Social Adaptive Module for Weakly-supervised Group Activity Recognition

Rui Yan[1], Lingxi Xie[2], Jinhui Tang[1*], Xiangbo Shu[1], and Qi Tian[2]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2] Huawei Inc., China
{ruiyan, jinhuitang, shuxb}@njust.edu.cn, 198808xc@gmail.com,
tian.qi1@huawei.com

Abstract. This paper presents a new task named weakly-supervised group activity recognition (GAR) which differs from conventional GAR tasks in that only video-level labels are available, yet the important persons within each frame are not provided even in the training data. This eases us to collect and annotate a large-scale NBA dataset and thus raise new challenges to GAR. To mine useful information from weak supervision, we present a key insight that key instances are likely to be related to each other, and thus design a social adaptive module (SAM) to reason about key persons and frames from noisy data. Experiments show significant improvement on the NBA dataset as well as the popular volleyball dataset. In particular, our model trained on video-level annotation achieves comparable accuracy to prior algorithms which required strong labels.

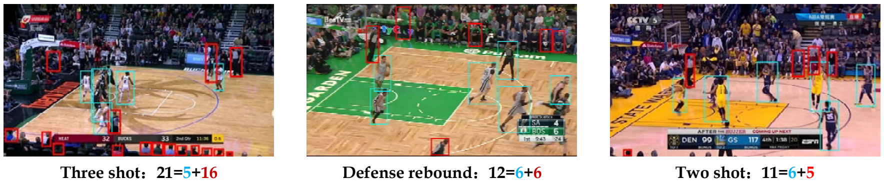
Keywords: Group Activity Recognition, Video Analysis, and Scene Understanding

1 Introduction

Group activity recognition (GAR) has a variety of applications in video understanding, such as sports analysis, video surveillance, and public security. Compared with traditional individual actions [30,38,23,14,27], group activities (*a.k.a*, collective activities) [10,18,45,42] are performed by multiple persons cooperating with each other. Thus, the models for GAR require to understand not only the individual behaviors but also the relationship between each person.

Previous fully-supervised methods that require person-level annotation (*i.e.* ground-truth bounding boxes and individual action label for each person, even interaction label for person-person pairs) have achieved promising performance on group activity recognition. Typically, these methods [18,45,40,39,35,42,3,4,44] extract feature for each people according to the corresponding bounding boxes supervised by individual action label, and then fuse person-level feature into a

[*] Corresponding author


Three shot：21=5+16

Defense rebound：12=6+6

Two shot：11=6+5

Fig. 1. *Best viewed in color*. Illustration of the uncertain input issue under weakly-supervised setting. For different activities, the off-the-shelf detector will generate varying numbers of proposals, most of which (in red boxes) are useless for recognizing group activities. For instance, "**Three shot: 21** = **5** + **16**" means that the detector generates a total of **21** proposals, but only **5** of them are players and other **16** proposals are outliers in an activity of three-shot

single representation for each frame. However, previous methods are sensitive to the varying number of people in each frame and require the explicit locations of them, which is limited in practical applications.

To this end, we investigate GAR in a weakly-supervised setting that only provides video-level labels for each video clip. This setting not only is practical to real-world scenarios but also provides a simpler and lower-cost way for the annotation of new benchmarks. Benefiting from it, we collect a larger and more challenging benchmark, NBA, consisting of 181 basketball games which involve more long-term temporal and fast-moving activities. Meanwhile, the weakly-supervised setting also brings uncertain input issue in each frame, as illustrated in Fig. 1. Under this setting, lots of useless proposals will be fed into the approach. Besides, numerous irrelevant frames will also appear in the video clip, if the temporal structure of activities (*e.g.*, in NBA) is long.

To tackle these issues, we further propose a simple yet effective module, namely Social Adaptive Module (SAM), which can adaptively select discriminative proposals and frames from the video for weakly-supervised GAR. SAM aims at assisting the weakly-supervised training by leveraging a social assumption that **key instances (people/frames) are highly related to each other**. Specifically, we firstly construct a dense relation graph on all possible input feature to measure the relatedness between each other, then pick the top ones according to their relatedness. Based on the selected feature, a sparse relation graph is built to perform relational embedding for them. Benefiting from SAM, our approach trained without fully-supervision still obtains the comparable performance to previous methods on the popular volleyball dataset [18].

Our contributions include: (a) The weakly-supervised setting that only provides video-level labels is introduced for GAR. (b) Thanks to this setting, a larger and more challenging benchmark, NBA, is collected from the web at a low cost. (c) To ease the weakly-supervised training, a SAM is proposed to adaptively find the effective person-level and frame-level representation based on the social assumption that key instances are usually closely related to each other.

2 Related Work

Group Activity Recognition. Initial approaches [18,45,40,42] for recognizing group activities adopted the two-stage pipeline. They pre-extracted feature for each person from a set of patch images and then fuse them into a single vector for each frame by various methods (*e.g.*, pooling strategies [18,35], attention mechanism [31,40,45], recurrent models [13,39,42,45], graphical models [2,25,24], and AND-OR grammar models [1,36]). Nevertheless, these two-stage methods separate feature aggregation from representation learning, which is not conducive to a deep understanding of group activities. To this end, Bagautdinov *et al.* [4] introduced an end-to-end framework to jointly detect multiple individuals, infer their individual actions, and estimate the group activity. Wu *et al.* [44] extended [4] by stacking multiple graph convolutional layer to infer the latent relation between each person. Azar *et al.* [3] constructed an activity map based on bounding boxes and explore the spatial relationships among people by iteratively refining the map. However, all of the above methods still require the action-level supervision (action labels and bounding boxes for each person), which is time-consuming to tag. Ramanathan *et al.* [32] detected events and key actors in multi-person videos without individual action labels, but they still needed to annotate the bounding boxes of all the players in a subset of $9,000$ frames for training a detector. This work introduces a more practical weakly-supervised setting that only provides video-level labels for group activity recognition.

