Self-supervised Multi-actor Social Activity Understanding in Streaming Videos

Shubham Trehan[1] and Sathyanarayanan N. Aakur[1]

CSSE Department, Auburn University, Auburn, AL, 36849
{szt0113,san0028}@auburn.edu

Abstract. This work addresses the problem of Social Activity Recognition (SAR), a critical component in real-world tasks like surveillance and assistive robotics. Unlike traditional event understanding approaches, SAR necessitates modeling individual actors' appearance and motions and contextualizing them within their social interactions. Traditional action localization methods fall short due to their single-actor, single-action assumption. Previous SAR research has relied heavily on densely annotated data, but privacy concerns limit their applicability in real-world settings. In this work, we propose a self-supervised approach based on multi-actor predictive learning for SAR in streaming videos. Using a visual-semantic graph structure, we model social interactions, enabling relational reasoning for robust performance with minimal labeled data. The proposed framework achieves competitive performance on standard group activity recognition benchmarks. Evaluation on three publicly available action localization benchmarks demonstrates its generalizability to arbitrary action localization.

Keywords: Group Activity Recognition · Action Localization

1 Introduction

Social activity recognition (SAR) is a key part of computer vision applications in the real world, such as surveillance and assistive robotic systems. It differs from traditional event understanding approaches [30,29,2,1,7] since it requires the modeling of individual actor's appearance and their motions, and contextualizing them within the scope of their social interactions. SAR brings a unique set of challenges. First, there is a need for actor localization, social relationship modeling, and social activity recognition. Second, the number of actors in each frame can change due to occlusion, camera range, or notice due to missed/false detection. Finally, a scene can have an arbitrary number of social groups. Traditional action localization approaches [1,2,29] cannot be directly extended to this problem since they assume a single action performed by a single actor.

The dominant approach has been to learn the social dynamics of a scene using attention-based or graph-based relational reasoning in a supervised learning setting. The key assumption has been the availability of densely annotated data for training and near-perfect actor localization. Hence, the literature has focused

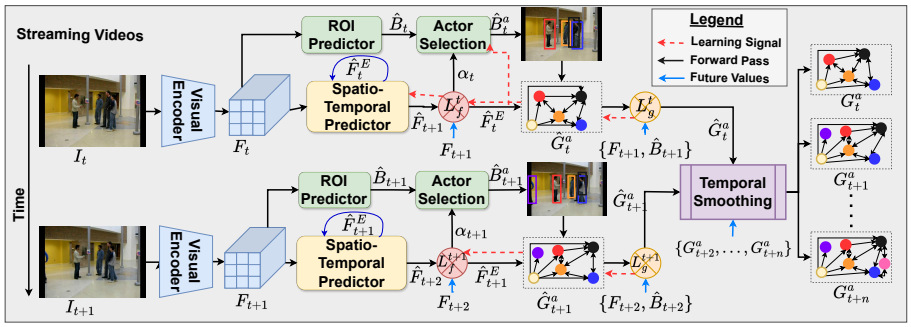

Fig. 1. Overall architecture. Using multi-actor predictive learning, we can localize actors and model their interactions as an action graph, which can be used for downstream event understanding tasks such as action and social activity detection.

on feature aggregation across time and social groups. While this has yielded tremendous progress, it is not always possible to expect densely annotated data for training, primarily due to the privacy concerns involved in collecting, storing, and annotating visual data in a social setting. There is a need to move away from over-reliance on labeled training data and towards self-supervised learning approaches that can learn in an open world, i.e., unconstrained training and test semantics, and in a streaming fashion, i.e., learning with a single pass through the data without storing it without loss of generalization.

In this work, we focus on addressing social activity understanding in streaming videos without labeled data. We propose the idea of multi-actor predictive learning for jointly modeling actor-level actions and contextual, group-level activities. We move away from the single-actor, single-action assumption from prior approaches [3,1,2,35] and propose to represent visual scenes in a social-contextualized action graph that provides an elegant, end-to-end trainable framework for social event understanding. The **contributions** of our approach are four-fold: (i) we are the first to tackle the problem of self-supervised social activity detection in streaming videos, (ii) we introduce a visual-semantic graph structure called an *action graph* to model the social interaction between actors in a group setting, (iii) we show that relational reasoning over this graph structure by spatial and temporal graph smoothing can help learn the social structure of cluttered scenes in a self-supervised manner requiring only a single pass through the training data to achieve robust performance, and (iv) we show that the framework can generalize to arbitrary action localization without bells and whistles to achieve competitive performance on publicly available benchmarks.

2 Related Work

Group activity recognition has been a widely studied area of social event understanding. The typical pipeline starts with actor detection, individual feature extraction, and social interaction modeling. Action features are extracted for

each actor using pre-trained action recognition models [7]. The primary mechanism has been to model the social interaction of actors within a group setting using attention-based mechanisms to generate social group features [31], model individual actor dynamics [10], to model the spatial and temporal dependencies [18] jointly, or for multi-view representation learning [27]. Others use transformers [36] to bypass object detection requirements [14], model keypoint dynamics [43], social relation modeling [9], and spatiotemporal multiscale feature aggregation [46] to reduce training requirements.

