

Joint Learning of Social Groups, Individuals Action and Sub-group Activities in Videos

Mahsa Ehsanpour^{1,3}, Alireza Abedin¹, Fatemeh Saleh^{2,3}, Javen Shi¹,
Ian Reid^{1,3}, and Hamid Rezaatofghi¹

¹ The University of Adelaide

² Australian National University

³ Australian Centre for Robotic Vision

`mahsa.ehsanpour@adelaide.edu.au`

Abstract. The state-of-the art solutions for human activity understanding from a video stream formulate the task as a spatio-temporal problem which requires joint localization of all individuals in the scene and classification of their actions or group activity over time. Who is interacting with whom, e.g. not everyone in a queue is interacting with each other, is often not predicted. There are scenarios where people are best to be split into sub-groups, which we call social groups, and each social group may be engaged in a different social activity. In this paper, we solve the problem of simultaneously grouping people by their social interactions, predicting their individual actions and the social activity of each social group, which we call the social task. Our main contributions are: i) we propose an end-to-end trainable framework for the social task; ii) our proposed method also sets the state-of-the-art results on two widely adopted benchmarks for the traditional group activity recognition task (assuming individuals of the scene form a single group and predicting a single group activity label for the scene); iii) we introduce new annotations on an existing group activity dataset, re-purposing it for the social task. The data and code for our method is publicly available ⁴.

Keywords: Collective behaviour recognition, Social grouping, Video understanding.

1 Introduction

Recognising individuals' action and group activity from video is a widely studied problem in computer vision. This is crucial for surveillance systems, autonomous driving cars and robot navigation in environments where humans are present [5,12,34]. In the last decade, most effort from the community has been dedicated to extract reliable spatio-temporal representations from video sequences. Dominantly, this was investigated in a simplified scenario where each video clip was trimmed to involve a single action, hence a classification problem [6,35,53,54]. Recently, the task of human activity understanding has been extended to more challenging and realistic scenarios, where the video is untrimmed

⁴ <https://github.com/mahsaep/Social-human-activity-understanding-and-grouping>



Fig. 1. Examples of our annotated data representing social groups of people in the scene and their common activity within their group. Realistically, individuals in the scene may perform their own actions, but they may also belong to a social group with a mutual activity, *e.g.*, *walking together*. In this figure, social groups and their corresponding common activities have been color-coded.

and may include multiple individuals performing different actions [22,23,28,56]. In parallel, there are works focusing on predicting a group activity label to represent a collective behaviour of all the actors in the scene [11,14,37]. There are also independent works aiming at only grouping individuals in the scene based on their interactions [7,20,25,47,57] or joint inferring of groups, events and human roles in aerial videos [49] by utilizing hand-crafted features.

In a real scenario, a scene may contain several people, individuals may perform their own actions while they might be connected to a social group. In other words, a real scene generally comprises several groups of people with potentially different social connections, *e.g.* contribution toward a common activity or goal (Fig. 1). To this end, in our work we focus on the problem of “*Who is with whom and what they are doing together?*”. Although the existing works mentioned in the previous paragraphs tackle some elements, we propose a holistic approach that considers the multi-task nature of the problem, where these tasks are not completely independent, and which can benefit each other. Understanding of this scene-wide social context would be conducive for many video understanding applications, *e.g.* anomalous behaviour detection in the crowd from a surveillance footage or navigation of an autonomous robot or car through a crowd.

To tackle this real-world problem, we propose an end-to-end trainable framework which takes a video sequence as input and learns to predict *a)* each individual’s action; *b)* their social connections and groups; and, *c)* a social activity for each predicted social group in the scene. We first introduce our framework for a relevant conventional problem, *i.e.* group activity recognition [11,14,37]. We propose an architecture design that incorporates: *i)* *I3D backbone* [23] as a state-of-the-art feature extractor to encode spatio-temporal representation of individuals in a video clip, *ii)* *Self-attention module* [59] to refine individuals’ feature representations, and *iii)* *Graph attention module* [60] to directly model the interactions and connections among the individuals. Our framework outperforms the state-of-the-art on two widely adopted group activity recognition datasets. We then introduce our extended framework that can elegantly handle social grouping and social activity recognition for each group. We also augment

an existing group activity dataset with enriched social group and social activity annotations for the social task. Our main contributions are:

1. We propose an end-to-end framework for the group activity recognition task through integration of I3D backbone, self-attention module and graph attention module in a well-justified architecture design. Our pipeline also outperforms existing solutions and sets a new state-of-the-art for the group activity recognition task on two widely adopted benchmarks.
2. We show that by including a graph edge loss in the proposed group activity recognition pipeline, we obtain an end-to-end trainable solution to the multi-task problem of simultaneously grouping people, recognizing individuals' action and social activity of each social group (social task).
3. We introduce new annotations, *i.e.* social groups and social activity label for each sub-group on a widely used group activity dataset to serve as a new benchmark for the social task.

2 Related Work

Action Recognition. Video understanding is one of the main computer vision problems widely studied over the past decades. Deep convolutional neural networks (CNNs) have shown promising performance in action recognition on short trimmed video clips. A popular approach in this line involves adoption of two-stream networks with 2D kernels to exploit the spatial and temporal information [18,19,53,61]. Recurrent neural networks (RNNs) have also been utilized to capture the temporal dependencies of visual features [15,41]. Unlike these approaches, [30,58] focused on CNNs with 3D kernels to extract features from a sequence of dense RGB frames. Recently, [6] proposed I3D, a convolutional architecture that is based on inflating 2D kernels pretrained on ImageNet into 3D ones. By pretraining on large-scale video datasets such as Kinetics [32], I3D outperforms the two-stream 2D CNNs on video recognition benchmark datasets.