Existing Datasets Related to GAR. Limited by the time-consuming tagging, there are currently only four datasets for understanding group activities, as shown in Table 1. Choi *et al.* [10] proposed the first dataset, Collective Activity Dataset (CAD), consisting of real-world pedestrian sequences. Then, Choi *et al.* [11] extended CAD to CAED by adding two new actions (*i.e.*, "Dancing" and "Jogging") and removing the ill-defined action (*i.e.*, "Walking"). There is no specific group activity defined in CAD and CAED, in which the scenarios are assigned group activities based on majority voting. Moreover, Choi and Savarese [9] collected a Choi's New Dataset (CND) composed of many artificial pedestrian sequences. Recently, Ibrahim *et al.* [18] introduced a sports video dataset, Volleyball Dataset (VD), which contains numerous volleyball games. However, as the largest and most popular dataset, VD contains quite a few wrong labels that directly affect the evaluation of proposed approaches. In addition, Ramanathan *et al.* [32] released NCAA but few researchers have used it for GAR since only YouTube video links are provided and many of them are dead now. Some activities (*e.g.*, "steal", "slam dunk *" and "free-throw *") in NCAA can be recognized using one key frame, which actually evades from some key challenges of GAR. Limited by the size and quality of the above datasets, the recent studies of group activity recognition have encountered the bottleneck. In this work, we collect a larger and more challenging dataset from the basketball games and do not provide any person-level information (*i.e.*, the bounding boxes and action labels for each person), thanks to the weakly-supervised setting. Moreover, compared with previous benchmarks, our NBA contains more activities that involve long-term temporal structure and are fast-moving.

Relational Reasoning. Recently, relationships among entities (*i.e.*, pixels, objects or persons) have been widely leveraged in various computer vision tasks, such as Visual Question Answering [34,20,5], Scene Graph Generation [21,26,46], Object Detection [17,8], and Video Understanding [47,43,29]. Santoro *et al.* [34] presented a relational network module to infer the potential relationships among objects for improving the performance of visual question answering. Hu *et al.* [17] embedded a relation module into existing object detection systems for simultaneously detecting a set of objects and interactions between their appearance and geometry. Besides the spatial relationship among objects in the image, some recent works also explored the temporal relational structure of the video. Liu *et al.* [29] proposed a novel neural network to learn video representations by capturing potential correspondences for each feature point. Moreover, some recent methods [13,31,44] explored the spatial relationships between each people in group activities. In this work, we apply relational reasoning to choose the most relevant people from a number of proposals for weakly-supervised GAR.

3 Weakly-supervised Group Activity Recognition

3.1 Weakly-supervised Setting

For a more practical group activity recognition, i) the number of people in the scene varies over different activities even time, and ii) the person-level annotations cannot be provided in real-world applications. Therefore, we introduce a weakly-supervised setting that *only video-level labels are available, yet the location and action label of each person are not provided.*

In this work, the task of recognizing group activity under this setting is called weakly-supervised GAR that aims to directly recognize the activity performed by multiple collectively from the video with only a video-level label during training. Apparently, weakly-supervised GAR can be applied to more complex and real-world applications (*e.g.*, real-time sports analysis and video surveillance) which cannot provide fine-grained supervision. Besides, the weakly-supervised setting eases the annotation of benchmarks for the task. Without annotating the person-level supervision, we only require $\frac{1}{2K+1}$ tagging labor[1] as before where K is the number of people in the scene.

3.2 The NBA Dataset for Weakly-supervised GAR

Under the weakly-supervised setting, we introduce a new video-based dataset, the NBA dataset. It describes the group activities that are common in basketball games. There is no annotation for each person and only a group activity label assigned to each clip. To the best of our knowledge, it is currently the largest and most challenging benchmark for group activity analysis, as shown in Table. 1.

[1] The fully-supervised setting requires K boxes, K actions, and 1 group activity, but the weakly-supervised setting only needs 1 group activity label. We roughly assumed the same labor for each annotation.

Table 1. Comparison of the existing datasets for group activity recognition

Dataset	# Videos	# Clips	# Individual Actions	# Group Activities	Activity Speed	Camera Moving
CAD [10]	44	$\approx 2,500$	5	5	slow	N
CAED [11]	30	$\approx 3,300$	6	6	slow	N
CND [9]	32	$\approx 2,000$	3	6	slow	N
VD [18]	55	$4,830$	9	8	medium	Y
NBA (ours)	181	$9,172$	-	9	fast	Y

We will introduce the NBA dataset from the following aspects: the source of the video data, the effective annotation strategy, and the statistics of this dataset.

Data Source. It is a natural choice to collect videos of team sports for studying group activity recognition. In this work, we collect a subset of the 181 NBA games of 2019 periods from the web. Compared with the activities in volleyball games [18], the ones in basketball games have more long-term temporal structure and fast moving-speed, which brings up new challenges to group activity analysis. For one thing, the number of players may vary over different frames. On the other hand, the activity is so fast that the single-frame based person-level annotation is useless to track these players. Therefore, it is difficult to label all people in these videos which differs from volleyball games, thus we annotate this benchmark under the weakly-supervised setting. Due to the copyright restriction, this dataset is available upon request.

