Spatio-Temporal Proximity-Aware Dual-Path Model for Panoramic Activity Recognition

Sumin Lee ⬤, Yooseung Wang, Sangmin Woo ⬤, and Changick Kim

Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
{suminlee94, yswang, smwoo95, changick}@kaist.ac.kr

Abstract. Panoramic Activity Recognition (PAR) seeks to identify diverse human activities across different scales, from individual actions to social group and global activities in crowded panoramic scenes. PAR presents two major challenges: 1) recognizing the nuanced interactions among numerous individuals and 2) understanding multi-granular human activities. To address these, we propose Social Proximity-aware Dual-Path Network (SPDP-Net) based on two key design principles. First, while previous works often focus on spatial distance among individuals within an image, we argue to consider the spatio-temporal proximity. It is crucial for individual relation encoding to correctly understand social dynamics. Secondly, deviating from existing hierarchical approaches (individual-to-social-to-global activity), we introduce a dual-path architecture for multi-granular activity recognition. This architecture comprises individual-to-global and individual-to-social paths, mutually reinforcing each other's task with global-local context through multiple layers. Through extensive experiments, we validate the effectiveness of the spatio-temporal proximity among individuals and the dual-path architecture in PAR. Furthermore, SPDP-Net achieves new state-of-the-art performance with 46.5% of overall F1 score on JRDB-PAR dataset.

Keywords: Panoramic activity recognition · Social group activity detection · Human activity understanding

1 Introduction

Understanding human activity in videos is a pivotal task in computer vision. It finds diverse real-world applications across various domains, including video sports analysis [20,34,40], surveillance [26,35], and social scene analysis [9,21,50]. Most primary research has focused on recognizing behaviors at different granularity levels, such as Human Activity Recognition (HAR) [22,25,42,45], Group Activity Recognition (GAR) [5,18,47,49], and Panoramic Activity Recognition (PAR) [3,13]. HAR systems concentrate on identifying the actions of a single individual in well-defined settings. The task of GAR extends the scope of HAR by focusing on understanding interactions and collective activities performed by multiple individuals in a shared space. Building on the concept of GAR, PAR tackles an even more comprehensive problem to understand human behavior in

Fig. 1: Importance of the spatio-temporal proximity for understanding social group dynamics. To distinguish between social groups, it is crucial to leverage positional relationships among individuals not just in space but also over time. Consider an initial scene where individuals marked with red, yellow, and green bounding boxes are close to each other, giving the impression that they belong to the same social group. However, as time goes on, it becomes evident that only the individuals in the red and yellow boxes move together, indicating shared social group membership, while the person in the green box does not.

complex environments. PAR aims to recognize social group activities as well as individual and collective activities in a panoramic scene. The objectives of PAR involve recognizing multi-granular activities through the three tasks: (i) individual action recognition, (ii) social group activity recognition coupled with social group detection, and (iii) global (panoramic) activity recognition.

PAR presents a unique set of challenges. Firstly, due to the intricate nature of panoramic scenes, those scenes often contain numerous individuals engaged in diverse activities and interactions within a wide field of view. For this reason, previous works [3,13] utilized spatial proximity in a single frame for determining social relations between individuals. However, we argue that relying solely on spatial proximity is insufficient; it is imperative to incorporate spatio-temporal proximity in PAR. For example, as depicted in Fig. 1, the three individuals in the first frame seem to belong to the same social group due to their closeness. A few frames later, the two in the yellow and red bounding boxes are moving together, whereas the individual in the green box is not. We can see that only two in the yellow and red boxes share social membership, while the one in the green box does not. Therefore, it is essential to consider spatial position variations over time to accurately assess the social proximity for precise understanding of social dynamics. Another challenge lies in comprehending the multi-granular activities ranging from simple individual actions to complex social group and global activities. Previous works [3, 13] hierarchically model three granular activities: 'from individuals to groups', and 'from groups to global'. From these studies, we observe that both global and social group activities require individual contextual information, while they also mutually influence each other. We empirically validate this observation (see Sec. 4.3).

To address these problems, we propose a novel network, called Social Proximity-aware Dual-Path Network (SPDP-Net). Our SPDP-Net consists of two stages: 1) individual relation encoding and 2) multi-granular activity recognition. In the individual relation encoding stage, spatio-temporal positional relationships among individuals are considered to refine their feature representations, enabling the precise measurement of social proximity. This allows SPDP-Net to capture ac-

curate social dynamics within a panoramic scene. Specifically, a spatio-temporal self-attention mechanism is employed on the features of each individual to emphasize crucial visual clues such as related objects and body posture. To further enhance this process and account for overall contextual information within a panoramic scene, we incorporate a panoramic positional embedding that encapsulates the spatio-temporal positional information of each individual across the entire scene. Additionally, we evaluate the social interactions among individuals for social group detection by measuring the feature similarity and the social proximity relation. To accurately measure social proximity among individuals, we extend the concept of GIoU to the temporal axis, named Temporal Generalized IoU (GIoU). For the multi-granular activity recognition, we introduce Dual-Path Activity Transformer (DPATr). Different from previous works that model multi-granular activities hierarchically [3, 13], each layer of DPATr consists of two paths, individual-to-global and individual-to-social paths. The individual-to-global path concurrently encodes individual and global activities, exploiting the global-local context of the given video. Then, the individual-to-social path explores social group activity information from globally attended individual features. Through multiple layers, DPATr mutually reinforces contextual understandings of multi-spatial activities, creating synergistic effects that enhance final predictions.

To evaluate SPDP-Net, we conduct extensive experiments on JRDB-PAR [13]. Throughout the performance improvements in both human activity recognition and social group detection, we demonstrate the effectiveness of utilizing spatio-temporal proximity to model social dynamics in a wide crowded scene. Moreover, we show that our DPATr mechanism cooperatively strengthens the tasks of PAR by adeptly capturing the global-local contextual information. Notably, the proposed SPDP-Net significantly outperforms the state-of-the-art methods by a large margin, achieving 46.4% of an overall F1 score for activity recognition and 56.4% of IoU@0.5 for social group detection.