Spatio-temporal Action Detection. Temporal action detection methods aim to recognize humans' actions and their corresponding start and end times in untrimmed videos [51,52,67,69]. By introducing spatio-temporal action annotation for each subject, *e.g.* as in AVA [23], spatio-temporal action detection received considerable attention [17,21,22,38,56,64]. In particular, [56] models the spatio-temporal relations to capture the interactions between human, objects, and context that are crucial to infer human actions. More recently, action transformer network [22] has been proposed to localize humans and recognize their actions by considering the relation between actors and context.

Group Activity Recognition. The aforementioned methods mostly focus on predicting individuals' actions in the scene. However, there are works focusing on group activity recognition where the aim is to predict a single group activity label for the whole scene. The early approaches typically extracted hand-crafted features and applied probabilistic graphical models [2,8,9,11,36,37,49] or AND-OR grammar models [1,49] for group activity recognition. In recent years, deep

learning approaches especially RNNs achieve impressive performance largely due to their ability of both learning informative representations and modelling temporal dependencies in the sequential data [4,13,14,26,27,40,44,45,62,50]. For instance, [27] uses a two-stage LSTM model to learn a temporal representation of person-level actions and pools individuals’ features to generate a scene-level representation. In [45], attention mechanism is utilized in RNNs to identify the key individuals responsible for the group activity. Later, authors of [48] extended these works by utilizing an energy layer for obtaining more reliable predictions in presence of uncertain visual inputs. Following these pipelines, [26] introduces a relational layer that can learn compact relational representations for each person. The method proposed in [4] is able to jointly localize multiple people and classify the actions of each individual as well as their collective activity. In order to consider the spatial relation between the individuals in the scene, an attentive semantic RNN has been proposed in [44] for understanding group activities. Recently, the Graph Convolutional Network (GCN) has been used in [65] to learn the interactions in an Actor Relation Graph to simultaneously capture the appearance and position relation between actors for group activity recognition. Similarly [3] proposed a CNN model encoding spatial relations as an intermediate activity-based representation to be used for recognizing group activities. There are also a number of attempts to simultaneously track multiple people and estimate their collective activities in multi-stage frameworks [8,39]. Although these approaches try to recognize the interactions between pairs of people by utilizing hand-crafted features, they are not capable of inferring social groups.

Despite the progress made towards action recognition, detection and group activity recognition [55], what still remains a challenge is simultaneously understanding of social groups and their corresponding social activity. Some existing works aim at only detecting groups in the scene [7,20] by relying on hand-crafted rules *e.g.* face orientations, which are only applicable to very specific tasks such as conversational group detection [25,43,47,57] and by utilizing small-scale datasets. In contrast to previous solutions which are task dependent and require domain expert knowledge to carefully design hand-crafted rules, we extend the concept of grouping to more general types of interactions. To this end, we propose an end-to-end trainable framework for video data to jointly predict individuals’ action as well as their social groups and social activity of each predicted group.

3 Social Activity Recognition

Social activity recognition seeks to answer the question of “*Who is with whom and what they are doing together?*”. Traditional group activity recognition can be seen as a simplified case of social activity recognition where all the individuals are assumed to form a single group in the scene and a single group activity label is predicted for the whole scene. For the ease of conveying our ideas, we present our framework first in the simpler setting of group activity recognition task, and then show how to augment it for the social task.

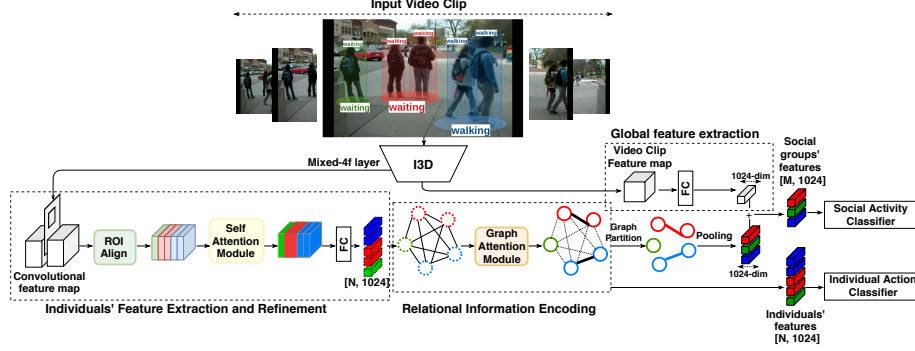


Fig. 2. Our network architecture for the social task. The set of aligned individuals’ features are initially refined by the self-attention module. Projected feature maps are then fed into the GAT module to encode relational information between individuals. During training, feature representation of each social group is pooled from the feature maps of its members according to the ground-truth social connections. At test time, we adopt a graph partitioning algorithm on the inter-node attention coefficients provided by GAT and accordingly infer the social groups and social activity of each sub-group. M and N refer to the number of social groups and number of individuals respectively.