Annotation. Given a video, the goal of annotation is to assign the group activities to the corresponding segments. It is time-consuming to manually label such a huge dataset with conventional annotation tools. To improve the annotation efficiency, we take full advantage of the logs provided by the NBA's official website and design a simple and automatic pipeline to label our dataset. There are three steps: i) Filter out some unwanted records in the log file corresponding with a video. ii) Identify the timer in each frame by Tesseract-OCR [37] and match it with the valid records generated from step i. iii) Save the segments with a fixed length according to the time points obtained from step ii.

Statistics. We collect a total of 181 videos with a high resolution of 1920×1080. Then we divide each video into 6-second clips by the above-mentioned annotation method and sub-sample them to 12fps. Besides, we remove some abnormal clips that contain close-up shots of players or instant replays. Ultimately, there are a total of $9,172$ video clips, each of which belongs to one of the 9 activities. Here, we drop some activities such as "dunk" and "turnover" due to the limited sample size, and do not use "free-throw" that is easy to be distinguished. We randomly select $7,624$ clips for training and $1,548$ clips for testing. Table 2 shows the sample distributions across different categories of group activities and the corresponding average number of people in the scene.

Table 2. Statistics of the group activity labels in NBA. “2p”, “3p”, “succ”, “fail”, “def” and “off” are abbreviations of “two points”, “three points”, “success”, “failure”, “defensive rebound” and “offensive rebound”, respectively

Group Activity		2p -succ.	2p -fail. -off.	2p -fail. -def.	2p -layup -succ.	2p -layup -fail. -off.	2p -layup -fail. -def.	3p -succ.	3p -fail. -off.	3p -fail. -def.
# clips	Train	798	434	1316	822	455	702	728	519	1850
	Test	163	107	234	172	89	157	183	83	360

4 Approach

4.1 Mining Key Instances via Social Relationship

In general, the key and difficult point in obtaining category information from visual input is to construct and learn their intermediate representation. For the task of group activity recognition, such intermediate representation made up of individual features and underlying relationships among them, refers to *social-representation* in this paper. The previous fully-supervised setting [18,42,45] provides a variety of extra fine-grained supervision information (*e.g.*, ground-truth bounding box and action label for each person, and even the interaction label for each person-person pair) to ensure that social-representation can be constructed and learned stably during training. However, under the weakly-supervised setting that only provides video-level labels, it is difficult for models to define and learn discriminative social-representation stably.

To this end, we propose a simple yet effective framework, as illustrated in Fig. 2, to stabilize the weakly-supervised training for GAR. The core idea of our approach is to firstly construct all possible social-representation and then find the effective ones based on the social assumption that **key instances (people/frames) are closely related to each other.** Formally, given a sequence of frames $(V_1, V_2, \cdots, V_T)$, our approach models them as follow:

$$O = \mathcal{O}(\mathcal{F}(V_1; \mathcal{D}(V_1); \mathbf{W}), \mathcal{F}(V_2; \mathcal{D}(V_2); \mathbf{W}), \cdots, \mathcal{F}(V_T; \mathcal{D}(V_T); \mathbf{W})). \tag{1}$$

Here, $\mathcal{D}(V_t)$ represents detecting N^{p} proposals from each frame. There are two choices to determine the value of N^{p} as follows, i) **Quantity-aware:** empirically select top-N^{p} boxes from numerous proposals; ii) **Probability-aware:** choose the boxes whose probability is larger than a threshold θ.

The spatial modeling function $\mathcal{F}(V_t; \mathcal{D}(V_t); \mathbf{W}))$ represents that i) adopt CNN with parameters $\mathbf{W}$ to extract the convolutional feature map for frame V_t, ii) apply RoIAlign [15] to extract person-level features according to the corresponding proposals from $\mathcal{D}(V_t)$, and iii) fuse person-level features into a single frame-level vector. However, without person-level annotation, it is unavoidable for $\mathcal{D}(\cdot)$ to get many useless proposals from each frame. Moreover, the number

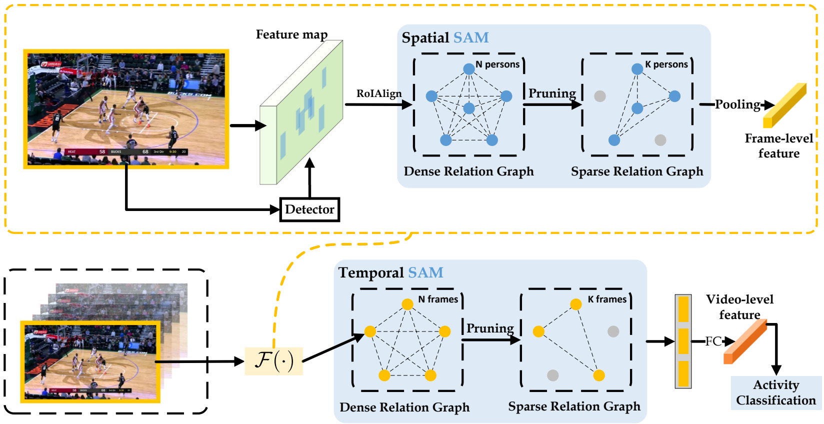


Fig. 2. Overview of our approach for weakly-supervised GAR. The inputs are a set of frames and the associating pre-detected bounding boxes for people. We apply SAM to concurrently select discriminative person-level features in the spatial domain and effective frame-level representations in the temporal domain (Best viewed in color)

of proposals (N^{p}) varies over samples in practical applications. Thus, $\mathcal{F}(\cdot)$ needs to be able to choose K^{p} discriminative person-level features in the spatial domain. $\mathcal{O}(\cdot)$ is a temporal modeling function that samples a set of N^{f} frames from the entire video sequence (T frames) as the input of our approach according to the sampling strategy used in [41]. However, the long temporal structure of the activities in our NBA dataset will bring numerous irrelevant frames that may affect the construction of social-representation. Therefore, we also hope $\mathcal{O}(\cdot)$ can select K^{f} effective frame-level representations in the temporal domain.