Relational reasoning is another line of work that focuses on aggregating actor-actor interactions for group activity understanding. These approaches aim to model the spatiotemporal dependencies by considering the spatial relationships between objects using a variety of mechanisms such as aggregating the relational contexts and scene information using transformers [24], using graph convolutional networks (GCNs) to capture the appearance and position relation between actors [38], or capture spatial coherence using recurrent neural networks (RNNs) [26,37], convolutional neural networks (CNNs) [4], graph-LSTMs [32], graph attention [22], factor graphs [40], knowledge distillation [33], tokenization [39], tracking [34], and contrastive learning [11], to name a few. The prevalent paradigm in the above approaches has been supervised learning to establish and learn social interactions, with varying levels of supervision, i.e., bounding box locations and labels of individual actors, group activity labels, and social group memberships, which requires immense human effort for annotations and may reduce their generalizability. Some works [42,14,46,43] have attempted to reduce the training requirements by relaxing assumptions about the availability of annotations but still require a large amount of labeled data.

Our work is one of the first to tackle this problem from a self-supervised learning perspective by modeling the actor dynamics from a multi-actor predictive learning perspective. *Predictive learning* has emerged as a powerful paradigm for visual event understanding. Proposed in cognitive science literature [44], the goal of predictive learning is to learn representations by anticipating the future and use the residuals for downstream tasks such as event segmentation [3], action localization [1,2], active object tracking [35], future frame generation [21], and hierarchical event perception [23], among others. All prior works have focused on single-actor settings, where only one global action is expected to be present. We offer a unique perspective on predictive learning by extending the idea to multi-actor predictions for group activity detection. We do not require any annotations and aim to learn robust representations at both the actor level and group level, while feature aggregation allows us to model social interactions.

3 Proposed Framework

Overview. We propose to tackle the problem of multi-actor, multi-action localization in streaming videos. The overall framework is illustrated in Figure 1. Given a sequence (stream) of video frames $\{I_0, I_1, \ldots, I_t\}$, we aim to localize actors of interests, characterized by their location (represented as bounding boxes)

$\hat{B}_i^a$ and visual features F_t^a. We then construct a graph structure called an *action graph* ($\hat{G}_t^a$) whose nodes are actors and edges are social interactions, along with an event node. A composite spatiotemporal graph is fed through a temporal smoothing layer that contextualizes event and action level features to construct a final action graph that can be used for various downstream tasks such as group activity understanding and social activity understanding and can be extended to arbitrary action localization. We present each module in detail below.

3.1 Visual Perception and ROI Prediction

Our framework begins with a visual perception module, aiming to extract visual features at both scene and object levels. Our primary visual perception module uses a DETR [6] model. For every frame I_t, we extract (i) global scene features, F_t, (ii) object regions of interests (ROI), $\hat{B}_t$, and (iii) object-level features, F_t^B. The ResNet backbone provides a lower-resolution, global feature map $F_t \in \mathbb{R}^{2048 \times H \times W}$, where W and H are the spatial resolution of the global feature map. DETR's detection heads are used to generate initial object ROI proposals $\hat{B}_t$, i.e., the search space for actor localization. and the decoder outputs for each ROI prediction are used for object-level features (F_t^B). Note that at this stage, we only generate actor candidates that will be refined using the actor selection module described in Section 3.2. We do not fine-tune DETR on the video datasets and use a model pre-trained on MS-COCO [20]. During training, all objects are considered as candidate actors in a class-agnostic manner, following prior works [1,2], while we filter out only "human" predictions during inference. This allows us to learn the dynamics between human and non-human actors during training while allowing us to focus only on social dynamics during inference.

3.2 Spatiotemporal Prediction for Actor Localization

We model the spatial-temporal dynamics of the scene using a spatiotemporal predictor module. The goal is to learn an event-level representation ($\hat{F}_t^E$) that captures how each object, represented by their location B_t and visual features F_t^B, change over time. We use a simple L-layer LSTM stack as our spatiotemporal predictor, which takes as the global scene-level feature F_t as input and anticipates the future global representation F_{t+1}. The goal is not to predict the future frame by pixel-level regression, but rather model how the scene changes over time. This event representation $\hat{F}_t^E$ is continuously updated at every time instant t as new frames (I_t) are observed in a streaming fashion using a predictive loss function given by

$$\mathcal{L}_{global} = \frac{1}{H * W} \sum_{H,W} M_t \odot ||F_{t+1} - \hat{F}_{t+1}||_2 \tag{1}$$

where M_t is the motion difference between frames I_t and I_{t+1} computed as the first-order hold between F_t and F_{t+1}; $\hat{F}_{t+1}$ is the anticipated global feature at

time t+1, obtained by projecting the event feature $\hat{F}_t^E$ back to the 2-D feature space. Hence, the predictive loss attempts to force the LSTM stack to learn a robust event representation ($\hat{F}_t^E$) that can anticipate the future scene's spatial features F_{t+1}. The overall prediction errors are summed and normalized based on the number of spatial features regressed.

