Towards More Practical Group Activity Detection: A New Benchmark and Model

Dongkeun Kim Youngkil Song Minsu Cho Suha Kwak

Pohang University of Science and Technology (POSTECH), South Korea
{kdk1563,songyk,mscho,suha.kwak}@postech.ac.kr
https://cvlab.postech.ac.kr/research/CAFE

Abstract. Group activity detection (GAD) is the task of identifying members of each group and classifying the activity of the group at the same time in a video. While GAD has been studied recently, there is still much room for improvement in both dataset and methodology due to their limited capability to address practical GAD scenarios. To resolve these issues, we first present a new dataset, dubbed Café. Unlike existing datasets, Café is constructed primarily for GAD and presents more practical scenarios and metrics, as well as being large-scale and providing rich annotations. Along with the dataset, we propose a new GAD model that deals with an unknown number of groups and latent group members efficiently and effectively. We evaluated our model on three datasets including Café, where it outperformed previous work in terms of both accuracy and inference speed.

Keywords: Group activity detection · Social group activity recognition

1 Introduction

Understanding group activities in videos plays a crucial role in numerous applications such as visual surveillance, social scene understanding, and sports analytics. The generic task of group activity understanding is complex and challenging since it involves identifying participants in an activity and perceiving their spatio-temporal relations as well as recognizing actions of individual actors. Due to these difficulties, most existing work on group activity understanding has been limited to the task of categorizing an entire video clip into one of predefined activity classes [2, 16, 25, 31, 37, 47, 53], which is called the group activity recognition (GAR) in the literature. The common setting of GAR assumes that only a single group activity appears per clip and actors taking part of the activity are identified manually in advance. However, these assumptions do not hold in many real crowd videos, which often exhibit multiple groups that perform their own activities and *outliers* who do not belong to any group. Moreover, it is impractical to manually identify the actors relevant to each group activity in advance. Hence, although GAR has served as a representative group activity understanding task for a decade, its practical value is largely limited.

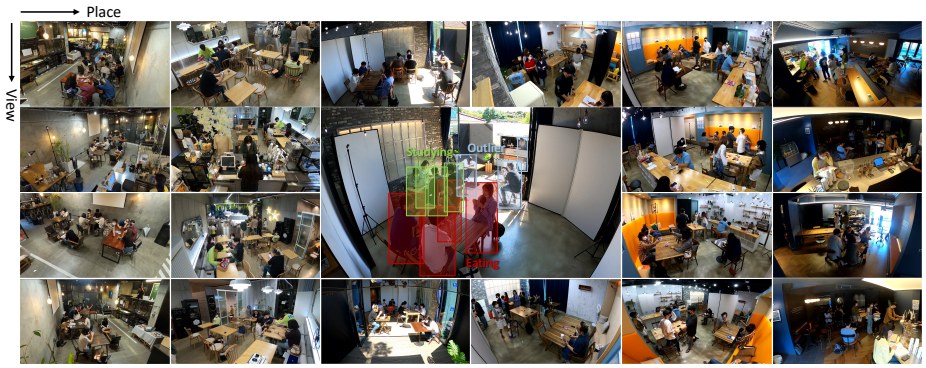

Fig. 1: Examples of videos in Café. The videos were taken at six different places and four cameras with different viewpoints in each place.

As a step toward more realistic group activity understanding, group activity detection (GAD) has recently been studied [12, 13, 43]. GAD aims to localize multiple groups in a video clip and classify each of the localized groups into one of predefined group activity classes, where the group localization means identifying actors of each group. Although a few prior work sheds light on this new and challenging task, there is still large room for improvement in both dataset and methodology due to their limited capability to address practical GAD scenarios. Existing datasets for GAD [12,13] are not constructed primarily for the task but are extensions of other datasets [7, 34] with additional group labels. Moreover, most of the groups in these datasets are singletons, which are individuals rather than groups. Meanwhile, most GAD models rely on off-the-shelf clustering algorithms for group localization, which are not only too heavy in computation but also not optimized for the task.

To address the dataset issue, we present a new dataset for GAD, dubbed Café. Examples of videos in Café are presented in Fig. 1. The videos were taken at six different cafes where people tend to gather in groups, capturing realistic daily group activities. Each video exhibits multiple groups performing various activities, along with outliers, presenting more practical scenarios for GAD. Café has several advantages over the existing GAD datasets. First, it is significantly larger, providing 10K clips and 3.5M actor bounding box labels, as summarized in Table 1. Second, Café poses a greater challenge for group localization since it capture more densely populated scenes than the others; group localization on Café demands an accurate understanding of semantic relations between actors as well as their spatial proximity. Finally, Café captures the same scene with up to four cameras from different viewpoints; these multi-view videos can be used to evaluate model's generalization on unseen views as well as unseen places.

In addition to the new dataset, we also propose a new model architecture for end-to-end GAD. Our model builds embedding vectors of group candidates and individual actors through the attention mechanism of Transformer [11, 44].

Table 1: Comparison between Café and other datasets for group activity understanding. '# Clips' and '# Boxes' represent the number of video clips and the number of annotated bounding boxes, respectively.