It is clear that $\mathcal{F}(\cdot)$ and $\mathcal{O}(\cdot)$ need to have similar properties that attending to effective person/frame-level features in the spatial and temporal domains, respectively. Therefore, this work aims at endowing the function $\mathcal{F}(\cdot)$ and $\mathcal{O}(\cdot)$ with the ability of feature selection according to the social assumption that key instances are highly related to each other.

4.2 Social Adaptive Module (SAM)

Inspired by relational reasoning [34,43,47], we build a generic module, namely Social Adaptive Module, to implement the idea of assisting weakly-supervised training with the social assumption. Specifically, we abstract $\mathcal{F}(\cdot)$ and $\mathcal{O}(\cdot)$ into a unified form as

$$\begin{aligned}\mathbf{Z} = \mathcal{M}(\mathbf{X}) = \{\mathbf{a} \mid \mathbf{a} \in \{\lambda_1\mathcal{E}(\mathbf{x}_1), \lambda_2\mathcal{E}(\mathbf{x}_2), \cdots, \lambda_N\mathcal{E}(\mathbf{x}_N)\},\\ \mathbf{a} \neq \mathbf{0}, \mathbf{x}_i \in \mathbf{X}, \lambda_i \in \{0,1\}, \|\boldsymbol{\lambda}\|_1 = K\},\end{aligned} \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{N \times D}$ and $\mathbf{Z} \in \mathbb{R}^{K \times D}$ are the input and output of $\mathcal{M}(\cdot)$, respectively, and $K \leq N$. Put simply, $\mathcal{M}(\cdot)$ aims to learn the parameter $\boldsymbol{\lambda} \in \mathbb{R}^N$, a zero-one vector used to select K discriminative ones from N input feature nodes. $\mathcal{E}(\cdot)$ is the embedding function for input and is optional. We hold that $\boldsymbol{\lambda}$ will be effective for feature selection only if driven by $\mathbf{X}$. Moreover, not only $N \neq K$ but also the value of N varies over samples. Therefore, directly replacing the function $\mathcal{F}(\cdot)$ and $\mathcal{O}(\cdot)$ in Eq.(1) with $\mathcal{M}(\cdot)$ is difficult for our approach to be optimized.

In this work, we approximate the solution of $\boldsymbol{\lambda}$ via pruning a Dense Relation Graph with N nodes to a Sparse Relation Graph with K nodes. Specifically, we build a dense relation graph on N input features to measure relationships between each other. During the process of pruning, we aim at maintaining the top-K feature nodes of the graph according to their relatedness. Based on the K selected features, a sparse relation graph is built to perform relational embedding for them. The details are described as follows.

Dense Relation Graph. We first build dense relationships between each input node, based on their visual features. More specifically, given a set of feature vectors as $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, we compute the directional relation between them as $r_{ij} = g(\mathbf{x}_i, \mathbf{x}_j)$ where i, j are indices and $g(\cdot, \cdot)$ is the relation function. There are several common implementations [29,34] of $g(\cdot, \cdot)$. For instance, we can measure the L_2 distance between each feature, but which is not a data-driven and learnable method. Besides that, we can treat the concatenation $[x_i, x_j]$ as the input of a multi-layer perceptron to get the relation score. However, as the number of pairs increases, this approach will consume a lot of memory and computation. In this work, we adopt a learnable and low-cost function to measure the relation between i-th and j-th feature node as $g(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^\top \Psi(\mathbf{x}_j)$, where $\Phi(\cdot)$ and $\Psi(\cdot)$ are two embeddings of i-th and j-th feature node, respectively. Based on this formulation, the calculation of relation matrices, $\mathbf{R} = \{r_{ij}\}^{N \times N}$, can be implemented by only two embedding processes and a matrix multiplication. We also apply a *softmax* computation along the dimension j of the matrix $\mathbf{R}$.

Pruning Operation. To approximate the solution of $\boldsymbol{\lambda}$ in Eq.(2), this paper select the K most relevant nodes from the above dense graph based on the social assumption that key instances are likely to be related to each other. Concretely, after obtaining the $N \times N$ relation matrix for all feature pairs, we construct the *relatedness* for each feature node as $\alpha_i = \sum_{j=1}^{N} (r_{ij} + r_{ji})$, where r_{i*} and r_{*i} denote the out-edges and in-edges of i-th feature node in the dense relation graph. Intuitively, the nodes with strong connections can be easily retained in the graph. Thus, we hold that the sum of a specific node's corresponding connections can depict the importance (*relatedness*) of itself. Based on the social assumption, we sort the values of $\boldsymbol{\alpha} \in \mathbb{R}^N$ in descending order and select the top-K values denoted as $\texttt{topk}(\boldsymbol{\alpha}) \in \mathbb{R}^K$. Thus, the satisfactory $\boldsymbol{\lambda}$ can be expressed as

$$\lambda_i = \begin{cases} 1, & \alpha_i \in \texttt{topk}(\boldsymbol{\alpha}), \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