Our main contributions are summarized as follows:

- In this paper, we propose a novel network, named SPDP-Net. For comprehensively understanding social dynamics in a wide crowded scene, SPDP-Net leverages spatio-temporal proximity among individuals by utilizing a panoramic positional embedding and TGIoU.

- Moreover, DPATr in SPDP-Net mutually reinforces collaboration among multi-granular activity tasks by utilizing a global-local context with a dual-path architecture.

- Throughout extensive experiments, we demonstrate the effectiveness of the proposed method. Furthermore, SPDP-Net significantly outperforms the state-of-the-art method, achieving 46.5% in the overall F1 score for activity recognition and 56.4% in IoU@0.5 for social group detection.

2   Related Work

Human Action Recognition. Comprehending human actions is a fundamental task in the field of video understanding. The task of action recognition aims to recognize the action classes performed by a single actor in a given trimmed video [51]. Earlier works focus on utilizing optical flow in two-stream architecture [11,42]. Thanks to the recently improved capacity of deep learning architecture, 3D convolution networks [4,10,38,43] or transformer-based networks [1,27] are proposed, which lead to significant performance improvements. ViViT [1] extracts visual appearance and pose features from tubelet embeddings of an input video with transformer layers [39]. Following these works, we also focus on understanding human. However, our focus lies in analyzing multi-granular human activity in videos with multiple individuals, necessitating consideration of social interactions among individuals.

Group Activity Recognition with Human Interaction. Group Activity Recognition (GAR) aims to identify the activity performed by a collective of individuals within a given video. Unlike recognizing actions of single individual, GAR requires exploring the aggregation of these actions to reveal the collective behavior of the entire group. Many studies incorporate the individual action labels as an auxiliary supervision [7, 12, 23] to model the relationships among multiple actors for richer representations of group dynamics [16–18,49,52]. Graph-based network [46, 53] and Transformer-based network [12, 15, 18, 24, 37] are two common approaches for constructing potential interactions among individuals. Actor-transformer [12] selectively exploits group activity information from actor-specific static and dynamic representations. Li *et al.* [24] introduce Groupformer, which is a spatio-temporal transformer with clustered attention to jointly augment the individual and group representations. These studies show the effectiveness of transformer architecture in capturing the relationships between individual actions and collective activities. In contrast to these works that focus on identifying collective activities, we propose Dual Path Activity Transformer (DPATr) for recognizing multi-granular activities including individual actions, social group activities, and global activities.

Social Activity and Group Detection. The goal of social activity and group detection is to divide crowds of people into distinct social groups as well as to understand the social activities of these groups. While initial methods [2,33] focus solely on detecting the social group, recent works have expanded to recognize the social activities with new datasets [7,8,28,32]. Some works [7,8] model relationships of individuals with adjacency matrices in a panoramic scene. In these works, individual visual features cropped from a panoramic image are used for graph clustering [54] and aggregated to social features for detecting social group activities. Contrary to these approaches solely relying on cropped visual features, we argue that the proximity among individuals over both space and time is vital for discerning group dynamics and individual interactions. Considering this, our
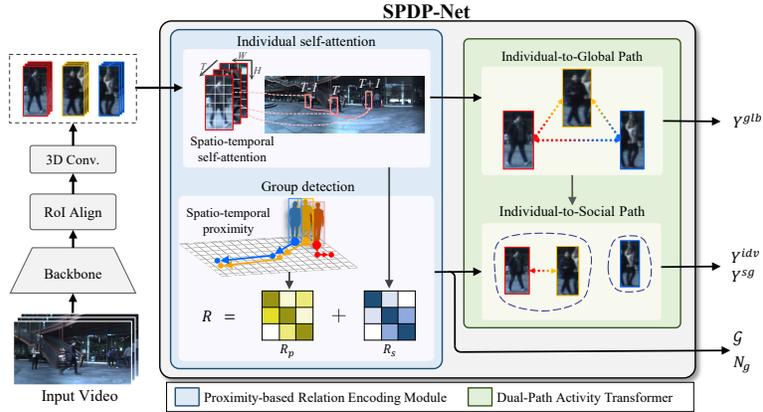
Fig. 2: Overview of the proposed SPDP-Net. SPDP-Net consists of two stages: 1) proximity-based relation encoding and 2) multi-granular activity recognition. T_0 indicates the center frame of a given video.

SPDP-Net adeptly utilizes spatio-temporal positional relationships among individuals to capture the positional context of cropped individual regions within an entire panoramic scene.

Panoramic Activity Recognition. Recently, Panoramic Activity Recognition (PAR) has gained attention in the field of video understanding. The PAR task involves recognizing what individuals are doing (individual actions), how groups of people interact (social group activities), and what is happening in an overall panoramic video (global activity). Han *et al.* [13] introduced a hierarchical graph network based on GCN. This allows for nodes at various levels to correspond to individual, group, and global activities, sequentially. Following this, Cao *et al.* [3] proposed Multi-granularity Unified Perception (MUP) framework, which is hierarchically applied to facilitate aggregation and interpretation across the spectrum of granularity. Those works utilize both individual and group-level features for global activity recognition, while a reverse stream that uses global features for social group-level and individual-level is employed in [13]. From these, we can see the importance of individual contextual information for both global and social group activities and their mutual influence on each other. To promote active interactions among multi-granular activities, DPATr consists of individual-to-global and individual-to-social paths. This architecture explores global-local context of a panoramic scene by using of individual context for social group and global activity recognition.

3 Proposed Method

Overview. SPDP-Net aims to exploit spatio-temporal proximity representations of individuals with global-local contexts in a panoramic scene for PAR

task. The overview of SPDP-Net is illustrated in Fig. 2. Given a panoramic video of length T, a 2D CNN backbone extracts frame-wise features. To crop the appearance representations of individuals, we apply RoIAlign [14] to each frame feature, followed by a 3D convolution for dimension reduction. We denote the resulting feature for the entire spatio-temporal individual features as $F^{idv} \in \mathbb{R}^{N_i \times T \times d \times h \times w}$. Here, N_i denotes the number of individuals, d is the hidden dimension, and h and w indicate the height and width of the cropped features, respectively.

