Context Aware Group Activity Recognition

Avijit Dasgupta
CVIT
IIIT Hyderabad, India
Email: avijit.dasgupta@research.iiit.ac.in

C. V. Jawahar
CVIT
IIIT Hyderabad, India
Email: jawahar@iiit.ac.in

Karteek Alahari
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK
38000 Grenoble, France
Email: karteek.alahari@inria.fr

***Abstract*—This paper addresses the task of group activity recognition in multi-person videos. Existing approaches decompose this task into feature learning and relational reasoning. Despite showing progress, these methods only rely on appearance features for people and overlook the available contextual information, which can play an important role in group activity understanding. In this work, we focus on the feature learning aspect and propose a two-stream architecture that not only considers person-level appearance features, but also makes use of contextual information present in videos for group activity recognition. In particular, we propose to use two types of contextual information beneficial for two different scenarios: *pose context* and *scene context* that provide crucial cues for group activity understanding. We combine appearance and contextual features to encode each person with an enriched representation. Finally, these combined features are used in relational reasoning for predicting group activities. We evaluate our method on two benchmarks, Volleyball and Collective Activity and show that joint modeling of contextual information with appearance features benefits in group activity understanding.**

I. Introduction

Group activity recognition is an important video understanding problem and it has several practical applications such as crowd monitoring, sports analytic, and behaviour understanding. To understand scenes involving multiple people (actors), the model needs to describe the individual activities as well as infer the activities of the groups they belong to.

Current group activity recognition approaches [4], [16], [18], [23], [32], [40] typically tackle this problem by decomposing it into two parts: *feature learning* and *relational reasoning*. The first part focuses on learning person-specific visual features important for understanding individual actions. In the second part, pairwise relations are modeled to infer the group activities. Despite recent advances, these approaches still confuse between visually similar group activities as they rely only on person-level appearance features in the feature learning part and ignore the contextual information present in videos. Consider the example shown in Fig. 1. It is challenging to differentiate *walking* activity in the first case from *crossing* activity in the second case using appearance features alone as in both the cases people are moving from one point to another. However, if we have additional cues that identify that a group of people is moving on a *sidewalk* in Fig. 1a vs. *road* in Fig. 1b, the model can learn to distinguish these group activities. We term these additional cues as *contextual* information and propose to integrate them with appearance features for group activity understanding.

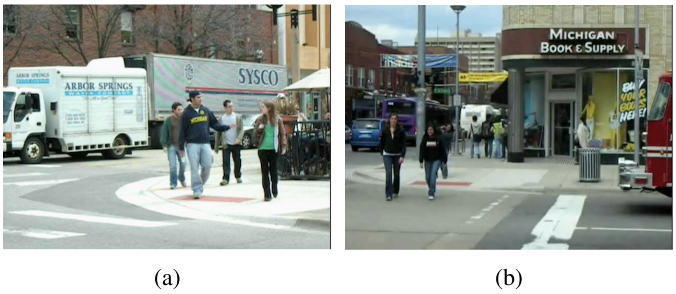

(a) (b)

Fig. 1: Inferring the group activities from a video is inherently complex and difficult task as it relates to (a) individual features of person, (b) relation among people, and (c) the context information. Context provides important cues about the environment (e.g. sidewalk vs. road). Such, additional information is exploited in our model to differentiate between visually similar (e.g. *walking* vs. in *crossing*) activities.

The influence of context on object recognition for human visual system is a well-studied topic in psychology [28]. Computer vision literature also suggests [12], [15], [27] that recognition algorithms can be improved by proper modeling of contextual information. Exploiting context with appearance information has proved to be beneficial for visual understanding tasks such as object detection [7], [26], trajectory prediction [24], [25], human-object-interaction detection [37] etc. However, contextual information about the scene for group activity recognition is relatively underexplored, where context such as scene labels can provide complementary information to standard person-centric appearance features.