Dataset	# Clips	Resolution	# Boxes	Source	Multi-group	Multi-view
CAD [7]	2,511	720×480	0.1M	Daily videos	✗	✗
Volleyball [24]	4,830	1280×720	1.2M	Sports videos	✗	✗
NBA [51]	9,172	1280×720	-	Sports videos	✗	✗
PLPS [39]	71	1920×1080	0.2M	Daily videos	✓	✗
Social-CAD [12]	2,511	720×480	0.1M	Daily videos	✓	✗
JRDB-Act [13]	3,625	3760×480	2.5M	Daily videos	✓	✗
Café	10,297	1920×1080	3.5M	Daily videos	✓	✓

Unlike GAR approaches using Transformer, which aggregate actor features to form a single group representation while capturing spatio-temporal relationship between actors, our model divides actors into multiple groups, each with its own group representation. Embedding vectors of an actor and a group are learned to be close to each other if the actor is a member of the group so that group localization is done by matching the actor and group embeddings. To deal with an unknown number of groups, we employ learnable group tokens whose number is supposed to be larger than the possible maximum number of groups in a video clip; the tokens are then transformed into group embeddings by Transformer, attending to actor embeddings. Each group embedding is also used as input to an activity classifier that determines its activity class. This mechanism allows to discover groups accurately without off-the-shelf clustering algorithms unlike most of previous work, leading to substantially faster inference.

We evaluated our model on three datasets, Café, Social-CAD [12], and JRDB-Act [13], where it outperformed previous work in terms of both accuracy and inference speed. In summary, our contribution is three-fold as follows:

- We introduce Café, a new challenging dataset for GAD. Thanks to its large-scale, rich annotations, densely populated scenes, and multi-view characteristics, it can serve as a practical benchmark for GAD.
- We present a novel GAD model based on Transformer that localizes groups based on the similarity between group embeddings and actor embeddings. Our model efficiently deals with an unknown number of groups and latent group members without off-the-shelf clustering algorithms.
- Our model outperformed previous work on Café and two other GAD benchmarks in terms of both GAD accuracy and inference speed.

2 Related Work

2.1 Group Activity Recognition

Group activity recognition (GAR) has been extensively studied as a representative group activity understanding task. With the advent of deep learning, recurrent neural networks have substantially improved GAR performance [3, 10,

23, 24, 33, 38, 41, 46, 49]. In particular, hierarchical long short-term memory networks [23, 46, 49] have been used to model the dynamics of individual actors and aggregate actors to infer the dynamics of a group.

A recent trend in GAR is modeling spatio-temporal relations between actors. To this end, graph neural networks (GNN) [12, 22, 47, 50, 53], have been placed on top of a convolutional neural network (CNN). Popular examples of such modules include graph convolutional networks [47], graph attention networks [12], dynamic relation graphs [53], and causality graphs [48, 57]. To employ global spatio-temporal dynamic relations between actors and contexts, Transformers [11, 44] have been adopted for GAR and shown significant performance improvement [16, 18, 25, 31, 32, 37, 52, 58]. They utilize the attention mechanism to employ spatio-temporal actor relations [16, 18, 32], relational contexts with conditional random fields [37], actor-specific scene context [52], intra- and inter-group contexts [31], and partial contexts of a group activity [25]. Although these methods have demonstrated outstanding performance, since GAR assumes that only one group is present in each video, their applicability in real-world scenarios is substantially limited.

2.2 Group Activity Detection

Group activity detection (GAD), which is closely related to social group activity recognition [12, 13, 43] and panoramic activity recognition (PAR) [19], has recently been studied to address the limitation of GAR. GNNs [12, 13, 19] have been utilized to model relations between actors and to divide them into multiple groups by applying graph spectral clustering [35, 54]. However, they require off-the-shelf clustering algorithms, which are not optimized for the task and resulting slow inference speed. Meanwhile, HGC [43], which is most relevant to our work, adopts Deformable DETR [59] for localization and matches a group and its potential members in 2D coordinate space. Unlike HGC, our model conducts such a matching in an embedding space to exploit semantic clues more explicitly, achieving better performance.

Along with these models, several datasets have been introduced. Social-CAD [12] extends CAD [7] by adding sub-group labels. JRDB-Act [13] and JRDB-PAR [19] extend annotations of JRDB [34], a multi-person dataset captured by a mobile robot with panoramic views. On these datasets, actors are divided into multiple groups, and the activity of each group is determined by majority voting of individual actions. However, most of the groups in these datasets are composed of a single actor, which is an individual who does not interact with other actors. Unlike these datasets, Café is constructed primarily for GAD. Also, in Café, people annotated as a group perform an activity together, and singleton groups are annotated as outliers.

3 Café Dataset

Café is a multi-person video dataset that aims to introduce a new challenging benchmark for GAD. The dataset contains more than 4 hours of videos taken at

six different cafes by four cameras with different viewpoints, and provides rich annotations including 3.5M bounding boxes of humans, their track IDs, group configurations, and group activity labels. In an untrimmed video, an actor can engage in varying group activities over time, which makes the task challenging and comparisons with existing methods infeasible. Thus, each of the videos is segmented into 6-second clips. In each clip, each actor is a member of a group that performs one of six different group activities or is an *outlier* who does not belong to any group (*i.e.*, a singleton group). Also, outliers are often located overly close to groups as shown in Fig. 1. Thus, for group localization in Café, it is required to grasp the properties of individual actors and their semantic relations as well as their spatial proximity.

3.1 Dataset Annotation and Splitting

Human annotators selected the key frame that clearly exhibited group activities in each video clip. Then, they annotated actor bounding boxes, group configurations, and group activity labels in the frame. Next, a multi-object tracker [56] was applied to extend the actor box labels from the key frame into tracklets across the frames of the clip. To improve the quality of estimated tracklets, the tracker utilized a person detector [17] pretrained on public datasets for person detection and tracking such as CrowdHuman [40], MOT17 [8], City Person [55], and ETHZ [14], which was further finetuned using the key frames of Café. Finally, the annotators manually fixed incorrect tracking IDs and box coordinates.

To examine both place and viewpoint generalization of tested models, we split the dataset in two different ways: *split by place* and *split by view*. The *split by view* setting demonstrates the multi-view characteristics of Café by evaluating the model on unseen views, a challenge absent in existing GAD benchmarks. Details of each dataset split is provided in the supplementary material (Sec. A.2).