SPDP-Net consists of two stages: 1) proximity-based relation encoding and 2) multi-granular activity recognition. In the proximity-based relation encoding stage, spatio-temporal positional relationships are leveraged to capture the positional context of cropped individual regions in a panoramic scene. Additionally, we measure the feature similarity and spatio-temporal proximity among individuals to understand social relationships among individuals. For the multi-granular activity recognition, we introduce Dual-Path Activity Transformer (DPATr), which models dependencies across three levels of granular activities with the individual-to-global and individual-to-social paths. From activity features computed in DPATr, dedicated classifiers for each granularity level predict activity scores: $Y^{idv} \in \mathbb{R}^{N_i \times C_{idv}}$ for individual activities, $Y^{sg} \in \mathbb{R}^{n_g \times C_{sg}}$, for social group activities, and $y^{glb} \in \mathbb{R}^{C_{glb}}$ for global activity, where C_i denotes the number of activity classes at each level.

3.1   Proximity-based Relation Encoding

In this stage, the goal is to extract rich action information by considering the spatio-temporal proximity among individuals and to accurately perform social group detection. The detailed mechanism of this stage is described in Fig. 3a. Firstly, we apply self-attention to the spatio-temporal individual feature F^{idv}. The initial step is to apply spatio-temporal self-attention to each individual feature. Motivated by [41], we employ a Multi-Head Self-Attention (MHSA) along the temporal, height, and width dimensions in sequence. A key challenge here is preserving the positional context of individuals within a wide panoramic scene. However, conventional positional embeddings are inadequate in this aspect, since they only encode positional information within cropped regions. To address these, we incorporate the panoramic positional embedding, denoted as e_{pn}, specifically designed to retain the positional integrity of individuals within the panoramic scene. We derive e_{pn} by extracting individual regions from the sinusoidal positional embedding of the entire scene. The self-attended features of individuals $\bar{F}^{idv}$ are calculated as follows:

$$\bar{F}^{idv} = A^w \left(A^h \left(A^t \left(F^{idv}, e_{pn} \right), e_{pn} \right), e_{pn} \right) + F^{idv}, \tag{1}$$

where A^t, A^h, and A^w indicate MHSA across the temporal, height, and width axes, respectively.

To detect social groups within the scene, our approach takes into account both the individual-specific features $\bar{F}^{idv}$ and the bounding box coordinates
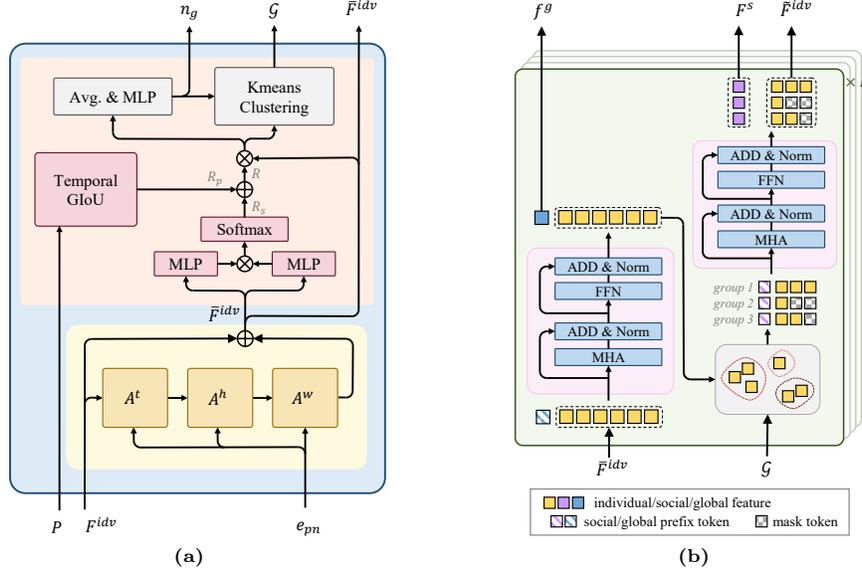
Fig. 3: Detailed architecture of two stages in SPDP-Net. (a) Proximity-based relation encoding and (b) multi-granular activity recognition (*i.e.*, DPATr).

$P \in \mathbb{R}^{N_i \times T \times 4}$. In practice, we measure the social connections between individuals by constructing a social relation matrix R based on the visual similarities, denoted as R_s, and the spatio-temporal proximity of individuals, represented as R_p. We define the visual similarity matrix as $R_s = \text{Softmax}\left(\mathbf{W}_\theta \bar{F}^{idv}(\mathbf{W}_\phi \bar{F}^{idv})^\top\right)$, where $\mathbf{W}_\theta$ and $\mathbf{W}_\phi$ are trainable matrices. In addition to the visual similarity, exploiting positional relationships among individuals not just in space but also over time is essential for accurately modeling social relationships. For this purpose, we introduce the social proximity relation matrix R_p, defined with Temporal Generalized IoU (TGIoU), as follows:

$$R_p(i,j) = \text{TGIoU}(P^i, P^j) = \frac{1}{T} \sum_{t=1}^{T} \text{GIoU}(p_t^i, p_t^j). \tag{2}$$

Here, $p_t^i \in [0,1]^4$ is a vector representing top-left and bottom-right coordinates of a bboxes and $B(p_t^i, p_t^j)$ means the smallest box containing p_t^i and p_t^j. Since GIoU [31] provides a thorough measure of spatial separation and overlap, TGIoU serves as an informative metric for evaluating physical individual relationships across time. The final social relation matrix is computed as $R = 1/2(R_s + R_p)$.