Actor Selection The unnormalized prediction errors (P_t=$M_t \odot ||F_{t+1} - \hat{F}_{t+1}||_2$) from Equation 1 are proportional to the predictability of each spatial location. Hence, higher prediction errors indicate the presence of a less predictable, foreground action(s), while lower prediction errors indicate a more predictable, background action. We formulate a prediction-driven attention mask α_t by passing P_t through a softmax activation function to increase focus on foreground actions while suppressing background actions. The top K attention "slots" are used to filter object ROIs $\hat{B}_t$ and select the *actor* ROIs $\hat{B}_t^a$. Note that actor ROIs $\hat{B}_t^a$ are predicted only if the prediction-based attention slots α_t^{ij} fall within any ROI $b_t^k \in \hat{B}_t$. Hence, the number of actors chosen from the list of candidate ROIs is much lower, allowing us to model actor-level dynamics better using action graphs, as described next.

Building an Action Graph To model actor-level interaction dynamics, we construct a graph $\hat{G}_t^a$ for every observed frame I_t, with actors as nodes $\mathcal{V}_t$. Each node $N_i \in \hat{G}_t^a$ is described by a feature vector ${}_i\hat{F}_t^a$=$[{}_iF_t^B; {}_i\hat{B}_t^a]$ that captures its geometry and visual features. The edges in this graph structure, $\mathcal{E}_t$, are defined by the spatial structure of the actors selected using the prediction-based attention α_t. Unlike previous graph-based approaches [9,38], we do not use a fully connected structure. Instead, we model their social connectivity using a distance-based formulation. An adjacency matrix A_t is constructed by computing the spatial proximity between each pair of nodes, given by the Euclidean distance ϕ between their locations and spatial geometry, and centering it by subtracting the mean distance between all nodes. The adjacency for each node N_i is normalized as A_t^i=$\sigma(\frac{A_t^i}{||A_t^i||_2})$, to ensure that the distances are scaled proportionally and σ is the Sigmoid function. The adjacency matrix is thresholded to get the final social structure by discarding all edges less than the average normalized distance in the adjacency. This formulation allows us to model the social interactions between the actors detected in the scene without the underlying assumption that all actors interact with each other, regardless of their social activity. Finally, an "action node", instantiated by the event features $\hat{F}_t^E$, is added to the graph and is connected to all actor nodes. This additional node allows us to propagate action features to relevant actors and the connections between actor nodes will enable us to capture contextual cues for modeling actions with interacting actors, as described in the next section. Empirically, in Section 4, we see that adding the action node and the subsequent contextualization using graph and temporal smoothing plays a big role in improving the performance of both group activity recognition and individual activity detection. The action graph formulation distinguishes us from prior unsupervised event understanding

approaches [3,1,2] since it allows us to model each actor individually without any prior assumptions about their role or interactions in a social group setting.

3.3 Contextualizing Cues with Graph and Temporal Smoothing

Recognizing social activity and individual actions in a group setting requires reasoning over the spatial interaction between actors at every instant and its evolution over time. To this end, given our action graph $\hat{G}_t^a$, the next step is contextualizing each person's action using a two-step spatial-temporal graph smoothing process. First, we use a message passing layer, as introduced in Graph Convolutional Networks (GCNs) [38], for spatial reasoning over the social interaction between actors as captured in $\hat{G}_t^a$. Formally, this is defined as

$$F_t^a = \sigma(A_t \hat{F}_t^a W_s) \tag{2}$$

where A_t is the adjacency matrix for G_t^a, $\hat{F}_t^a$ is the feature representation for each node of the action graph, W_s is the learnable parameter matrix for the GCN layer, and σ is the ReLU activation function. The resulting features F_t^a are contextualized across actors, conditioned on their social structure (specified by weighted edges $\mathcal{E}_t$), and the event-level features $\hat{F}_t^E$ represented by the action node in G_t^a. While this reasoning layer can be repeated, additional layers harm the model's performance (see Section 4) due to the homogenization of features.

For temporal contextualization, we construct a composite spatial-temporal graph by establishing temporal edges between actor nodes in G_t^a with their corresponding nodes in the subsequent graph G_{t+1}^a. While straightforward in theory, we must address two critical challenges for implementation. First, we do not have the ground truth tracking annotations that would enable us to establish actor-actor correspondences across frames. Second, the number of detected actors is not constant across time, requiring comparing graphs of different sizes. Hence, registering nodes across actor graphs between consecutive frames requires us (i) to establish a permutation matrix $\mathcal{P}$ to account for varying node ordering across graphs and (ii) to add null nodes (representing missed/false detections) to the graph with fewer nodes to ensure every node is registered to one node across time. The optimal permutation matrix $\mathcal{P}$ is obtained by computing a one-to-one match between two graphs G_t^a and G_{t+1}^a using the Hungarian matching algorithm to minimize the distance between the two graphs. Formally, this is the optimization for

$$\arg\min_{\mathcal{P} \in \mathcal{P}_n} \sum_{i=1}^{N} w_1 ||_i F_t^a - \mathcal{P}(_i F_{t+1}^a)||_2 + w_2 \phi(_i B_t^a - \mathcal{P}(_i B_{t+1}^a)) \tag{3}$$