3.2 Dataset Statistics

Important statistics that characterize Café are summarized in Fig. 2. Fig. 2a shows group population versus group size (*i.e.*, the number of group members) for each activity class. The class distribution of Café is imbalanced: The least frequent group activity *Queueing* appears about seven times fewer than the most frequent group activity *Taking Selfie*. Such an imbalance is natural in the real world, and may deteriorate activity classification accuracy.

As shown in Fig. 2b, the number of actors in each video clip varies from 3 to 14, and most clips contain 10 or 11 actors. We thus argue that videos in Café well simulate real crowd scenes. Also, about half of the actors are outliers in each clip, which suggests that, on Café, group localization is more challenging.

3.3 Comparison with Existing GAD Datasets

To show the unique challenges and practical aspects of Café, we compare Café with the existing GAD datasets, Social-CAD [12] and JRDB-Act [13]. Fig. 3a

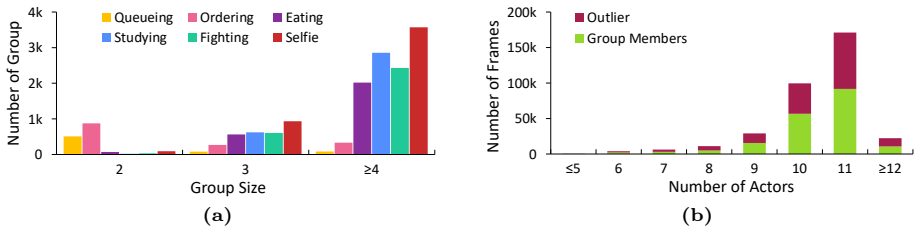

Fig. 2: A summary statistics of Café. (a) Group population versus group size per activity class. (b) Distribution of the number of actors in each video frame.

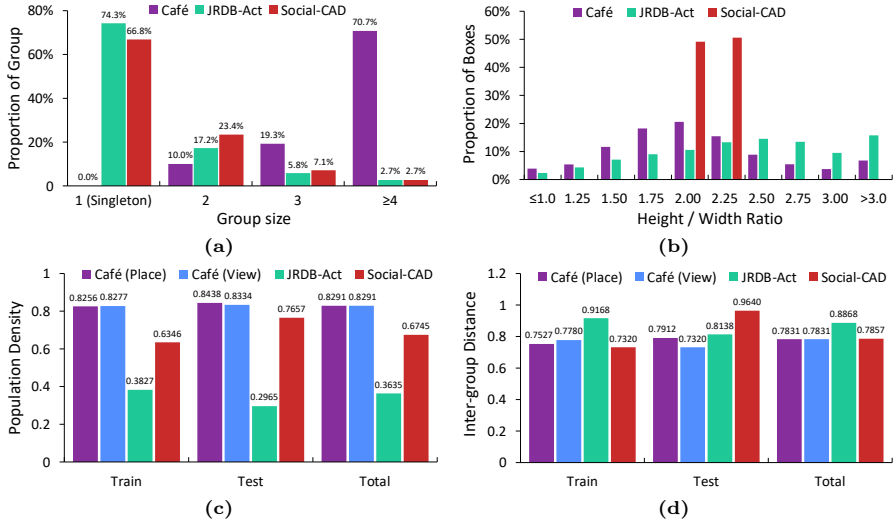

Fig. 3: Comparison between Café and existing GAD datasets in terms of (a) group size, (b) aspect ratios of actor boxes, (c) population density, and (d) inter-group distance.

shows that most groups of existing datasets comprise only a single actor, which are not actually groups but individuals. On the other hand, all groups in Café have at least two actors, and mostly contain more than or equal to four actors.

Fig. 3b illustrates the aspect ratio distribution of actor bounding boxes. In Social-CAD, actors are predominantly pedestrians moving or standing, resulting in nearly all aspect ratios being around 1 : 2. In contrast, Café and JRDB-Act present diverse group activities, resulting in significant pose variation and diverse aspect ratios. Particularly, activities like *Fighting* and *Taking Selfie* in Café necessitates capturing fine-grained pose information, making it a more challenging.

We also compare the datasets in terms of population density, which we define as the ratio between the union area of actors participating in group activities and the area of their convex hull. As shown in Fig. 3c, Café exhibits a higher population density compared to the others in both the *split by view* and *split by place*, making it more challenging for detecting group activities.

Finally, Fig. 3d compares the datasets in the inter-group distance, which calculated by computing the distance between each group and its nearest group or outlier and taking the average of such distances; the exact formulation of the inter-group distance can be found in the supplement. A lower inter-group distance indicates that groups are harder to be localized only by spatial proximity, making the benchmark more challenging than the other two datasets.

3.4 Evaluation Metrics

A proper evaluation metric for GAD should consider following two aspects of predictions: (1) group localization, *i.e.*, identification of members per group, and (2) activity classification per group. While a few evaluation metrics such as social accuracy, social mAP, and G-Act mAP were already proposed in previous work [12, 13], they evaluate group localization based on individual actors rather than groups, which makes them less strict criterion for evaluating group localization quality.

Hence, we propose new evaluation metrics for GAD: *Group mAP* and *Outlier mIoU*. Group mAP is a modification of mAP that has been widely used as the standard performance metric for object detection. On the other hand, Outlier mIoU evaluates how much correctly a model identifies outliers of input video.