To predict the normalized number of social groups within the scene, we apply a linear layer on the individual features attended by the relation matrix, as $n_g = MLP\left(Avg(R\bar{F}^{idv})\right)$. The actual number of social groups N_g is then calculated by multiplying N_i with n_g. Finally, the social group identifier $\mathcal{G} \in [1, N_g]^{N_i}$

are determined via K-means clustering on the relation-augmented features, as $\mathcal{G} = \text{KMeans}\left(R\bar{F}^{idv}, N_g\right)$.

3.2   Multi-Granular Activity Recognition

For this stage, DPATr is designed to effectively capture human activity features across multiple granularities by modeling the dependencies among individual, social group, and global contextual information. As illustrated in Fig. 3b, DPATr has L layers. Each layer based on transformer encoder blocks [39] has a dual-path architecture: individual-to-global activity and individual-to-social activity paths. The first path takes the individual features $\bar{F}^{idv}$ as an input. To this input, a learnable global token is prepended. By computing local individual patches and the global token, it facilitates explorations of not just inter-individual interactions among individuals but also global context within the overall scene. To consider social group activities, the individual features obtained from the first path are grouped based on $\mathcal{G}$ and organized into sequences. Then, each cluster is prefixed with a learnable social token to capture the dynamics of each social group. In the individual-to-social activity path, these social tokens are specifically trained to accurately predict their respective social group activities by exploring global-local context from the individual features. Through multiple layers, DPATr cooperatively promotes three tasks to create synergistic effects that enhance final activity features of individuals ($\tilde{F}^{idv} \in \mathbb{R}^{N_i \times d}$), social groups ($F^{sg} \in \mathbb{R}^{N_g \times d}$), and global ($F^{glb} \in \mathbb{R}^d$).

3.3   Training

For training SPDP-Net, we use a total loss function defined as follows:

$$\mathcal{L} = \mathcal{L}_{idv} + \mathcal{L}_R + \mathcal{L}_{aux} + \lambda_{sg}\mathcal{L}_{sg} + \lambda_{glb}\mathcal{L}_{glb} + +\lambda_n\mathcal{L}_n, \tag{3}$$

where λ represents the balancing hyperparameters. Here, $\mathcal{L}_{idv}$ indicates individual action loss function, $\mathcal{L}_R$ is the loss function for the social relation matrix in group detection, and $\mathcal{L}_{aux}$ is an auxiliary loss for individual action recognition based on $\bar{F}^{idv}$. $\mathcal{L}_{sg}$ and $\mathcal{L}_{glb}$ indicate the losses for the social group and global activity recognition, respectively. We apply $L2$ loss for $\mathcal{L}_n$ term, which addresses the estimation of the number of social groups, and use binary cross-entropy loss for the rest. The ratio of the losses is $\lambda_{sg} : \lambda_{glb} : \lambda_n = 3 : 2 : 5$.

4   Experiments

4.1   Dataset and Metrics

We evaluated SPDP-Net on JRDB-PAR dataset [13]. This dataset is an extension of JRDB [28] and JRDB-act [8] datasets, which contain multi-person scenes in crowded environments captured by a mobile robot. It consists of 27 videos, of which 20 are for training purposes and 7 are for evaluation. The dataset

Table 1: Ablation experiments on the Panoramic Positional Embedding (PPE) in the proximity-based relation encoding w.r.t self-attention operation. 'S' and 'T' indicate spatial and temporal attention, respectively.

Axis	PPE	Individual Action			Social Activity			Global Activity			Overall
		$\mathcal{P}_i$	$\mathcal{R}_i$	$\mathcal{F}_i$	$\mathcal{P}_p$	$\mathcal{R}_p$	$\mathcal{F}_p$	$\mathcal{P}_g$	$\mathcal{R}_g$	$\mathcal{F}_g$	$\mathcal{F}_a$
T	✗	51.4	45.1	45.7	32.2	31.4	30.5	58.8	48.0	51.4	42.5
	✓	55.7	49.8	50.1	34.3	34.6	33.0	58.2	42.8	47.8	43.6
S	✗	55.4	47.7	48.9	31.9	33.2	31.2	58.3	45.3	49.5	43.2
	✓	56.3	51.4	51.2	34.4	35.4	33.4	60.4	48.9	52.6	45.7
S+T	✗	56.2	50.9	50.7	31.6	33.2	30.9	56.7	44.8	48.7	43.4
	✓	**59.4**	**49.7**	**51.8**	**36.5**	**34.7**	**34.1**	**63.4**	**48.8**	**53.5**	**46.5**

includes 27 individual action classes (*e.g.*, *skating* and *holding something*), 11 categories of social group activities (*e.g.*, *sitting closely* and *working together*), and 7 categories of global activities (*e.g.*, *commuting* and *conversing*). JRDB-PAR contains 27,920 frames with over 628k human bounding boxes.

Following [13], we adopted three evaluation metrics for activity recognition: precision, recall, and F_1 scores for activity prediction, denoted as $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ respectively. The overall metric for human activity recognition is represented by $\mathcal{F}_a = (\mathcal{F}_i + \mathcal{F}_p + \mathcal{F}_g)/3$. To evaluate the results of social group detection, we also follow the protocols of [13], using the classical Half metric [44] (IoU@0.5), the Area Under the Curve (IoU@AUC), and the matrix IoU score (Mat.IoU).

4.2   Implementation Details

We utilized Inception-v3 [36] as a frozen backbone pretrained on CAD dataset [6], following [13]. We used 4 DPATr layers with 4 attention heads and 256 channels. The size of the input frame is 480×3760, and the video length T is set to 3. We implemented the proposed network using 4 NVIDIA GTX 3090 GPUs with PyTorch framework [29]. The batch size of each GPU is set to 4. For training, we used Adam optimizer [19] for 60 epochs. During the initial 15 epochs, we applied a linear warm-up strategy for the learning rate. For the remainder of the training period, we maintained a constant learning rate of 4×10^{-5} with the weight decay of 10^{-2}.