Sparse Relation Graph. According to $\boldsymbol{\lambda}$, we can get the corresponding K selected feature, $\hat{\mathbf{X}} \in \mathbb{R}^{K \times D}$, namely sparse feature. However, $\boldsymbol{\lambda}$ is driven by $\mathbf{R}$, but $\hat{\mathbf{X}}$ is unrelated to it. Therefore, $\boldsymbol{\lambda}$ will be unlearnable if we directly regard $\hat{\mathbf{X}}$ as the output of this module. To tackle this problem, we construct a relational embedding $\mathcal{E}(\cdot)$ for the sparse feature $\hat{\mathbf{X}}$ by combing with relation matrix $\mathbf{R}$. Similarly, we obtain a sparse relation matrix $\hat{\mathbf{R}} = \{\hat{r}_{ij}\}^{K\times K}$ associating to the K selected feature, and then perform relational embedding as

$$\mathbf{z}_i = \mathcal{E}(\hat{\mathbf{x}}_i) = \mathbf{W}_z\Big(\sum_{j=1}^{K}\hat{r}_{ij}\mathit{\Omega}(\hat{\mathbf{x}}_j)\Big) + \hat{\mathbf{x}}_i. \qquad (4)$$

Here "+" denotes a residual connection, $\mathit{\Omega}(\cdot)$ is the embedding of sparse feature $\hat{\mathbf{x}}_j$, and $\mathbf{W}_z$ is a weight vector that projects the relational feature to the new representation with the same dimension as the sparse feature $\hat{\mathbf{x}}_i$.

SAM is the first to introduce the "social assumption" that helps a lot in the GAR scenario where many uncertain inputs are involved. More importantly, this module makes our method more appropriate to work in the weakly-supervised setting and can also be used upon previous methods [45,40,13,31,39].

4.3 Implementation details

Person Detection & Feature Extraction. For each frame, we first adopt Faster-RCNN [33] pre-trained on the MS-COCO [28] to detect possible persons in the scene, based on the mmdetection toolbox [7]. Then, we track them over all frames by correlation tracker [12] implemented by Dlib [22]. After that, we adopt ResNet-18 [16] as the backbone to extract the convolutional feature map for each frame. Finally, we get the aligned feature for each proposal from the map by RoIAlign [15] with the crop size of 5×5 and embed it to 1024 dimensional feature vector by a fully connected layer.

Social Adaptive Module. This module is designed to select out K effective feature from N input ones. However, the values of N and K depend on the situation and will be explained in experiments. If N varies over samples (*e.g.*, different numbers of proposals are generated by the Probability-aware strategy mentioned in Section 4.1), we feed data into this module with a batch-size of 1 but do not change the batch-size of the entire framework. The $\mathit{\Phi}(\cdot)$, $\mathit{\Psi}(\cdot)$, and $\mathit{\Omega}(\cdot)$ used to embed input feature are implemented by 1×1 convolutional layers.

Optimization. We adopt the ADAM to optimize our approach with fixed hyper-parameters ($\beta_1 = \beta_2 = 0.9$, $\varepsilon = 10^{-4}$) and train it in 30 epochs with an initial learning rate of 0.0001 that is reduced to 1/10 of the previous value for every 5 epochs. Compared with SSU [4] and ARG [44] that require pre-training the CNN backbone and fine-tuning the top model separately, our approach excluding detection can be optimized in an end-to-end fashion.

5 Experiments

5.1 Quantitative Analysis on the NBA Dataset

We first evaluate our approach on the new benchmark by compared with several variants and baseline methods. For this dataset, we sample $N^{\mathrm{f}} = 20$ frames from

Table 3. Ablation studies on NBA. Quan-N^{p} and Prob-N^{p} are two different strategies of deciding the number of input proposals, as described in Section 4.1. θ is the probability threshold used in Prob-N^{p}, N^{f} is the number of input frames, and K^* denote the number of feature selected by our SAM

Type	Options of Our Approach	Acc (%)	Mean Acc (%)
Quan-N^{p}	B1: w/o SAM ($N^{\mathrm{p}} = 8$)	44.6	39.5
	B2: w/ Spatial-SAM ($N^{\mathrm{p}} = 14, K^{\mathrm{p}} = 14$)	46.8	41.3
	B3: w/ Spatial-SAM ($N^{\mathrm{p}} = 14, K^{\mathrm{p}} = 8$)	**50.3**	43.6
	B4: w/ Spatial-SAM ($N^{\mathrm{p}} = 8, K^{\mathrm{p}} = 8$)	47.4	41.4
	B5: w/ Spatial-SAM ($N^{\mathrm{p}} = 14, K^{\mathrm{p}} = 8$) + w/ Temporal-SAM ($N^{\mathrm{f}} = 20, K^{\mathrm{f}} = 6$)	49.1	**47.5**
Prob-N^{p}	B6: w/ Spatial-SAM ($\theta = 0.9, K^{\mathrm{p}} = 8$)	47.5	42.6

the entire video clip as the input for all methods and train them with a batch-size of 16. Because of the fast speed of activities in this benchmark, we do not track pre-detected proposals over frames. Moreover, we do not apply any strategy to handle the class-imbalance issue in this benchmark.

Ablation Study. To evaluate the effectiveness of SAM, different variants of our approach are performed on NBA and the results are reported in Table 3. B1 that does not use the proposed SAM achieves the base accuracy of 44.6% and 39.5% on Acc and Mean Acc, respectively. Compared with B1, B2 that employs the SAM to build relational embeddings among $N^{\mathrm{p}} = 14$ proposals in the spatial domain but does not prune useless ones, only obtains 2.2% and 1.8% improvement on Acc and Mean Acc. Similarly, B4 that directly adapts SAM to generate relational representations from $N^{\mathrm{p}} = 8$ proposals has small improvement. However, by selecting $K^{\mathrm{p}} = 8$ persons from $N^{\mathrm{p}} = 14$ proposals and modeling relationships among them, B3 improves Acc and Mean Acc by 5.7% and 4.1%, respectively, compared with B1. Moreover, our Quan-N^{p} based approach (B6) suffering an uncertain number of proposals also gets a satisfactory Mean Acc of 42.6%. Based on B3, B5 obtains the best Mean Acc by applying SAM on the temporal domain, suggesting that the ability of feature selection of SAM can also be used to capture the long temporal structure. The further analysis on the parameters of N^* and K^* are present in Section 5.2.