where N is the total number of nodes in the graphs G_t^a and G_{t+1}^a, $\mathcal{P}_n$ is the space over all permutation matrices, ϕ is the Intersection over Union (IoU) distance between two bounding boxes, the function $\mathcal{P}(\cdot)$ results in the transformation of a given set of nodes after applying a permutation matrix, i.e., $v \mapsto \mathcal{P}v$, and w_1 and w_2 are scaling factors to balance the two difference distances (i.e., between

feature distance and IoU distance across nodes, respectively). Finally, based on this learned permutation matrix, we establish temporal edges between the nodes registered across time. The composite adjacency matrix $\mathcal{A}_G$ and the corresponding action feature matrix $\mathcal{F}_G$ are used to construct the spatial-temporal graph ($\mathcal{G}_a$), representing the entire video $\mathcal{V}$=$I_1, I_2, \ldots I_T$. A temporal smoothing is performed on $\mathcal{G}_a$ to get the final actor-level features, as defined by

$$\hat{\mathcal{F}}_G = \sigma(\mathcal{A}_G \mathcal{F}_G W_t) \tag{4}$$

where W_t is the set of learnable parameters for a GCN layer. Similar to the spatial smoothing process, this process can be repeated, but empirically, we find one layer is ideal for our experiments. As seen in Section 4, temporal smoothing provides substantial gains in group activity recognition and action detection.

3.4 Social Modeling with Multi-actor Predictive Learning

In addition to the global, event-level predictive learning introduced in Equation 1, we introduce the notion of multi-actor predictive learning. This allows us to model the spatial-temporal dynamics of all actors, conditioned on their social interactions and the overall event dynamics of the scene. We model this using a multi-actor prediction loss given by

$$\mathcal{L}_{actor} = \frac{1}{N} \sum_{i=0}^{N} ||_i \hat{F}_t^a + \mathcal{P}(_i \hat{F}_t^a)||_2 + ||_i B_t^a + \mathcal{P}(_i B_{t+1}^a)||_2 \tag{5}$$

where the first term minimizes the differences between the anticipated actor-level features and the actual actor-level features between consecutive frames, and the second minimizes their respective geometry. We anticipate the future feature and geometry of each actor using two fully connected neural networks defined by $_i F_{t+1}^a$=$W_{act} *_i F_t^a$ and $_i B_{t+1}^a$=$W_{bb} *_i F_t^a$, respectively. This allows us to train our overall spatial-temporal prediction stack (defined in Equations 1 and 5) and the smoothing layers (Equations 2 and 4) by minimizing the overall prediction errors given by

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{actor} \tag{6}$$

where λ_1 and λ_2 allow us to balance the global event-level prediction loss and the actor-level multi-actor prediction loss.

Inferring Labels. For group activity recognition, we do mean average pooling over all actor-level features $\hat{\mathcal{F}}_G$ defined in the composite spatial-temporal action graph $\mathcal{G}_a$. K-means clustering is performed on the mean-pooled features to obtain the final labels. K-means clustering over actor-level features $\hat{\mathcal{F}}_G$ provides actor-level action labels. Following prior work [3,1,2] Hungarian matching is performed between the predicted labels and groundtruth labels to compute the quantitative metrics, as defined in Section 4. A Spectral Clustering model is fit on the adjacency matrix $\mathcal{A}_G$ to find social communities for social activity recognition, following the protocol from the prior work [9].

Table 1. Group Activity Recognition results evaluated on the Collective Activities dataset. Accuracy is reported for group activity recognition and mAP for individual action detection. Note: "-" indicates the model does not *detect* individual actions.

Approach	Training Requirements			Bbox for eval	Group Activity (Acc.)	Indiv.Action (mAP)
	Bboxes	Actor Labels	Grp.Labels			
HDTM[15]	✓	✓	✓	✓	81.5	-
HANs+HCNs[17]	✓	✓	✓	✓	84.3	-
CCGL[32]	✓	✓	✓	✓	90.0	-
CERN [28]	✓	✓	✓	✓	87.2	-
stagNet [25]	✓	✓	✓	✓	89.1	-
GAIM [16]	✓	✓	✓	✓	90.6	-
AT [10]	✓	✓	✓	✓	<u>90.8</u>	-
GroupFormer [18]	✓	✓	✓	✓	**93.6**	-
HIGCIN [41]	✓	✗	✓	✓	92.5	-
CRM [4]	✓	✓	✓	✗	83.4	-
SBGAR [19]	✗	✓	✓	✗	83.7	-
Zhang et al [45]	✓	✗	✓	✗	83.7	-
ARG [38]	✓	✓	✓	✗	86.10	49.60
Ehsanpour et. al.[9]	✓	✓	✓	✗	<u>89.40</u>	<u>55.90</u>
HGC-Former[31]	✓	✓	✓	✗	**96.50**	**64.90**
PredLearn($K = K_{GT}$)	✗	✗	✗	✗	62.83	2.82
AC-HPL($K = K_{GT}$)	✗	✗	✗	✗	<u>72.20</u>	<u>10.68</u>
Ours ($K = K_{GT}$)	✗	✗	✗	✗	**75.95**	**26.75**
Ours ($K = K_{OPT}$)	✗	✗	✗	✗	**90.41**	**33.02**