Group mAP. Before introducing the definition of Group mAP, we first define Group IoU [6], analogous to IoU used in computation of mAP for object detection. Group IoU measures group localization accuracy by comparing a ground-truth group and a predicted group as follows:

$$\text{Group IoU}(G, \hat{G}) = \frac{|G \cap \hat{G}|}{|G \cup \hat{G}|}, \tag{1}$$

where G is a ground-truth group and $\hat{G}$ is a predicted group; both groups are sets of actors. Group IoU is 1 if all members of $\hat{G}$ are exactly the same with those of G and 0 if no member co-occurs between them. $\hat{G}$ is considered as a correctly localized group if there exists a ground-truth group G that holds $\text{Group IoU}(G, \hat{G}) \geq \theta$, where θ is a predefined threshold. Note that we use two thresholds, $\theta = 1.0$ and $\theta = 0.5$, for evaluation. Group mAP is then defined by using Group IoU as a localization criterion along with activity classification scores. To be specific, we utilize the classification score of the ground-truth activity class as the detection confidence score of the predicted group, and calculate average precision (AP) score per activity class through all-point interpolation [15]. Finally, AP scores of all classes are averaged to produce Group mAP.

Outlier mIoU. It is important for GAD in the real world videos to distinguish groups and outliers (*i.e.*, singletons). We thus propose Outlier mIoU to evaluate outlier detection. Similar to Group IoU, its format definition is given by

$$\text{Outlier mIoU} = \frac{1}{|V|} \sum_{v \in V} \frac{|O_v \cap \hat{O}_v|}{|O_v \cup \hat{O}_v|}, \tag{2}$$

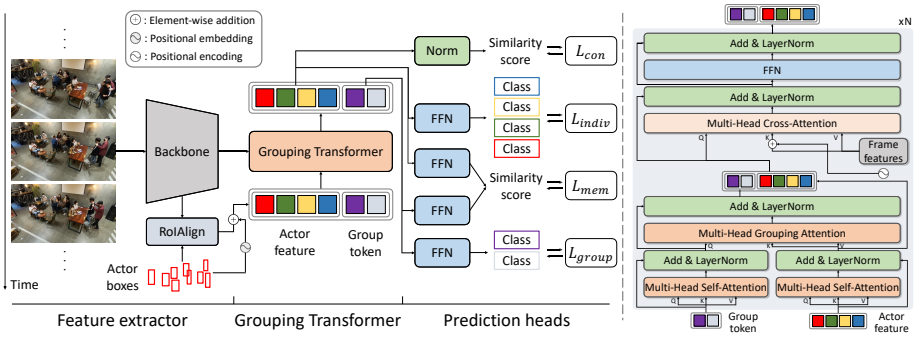

Fig. 4: (*Left*) Overall architecture of our model. (*Right*) Detailed architecture of the Grouping Transformer.

where $|V|$ is the set of video clips for evaluation, O_v is the set of ground-truth outliers in clip v, and $\hat{O}_v$ is the set of predicted outliers in clip v.

4 Proposed Model for GAD

The purpose of GAD is to identify members of each group (*i.e.*, group localization) and classify the activity conducted by each group simultaneously. The task is challenging since both the number of groups and their members are unknown. We present a new model based on Transformer [11,44] to deal with these difficulties; its overall architecture is illustrated in Fig. 4.

The key idea at the heart of our model is that embedding vectors of an actor and a group should be close if the actor is a member of the group. To compute embedding vectors of groups and individual actors, we adopt attention mechanism of Transformer. To deal with a varying number of groups in each video clip, our model defines and utilizes learnable group tokens, whose number is supposed to be larger than the possible maximum number of groups in a clip. The group tokens along with actor features obtained by RoIAlign [20] are fed to a Transformer called Grouping Transformer to become the embedding vectors.

4.1 Model Architecture

Our model consists of three parts: feature extractor, Grouping Transformer, and prediction heads.

Feature extractor. As in recent GAR models [12,13,16,31,47,52], our model extracts frame-level features using a CNN backbone, and extracts actor features from the frame features by RoIAlign given actor bounding boxes. To be specific, we adopt an ImageNet [9] pretrained ResNet-18 [21] for the feature extraction, and actor features extracted by RoIAlign are of 5×5 size. Additionally, to incorporate spatial cues when identifying group members, learnable positional embeddings of actor box coordinates are added to their associated actor features.

Grouping Transformer. Grouping Transformer takes learnable group tokens, actor features, and frame features as input, and produces embedding vectors of group candidates and actors in a frame-wise manner. As illustrated in the right-hand side of Fig. 4, it comprises three types of multi-head attention layers: (1) multi-head self-attention layers that capture relations between actors and those between groups separately, (2) multi-head grouping attention layers where group tokens as queries attend to actor features serving as keys and values, and (3) multi-head cross-attention layers where actor features and group tokens draw attentions on frame features to capture contextual information. The core of the Grouping Transformer lies in the grouping attention layer. Each group token produces group representation by attending to actor features potentially belonging to its group, based on the similarity in the embedding space. In addition, to exploit spatial cues, we apply a distance mask to the multi-head self-attention layers for actor features: Following ARG [47], a pair of actors whose distance is greater than a threshold μ do not attend to each other.

Prediction heads. Two types of prediction heads in the form of feed-forward networks (FFNs) are attached to individual outputs of the Grouping Transformer, actor embeddings and group embeddings. The first prediction heads are for group activity classification, and the second prediction heads further project the actor/group embeddings so that the results are used for identifying group members: An actor embedding and a group embedding projected separately are dot-producted to compute their semantic affinity, which is used as the membership score of the actor for the group. At inference, each actor is assigned to the group with the highest membership score among all predicted groups.