3.1 Group Activity Recognition Framework

A group activity recognition framework should be capable of: *a)* generating a holistic and enriched spatio-temporal representation from the entire video clip; *b)* extracting fine-detailed spatial features from bounding box of each person to accurately predict the individual actions; and *c)* learning an aggregated representation from all individuals for precise realization of their collective activities. Illustrated in Fig. 2, we carefully design effective components in our framework to achieve the above desirable properties. We elaborate the components as follows.

I3D Backbone. We use the Inflated 3D ConvNet (I3D) [6] (based on Inception architecture [29]) as the backbone to capture the spatio-temporal context of an input video clip. In I3D, ImageNet pre-trained convolutional kernels are expanded into 3D, allowing it to seamlessly learn effective spatio-temporal representations. Motivated by the promising performance of the pre-trained I3D models in a wide range of action classification benchmarks, we exploit the feature representations offered by this backbone at multiple resolutions. More specifically, we use the deep spatio-temporal feature maps extracted from the final convolutional layer as a rich semantic representation describing the entire video clip. These deeper features provide low-resolution yet high-level representations that encode a summary of the video. In addition, accurate recognition of individuals’ action rely on finer details which are often absent in very deep coarse representations. To extract fine spatio-temporal representations for the individuals, we use the higher resolution feature maps from the intermediate Mixed-4f layer of I3D. As depicted in Fig. 2, from this representation we extract the

temporally-centered feature map corresponding to the centre frame of the input video clip. Given the bounding boxes in the centre frame, we conduct ROIAlign [24] to project the coordinates on the frame’s feature map and slice out the corresponding features for each individual’s bounding box.

Self-attention Module. Despite being localized to the individual bounding boxes, these representations still lack emphasis on visual clues that play a crucial role in understanding the underlying actions *e.g.* a person’s key-points and body posture. To overcome this, we adopt self-attention mechanism [59,63] to directly learn the interactions between any two feature positions of an individual’s feature representation and accordingly leverage this information to refine each individual’s feature map. In our framework the self-attention module functions as a non-local operation and computes the response at each position by attending to all positions in an individual’s feature map. The output of the self-attention module contextualizes the input bounding box feature map with visual clues and thus, enriches the individual’s representation by highlighting the most informative features. As substantiated by ablation studies in Section 5.1, capturing such fine details significantly contribute to the recognition performance.

Graph Attention Module. Uncovering subtle interactions among individuals present in a multi-person scenario is fundamental to the problem of group activity recognition; each person individually performs an action and the set of inter-connected actions together result in the underlying global activity context. As such, this problem can elegantly be modeled by a graph, where the nodes represent refined individuals’ feature map and the edges represent the interactions among individuals. We adopt the recently proposed Graph Attention Networks (GATs) [60] to directly learn the underlying interactions and seamlessly capture the global activity context. GATs flexibly allow learning attention weights between nodes through parameterized operations based on a self-attention strategy and have successfully demonstrated state-of-the-art results by outperforming existing counterparts [33]. GATs compute attention coefficients for every possible pair of nodes, which can be represented in an adjacency matrix \hat{O}^a .

Training. In our framework, the GAT module consumes the individuals’ feature map obtained from the self-attention component, encodes inter-node relations, and generates an updated representation for each individual. We acquire the group representation by max-pooling the enriched individuals’ feature map and adding back a linear projection of the holistic video’s features obtained from the I3D backbone. A classifier is then applied on this representation to generate group activity scores denoted by \hat{O}^G . Similarly, another classifier is applied on the individuals’ representation to govern the individual action scores denoted by \hat{O}_n^I . The associated operations provide a fully differentiable mapping from the input video clip to the output predictions, allowing the framework to be trained in an end-to-end fashion by minimizing the following objective function,

$$\mathcal{L} = \mathcal{L}_{\text{gp}}(O^G, \hat{O}^G) + \lambda \sum_n \mathcal{L}_{\text{ind}}(O_n^I, \hat{O}_n^I), \quad (1)$$

where \mathcal{L}_{gp} and \mathcal{L}_{ind} respectively denote the cross-entropy loss for group activity and individual action classification. Here, O^G and O_n^I represent the ground-truth

group activities and individual actions, n identifies the individual and λ is the balancing coefficient for the loss functions.

3.2 Social Activity Recognition Framework

In a real-world multi-person scene, a set of social groups each with different number of members and social activity labels often exist. We refer to this challenging problem as the *social activity recognition* task. In this section, we propose a novel yet simple modification to our group activity recognition framework that naturally allows understanding of social groups and their corresponding social activities in a multi-person scenario. The *I3D backbone* and the *self-attention module* remain exactly the same as elucidated in Section 3.1. We explain the required modifications for the *graph attention module* as follows.