4 Experimental Evaluation

Data. We use the Collective Activities Dataset (CAD) [8] and its annotations-augmented version, SocialCAD [9], to evaluate our framework. CAD consists of 44 videos taken in unconstrained real-world scenarios of people performing 6 individual actions across 5 group activities. SocialCAD augmented CAD with additional information, such as social group identification of each person and their collective social activity. We follow prior work [9] and use 31 videos for training and 11 for evaluation. To evaluate the generalization capabilities of our framework to arbitrary action localization, we use three publicly available benchmarks - UCF Sports [30], JHMDB [12], and THUMOS'13 [13]. Each dataset contains a varying number of actions (10 in UCF Sports, 21 in JHMDB, and 24 in THUMOS'13) across different domains (sports and daily activities). Each dataset offers a unique challenge for action localization, such as cluttered scenes, highly similar action classes, large camera motion, and object occlusion. We follow prior work [2,29] and use official train-test splits for all datasets.

Metrics. We use different metrics for evaluating the performance on each task. We use the mean multi-class classification accuracy (MCA) for group activity recognition. For individual action detection, we follow prior work [9] and use the mean average precision (mAP) as the evaluation metric to account for missed and false detections. To evaluate social activity understanding, we use two different metrics - membership accuracy and social activity recognition, as defined in SocialCAD. The former measures the accuracy of recognizing a per-

son's social group in the video. The latter measures the ability to jointly predict a person's membership and the social activity label. We report the video-level mAP at 0.5 IOU threshold for arbitrary action localization, i.e., the bounding box predictions must have 0.5 spatial IOU in at least 50% of the frames.

Baselines. We compare against a variety of supervised, weakly supervised, and unsupervised learning approaches for both group activity understanding and action localization. The supervised [15,17,32,28,25,16,10,18] and weakly supervised learning baselines [41,4,19,38,9,31] provide solid baselines for comparing the representation learning capabilities of our framework. Unsupervised learning approaches, particularly closely related approaches such as AC-HPL [2] and PredLearn [1], allow us to benchmark our approach with others trained under the same settings. We use Hungarian matching for all unsupervised learning baselines to align their predictions with the ground truth labels, following prior work [1,2]. Note that all baselines, except AC-HPL and PredLearn are not trained in a streaming fashion and require strong visual encoders pre-trained on large amounts of video data (e.g., I3D [7] on Kinetics [7]) and fine-tuned for a large number of epochs (> 50). We do not require either and only use DETR [6] pre-trained on MS-COCO for person detection and train in a streaming fashion, requiring only one pass through all the videos.

4.1 Group Activity Recognition

We first evaluate our approach on the group activity *recognition* task, where the goal is to identify the activity in which the *majority* of the people are involved. Table 1 summarizes the results. As can be seen, we perform competitively with supervised learning approaches and significantly outperform prior unsupervised learning approaches such as PredLearn [1] (by 13.12%) and AC-HPL [2] (by 3.75%). We observe that some activity classes, such as "walking" and "crossing", exhibit high intra-class variation in the clustering. Hence, we increase the number of clusters for recognition to its optimal number (using the elbow method with intra-cluster variation as the metric) and devise a baseline indicated by $K = K_{OPT}$. We observe that the accuracy increases significantly to 90.41%, outperforming many of the supervised and weakly supervised approaches. It is to be noted that the supervised learning approaches (at the top of Table 1) require ground truth bounding boxes during inference for efficient recognition. Weakly supervised approaches [38,31,9] do not require bounding boxes during inference but require supervision from dense annotations. Interestingly, we observe that the mean per-class accuracy (MPCA) is 81.25%, with the class "Waiting" being the worst-performing one at 35.51%. We attribute it to the predictive learning paradigm, which naturally focuses on actors with the least predictive motions. It has actors with highly predictable motion, which reduces the model's attention on them and leads to poorer recognition accuracy. However, other classes have a recognition accuracy above 90%, indicating the model's effectiveness in recognizing actions that involve reasoning over actor appearance and motion.

In addition to group activity recognition, we also report the mAP score for individual action detection (last column of Table 1), where the goal is to localize

Table 2. Social Activity Understanding results on the SocialCAD dataset [9]. Note: All results are in the detection setting, i.e., without GT bounding boxes.

Approach	Training Requirements			Membership Recognition	Social Activity Recognition
	Bboxes	Labels	Member		
GT [Group] (Upper Bound)	-	-	-	54.4	51.6
HGC-Former [31]	✓	✓	✓	-	46.0
ARG [Group] [38]	✓	✓	✓	49.0	34.8
Ehsanpour et al [Group] [9]	✓	✓	✗	49.0	35.6
Ours	✗	✗	✗	**32.33**	**25.07**

and recognize every actor's actions. As can be seen, we once again outperform prior unsupervised learning approaches significantly while offering competitive performance to supervised learning approaches [38,9,31]. We do not finetune our ROI prediction (DETR) on the CAD dataset as with the supervised learning approaches. This significantly increases the number of actors detected in the scene, which is not always reflected in the ground truth. One such instance is highlighted in Figure 2, where it can be seen that we correctly localize and recognize the individual actions of *all* actors in the scene and not just those in the ground truth. Note that prior unsupervised learning approaches (PredLearn and AC-HPL) do not predict distinct actions for each actor, but rather, a collective group activity is assigned to each person. This reduces their utility in action detection and stems from their inherent assumption that there is one action per video and that all actors participate in this global action. We do not make such assumptions and hence can effectively recognize and localize multiple, simultaneous actions performed by multiple actors.