4.2 Training Objectives

Group matching loss. Motivated by DETR [4], we first establish the optimal bipartite matching between ground-truth groups and predicted groups using Hungarian algorithm [28]. Since our model produces K predicted groups, where K is the number of group tokens and is supposed to be larger than the number of ground-truth groups, we add empty groups with no activity class, denoted by $\varnothing$, to the set of ground-truth groups so that the number of ground-truth groups becomes K and they are matched with the predicted groups in a bipartite manner accordingly. Then, among all possible permutations of K predicted groups, denoted by $\mathfrak{S}_K$, Hungarian algorithm finds the permutation with the lowest total matching cost:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_K}{\arg\min} \sum_{i}^{K} C_{i,\sigma(i)}. \tag{3}$$

$C_{i,\sigma(i)}$ in Eq. (3) is the matching cost of the ground-truth group i and the predicted group $\sigma(i)$ and is given by

$$C_{i,\sigma(i)} = -\mathbb{1}_{\{y_i \neq \varnothing\}} \hat{p}_{\sigma(i)}(y_i) + \mathbb{1}_{\{y_i \neq \varnothing\}} \|\mathbf{m}_i - \hat{\mathbf{m}}_{\sigma(i)}\|_2, \tag{4}$$

where y_i is the activity class label of the ground-truth group i and $\hat{p}_{\sigma(i)}(y_i)$ is the predicted class probability for y_i. Also, $\mathbf{m}_i = [m_i^1, \ldots, m_i^N]^\mathsf{T}$ indicates ground-truth membership relations between actors and group i, and $\hat{\mathbf{m}}_{\sigma(i)} = [\hat{m}_{\sigma(i)}^1, \ldots, \hat{m}_{\sigma(i)}^N]^\mathsf{T}$ is a collection of predicted membership scores of actors for predicted group $\sigma(i)$, where N is the number of actors in the input clip; each dimension of the two vectors is computed by

$$m_i^j = \begin{cases} 1, & \text{if actor } j \text{ is a member of group } i, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

$$\hat{m}_{\sigma(i)}^j = {\psi_j}^\mathsf{T} \phi_{\sigma(i)}, \tag{6}$$

where ψ_j is the output of the second prediction head for actor j and $\phi_{\sigma(i)}$ is the output of the second prediction head for predicted group $\sigma(i)$. The group activity classification loss L_{group} and the membership loss L_{mem} are calculated for all matched pairs. To be specific, we adopt the standard cross-entropy loss for $L_{\text{group}} = L_{\text{group}}(i, \sigma(i))$:

$$L_{\text{group}} = -\log \frac{\exp(\hat{p}_{\sigma(i)}(y_i))}{\sum_{c=1}^{C} \exp(\hat{p}_{\sigma(i)}(c))}, \tag{7}$$

where C is the number of group activity classes, and the binary cross-entropy loss for $L_{\text{mem}} = L_{\text{mem}}(i, \sigma(i))$:

$$L_{\text{mem}} = -\frac{1}{N} \sum_{j=1}^{N} \big(m_i^j \cdot \log \hat{m}_{\sigma(i)}^j + (1 - m_i^j) \cdot \log(1 - \hat{m}_{\sigma(i)}^j) \big). \tag{8}$$

Group consistency loss. It has been known that supervisory signals given by the bipartite matching of Hungarian algorithm may fluctuate and thus lead to slow convergence [29]. To alleviate this issue, we additionally introduce a group consistency loss, which is a modification of InfoNCE [36] and enhances the quality of group localization while bypassing the bipartite matching. The loss is formulated by

$$L_{\text{con}} = -\sum_{g_i} \sum_{j \in g_i} \log \frac{\sum_{k \in g_i, k \neq j} \exp\big(\cos(f_j, f_k)/\tau\big)}{\sum_{k \neq j} \exp\big(\cos(f_j, f_k)/\tau\big)}, \tag{9}$$

where g_i means the i-th ground-truth group, τ is the temperature, f_j stands for j-th actor embeddings and cos indicates the cosine similarity function. This loss provides a consistent group supervision to actors that belong to the same group.

Individual action classification loss. We adopt a standard cross-entropy loss for individual action loss L_{ind}. The individual action class of an actor who belongs to a group is regarded as the group activity class of the group, and the action class of an outlier is no activity class, denoted by $\varnothing$.

Total loss. Our model is trained with four losses simultaneously in an end-to-end manner. Specifically, the total training objective of our proposed model is a

linear combination of the four losses as follows:

$$L = L_{\text{ind}} + \sum_i L_{\text{group}} + \lambda_m \sum_i L_{\text{mem}} + \lambda_c L_{\text{con}}. \tag{10}$$

5 Experiments

5.1 Implementation Details

Hyperparameters. We use an ImageNet pretrained ResNet-18 as a backbone network. Ground-truth actor tracklets are used to extract actor features with 256 channels by applying RoIAlign with crop size 5×5. For the Grouping Transformer, we stack 6 Transformer layers with 4 attention heads for Café and JRDB-Act, 3 Transformer layers with 8 attention heads for Social-CAD. The number of group tokens K is set to 12 for Café and JRDB-Act, 10 for Social-CAD.
Training. We sample T frames using the segment-based sampling [45], where T is 5, 1, and 2 for Café, Social-CAD, and JRDB-Act, respectively. We train our model with Adam optimizer [26] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\text{e}{-8}$ for 30 epochs. Learning rate is initially set to $1\text{e}{-5}$ with linear warmup to $1\text{e}{-4}$ for 5 epochs, and linearly decayed for remaining epochs. Mini-batch size is set to 16. Loss coefficients are set to $\lambda_m = 5.0$, and $\lambda_c = 2.0$. The temperature τ is set to 0.2 for the group consistency loss.