Training. Previously, the GAT’s inter-node attention coefficients were updated with the supervision signal provided by the classification loss terms in Eq. 1. To satisfy the requirements of the new problem, *i.e.* to generate social groups and their corresponding social activity label, we augment the training objective with a graph edge loss \mathcal{L}_c that incentivizes the GAT’s self-attention strategy to converge to the individuals’ social connections

$$\mathcal{L} = \sum_s \mathcal{L}_{\text{sgp}}(O_s^{\text{SG}}, \hat{O}_s^{\text{SG}}) + \lambda_1 \sum_n \mathcal{L}_{\text{ind}}(O_n^{\text{I}}, \hat{O}_n^{\text{I}}) + \lambda_2 \mathcal{L}_c(O^\alpha, \hat{O}^\alpha), \quad (2)$$

where, \mathcal{L}_c is the binary cross-entropy loss to reduce the discrepancy between GAT’s learned adjacency matrix \hat{O}^α and the ground-truth social group connections O^α . Further, \mathcal{L}_{sgp} and \mathcal{L}_{ind} respectively denote the cross-entropy loss for social activity of each social group and individual actions classification. Notably, given the ground-truth social groupings during training, we achieve the representation for each social group by max-pooling its corresponding nodes’ feature-map and adding back a linear projection of the video features obtained from the I3D backbone (similar to learning group activity representations in our simplified group activity framework). A classifier is then applied on top to generate the social activity scores \hat{O}_s^{SG} . At inference time however, we require a method to infer the social groups in order to compute the corresponding social representations. To this end, we propose to utilize graph spectral clustering [68] on the learned attention coefficients by GAT, \hat{O}^α and achieve a set of disjoint partitions representing the social groups. In the above formulation s is the social group identifier and (λ_1, λ_2) are the loss balancing coefficients.

4 Datasets

We evaluate our group activity recognition framework on two widely adopted benchmarks: Volleyball dataset and Collective Activity dataset (CAD). We also perform evaluation of our social activity recognition framework on our provided social task dataset by augmenting the CAD with social groups and social activity label for each group annotations.

4.1 Group Activity Recognition Datasets

Volleyball Dataset [27] contains 4830 videos from 55 volleyball games partitioned into 3493 clips for training and 1337 clips for testing. Each video has a group activity label from the following activities: *right set*, *right spike*, *right pass*, *right win-point*, *left set*, *left spike*, *left pass*, *left win-point*. The centered frame of each video is annotated with players’ bounding boxes and their individual action including *waiting*, *setting*, *digging*, *failing*, *spiking*, *blocking*, *jumping*, *moving*, *standing*.

Collective Activity Dataset(CAD) [10] has 44 video sequences captured from indoor and street scenes. In each video, actors’ bounding box, their actions and a single group activity label are annotated for the key frames (*i.e.* 1 frame out of every 10 frames). Individual actions include *crossing*, *waiting*, *queuing*, *walking*, *talking*, *N/A*. The group activity label associated with each key frame is assigned according to the majority of individuals’ actions in the scene. We adopt the same train/test splits as previous works [3,44,65].

4.2 Social Activity Recognition Dataset

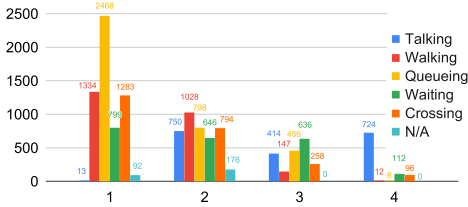


Fig. 3. The histogram of social activities for varying social group sizes (1 to 4 people per social group).

In order to solve the problem of social activity recognition, a video dataset containing scenes with different social groups of people, each performing a social activity is required. Thus, we decided to utilize CAD which is widely used in the group activity recognition task and its properties suit well our problem. Other video action datasets [32,23,31] could not be used in this problem since they mostly consist of scenes with only one social group or a number of singleton groups.

We provide enriched annotations on CAD for solving the social task, which we call *Social-CAD*. As such, for each key frame, we maintain the exact same bounding box coordinates and individual action annotations as the original dataset. However, rather than having a single group activity label for the scene, we annotate different social groups and their corresponding social activity labels. Since there may only exist a subtle indicator in the entire video sequence, *e.g.* an eye contact or a hand shake, suggesting a social connection between the individuals, determining “*who is with whom*” can be a challenging and subjective task. Therefore, to generate the social group annotations as reliable as possible: *a)* we first annotate the trajectory of each person by linking his/her bounding boxes over the entire video, *b)* Given the trajectories, we asked three annotators to independently divide the tracks into different sub-groups according to their social interactions, and *c)* we adopted

majority voting to confirm the final social groups. Similar to the CAD annotation, the social activity label for each social group is defined by the dominant action of its members. We use the same train/test splits in Social-CAD as in CAD. Detailed activity distributions of Social-CAD are given in Fig 3.

5 Experimental Results

To evaluate our proposed framework, we first show that our group activity recognition pipeline outperforms the state-of-arts in both individual action and group activity recognition tasks on two widely adopted benchmarks. Then we evaluate the performance of our social activity recognition framework on Social-CAD.

5.1 Group Activity Recognition

Implementation Details. In our model, we use an I3D backbone which is initialized with Kintetics-400 [32] pre-trained model. We utilize ROI-Align with crop size of 5×5 on extracted feature-map from Mixed-4f layer of I3D. We perform self-attention on each individuals' feature map with query, key and value being different linear projections of individuals' feature map with output sizes being $1/8, 1/8, 1$ of the input size. We then learn a 1024-dim feature map for each individual features obtained from self-attention module. Aligned individuals' feature maps are fed into our single-layer, multi-head GAT module with 8 heads and input-dim, hidden-dim, output-dim being 1024 and dropout probability of 0.5 and $\alpha = 0.2$ [60]. We utilize ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ following [65]. We train the network in two stages. First, we train the network without the graph attention module. Then we fine-tune the network end-to-end including GAT. For the Volleyball dataset, we train the network in 200 epochs with a mini-batch size of 3 and a learning rate ranging from 10^{-4} to 10^{-6} and $\lambda_1 = 8$. For the CAD, we train the network in 150 epochs with a mini-batch size of 4 and a learning rate ranging from 10^{-5} to 10^{-6} and $\lambda_1 = 10$.