4.2 Social Activity Understanding

In addition to group activity recognition and individual action detection, we also evaluate the representation learning capabilities of our framework to social activity understanding tasks such as membership recognition and social activity recognition. For membership recognition, we follow Ehsanpour *et al.* [9] and use graph spectral clustering to segment the individual actors into social groups to compute the membership recognition accuracy. Table 2 summarizes the results. We perform competitively with supervised learning approaches such as HGC-Former [31] and ARG [38], which require prior knowledge of memberships during training. We also perform competitively with Ehsanpour *et al.* [9], which does not require membership labels during training but does require other annotations, such as individual and group labels, along with their bounding box annotations during training. The baselines GT[Group], taken from Ehsanpour *et al.* [9], provides the upper bound for detection-based models when the member locations and actions are provided, and an I3D model [7] is used for labeling the membership and social activity of the person. On inspecting the results, we find that much of the reduction in membership recognition accuracy is because

Table 3. Generalization to arbitrary action localization. We report the video-mAP and compare it with unsupervised *action* localization baselines. OOD refers to the evaluation on data other than the training domain (CAD).

Approach	OOD Eval	UCF Sport	JHMDB	THUMOS13
Ours	✓	**0.40**	**0.22**	**0.15**
AC-HPL[2]	✗	**0.59**	0.15	<u>0.20</u>
PredLearn[1]	✗	0.32	0.10	0.10
Soomro[29]	✗	0.30	<u>0.22</u>	0.06
Ours	✗	<u>0.49</u>	**0.25**	**0.21**

we predict and localize more actions than provided in the ground truth and, hence, make more predictions per frame. For example, in Figure 2, we detect the membership and actions of all people, not just those in the annotations. We anticipate that fine-tuning DETR on the ground truth annotations and reducing the number of detected people will reduce the false alarms and improve the performance of our approach on these metrics at the cost of generalization.

Generalization to Arbitrary Action Localization Since our approach does not make any assumptions on the number of actions or type of action, we evaluate its capability to generalize to arbitrary action localization in videos. We evaluate on the UCF Sports [30], JHMDB [12], and THUMOS'13 [13] datasets, where there is a single action in the scene with a varying number of actors. Table 3 summarizes the results, comparing our approach against other unsupervised learning baselines. We outperform the baselines on all benchmarks, except UCF Sports, when trained on videos from the same domain. The most interesting result is the top row, which shows the performance of our model, trained on CAD and evaluated on out-of-domain videos. We perform well in arbitrary action localization without explicit training, showcasing its generalization capabilities.

4.3 Ablation Studies

We systematically analyze the contributions of each part of our framework and quantify their effects in Table 4. We examine the presence and absence of graph smoothing, temporal smoothing, and the use of action nodes in our action graphs. We see that removing action graphs causes a dramatic decrease in group activity recognition while having minimal effect on individual action recognition. Temporal smoothing has the most impact on both metrics, which could be attributed to the fact that information from the entire video is propagated through the temporal edges and enables better contextualization of group dynamics. Graph smoothing, which enables nodes within the same frame to share information, is essential in propagating information from the action node to each person node. Adding additional layers of temporal and graph smoothing reduces the performance of the approach since it makes the node representations uniform and, hence, loses information about the changes in actor appearances and locations.

Table 4. Ablation study results on the collective activities dataset. We report accuracy for group activity recognition and the mAP for individual action detection tasks.

Approach	Group Activity	Indiv. Action
Ours (full model)	75.95	26.75
w/ 2 layers of temporal smoothing	72.79	23.15
w/ 2 layers of graph smoothing	73.28	22.84
w/o temporal smoothing	59.13	14.27
w/o graph smoothing	68.39	11.56
w/o action nodes	63.46	18.28

4.4 Qualitative Evaluation

Figure 2 presents some qualitative visualization of the output from our framework. The top half presents successful social activity detection results. The first row is the ground truth annotations, while the second row shows our corresponding predictions. As can be seen, we can localize and recognize both the social membership (indicated by the color of the shaded region) and the social action (indicated by the bounding box color) of each actor in the scene. Interestingly, we see that we detect and recognize the social activities of people not in the groundtruth (bottom left) and consistently maintain prediction throughout the sequence. The bottom half of Figure 2 shows some unsuccessful results where the membership was misclassified, although the social action is correct. We attribute this to our framework's additional action detections that provide "distractors" for the membership classification. This effect is also reflected in the individual action mAP score (26.75), where the number of false alarms (due to detections not in the groundtruth annotations) plays a major role. We see that the average recall (across classes) of the groundtruth bounding boxes is 67%, indicating that we can recover and label many of the actors correctly.