5.2 Comparison with the State of the Art

Compared methods. We compare our method with three clustering-based methods [12, 13, 47] and one Transformer-based method [43]. Since most GAD methods do not provide the official source code, we try our best to implement these previous work with necessary modifications for dealing with different problem settings, in particular the existence of outliers in Café. Specifically, we adopt a fixed cluster size [35] for clustering-based methods since estimating the number of clusters did not perform well in Café due to the presence of spatially close outliers. Note that no adjustment was made for datasets other than Café.
• *ARG* [47]: ARG utilizes graph convolutional networks [27] to model relations between actors in terms of position and appearance similarity. We apply spectral clustering [35] on a relation graph to divide actors into multiple groups.
• *Joint* [12] and *JRDB-base* [13]: These models utilize GNNs to model relations between actors, and train actor representations to partition graphs by adopting a graph edge loss. JRDB-base further adopts geometric features. Then, spectral clustering [35] is applied on the graph.
• *HGC* [43]: Similar to our method, HGC employs a Transformer for GAD. However, unlike our method, HGC identifies group members by point matching between groups and actors on the 2D coordinate space. For a fair comparison, we utilize ground-truth actor tracklets to obtain actor features for HGC.
Café dataset. For a fair comparison, we use ImageNet pretrained ResNet-18 as the backbone and apply distance mask for all the methods including ours.

Table 2: Comparison with the previous work on Café. All methods are built on the same ResNet-18 backbone. The second column means the number of tokens for Transformer-based models and the number of clusters for clustering-based models. '# G' and '# O' indicates the true number of groups and that of outliers in each video clip, respectively. The third column is the wall-clock inference time for a single video clip measured on a Titan XP GPU. The subscripts of Group mAP mean Group IoU thresholds (θ in Sec. 3.4). We mark the best and the second-best performance in **bold** and underline, respectively.

Method	# Token (# Cluster)	Inference time (s)	Split by view			Split by place		
			Group $mAP_{1.0}$	Group $mAP_{0.5}$	Outlier mIoU	Group $mAP_{1.0}$	Group $mAP_{0.5}$	Outlier mIoU
ARG [47]	4	0.22	11.03	34.50	56.61	6.87	28.44	46.72
	5	0.26	5.46	30.34	58.89	5.79	24.25	49.25
	6	0.28	1.27	27.69	60.41	2.59	22.33	51.00
	#G+#O	0.30	2.64	28.98	58.21	2.29	22.33	50.01
Joint [12]	4	0.23	13.86	34.68	53.67	6.69	27.76	49.50
	5	0.25	14.05	36.08	60.09	8.39	26.26	55.95
	6	0.28	5.94	33.14	60.63	5.11	24.55	56.94
	#G+#O	0.32	4.54	31.24	59.78	2.87	21.35	56.68
JRDB-base [13]	4	0.23	15.43	34.81	60.43	9.42	25.75	48.00
	5	0.25	13.26	37.40	63.91	9.42	26.19	51.30
	6	0.28	6.77	35.22	63.85	6.37	26.23	51.53
	#G+#O	0.32	4.49	34.40	61.46	3.15	25.80	49.71
HGC [43]	12	0.10	5.18	23.02	57.23	3.50	17.92	57.42
	24	0.10	5.60	21.44	54.57	3.00	14.48	53.64
	50	0.10	6.55	26.29	56.84	3.47	18.46	52.56
	100	0.10	3.63	15.42	54.59	3.07	19.97	56.80
Ours	4	0.10	16.02	**40.22**	64.06	8.97	27.33	62.35
	8	0.10	18.10	37.51	65.49	9.79	29.23	**63.93**
	12	0.10	**18.84**	37.53	**67.64**	**10.85**	**30.90**	63.84
	16	0.10	15.03	37.03	65.31	7.57	25.08	58.66

We test every model on two different dataset splits as explained in Sec. 3.1: *split by view* and *split by place*. Table 2 summarizes the results. Our model outperforms all the other methods by substantial margins on both splits in terms of both Group mAP and Outlier mIoU. Note that the performance of clustering-based methods largely depends on the number of clusters, which is hard to determine or predict when there are outliers in a video clip. On the other hand, our model is less sensitive to the number of group tokens, 12 tokens shows the best performance on both settings though. Our model outperforms HGC, demonstrating the effectiveness of our group-actor matching in an embedding space, as opposed to the point matching strategy used in HGC. We also conduct experiments in a detection-based setting for all methods, and the results can be found in the supplementary material (Sec. C.2).

JRDB-Act dataset. Table 3 presents results on JRDB-Act. Our model achieves 59.8 mAP, surpassing all other methods. This indicates that our method effectively detects group activities across varying group sizes, especially when the group size is larger than 2. Notably, our model with the ResNet-18 backbone outperforms Joint and JRDB-base with the substantially heavier I3D backbone.

Table 3: Quantitative results on JRDB-Act validation-set.

Method	Backbone	G1 AP	G2 AP	G3 AP	G4 AP	G5$^+$ AP	mAP
SHGD [30]	Unipose [1]	3.1	25.0	17.5	45.6	25.2	23.3
Joint [12]	I3D [5]	8.0	29.3	37.5	65.4	67.0	41.4
PAR [19]	Inception-v3 [42]	52.0	59.2	46.7	46.6	31.1	47.1
JRDB-base [13]	I3D	81.4	64.8	49.1	63.2	37.2	59.2
Ours	ResNet-18 [21]	70.1	56.3	50.4	71.7	50.8	**59.8**

Table 4: Quantitative results on Social-CAD.

Method	Backbone	# frames	Social Accuracy
ARG [47]	Inception-v3	17	49.0
Joint [12]	I3D	17	69.0
Ours	ResNet-18	1	**69.2**

Table 5: Ablation study on the group consistency loss.