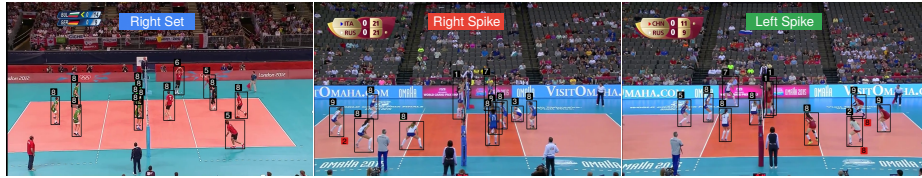


Fig. 4. Visual results of our method on Volleyball dataset for the individual/group activity recognition task (better viewed in color). The bounding boxes around players are produced by our detection-based approach. The numbers above the boxes denote the predicted action IDs. The ground-truth action ID for each player is indicated in red when the predicted action ID is wrong. The label on top of each key frame shows the predicted group activity. Please refer to the dataset section to map IDs to their corresponding actions.

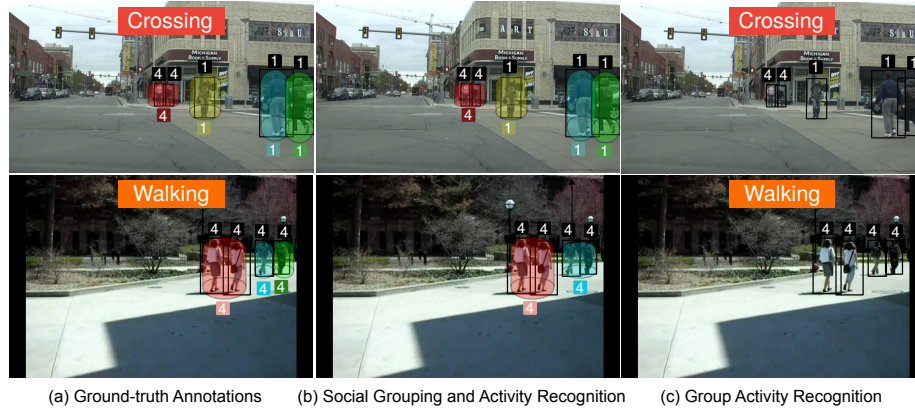


Fig. 5. Visual results of our method on CAD for both social and group activity recognition tasks. Column(a) shows the ground-truth annotation for both tasks. Column(b) represents our prediction for the social task. Column(c) is our predictions for the group task. Note that social groups are denoted by a colored cylinder with their social group labels underneath. The numbers on top of bounding boxes denote the individual action IDs and the label tag above each key frame is the group activity label for the whole scene. 1 and 4 refer to crossing and walking activities respectively.

Input video clips to the model are 17 frames long, with the annotated key frame in the centre. At test time, we perform our experiments based on two widely used settings in group activity recognition literature namely groundtruth-based and Detection-based settings. In the groundtruth-based setting, ground-truth bounding boxes of individuals are given to the model to infer the individual action for each box and the group activity for the whole scene. In the detection-based setting, we fine-tune a Faster-RCNN [46] on both datasets and utilize the predicted boxes for inferring the individuals’ action and group activity.

Evaluation. In order to evaluate the performance of our model for the group activity recognition task, we adopt the commonly used metric, *i.e.* average accuracy, reported in all previous works [3,26,44,65]. To report the performance of individuals’ action in GT-based setting, similar to [65], we used the average accuracy as the measure. In the case of Detection-based setting, average accuracy for evaluating the individuals action is not a valid measure due to the presence of false and missing detections. Instead, we report the commonly used measure for object detection, *i.e.* mean average precision (mAP) [16].

Ablation Study. We justify the choice of each component in our framework with detailed ablation studies. The results on volleyball and CAD are shown in Table 1. As the simplest baseline denoted by Ours[group-only], we use a Kinetics-400 pre-trained I3D backbone and fine-tune it by utilizing the input videos’ feature representation obtained from the last layer of I3D and using a cross-entropy loss to learn the group activity without considering individuals’ action. It is worth mentioning that surprisingly, our simplest baseline already

Table 1. Ablation study of our method for group activity recognition. w/o SA: without self-attention module. w/o GAT: without graph attention module.