Comparison with the baselines. We also compare our approach with recent works in the video classification domain, including TSN [41], TRN [47], I3D [6], I3D+NLN [43]. To be fair, all these baseline methods are built on ResNet-18 and the input modality is RGB. The results are reported in Table 4. We see that "Ours w/o SAM" is hardly improved or worse due to noisy input (irrelevant pre-detected proposals), compared with methods ("TSN" and "TRN") only using frame-level information. By introducing SAM to select discriminative proposals in the spatial domain, "Ours w/ SAM (S)" achieves significant improvement on Mean Acc but still overfits on some classes. As expected, "Ours

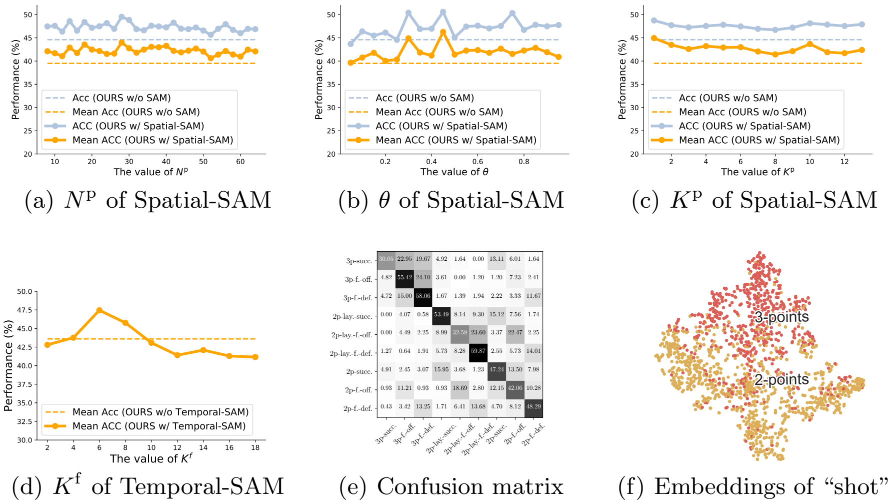
Table 4. Comparison on NBA. "Ours w/o SAM", "Ours w/ SAM (S)", and "Ours w/ SAM (S+T)" are the B1, B3, and B5 reported in Table 3, respectively

Group Activity	Frame Classification				Our Approach		
	TSN [41]	TRN [47]	I3D [6]	I3D+NLN [43]	w/o SAM	w/ SAM (S)	w/ SAM (S+T)
2p-succ.	38.7	44.8	33.1	22.1	46.6	39.3	**47.2**
2p-fail.-off.	30.8	23.4	14.0	20.6	28.0	25.2	**42.1**
2p-fail.-def.	49.1	50.0	39.3	45.3	49.6	**53.4**	48.3
2p-layup-succ.	52.9	54.7	50.6	48.8	44.2	**57.6**	53.5
2p-layup-fail.-off.	10.1	22.5	22.5	22.5	20.2	19.1	**32.6**
2p-layup-fail.-def.	44.6	46.5	43.3	31.2	44.6	51.6	**59.9**
3p-succ.	39.3	37.7	31.1	26.8	39.9	**41.0**	30.1
3p-fail.-off.	10.8	20.5	4.8	12.0	24.1	38.6	**55.4**
3p-fail.-def.	63.9	62.8	55.3	61.7	58.6	**66.9**	58.1
Mean Acc (%)	37.8	40.3	32.7	32.3	39.5	43.6	**47.5**

w/ SAM (S+T)" outperforms all baselines by a good margin and obtained the best Mean Acc by applying SAM to both the spatial and temporal domains. Nevertheless, "Ours w/ SAM (S+T)" performs poorly on the activity of "3p-succ." that does not have long-term temporal structure. Moreover, "I3D" and "I3D+NLN" that depend on dense frames perform poorly on this benchmark.

5.2 Qualitative Analysis on the NBA dataset

Analysis of parameters. We first diagnose N, the number of nodes of the dense relation graph. Limited by the computation resource, we only analyze the N^{p} of Spatial-SAM and it indicates how many pre-detected proposals should be fed into our approach. It can be decided by two strategies as mentioned in Section 4.1. Thus, we first run our Quan-N^{p} based approach on the NBA dataset by fixing $K^{\mathrm{p}} = 8$ and changing N^{p} from 8 to 64 with a step of 4. As shown in Fig. 3(a), although N^{p} is increasing, the performance of our approach has been persistently higher than the baseline. Moreover, we also conduct our Prob-N^{p} based approach on NBA by using fixed $K^{\mathrm{p}} = 8$ and adjust θ from 0.05 to 0.95 with a mini-step of 0.05. As shown in Fig. 3(b), our approach can achieve promising results when $\theta \geq 0.3$ and is more likely to get high performance when θ around 0.4. Overall, our Spatial-SAM is not sensitive to N^{p} whether decided by Quan-N^{p} or Prob-N^{p}. We also diagnose K, the number of nodes of the sparse relation graph, and it decides how many feature nodes need to be selected for modeling. As shown in Fig. 3(c), the performance of Spatial-SAM maintains over the baseline and it obtains the best result at $K^{\mathrm{p}} = 1$. Therefore, we hold that Spatial-SAM is not sensitive to K^{p}. By contrast, the performance of Temporal-SAM cannot get satisfactory performance when the K^{f} is too small or large, due