L_{con}	Group mAP$_{1.0}$	Outlier mIoU
✗	15.06	63.35
✓	**18.84**	**67.64**

Table 6: Ablation on the attention layers of the Grouping Transformer.

Method	Group mAP$_{1.0}$	Outlier mIoU
Ours	**18.84**	**67.64**
w/o self-attention	13.53	65.62
w/o cross-attention	13.12	64.19
w/o grouping-attention	12.86	64.65

Table 7: Ablation on the use of distance mask and its threshold.

Distance threshold (μ)	Group mAP$_{1.0}$	Outlier mIoU
0.1	14.46	62.75
0.2	**18.84**	**67.64**
0.3	15.09	63.49
No threshold	14.96	67.03

Social-CAD dataset. Table 4 summarizes the results on Social-CAD. Our model surpasses the previous methods by using ResNet-18 with a single frame as backbone, which is significantly lighter than I3D backbone taking 17 frames as input in Joint model [12]. Due to the short length of video clips and small variations within clips, our model achieves the best even with a single frame.

5.3 Ablation Studies

We also verify the effectiveness of our proposed model through ablation studies on Café, *split by view* setting.

Impact of the proposed loss function. Table 5 shows the effectiveness of the group consistency loss, which improves Group mAP by a substantial margin. This result demonstrates that the group consistency loss, which brings actor embeddings within the same group closer, has a significant impact for GAD. We do not ablate the other losses since they are inevitable for the training.

Effects of the attention layers. Table 6 summarizes the effects of multi-head attention layers in Grouping Transformer. Note that self-attention in this table stands for both multi-head self-attention layers that captures relationship between actors and those between groups, cross-attention means multi-head cross-attention layers that actor features attends frame-level features to capture contextual information, and grouping-attention refers to the layers that group tokens attends actor features to form group representation. The results demonstrate that all three attention layers contribute to the performance. Particularly, removing grouping-attention layer results in the largest performance drop in Group mAP since grouping-attention layer learns the relationship between group embeddings and actor embeddings, aiding in group localization.

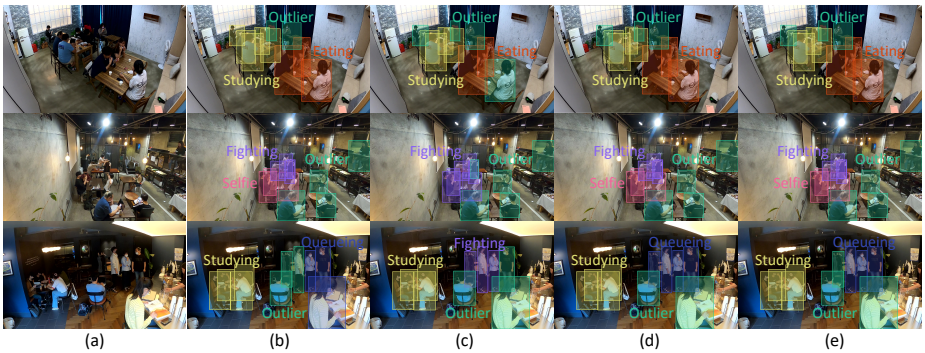

Fig. 5: Qualitative results on Café test-set, *split by view* setting. Boxes with the same color belong to the same group. (a) Input frame. (b) Prediction of JRDB-base. (c) Prediction of HGC. (d) Prediction of our model. (e) Ground-truth.

Effects of the distance mask. We investigate the efficacy of utilizing the distance mask. Distance mask inhibits self-attention between a pair of actors whose distance is greater than the distance threshold μ. As shown in Table 7, applying distance mask between actors is effective in most cases but too small threshold, 0.1 in this table, degrades the performance. It is because actors can interact only with nearby actors at small distance threshold, which might mask the interaction between actors of the same group. Distance threshold of 0.2 reaches the best result while slightly degrades at 0.3.

5.4 Qualitative Analysis

Fig. 5 visualizes the predictions of JRDB-base, HGC, and our model. The results show that our model is able to localize multiple groups and predict their activity classes at the same time, and more reliably than the others, even in challenging densely populated scenes with a lot of outliers.

6 Conclusion

We have introduced a new challenging benchmark, dubbed Café, and a new model based on Transformer to present a direction towards more practical GAD. As Café exhibits multiple non-singleton groups per clip and provides rich annotations of actor bounding boxes, track IDs, group configurations, and group activity labels, it can serve as a new, practical, and challenging benchmark for GAD. Also, the proposed model can deal with a varying number of groups as well as predicting members of each group and its activity class. Our model outperformed prior arts on three benchmarks including Café. We believe that our dataset and model will promote future research on more practical GAD.
Limitation: Our model does not consider much about temporal and multi-view aspects of the proposed dataset. Improving upon these aspects will be a valuable direction to explore.

Acknowledgement. This work was supported by the NRF grant and the IITP grant funded by Ministry of Science and ICT, Korea (RS-2019-II191906, IITP-2020-0-00842, NRF-2021R1A2C3012728, RS-2022-II220264). We thank Deeping Source for their help with data collection.