	Volleyball	Collective Activity
	Group (Individual)	Group (Individual)
	Acc.% (Acc.%)	Acc.% (Acc.%)
Ours[group-only]	91.0 (-)	84.6 (-)
Ours[w/o SA- w/o GAT]	92.0 (82.0)	88.2 (73.4)
Ours[w/o GAT]	92.5 (83.2)	88.3 (75.3)
Ours[final]	93.1(83.3)	89.4 (78.3)

outperforms many existing frameworks on group activity recognition (see the group accuracy in Table 2). This shows the importance of extracting spatio-temporal features simultaneously using 3D models as well as taking into account the whole video clip rather than solely focusing on individuals and their relations. To consider the effect of jointly training the model on group activity and individual action tasks, we add a new cross-entropy loss to our simplest baseline for training the individuals’ action. As Ours[w/o SA- w/o GAT] experiment shows, training both tasks jointly helps improve the group activity recognition performance. In Ours[w/o GAT], we add the self-attention module performing on each individual’s feature-map in order to highlight the most important features and improve the individual action recognition performance. As shown in Ours[w/o GAT], utilizing self-attention module improves the individual action accuracy by 1.2% on volleyball dataset and by 2.1% on CAD which also contributes to a slight improvement in the group activity accuracy on both datasets. Finally, we add the GAT module to capture the interactions among individuals which is essential in recognizing group activity. As shown in the Ours[final] experiment, utilizing GAT increases the group activity accuracy by 0.6% on volleyball and by 1.1% on collective activity dataset. GAT also improves the individual action performance on both volleyball and collective active datasets by 0.1% and 3% respectively. The higher boost in individual action performance on CAD compared to the one in the volleyball dataset shows the effectiveness of GAT in highlighting social sub-groups and updating individual feature representations accordingly as it is benefiting from a self attention strategy between nodes.

Comparison with the State-of-the-arts. We compare our results on Volleyball and CAD with the state-of-the-art methods in Table 2, using group activity accuracy and individual action accuracy as the evaluation metrics. The top section of the table demonstrates the performance of the approaches in the groundtruth-based setting, where ground-truth bounding box of each person is used for prediction of the individual action as well as group activity of the whole scene. However, in the detection-based settings (indicated by (Det) in Table 2), a Faster-RCNN is fine-tuned on both datasets and predicted bounding boxes for individuals are used during inference (which is more realistic in practice). In group activity recognition using predicted bounding boxes, our model has the least performance drop compared to other methods. In Table 2, ARG[†] is the

Table 2. Comparison with the state-of-the-arts on Volleyball dataset and CAD for group activity recognition.

	Volleyball	Collective Activity
	Group (Individual)	Group (Individual)
	Acc.% (Acc.%)	Acc.% (Acc.%)
HDTM [27]	81.9 (-)	81.5 (-)
CERN [48]	83.3 (-)	87.2 (-)
StagNet [44]	89.3 (-)	89.1 (-)
HRN [26]	89.5 (-)	- (-)
SSU [4]	90.6 (81.8)	- (-)
CRM [3]	93.0 (-)	85.8 (-)
ARG [†] [65]	92.5 (83.0)	88.1 (77.3)
Ours	93.1 (83.3)	89.4 (78.3)
	Acc.% (mAP%)	Acc.% (mAP%)
StagNet(Det) [44]	87.6 (-)	87.9 (-)
SSU(Det) [4]	86.2 (-)	- (-)
ARG(Det) [†] [65]	91.5 (39.8)	86.1 (49.6)
Ours(Det)	93.0 (41.8)	89.4 (55.9)

Table 3. The mean per class group activity accuracies (MPCA) and per class group activity accuracies of our model compared to the existing models on CAD. M, W, Q and T stand for Moving, Waiting, Queuing and Talking respectively. Note that Crossing and Walking are merged as Moving.

Method	M	W	Q	T	MPCA
HDTM [27]	95.9	66.4	96.8	99.5	89.7
SBGAR [40]	90.08	81.4	99.2	84.6	89.0
CRM [3]	91.7	86.3	100.0	98.91	94.2
ARG [†] [65]	100.0	76.0	100.0	100.0	94.0
Ours	98.0	91.0	100.0	100.0	97.2

result that we obtained by running [65]’s released code with the same setting mentioned for each dataset, and the reproduced results perfectly matched the reported results on volleyball dataset. However, we could not reproduce their reported results on CAD. Having their source code available, in order to have a fair comparison with our framework, we also reported their performance on individuals’ action on CAD datasets using both groundtruth-based and detection-based settings (not reported in the original paper). Our framework outperforms all existing methods in all settings on both datasets. We observe that a common wrong prediction in all the existing methods on CAD is the confusion between *crossing* and *walking* in the previous setting. *crossing* and *walking* are essentially same activities being performed at different locations. Thus, we merge these two classes into a single *moving* class and report the Mean Per Class Accuracy (MPCA) in Table 3 as in [3]. ARG[†] outperforms our method in one class *moving*, and our model performs the best in all other 3 classes and the overall metric MPCA.

5.2 Social Activity Recognition

Implementation Details. Our model is trained end-to-end for 200 epochs with a mini-batch size of 4 and a learning rate of 10^{-5} and $\lambda_1 = 5$ and $\lambda_2 = 2$. Other hyper-parameters have the same values as in the group activity recognition experiments on CAD. For graph partitioning at test time, we used graph spectral clustering technique [42,68].

Evaluation. Similar to the group activity recognition task, we perform experiments in two groundtruth-based and detection-based settings. For the social task, we evaluate three sub-tasks: 1) *social grouping*, 2) *social activity recognition of each group* and 3) *individuals' action recognition*. In the GT-based setting, for (1) we calculate the *membership accuracy* as the accuracy of predicting each person's assignment to a social group (including singleton groups). This accuracy is also known as unsupervised clustering accuracy [66]. For (2), we evaluate if both the membership and the social activity label of a person are jointly correct. If so, we consider this instance as a true positive, otherwise, it is assumed a false positive. The final measure, *i.e. social accuracy* is attained as a ratio between the number of true positives and the number of predictions. For (3), we evaluate if the individual's action label is correctly predicted and report *individual action accuracy*. In the detection-based setting, mAP is reported in order to evaluate predicted sub-groups, social activity of each group and individuals' action. For this experiment, bounding boxes with N/A groundtruth are excluded.