4.3   Ablations

In this section, we conducted extensive experiments to demonstrate the effectiveness of the proposed method. We evaluated the experiments for Individual Action Recognition (IAR), Social Group Activity Recognition (SGAR), GloBal Activity Recognition (GBAR), and Social Group Detection (SGDet).

Spatio-Temporal Proximity in PAR. We investigate the effects of spatio-temporal positional relationships in the PAR task. Specifically, we validate the impact of spatio-temporal proximity in the proximity-based relation encoding stage through ablation experiments on the panoramic positional embedding and

Table 2: Ablation experiments on measuring spatial and temporal social proximity. 'S' and 'T' indicate spatial and temporal axes, respectively.

Distance	Axis	Activity Recognition			Social Group Detection		
		$\mathcal{F}_i$	$\mathcal{F}_p$	$\mathcal{F}_g$	IoU@0.5	IoU@AUC	Mat.IoU
Euclidean	S	43.6	20.7	52.4	43.4	29.6	26.5
Euclidean	S+T	48.2	25.7	52.5	44.7	31.4	26.0
GIoU	S	43.7	28.4	**57.0**	48.7	37.1	28.1
TGIoU	S+T	**51.8**	**34.2**	53.5	**56.4**	**42.5**	**34.3**

Table 3: Ablation study on the similarity relation R_s and the physical relation R_p.

R_s	R_p	Activity Recognition			Social Group Detection		
		$\mathcal{F}_i$	$\mathcal{F}_p$	$\mathcal{F}_g$	IoU@0.5	IoU@AUC	Mat.IoU
		46.2	22.7	53.1	32.2	20.6	17.9
✓		48.1	26.1	52.2	37.6	24.6	21.9
	✓	51.3	33.2	48.8	55.8	**44.0**	33.9
✓	✓	**51.8**	**34.2**	**53.5**	**56.4**	<u>42.5</u>	**34.3**

TGIoU. In Table 1, we summarize the results of SPDP-Net with and without panoramic positional embedding in spatial and temporal axes. The results show that the panoramic positional embedding outperforms conventional embedding in all dimension experiments in terms of all metrics, including precision, recall, and F1 score. Particularly, compared to using the panoramic positional embedding in either spatial or temporal dimension, applying it in both spatial and temporal dimensions yields improvements of 0.8% and 2.9%, respectively, in terms of F_a.

Table 2 shows the experimental results of using spatial and spatio-temporal Euclidean, GIoU, and TGIoU as a proximity metric to demonstrate the impact of the spatio-temporal proximity for PAR. For spatio-temporal Euclidean distance, we average the spatial Euclidean distances across frames. From the results, it is evident that spatio-temporal proximity (S+T) outperforms spatial proximity (S). Specifically, employing Euclidean distance in the spatio-temporal axis leads to a 1.3% improvement in IoU@0.5 compared to using Euclidean distance. Similary, TGIoU yields 56.4% in terms of IoU@0.5, while GIoU achieves 48.7%. Moreover, we see that TGIoU achieves superior performance than the spatio-temporal Euclidean distance in both social group detection and activity recognition. These results indicate that incorporating TGIoU enables SPDP-Net to correctly understand social dynamics in crowded wide scenes.

Social Group Relation. To demonstrate the effectiveness of the similarity relation R_s and the social proximity relation R_p in the social relation R, we ablate them and the results are presented in Table 3. Compared to the baseline not using any social relation, either utilizing R_s or R_p results in higher group detection performance by 5.4% and 23.6% in terms of IoU@0.5, respectively. Moreover, the performances of human activity recognition are also improved.
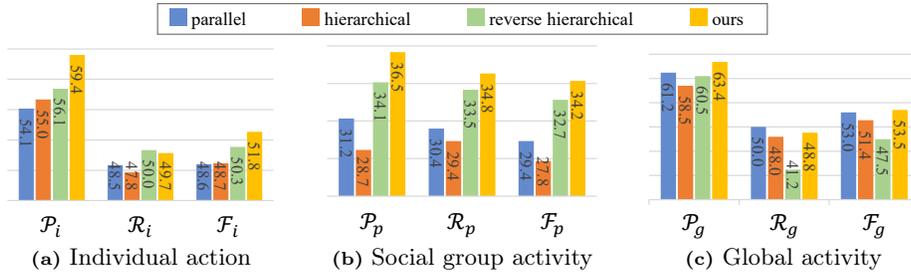
Fig. 4: The results of ablation experiments of a dual-path architecture in DPATr.

Especially, $\mathcal{F}_p$ is increased by 3.4%p and 10.5%p with R_s and R_p, respectively. Notably, while R_s is effective, spatio-temporal proximity from R_p is a crucial factor for defining individual relationships within a panoramic scene. Ultimately, employing both R_s and R_p achieves 56.4% of IoU@0.5 for SGDet and 34.2% of $\mathcal{F}_p$ for SGAR.

Dual-Path Activity Transformer. We analyze the effectiveness of a dual-path architecture for multi-granular activity recognition. For comparisons, we model three types of transformer structure: parallel, hierarchical, and reverse hierarchical. Each of these models comprises three Transformer encoder blocks to refine features of individual, social group, and global activities. As depicted in Fig. 5, the parallel structure independently extracts specific granular activities, while the hierarchical structure sequentially captures activity information from smaller to larger spatial granularity. The reverse hierarchical structure operates inversely to the hierarchical, from larger to smaller spatial granularity.
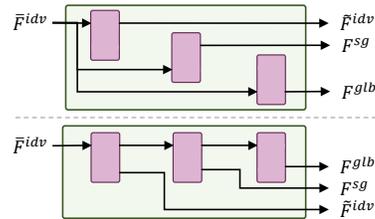
Fig. 5: Simple illustration of a parallel (upper) and hierarchical (lower) structures.