5 Discussion, Limitations, and Future Work

In this paper, we presented a framework for unsupervised multi-actor, multi-action localization in streaming videos. We showed that it can be adapted to perform group activity recognition, action detection, social membership identification, and social action detection tasks in multi-actor settings. We also demonstrated its potential for localizing an arbitrary number of actions in streaming videos and showed its generalization capabilities by evaluating on out-of-domain data. While it outperformed unsupervised baselines and was competitive with supervised learning approaches, we observe some limitations that offer potential for future work. First, the actor selector module focuses on actions with unpredictable motion. Hence, it fails to consistently localize those with limited predictability, such as "waiting." Similarly, it is sensitive to missed detections. It relies heavily on the ROI detector to provide quality region proposals. Finally, imposing constraints on group formations in frame-level action graphs will likely

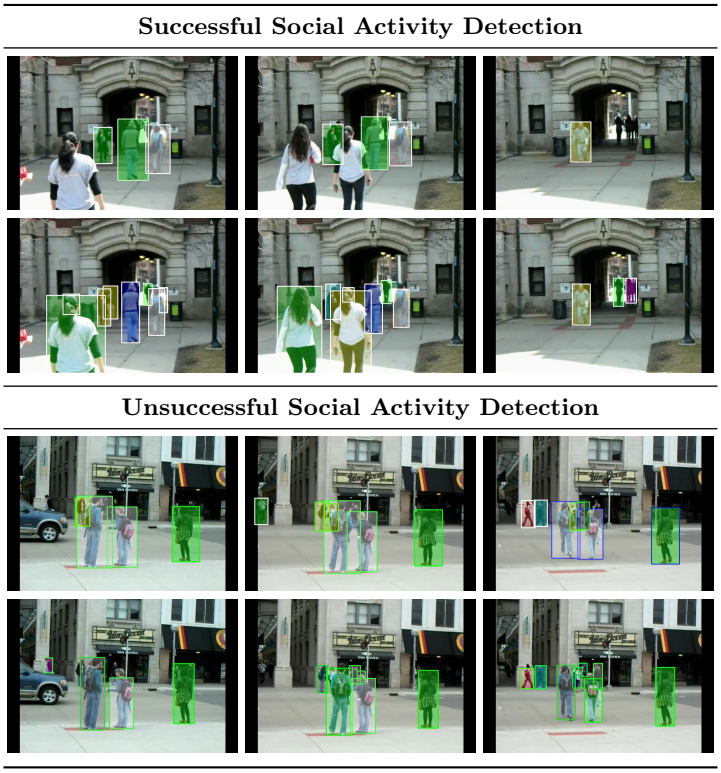

Fig. 2. Qualitative visualization successful (top) and unsuccessful (bottom) activity detection on the Collective Activities dataset. People from the same social group are highlighted in the same color, and the bounding box color indicates their social activity.

yield more robust social membership recognition performance. Our future work is focused on improving social action detection by dynamic graph modeling [5].

Acknowledgements. This work was partially supported by the US National Science Foundation Grants IIS 2348689 and IIS 2348690 and the US Department of Agriculture grant 2023-69014-39716-1030191.

References

1. Aakur, S., Sarkar, S.: Action localization through continual predictive learning. In: ECCV. pp. 300–317 (2020)
2. Aakur, S., Sarkar, S.: Actor-centered representations for action localization in streaming videos. In: ECCV. pp. 70–87 (2022)
3. Aakur, S.N., Sarkar, S.: A perceptual prediction framework for self supervised event segmentation. In: CVPR. pp. 1197–1206 (2019)
4. Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: CVPR. pp. 7892–7901 (2019)

5. Bal, A.B., Mounir, R., Aakur, S., Sarkar, S., Srivastava, A.: Bayesian tracking of video graphs using joint kalman smoothing and registration. In: ECCV. pp. 440–456 (2022)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
8. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: ICCV Workshops. pp. 1282–1289 (2009)
9. Ehsanpour, M., Abedin, A., Saleh, F., Shi, J., Reid, I., Rezatofighi, H.: Joint learning of social groups, individuals action and sub-group activities in videos. In: ECCV. pp. 177–195 (2020)
10. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.: Actor-transformers for group activity recognition. In: CVPR. pp. 839–848 (2020)
11. Han, M., Zhang, D.J., Wang, Y., Yan, R., Yao, L., Chang, X., Qiao, Y.: Dual-ai: Dual-path actor interaction learning for group activity recognition. In: CVPR. pp. 2990–2999 (2022)
12. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV. pp. 3192–3199 (Dec 2013)
13. Jiang, Y.G., Liu, J., Zamir, A.R., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: Thumos challenge: Action recognition with a large number of classes (2014)
14. Kim, D., Lee, J., Cho, M., Kwak, S.: Detector-free weakly supervised group activity recognition. In: CVPR. pp. 20083–20093 (2022)
15. Kim, J., Lee, M., Heo, J.P.: Self-feedback detr for temporal action detection. In: ICCV. pp. 10286–10296 (2023)
16. Kong, L., Pei, D., He, R., Huang, D., Wang, Y.: Spatio-temporal player relation modeling for tactic recognition in sports videos. IEEE T-CSVT **32**(9), 6086–6099 (2022)
17. Kong, L., Qin, J., Huang, D., Wang, Y., Van Gool, L.: Hierarchical attention and context modeling for group activity recognition. In: ICASSP. pp. 1328–1332 (2018)
18. Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: Groupformer: Group activity recognition with clustered spatial-temporal transformer. In: ICCV. pp. 13668–13677 (2021)
19. Li, X., Choo Chuah, M.: Sbgar: Semantics based group activity recognition. In: ICCV. pp. 2876–2885 (2017)
20. 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. pp. 740–755 (2014)
21. Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104 (2016)
22. Lu, L., Lu, Y., Yu, R., Di, H., Zhang, L., Wang, S.: Gaim: Graph attention interaction model for collective activity recognition. IEEE T-MM **22**(2), 524–539 (2020)
23. Mounir, R., Vijayaraghavan, S., Sarkar, S.: Streamer: Streaming representation learning and event segmentation in a hierarchical manner. NeurIPS **36** (2024)
24. Pramono, R.R.A., Chen, Y.T., Fang, W.H.: Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In: ECCV. pp. 71–90 (2020)