Results and Comparison. The performance of our method on social task is reported in Table 4, using membership accuracy, social activity accuracy and individual action accuracy in GT-based setting and mAP for each sub-task in detection-based setting as evaluation metrics. We consider three scenarios of baselines for evaluating this task:

- 1) *Single group setting*: forcing all individuals form a single social group, and then assessing algorithms' performance. ARG[group] [65] and ours[group] essentially are the approaches in Table 2, but are evaluated in membership and social activity metrics. GT[group] uses ground-truth activity labels serving as the upper bound performance for group activity recognition frameworks;
- 2) *Individuals setting*: forcing each individual as a unique social group, *e.g.*, if there are 10 people in the scene, they will be considered as 10 social groups. GT[individuals] uses ground-truth action labels serving as the upper bound performance for group activity recognition frameworks;
- 3) *Social group setting*: Partitioning individuals into multiple social groups. Our first approach uses group activity recognition pipeline in training and uses graph spectral clustering technique to find social groups at inference, named as Ours[cluster] (third part of Table 4). This produces better performance compared to the other baselines, but it is outperformed by our final framework denoted by Ours[learn2cluster], where we learn representations via the the additional graph edge loss.

Discussion. As mentioned in Section 4.2, there might exist only a single frame with a subtle gesture in the entire video, demonstrating a social connection between people in the scene. Therefore as future direction, incorporating in-

Table 4. Social activity recognition results. In each column we report accuracy and mAP for the groundtruth-based and detection-based settings respectively.

	Membership	Social Activity	Individual Action
	GT(Det)	GT(Det)	GT(Det)
	Acc.%,(mAP%)	Acc.%, (mAP%)	Acc.%, (mAP%)
ARG[group] [65]	54.4(49.0)	47.2(34.8)	78.4(62.6)
Ours [group]	54.4(49.0)	47.7(35.6)	79.5(64.2)
GT[group]	54.4(-)	51.6(-)	-(-)
ARG[individuals] [65]	62.4(52.4)	49.0(41.1)	78.4(62.6)
Ours[individuals]	62.4(52.4)	49.5(41.8)	79.5(64.2)
GT[individuals]	62.4(-)	54.9(-)	-(-)
Ours[cluster]	78.2(68.2)	52.2(46.4)	79.5(64.2)
Ours[learn2cluster]	83.0(74.9)	69.0(51.3)	83.3(66.6)

dividuals’ track and skeleton pose might disambiguate some challenging cases for social grouping. Moreover, the performance of our proposed social grouping framework heavily relies on the performance of the graph spectral clustering technique, which is not part of the learning pipeline. Improving this step by substituting it with a more reliable graph clustering approach, or making it a part of learning pipeline can potentially ameliorate the final results.

6 Conclusion

In this paper, we propose the novel social task which requires jointly predicting of individuals’ action, grouping them into social groups based on their interactions and predicting the social activity of each social group. To tackle this problem, we first considered addressing the simpler task of group activity recognition, where all the individuals are assumed to form a single group and a single group activity label is predicted for the scene. As such, we proposed a novel deep framework incorporating well-justified choice of state-of-the-art modules such as I3D backbone, self-attention and graph attention network. We demonstrate that our proposed framework achieves state-of-the-art results on two widely adopted datasets for the group activity recognition task. Next, we introduced our social task dataset by providing additional annotations and re-purposing an existing group activity dataset. We discussed how our framework can readily be extended to handle social grouping and social activity recognition of groups through incorporation of a graph partitioning loss and a graph partitioning algorithm *i.e.* graph spectral clustering. In the future, we aim to use the social activity context for development of better forecasting models, *e.g.* the task of social trajectory prediction, or social navigation system for an autonomous mobile robot.

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: Proceedings of

- the European Conference on Computer Vision. pp. 187–200 (2012)
2. Amer, M.R., Lei, P., Todorovic, S.: Hirf: Hierarchical random field for collective activity recognition in videos. In: Proceedings of the European Conference on Computer Vision. pp. 572–585 (2014)
 3. Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7892–7901 (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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4315–4324 (2017)
 5. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Nibbles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
 6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
 7. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Discovering groups of people in images. In: Proceedings of the European conference on computer vision. pp. 417–433 (2014)
 8. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: Proceedings of the European Conference on Computer Vision. pp. 215–230 (2012)
 9. Choi, W., Savarese, S.: Understanding collective activities of people from videos. *IEEE transactions on pattern analysis and machine intelligence* **36**(6), 1242–1257 (2013)
 10. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: Proceedings of the IEEE 12th International Conference on Computer Vision Workshops. pp. 1282–1289 (2009)
 11. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3273–3280 (2011)
 12. Collins, R.T., Lipton, A.J., Kanade, T.: Introduction to the special section on video surveillance. *IEEE Transactions on pattern analysis and machine intelligence* **22**(8), 745–746 (2000)
 13. Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4772–4781 (2016)
 14. Deng, Z., Zhai, M., Chen, L., Liu, Y., Muralidharan, S., Roshtkhari, M.J., Mori, G.: Deep structured models for group activity recognition. *arXiv preprint arXiv:1506.04191* (2015)
 15. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
 16. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *International journal of computer vision* **88**(2), 303–338 (2010)

17. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6202–6211 (2019)
18. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: Advances in neural information processing systems. pp. 3468–3476 (2016)
19. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016)
20. Ge, W., Collins, R.T., Ruback, R.B.: Vision-based analysis of small groups in pedestrian crowds. *IEEE transactions on pattern analysis and machine intelligence* **34**(5), 1003–1016 (2012)
21. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: A better baseline for ava. arXiv preprint arXiv:1807.10066 (2018)
22. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 244–253 (2019)
23. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6047–6056 (2018)
24. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
25. Hung, H., Kröse, B.: Detecting f-formations as dominant sets. In: Proceedings of the 13th international conference on multimodal interfaces. pp. 231–238 (2011)
26. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: Proceedings of the European Conference on Computer Vision. pp. 721–736 (2018)
27. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1971–1980 (2016)
28. Inc, K.: The multiview extended video with activities (meva) dataset. <https://mevadata.org/>
29. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
30. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* **35**(1), 221–231 (2012)
31. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I., et al.: Panoptic studio: A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence* **41**(1), 190–204 (2017)
32. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017)
33. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
34. Kruse, T., Pandey, A.K., Alami, R., Kirsch, A.: Human-aware robot navigation: A survey. *Robotics and Autonomous Systems* **61**(12), 1726–1743 (2013)

35. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2556–2563 (2011)
36. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1354–1361 (2012)
37. Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(8), 1549–1562 (2011)
38. Li, D., Qiu, Z., Dai, Q., Yao, T., Mei, T.: Recurrent tubelet proposal and recognition networks for action detection. In: Proceedings of the European conference on computer vision. pp. 303–318 (2018)
39. Li, W., Chang, M.C., Lyu, S.: Who did what at where and when: simultaneous multi-person tracking and activity recognition. *arXiv preprint arXiv:1807.01253* (2018)
40. Li, X., Choo Chuah, M.: Sbgar: Semantics based group activity recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2876–2885 (2017)
41. Li, Z., Gavriluk, K., Gavves, E., Jain, M., Snoek, C.G.: Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding* **166**, 41–50 (2018)
42. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in neural information processing systems. pp. 849–856 (2002)
43. Patron-Perez, A., Marszalek, M., Zisserman, A., Reid, I.D.: High five: Recognising human interactions in tv shows. In: BMVC. vol. 1, p. 33 (2010)
44. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Van Gool, L.: stagnet: An attentive semantic rnn for group activity recognition. In: Proceedings of the European Conference on Computer Vision. pp. 101–117 (2018)
45. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3043–3053 (2016)
46. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
47. Setti, F., Lanz, O., Ferrario, R., Murino, V., Cristani, M.: Multi-scale f-formation discovery for group detection. In: Proceedings of the IEEE International Conference on Image Processing. pp. 3547–3551 (2013)
48. Shu, T., Todorovic, S., Zhu, S.C.: Cern: confidence-energy recurrent network for group activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5523–5531 (2017)
49. Shu, T., Xie, D., Rothrock, B., Todorovic, S., Chun Zhu, S.: Joint inference of groups, events and human roles in aerial videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4576–4584 (2015)
50. Shu, X., Tang, J., Qi, G., Liu, W., Yang, J.: Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE transactions on pattern analysis and machine intelligence* (2019)
51. Sigurdsson, G.A., Divvala, S., Farhadi, A., Gupta, A.: Asynchronous temporal fields for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 585–594 (2017)

52. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: *Proceedings of the European Conference on Computer Vision*. pp. 510–526 (2016)
53. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: *Advances in neural information processing systems*. pp. 568–576 (2014)
54. 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)
55. Stergiou, A., Poppe, R.: Analyzing human–human interactions: A survey. *Computer Vision and Image Understanding* **188**, 102799 (2019)
56. Sun, C., Shrivastava, A., Vondrick, C., Murphy, K., Sukthankar, R., Schmid, C.: Actor-centric relation network. In: *Proceedings of the European Conference on Computer Vision*. pp. 318–334 (2018)
57. Swofford, M., Peruzzi, J.C., Vázquez, M., Martín-Martín, R., Savarese, S.: DANTE: Deep affinity network for clustering conversational interactants. *arXiv preprint arXiv:1907.12910* (2019)
58. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: *Proceedings of the IEEE international conference on computer vision*. pp. 4489–4497 (2015)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
60. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks (2017)
61. 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: *Proceedings of the EProceedings of the European conference on computer vision*. pp. 20–36 (2016)
62. Wang, M., Ni, B., Yang, X.: Recurrent modeling of interaction context for collective activity recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3048–3056 (2017)
63. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7794–7803 (2018)
64. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 284–293 (2019)
65. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9964–9974 (2019)
66. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: *International conference on machine learning*. pp. 478–487 (2016)
67. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: *Proceedings of the IEEE international conference on computer vision*. pp. 5783–5792 (2017)
68. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: *Advances in neural information processing systems*. pp. 1601–1608 (2005)
69. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: *Proceedings of the European Conference on Computer Vision*. pp. 803–818 (2018)