The performances of DPATr with different structures are illustrated in Fig. 4. The parallel architecture outperforms the hierarchical structure in both SGAR and GBAR, suggesting that individual action information is essential for achieving high performance in these tasks. Additionally, the reverse hierarchical structure, which leverages global context for encoding social group activity features, exhibits superior performance in SGAR compared to the parallel and forward hierarchical structures that only use individual information. These results indicate the significance of global context in defining social group activity. DPATr outperforms others in all multi-granular activity recognition tasks by creating synergistic effects of the individual-to-global and individual-to-social paths. From these results, we demonstrate that

Table 4: Comparative results of the panoramic activity recognition (%). The best scores are marked in **bold** and the second best ones are underlined.

Method	Individual Action			Social Activity			Global Activity			Overall
	$\mathcal{P}_i$	$\mathcal{R}_i$	$\mathcal{F}_i$	$\mathcal{P}_p$	$\mathcal{R}_p$	$\mathcal{F}_p$	$\mathcal{P}_g$	$\mathcal{R}_g$	$\mathcal{F}_g$	$\mathcal{F}_a$
ARG [46]	39.9	30.7	33.2	8.7	8.0	8.2	**63.6**	44.3	50.7	30.7
SA-GAT [7]	44.8	40.4	40.3	8.8	8.9	8.8	36.7	29.9	31.4	26.8
JRDB-Base [8]	19.1	34.4	23.6	14.3	12.2	12.8	44.6	46.8	45.1	27.2
JRDB-PAR [13]	51.0	40.5	43.4	24.7	26.0	24.8	52.8	31.8	38.8	35.6
MUP [3]	55.4	44.8	47.7	25.4	26.6	25.1	58.0	**49.0**	51.8	41.5
Ours	**59.4**	**49.7**	**51.8**	**36.5**	**34.8**	**34.2**	63.4	48.8	**53.5**	**46.5**

Table 5: Performance comparison with ground truth group detection results for the panoramic activity recognition.

Method	Individual Action			Social Activity			Global Activity			Overall
	$\mathcal{P}_i$	$\mathcal{R}_i$	$\mathcal{F}_i$	$\mathcal{P}_p$	$\mathcal{R}_p$	$\mathcal{F}_p$	$\mathcal{P}_g$	$\mathcal{R}_g$	$\mathcal{F}_g$	$\mathcal{F}_a$
AT [12]	38.9	33.9	34.6	32.5	32.3	32.0	21.2	19.1	19.8	28.8
SACRF [30]	31.3	23.6	25.9	25.7	24.5	24.8	42.9	35.5	37.6	29.5
Dynamic [52]	40.7	33.4	35.1	33.5	30.1	30.9	37.5	27.1	30.6	32.2
HiGCIN [48]	34.6	26.4	28.6	34.2	31.8	32.2	39.3	30.1	33.1	31.3
ARG [46]	42.7	34.7	36.6	27.4	26.1	26.2	26.9	21.5	23.3	28.8
SA-GAT [7]	39.6	34.5	35.0	32.5	32.5	30.7	28.6	24.0	25.5	30.4
JRDB-Base [8]	21.5	44.9	27.7	54.3	45.9	48.5	38.4	33.1	34.8	37.0
JRDB-PAR [13]	54.3	44.2	46.9	50.3	52.5	50.1	42.1	24.5	30.3	42.4
MUP [3]	56.8	45.6	48.6	55.7	49.7	51.3	57.0	46.2	47.3	49.2
Ours	**60.4**	**50.5**	**52.7**	**56.5**	**54.8**	**53.5**	**62.9**	**48.4**	**53.1**	**53.1**

mutually interacting with contextual understanding of multi-spatial activities is imperative for PAR.

4.4   Comparison with the State-of-the-arts

Human Activity Recognition. In Table 4, we compare the proposed method with other comparative methods for PAR. SPDP-Net surpasses the comparative methods in the overall performance. Particularly, SPDP-Net achieves huge performance improvement in social group activity recognition, attaining 36.5% in $\mathcal{P}_p$, 34.8% in $\mathcal{R}_p$, and 34.2% in $\mathcal{F}_p$. This represents significant improvements of 10%, 8.2%, and 9.1%, respectively, compared to MUP [3], which hierarchically recognizes multi-granular activities.

In Table 5, we further compare the proposed SPDP-Net with other state-of-the-art methods using ground-truth social group detection. This comparison is intended to solely evaluate the capability of recognizing multi-granular activity. Specifically, SPDP-Net with the ground-truth group detection results achieves 19.3% improvement in $\mathcal{F}_p$ compared to not using it. Notably, SPDP-Net consistently outperforms the comparison methods even when utilizing ground-truth social group detection results. This shows the superiority of the proposed method in capturing the dependencies among multi-spatial granular activities and accurately predicting them.

Social Group Detection. We evaluate SPDP-Net against state-of-the-art methods for SGDet as presented in Table 6. Compared to the other meth-

Table 6: Performance comparisons of group detection (%). * indicates that we use the ground-truth number of social groups as input. The best scores are marked in **bold** and the second best ones are <u>underlined</u>.

Method	Activity Recognition			Social Group Detection		
	$\mathcal{F}_i$	$\mathcal{F}_p$	$\mathcal{F}_g$	IoU@0.5	IoU@AUC	Mat. IoU
ARG [46]	33.2	8.2	50.7	35.2	21.6	19.3
SA-GAT [7]	40.3	8.8	31.4	29.1	20.4	16.6
JRDB-Base [8]	23.6	12.8	45.1	38.4	26.3	20.6
JRDB-PAR [13]	43.4	24.8	38.8	53.9	38.1	30.6
Ours	<u>51.8</u>	<u>34.2</u>	**53.5**	<u>56.4</u>	<u>42.5</u>	<u>34.3</u>
Ours*	**52.2**	**35.8**	<u>52.9</u>	**59.7**	**45.5**	**40.2**