25. Qi, M., Wang, Y., Qin, J., Li, A., Luo, J., Van Gool, L.: Stagnet: An attentive semantic rnn for group activity and individual action recognition. IEEE T-CSVT **30**(2), 549–565 (2019)
26. Qi, M., Wang, Y., Qin, J., Li, A., Luo, J., Van Gool, L.: stagnet: An attentive semantic rnn for group activity and individual action recognition. IEEE T-CSVT **30**(2), 549–565 (2020)
27. Raviteja Chappa, N.V., Nguyen, P., Nelson, A.H., Seo, H.S., Li, X., Dobbs, P.D., Luu, K.: Sogar: Self-supervised spatiotemporal attention-based social group activity recognition. arXiv e-prints pp. arXiv–2305 (2023)
28. Shu, T., Todorovic, S., Zhu, S.C.: Cern: confidence-energy recurrent network for group activity recognition. In: CVPR. pp. 5523–5531 (2017)
29. Soomro, K., Shah, M.: Unsupervised action discovery and localization in videos. In: ICCV. pp. 696–705 (2017)
30. Soomro, K., Zamir, A.R.: Action recognition in realistic sports videos. In: Computer Vision in Sports, pp. 181–208. Springer (2015)
31. Tamura, M., Vishwakarma, R., Vennelakanti, R.: Hunting group clues with transformers for social group activity recognition. In: ECCV. pp. 19–35 (2022)
32. Tang, J., Shu, X., Yan, R., Zhang, L.: Coherence constrained graph lstm for group activity recognition. IEEE T-PAMI **44**(2), 636–647 (2019)
33. Tang, Y., Lu, J., Wang, Z., Yang, M., Zhou, J.: Learning semantics-preserving attention and contextual interaction for group activity recognition. IEEE T-IP **28**(10), 4997–5012 (2019)
34. Thilakarathne, H., Nibali, A., He, Z., Morgan, S.: Group activity recognition using unreliable tracked pose. arXiv preprint arXiv:2401.03262 (2024)
35. Trehan, S., Aakur, S.N.: Towards active vision for action localization with reactive control and predictive learning. In: WACV. pp. 783–792 (2022)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
37. Wang, M., Ni, B., Yang, X.: Recurrent modeling of interaction context for collective activity recognition. In: CVPR. pp. 3048–3056 (2017)
38. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: CVPR. pp. 9964–9974 (2019)
39. Wu, L., Lang, X., Xiang, Y., Chen, C., Li, Z., Wang, Z.: Active spatial positions based hierarchical relation inference for group activity recognition. IEEE T-CSVT (2022)
40. Xie, Z., Jiao, C., Wu, K., Guo, D., Hong, R.: Active factor graph network for group activity recognition. IEEE T-IP (2024)
41. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Higcin: Hierarchical graph-based cross inference network for group activity recognition. IEEE T-PAMI **45**(6), 6955–6968 (2020)
42. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Social adaptive module for weakly-supervised group activity recognition. In: ECCV. pp. 208–224 (2020)
43. Yuan, H., Ni, D.: Learning visual context for group activity recognition. In: AAAI Conference on Artificial Intelligence. vol. 35, pp. 3261–3269 (2021)
44. Zacks, J.M., Tversky, B., Iyer, G.: Perceiving, remembering, and communicating structure in events. Journal of Experimental Psychology: General **130**(1), 29 (2001)
45. Zhang, P., Tang, Y., Hu, J.F., Zheng, W.S.: Fast collective activity recognition under weak supervision. IEEE T-IP **29**, 29–43 (2019)

46. Zhou, H., Kadav, A., Shamsian, A., Geng, S., Lai, F., Zhao, L., Liu, T., Kapadia, M., Graf, H.P.: Composer: compositional reasoning of group activity in videos with keypoint-only modality. In: ECCV. pp. 249–266 (2022)