In this work, we leverage these contextual information that exist in scenes. We present a two-stream network for group activity recognition as shown in Fig. 2. The contextually enriched visual features are extracted from two-streams - the appearance stream which describes the static features of people and the context stream which encodes the context around every person present in the scene. The appearance branch and the context branch provide complementary cues for group activity recognition. These appearance and contextual features are combined together to represent each person in a group with visually enriched features. These features are then used for relational reasoning. Following [40], we model the human-human relation using a graph convolutional network (GCN) [20]. However, our proposal to exploit context is general

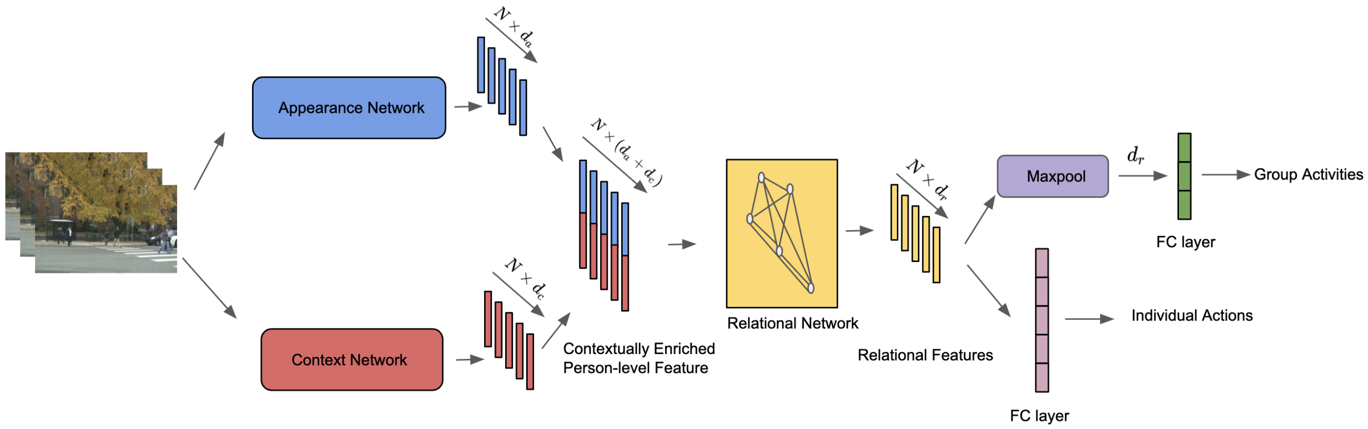


Fig. 2: Overview of our proposed model. Given a sequence of frames and person bounding boxes, each of them are processed by an appearance network (blue) and a context network (red) to extract intermediate visually enriched two-stream representation. The graph convolution network (yellow) is used to model the relationship among the people present in the scene. We use these representations to predict final individual actions and group activities.

enough to be used with any type of relational models.

In summary, we first argue that the contextual information can provide complimentary cues for group activity understanding. Two types of contextual cues helpful for recognizing group activities are proposed: namely *pose* for Volleyball [17] and *scene labels* for Collective Activity dataset [9]. We model a context network to encode contextual information that are present in scenes. We evaluate all variants of our model on two publicly available datasets - Collective Activity [9] and Volleyball [18] to empirically validate the efficacy of contextual information over appearance only models (see Sec. IV). Additionally, we provide an extensive experimental analysis, with ablation studies to demonstrate the influence of all the components in our proposed network.

II. Related Work

Group activity recognition. Previous approaches [2], [3], [21], [22] for group activity recognition focus on designing suitable features and modeling relation among the actors using probabilistic graphical models or AND-OR grammars. Recently, significant progress has been made in the domain of group activity recognition [5], [13], [16]–[18], [23], [29], [32], [40], mainly due to the advent of convolutional neural networks (CNNs). Ibrahim *et al.* [18] propose a two-stage deep temporal model to capture temporal dynamics. Shu *et al.* [32] extend [18] with an energy-based model to remove brittleness in the predictions of the temporal model. Hierarchical relation network is used to build relation representation among the actors in a scene in [17]. Bagautdinov *et al.* [5] propose a unified model to jointly detect actors present in a scene and recognize their group activities. Li *et al.* [23] propose a semantic based approach to generate captions for videos. Later, these generated captions are used to predict group activities. Wu *et al.* [40] build an actor-relation graph using a GCN to model the relational feature among the actors. Gavrilyuk *et al.* [13] use self attention mechanism to model the dependency among the people present in a scene. These approaches mainly focus on designing appropriate models to understand the interaction pattern involving people present in a scene. Unlike these approaches, we focus on designing appropriate contextual information and adapting existing relational models to show the efficacy of utilizing context for group activity recognition task.