	GT	R	R_p	R_s		GT	R	R_p	R_s
		(a)					(b)		
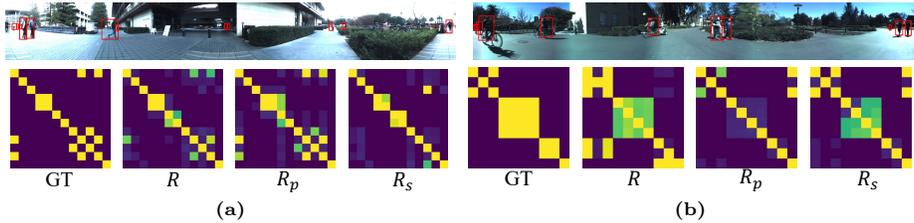
Fig. 6: Visualization of GT and predicted relation matrix R, the proximity relation matrix R_p, and the similarity matrix R_s. Best viewed zoomed-in on screen.

ods [7,8,13,46], SPDP-Net achieves the best performances of 56.4% in IoU@0.5, 42.5% in IoU@AUC, and 34.3% in Mat.IoU. By the enhancements in social group detection performance, SPDP-Net achieves improvements not only in F_p but also in F_i and F_g. We further evaluate SPDP-Net with the ground-truth number of social groups N_g, denoted as 'Ours*' in Table 6. We note that the utilization of ground-truth N_g in SPDP-Net enhances the performances of human activity recognition and social group detection, particularly in SGAR.

4.5   Visualization

Relation Matrix. In Fig. 6, we visualize the relation matrix R, the social proximity relation matrix R_p, and the similarity matrix R_s, which are $N_i \times N_i$ matrices. The ground-truth social relation matrix takes a value of 1 for individuals belonging to the same social group and 0 for otherwise. In Fig. 6a, it is evident that R_p closely aligns with the ground-truth social relation compared to R_s. This emphasizes the significant contribution of the spatio-temporal proximity (R_s) for discerning social dynamics in a crowded scene. Conversely, in scenarios depicted in Fig. 6b, where bounding boxes are relatively large and fewer people are present, we observe that the visual similarity (R_s) is effective.

Qualitative Result. Figure 7 shows the visual comparisons between the ground-truth, SPDP-Net with and without the social proximity relation R_p on JRDB-PAR dataset [13]. We observe that not utilizing R_p results in inaccurate or

(a) GT

(b) without the proximity relation matrix R_p

(c) SPDP-Net(Ours)
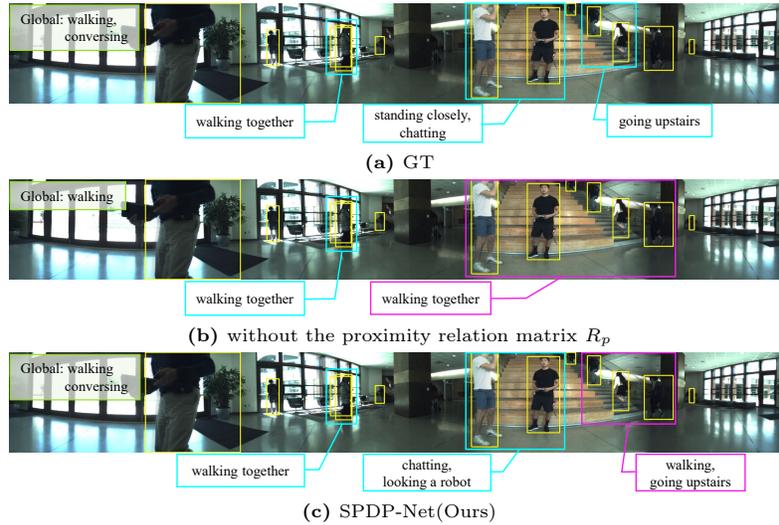
Fig. 7: The visual comparison of the social group activity detection and global activity recognition on JRDB-PAR dataset [13]. The individual bounding boxes are marked in yellow. Misclassified social group bounding boxes are indicated in magenta, while ground-truth and correctly predicted bounding boxes are in cyan.

missed social group detections. The spatio-temporal positional information enables SPDP-Net to infer that the two men standing in front of the stairs continue to converse while the three individuals on the stairs are still ascending, indicating no social relationship between those two groups. However, the person at the right bottom of the stairs, despite not engaging in any social group interaction with others, is misclassified as part of a social group due to exhibiting similar behavior and physical closeness.

5   Conclusion

In this paper, we have proposed a novel network for Panoramic Activity Recognition (PAR), named Social Proximity-aware Dual-Path Network (SPDP-Net). By incorporating spatio-temporal positional relationships among individuals throughout the panoramic positional embedding and spatio-temporal proximity relations, SPDP-Net accurately captures social and global dynamics within a crowded panoramic scene. Furthermore, we have introduced Dual-Path Activity Transformer (DPATr), which consists of individual-to-global and individual-to-social paths. DPATr synergistically enhances final predictions by mutually reinforcing contextual understandings of multi-spatial activities. In this end, the proposed SPDP-Net sets new state-of-the-art records for PAR. We discuss limitations in Sec. C of the supplementary material.

Acknowledgements

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: ICCV. pp. 6836–6846 (2021)
2. Bazzani, L., Cristani, M., Murino, V.: Decentralized particle filter for joint individual-group tracking. In: CVPR. pp. 1886–1893. IEEE (2012)
3. Cao, M., Yan, R., Shu, X., Zhang, J., Wang, J., Xie, G.S.: Mup: Multi-granularity unified perception for panoramic activity recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7666–7675 (2023)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
5. Chappa, N.V., Nguyen, P., Nelson, A.H., Seo, H.S., Li, X., Dobbs, P.D., Luu, K.: Sogar: Self-supervised spatiotemporal attention-based social group activity recognition. arXiv preprint arXiv:2305.06310 (2023)
6. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: 2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops. pp. 1282–1289. IEEE (2009)
7. Ehsanpour, M., Abedin, A., Saleh, F., Shi, J., Reid, I., Rezatofighi, H.: Joint learning of social groups, individuals action and sub-group activities in videos. In: ECCV. pp. 177–195. Springer (2020)
8. Ehsanpour, M., Saleh, F., Savarese, S., Reid, I., Rezatofighi, H.: Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In: CVPR. pp. 20983–20992 (2022)
9. Elbishlawi, S., Abdelpakey, M.H., Eltantawy, A., Shehata, M.S., Mohamed, M.M.: Deep learning-based crowd scene analysis survey. Journal of Imaging **6**(9), 95 (2020)
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019)
11. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR. pp. 1933–1941 (2016)
12. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.: Actor-transformers for group activity recognition. In: CVPR. pp. 839–848 (2020)
13. Han, R., Yan, H., Li, J., Wang, S., Feng, W., Wang, S.: Panoramic human activity recognition. In: ECCV. pp. 244–261. Springer (2022)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017)
15. Hu, B., Cham, T.J.: Entry-flipped transformer for inference and prediction of participant behavior. In: ECCV. pp. 439–456. Springer (2022)
16. Hu, G., Cui, B., He, Y., Yu, S.: Progressive relation learning for group activity recognition. In: CVPR. pp. 980–989 (2020)
17. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: CVPR. pp. 1971–1980 (2016)