(a) N^{p} of Spatial-SAM (b) θ of Spatial-SAM (c) K^{p} of Spatial-SAM

(d) K^{f} of Temporal-SAM (e) Confusion matrix (f) Embeddings of "shot"

Fig. 3. (a)-(d) Experimental analysis on parameters. (e) The confusion matrix of OURS w/ Spatial-SAM and Temporal-SAM. (f) t-SNE visualization of embeddings of 2/3-points based activities. These experiments are carried out on the NBA dataset

to the different temporal length of activities in NBA. However, our approach with Temporal-SAM significantly improves Mean Acc when $4 < K^{\mathrm{f}} < 10$.

Confusion matrix. To figure out the confusion between each activity in the NBA dataset, we report the confusion matrix of our approach in Fig.3(d). We can see that the activities involving "defense" and "offense" are easily confused, due to the class-imbalance issue between these two kinds of activities. However, it is relatively easy to distinguish 2-points and 3-points, as embeddings shown in Fig.3(f). Because 3-point players usually jump to shot behind the 3-point line without blocking. By contrast, 2-point players are often blocked by others.

Visualization. To further understand the discriminative learning process of SAM, we show some typical cases of NBA in Fig.4. The group activities in NBA have long-term temporal information, thus top-K proposals vary over time. Take the rightmost one as an example, a "3p-failure-defense" has 3 parts: (1) preparation, (2) shooting, (3) defensive rebound. For (1) and (2), the players controlling the ball are the key instances, but for (3), the players that quickly turn back are the key instances. It is not hard to find that SAM aims at focusing on the players who are controlling the basketball or close to it and these people can form a group semantically.

5.3 Quantitative Analysis on Volleyball Dataset

We also evaluate our approach on the existing largest and most widely-used benchmark, Volleyball Dataset (VD) [18] consisting of 4830 volleyball game sequences. The middle frame of each sequence is labeled with 9 action labels (not

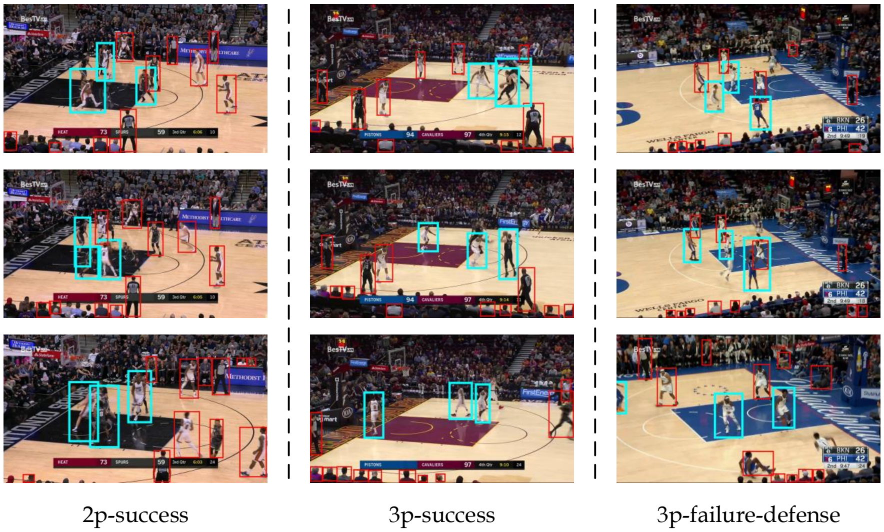


Fig. 4. A visualization of the top-K proposals focused by SAM over time on the NBA dataset, where $K = 3$. Each column shows three different frames of an activity. We highlight the top-K players (in cyan boxes) at three time steps of different activities. The people in red boxes are treat as noisy data by our model

used in our approach) and 8 group activity labels. However, we find that there are many wrong annotations between "pass" and "set", which seriously affects the evaluation for models, thus we merged them into "pass-set". To be fair, we follow the train/test split provided in [18] and sample $N^{\mathrm{f}} = 3$ from the video clip similar to [44]. Because the activities in VD always occur in the middle frame, we do not apply our SAM to the temporal domain for this benchmark.

Ablation Study. We also perform ablation study on VD and the experimental results are reported in Table 5(a). All these variants do not use person-level supervision information (bounding boxes and action labels) provided by [18] and are built on ResNet-18. Compared with the baseline method B1, our B2 and B3 that only apply SAM to generate relational embedding for proposals but do not prune the irrelevant ones, only improve the accuracy by 0.9% and 0.4%, respectively. Besides, by using SAM to build relationships among $N = 16$ proposals and choosing $K = 12$ effective proposals from them, B3 and B5 improve the accuracy of 1.6% based on whether Quan-N or Prob-N. This observation indicates again that useless proposals will affect the weakly-supervised training and SAM is effective for pruning them.