References

1. Artacho, B., Savakis, A.: Unipose: Unified human pose estimation in single images and videos. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7035–7044 (2020)
2. Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7892–7901 (2019)
3. Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4315–4324 (2017)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proc. European Conference on Computer Vision (ECCV). pp. 213–229. Springer (2020)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
6. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Discovering groups of people in images. In: Proc. European Conference on Computer Vision (ECCV). pp. 417–433. Springer (2014)
7. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: Proc. IEEE International Conference on Computer Vision (ICCV) Workshops. pp. 1282–1289. IEEE (2009)
8. Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., Leal-Taixé, L.: Motchallenge: A benchmark for single-camera multiple target tracking. International Journal of Computer Vision (IJCV) **129**(4), 845–881 (2021)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
10. Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4772–4781 (2016)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=YicbFdNTTy
12. 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: Proc. European Conference on Computer Vision (ECCV). pp. 177–195. Springer (2020)

13. 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: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20983–20992 (2022)
14. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8. IEEE (2008)
15. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision (IJCV) **111**, 98–136 (2015)
16. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.: Actor-transformers for group activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 839–848 (2020)
17. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
18. Han, M., Zhang, D.J., Wang, Y., Yan, R., Yao, L., Chang, X., Qiao, Y.: Dual-ai: Dual-path actor interaction learning for group activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2990–2999 (2022)
19. Han, R., Yan, H., Li, J., Wang, S., Feng, W., Wang, S.: Panoramic human activity recognition. In: Proc. European Conference on Computer Vision (ECCV). pp. 244–261. Springer (2022)
20. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proc. IEEE International Conference on Computer Vision (ICCV). pp. 2961–2969 (2017)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
22. Hu, G., Cui, B., He, Y., Yu, S.: Progressive relation learning for group activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 980–989 (2020)
23. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: Proc. European Conference on Computer Vision (ECCV). pp. 721–736 (2018)
24. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1971–1980 (2016)
25. Kim, D., Lee, J., Cho, M., Kwak, S.: Detector-free weakly supervised group activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20083–20093 (2022)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. International Conference on Learning Representations (ICLR) (2015)
27. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. Proc. International Conference on Learning Representations (ICLR) (2017)
28. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
29. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13619–13627 (2022)

30. Li, J., Han, R., Yan, H., Qian, Z., Feng, W., Wang, S.: Self-supervised social relation representation for human group detection. In: Proc. European Conference on Computer Vision (ECCV). pp. 142–159. Springer (2022)
31. Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: Groupformer: Group activity recognition with clustered spatial-temporal transformer. In: Proc. IEEE International Conference on Computer Vision (ICCV). pp. 13668–13677 (2021)
32. Li, W., Yang, T., Wu, X., Du, X.J., Qiao, J.J.: Learning action-guided spatio-temporal transformer for group activity recognition. In: Proc. ACM Multimedia Conference (ACMMM). pp. 2051–2060 (2022)
33. Li, X., Choo Chuah, M.: Sbgar: Semantics based group activity recognition. In: Proc. IEEE International Conference on Computer Vision (ICCV). pp. 2876–2885 (2017)
34. 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2021)
35. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. Proc. Neural Information Processing Systems (NeurIPS) **14** (2001)
36. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
37. 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: Proc. European Conference on Computer Vision (ECCV). pp. 71–90. Springer (2020)
38. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Van Gool, L.: stagnet: An attentive semantic rnn for group activity recognition. In: Proc. European Conference on Computer Vision (ECCV). pp. 101–117 (2018)
39. Qing, L., Li, L., Xu, S., Huang, Y., Liu, M., Jin, R., Liu, B., Niu, T., Wen, H., Wang, Y., et al.: Public life in public space (plps): A multi-task, multi-group video dataset for public life research. In: Proc. IEEE International Conference on Computer Vision (ICCV) Workshops. pp. 3618–3627 (2021)
40. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
41. Shu, T., Todorovic, S., Zhu, S.C.: Cern: confidence-energy recurrent network for group activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5523–5531 (2017)
42. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
43. Tamura, M., Vishwakarma, R., Vennelakanti, R.: Hunting group clues with transformers for social group activity recognition. In: Proc. European Conference on Computer Vision (ECCV) (2022)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc. Neural Information Processing Systems (NeurIPS). pp. 5998–6008 (2017)
45. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: Proc. European Conference on Computer Vision (ECCV). pp. 20–36. Springer (2016)

46. Wang, M., Ni, B., Yang, X.: Recurrent modeling of interaction context for collective activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3048–3056 (2017)
47. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9964–9974 (2019)
48. Xie, Z., Gao, T., Wu, K., Chang, J.: An actor-centric causality graph for asynchronous temporal inference in group activity. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6652–6661 (2023)
49. Yan, R., Tang, J., Shu, X., Li, Z., Tian, Q.: Participation-contributed temporal dynamic model for group activity recognition. In: Proc. ACM Multimedia Conference (ACMMM). pp. 1292–1300 (2018)
50. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Higcin: hierarchical graph-based cross inference network for group activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
51. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Social adaptive module for weakly-supervised group activity recognition. In: Proc. European Conference on Computer Vision (ECCV). pp. 208–224. Springer (2020)
52. Yuan, H., Ni, D.: Learning visual context for group activity recognition. In: Proc. AAAI Conference on Artificial Intelligence (AAAI). vol. 35, pp. 3261–3269 (2021)
53. Yuan, H., Ni, D., Wang, M.: Spatio-temporal dynamic inference network for group activity recognition. In: Proc. IEEE International Conference on Computer Vision (ICCV). pp. 7476–7485 (2021)
54. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. Proc. Neural Information Processing Systems (NeurIPS) **17** (2004)
55. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3213–3221 (2017)
56. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: Proc. European Conference on Computer Vision (ECCV) (2022)
57. Zhang, Y., Liu, W., Xu, D., Zhou, Z., Wang, Z.: Bi-causal: Group activity recognition via bidirectional causality. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1450–1459 (2024)
58. Zhou, H., Kadav, A., Shamsian, A., Geng, S., Lai, F., Zhao, L., Liu, T., Kapadia, M., Graf, H.P.: Composer: Compositional learning of group activity in videos. In: Proc. European Conference on Computer Vision (ECCV) (2022)
59. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: Proc. International Conference on Learning Representations (ICLR) (2021)