18. Kim, D., Lee, J., Cho, M., Kwak, S.: Detector-free weakly supervised group activity recognition. In: CVPR. pp. 20083–20093 (2022)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014)
20. Kong, L., Huang, D., Qin, J., Wang, Y.: A joint framework for athlete tracking and action recognition in sports videos. IEEE TCSVT **30**(2), 532–548 (2019)
21. Lazaridis, L., Dimou, A., Daras, P.: Abnormal behavior detection in crowded scenes using density heatmaps and optical flow. In: 2018 26th European Signal Processing Conference (EUSIPCO). pp. 2060–2064. IEEE (2018)
22. Lee, S., Woo, S., Park, Y., Nugroho, M.A., Kim, C.: Modality mixer for multi-modal action recognition. In: IEEE Winter Conf. on Applications of Comput. Vis. pp. 3298–3307 (2023)
23. Li, J., Han, R., Yan, H., Qian, Z., Feng, W., Wang, S.: Self-supervised social relation representation for human group detection. In: ECCV. pp. 142–159. Springer (2022)
24. Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: Groupformer: Group activity recognition with clustered spatial-temporal transformer. In: ICCV. pp. 13668–13677 (2021)
25. Liang, B., Zheng, L.: A survey on human action recognition using depth sensors. In: 2015 International conference on digital image computing: techniques and applications (DICTA). pp. 1–8. IEEE (2015)
26. Lin, W., Sun, M.T., Poovandran, R., Zhang, Z.: Human activity recognition for video surveillance. In: IEEE International Symposium on Circuits and Systems (ISCAS). pp. 2737–2740. IEEE (2008)
27. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: CVPR. pp. 3202–3211 (2022)
28. Martin-Martin, R., Patel, M., Rezatofighi, H., Shenoi, A., Gwak, J., Frankel, E., Sadeghian, A., Savarese, S.: Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. IEEE TPAMI (2021)
29. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
30. Pramono, R.R.A., Chen, Y.T., Fang, W.H.: Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In: ECCV. pp. 71–90. Springer (2020)
31. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR. pp. 658–666 (2019)
32. Shenoi, A., Patel, M., Gwak, J., Goebel, P., Sadeghian, A., Rezatofighi, H., Martin-Martin, R., Savarese, S.: Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset. In: IEEE/RSJ Int. Conf. on Intell. Robots and Syst. pp. 10335–10342. IEEE (2020)
33. Solera, F., Calderara, S., Cucchiara, R.: Socially constrained structural learning for groups detection in crowd. IEEE TPAMI **38**(5), 995–1008 (2015)
34. Soomro, K., Zamir, A.R.: Action recognition in realistic sports videos pp. 181–208 (2015)
35. Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J.: Human action recognition from various data modalities: A review. IEEE TPAMI (2022)
36. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
37. Tamura, M., Vishwakarma, R., Vennelakanti, R.: Hunting group clues with transformers for social group activity recognition. In: ECCV. pp. 19–35. Springer (2022)

38. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR. pp. 6450–6459 (2018)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
40. Vinyes Mora, S., Knottenbelt, W.J.: Deep learning for domain-specific action recognition in tennis. In: CVPRW. pp. 114–122 (2017)
41. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: ECCV. pp. 108–126. Springer (2020)
42. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. pp. 20–36. Springer (2016)
43. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)
44. Wang, X., Zhang, X., Zhu, Y., Guo, Y., Yuan, X., Xiang, L., Wang, Z., Ding, G., Brady, D., Dai, Q., et al.: Panda: A gigapixel-level human-centric video dataset. In: CVPR. pp. 3268–3278 (2020)
45. Woo, S., Lee, S., Park, Y., Nugroho, M.A., Kim, C.: Towards good practices for missing modality robust action recognition. In: AAAI (2023)
46. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: CVPR. pp. 9964–9974 (2019)
47. Wu, L.F., Wang, Q., Jian, M., Qiao, Y., Zhao, B.X.: A comprehensive review of group activity recognition in videos. International Journal of Automation and Computing **18**, 334–350 (2021)
48. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Higcin: Hierarchical graph-based cross inference network for group activity recognition. IEEE TPAMI (2020)
49. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Social adaptive module for weakly-supervised group activity recognition. In: ECCV. pp. 208–224. Springer (2020)
50. You, Q., Jiang, H.: Action4d: Online action recognition in the crowd and clutter. In: CVPR. pp. 11857–11866 (2019)
51. Yu, G., Yuan, J., Liu, Z.: Propagative hough voting for human activity recognition. In: ECCV. pp. 693–706. Springer (2012)
52. Yuan, H., Ni, D.: Learning visual context for group activity recognition. In: AAAI. vol. 35, pp. 3261–3269 (2021)
53. Yuan, H., Ni, D., Wang, M.: Spatio-temporal dynamic inference network for group activity recognition. In: ICCV. pp. 7476–7485 (2021)
54. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: CVPRW. pp. 28–35 (2012)