 122 (2022) 108360.
- [37] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.
- [38] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
- [40] E. Ranjan, S. Soumya, P. P. Talukdar, Asap: Adaptive structure aware pooling for learning hierarchical graph representations, in: 2020 AAAI Conference on Artificial Intelligence, 2020, pp. 5470–5477.
- [41] J. Lee, I. Lee, J. Kang, Self-attention graph pooling, in: *International Conference on Machine Learning*, 2019, pp. 3734–3743.
- [42] M. Tamura, R. Vishwakarma, R. Vennekanti, Hunting group clues with transformers for social group activity recognition, in: *Computer Vision – ECCV 2022*, Cham, 2022, pp. 19–35.
- [43] M. Ehsanpour, F. Saleh, S. Savarese, I. Reid, H. Rezaatofghi, Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20983–20992.