Comparison with the state-of-the-art. Referring to [42], we report the results of HTDM [18,19], PCTDM [45], CCGL [39], and StagNet [31] by computing their corresponding confusion matrices. We reproduce the state-of-the-art method, ARG [44], with fully-supervised and weakly-supervised settings, respectively. As shown in Table 5(b), our weakly-supervised approach with the back-

Table 5. Results on VD. (a) Ablation studies. (b) Comparison with SOTA. “Ours” represents “Ours w/ Spatial-SAM” with $N^{\mathrm{p}} = 16$ and $K^{\mathrm{p}} = 12$ based on Quan-N^{p}

(a)

Type	Our Approach	Acc (%)
Quan-N^{p}	B1: w/o SAM	91.5
	B2: w/ Spatial-SAM (N^{p}=16, K^{p}=16)	92.4
	B3: w/ Spatial-SAM (N^{p}=16, K^{p}=12)	**93.1**
	B4: w/ Spatial-SAM (N^{p}=12, K^{p}=12)	91.9
Prob-N^{p}	B5: w/ Spatial-SAM ($\theta = 0.9, K^{\mathrm{p}} = 12$)	**93.1**

(b)

Method	Supervision	Acc (%)
HTDM	Fully	89.7
PCTDM	Fully	90.2
CCGL	Fully	91.0
StagNet	Fully	90.0
‡ARG	Fully	**94.0**
‡ARG	Weakly	90.7
†Ours	Weakly	93.1
‡Ours	Weakly	**94.0**

† ResNet-18
‡ Inception-v3

bone of ResNet-18 is superior to almost all previous fully-supervised methods, except ARG that is built on Inception-v3. But our approach goes far beyond ARG under the weakly-supervised setting, suggesting that useless pre-detected proposals seriously affect the construction of relation graphs in ARG. Furthermore, our approach with Inception-v3 can achieve the best performance.

6 Conclusions

In this work, we introduce a weakly-supervised setting for GAR, which is more practical and friendly for real-world scenarios. To investigate this problem, we collect a larger and more challenging dataset from high-resolution basketball videos of NBA. Furthermore, we propose a social adaptive module (SAM) for assisting the weakly-supervised training by leveraging the social assumption that discriminative features are highly related to each other. As demonstrated on two datasets, our approach achieves state-of-the-art results while it can attend to key proposals/frames automatically.

This work reveals that social relationship among visual entities is helpful for high-level semantic understanding. We look forward to applying this method to more challenging scenarios, in particular, for mining semantic knowledge from weakly-annotated or un-annotated visual data.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2018AAA0102002, the National Natural Science Foundation of China under Grants 61732007, 61702265, and 61932020.

References

1. Amer, M.R., Xie, D., Zhao, M., Todorovic, S., Zhu, S.C.: Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In: ECCV (2012)
2. Amer, M.R., Lei, P., Todorovic, S.: Hirf: Hierarchical random field for collective activity recognition in videos. In: ECCV (2014)
3. Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: CVPR (2019)
4. 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: CVPR (2017)
5. Cadene, R., Ben-Younes, H., Cord, M., Thome, N.: Murel: Multimodal relational reasoning for visual question answering. In: CVPR (2019)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
7. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
8. Chen, X., Gupta, A.: Spatial memory for context reasoning in object detection. In: ICCV (2017)
9. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: ECCV (2012)
10. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: ICCV Workshops (2009)
11. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR (2011)
12. Danelljan, M., Häger, G., Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: BMVC (2014)
13. Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: CVPR (2016)
14. Gan, C., Wang, N., Yang, Y., Yeung, D.Y., Hauptmann, A.G.: Devnet: A deep event network for multimedia event detection and evidence recounting. In: CVPR (2015)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
17. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: CVPR (2018)
18. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: CVPR (2016)
19. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: Hierarchical deep temporal models for group activity recognition. arXiv preprint arXiv:1607.02643 (2016)
20. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: CVPR (2017)

21. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR (2015)
22. King, D.E.: Dlib-ml: A machine learning toolkit. JMLR (2009)
23. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: ICCV (2011)
24. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: CVPR (2012)
25. Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing contextual group activities. TPAMI (2012)
26. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and caption regions. In: ICCV (2017)
27. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: ICCV (2019)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
29. Liu, X., Lee, J.Y., Jin, H.: Learning video representations from correspondence proposals. In: CVPR (2019)
30. Long, X., Gan, C., De Melo, G., Wu, J., Liu, X., Wen, S.: Attention clusters: Purely attention based local feature integration for video classification. In: CVPR (2018)
31. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Van Gool, L.: stagnet: An attentive semantic rnn for group activity recognition. In: ECCV (2018)
32. Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., Fei-Fei, L.: Detecting events and key actors in multi-person videos. In: CVPR. pp. 3043–3053 (2016)
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
34. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: NeurIPS (2017)
35. Shu, T., Todorovic, S., Zhu, S.C.: Cern: confidence-energy recurrent network for group activity recognition. In: CVPR (2017)
36. Shu, T., Xie, D., Rothrock, B., Todorovic, S., Chun Zhu, S.: Joint inference of groups, events and human roles in aerial videos. In: CVPR (2015)
37. Smith, R.: An overview of the tesseract ocr engine. In: ICDAR (2007)
38. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
39. Tang, J., Shu, X., Yan, R., Zhang, L.: Coherence constrained graph lstm for group activity recognition. TPAMI (2019)
40. Tang, Y., Wang, Z., Li, P., Lu, J., Yang, M., Zhou, J.: Mining semantics-preserving attention for group activity recognition. In: ACM MM (2018)
41. 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: ECCV (2016)
42. Wang, M., Ni, B., Yang, X.: Recurrent modeling of interaction context for collective activity recognition. In: CVPR (2017)
43. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
44. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: CVPR (2019)
45. Yan, R., Tang, J., Shu, X., Li, Z., Tian, Q.: Participation-contributed temporal dynamic model for group activity recognition. In: ACM MM (2018)

46. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: ECCV (2018)
47. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV (2018)