Contextual information in computer vision. Images and videos contain a rich set of contextual information about the scenes they represent. In computer vision, a number of approaches, e.g., [1], [12], [24]–[27], [31], [37], have exploited this information to improve recognition performance. Local and global context information is exploited in [27] for object detection. The prediction of an object in irrelevant scenes acts as a penalty in [36]. Shrivastava *et al.* [31] use segmentation to guide region proposal generation. Scene contextual information has proved to be beneficial in trajectory prediction task [24], [25]. Specifically, [24] designs a person-scene interaction module that encodes nearby scene of a person which in turn helps forecast the trajectory movement of the person. Pose information has also been leveraged in previous approaches [24], [45], for instance to encode person behaviour [24] or as a contextual information for daily activity recognition [45]. Ulutan *et al.* [37] use scene context for improved detection of human-object-interaction. Contextual information is known to be essential for semantic segmentation [11], [43].

Context for group activity understanding. While most of the works on group activity recognition focus on relational modeling of the people present in a scene, relatively less progress has been made in using contextual information. Deng *et al.* [10] use the CNN representation of the whole scene as context. Wang *et al.* [39] model interaction context to encode higher order interactions that happen between people present in a scene.

Different from the previous approaches [10], [39], we propose to exploit a more fine-level contextual information, namely scene context and pose context that are readily present in a scene and we show that these context information can provide us with complementary information when combined with appearance features.

III. Approach

Our model takes video frames as input, and predicts group activities along with the individual actions of people present in the video, as shown in Fig. 2. For example, group activities can be *right set*, *right spike* etc., whereas individual actions can be *waiting*, *standing* etc., in case of Volleyball dataset. Following [18], [40], we assume each multi-person video is first processed to obtain bounding box coordinates of every person (actor) present in all frames. These bounding box coordinates are then used to extract person-level appearance and contextual information from the input video frames. We use these features to build a fully-connected graph to model pair-wise relationship among the actors. This graph structure is then processed by a graph convolution network (GCN). Finally, the person-level relational features obtained from GCN are passed through two classifiers to predict group activities and individual actions. The whole framework for our method as shown in Fig 2 will be detailed in the following sections.

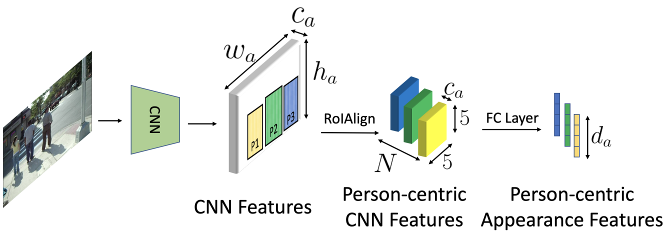


Fig. 3: **The appearance network**. It encodes visual information of every person present in a scene into a feature representation of dimension d_a. See section III-B for more details.

A. Network Architecture

Fig. 2 shows the overall framework of our proposed model. Unlike previous works, our model encodes contextual information of each person. Our model has the following key components:

Appearance network extracts visual information using the bounding boxes around each person present in a scene.

Context network extracts contextual information around a person. Two kinds of contextual information is proposed in this paper described in III-C.

Relational network models interactions that happen among the actors in a scene. Following [40], [41], we also use GCN as our relational network.

B. Appearance Network

This module encodes the visual information of every individual present in a scene. We first pass video frames of dimension $H \times W \times 3$ through a CNN network and extract convolutional features from the intermediate layers of dimension $h_a \times w_a \times c_a$, where h_a, w_a, and c_a denotes the height, width and depth of the convolutional feature map respectively and subscript a denotes the appearance network. Given the person bounding box coordinates, we extract fixed size person-level convolutional features of dimension $N \times 5 \times 5 \times c_a$ using RoIAlign [14] layer, where N is the total number of people present in a scene . These features are then passed through a fully connected (FC) layer to obtain appearance representation of dimension $x_a \in \mathbb{R}^{N \times d_a}$, where d_a is the appearance feature dimension (see Fig. 3).

C. Context Network

As mentioned earlier, we propose to exploit the contextual cues that are already present in the scene to alleviate the confusion that arises in differentiating visually similar activities. For volleyball dataset, posture plays an important role to understand the individual actions as well as group activities as shown in Fig. 4. For collective activity dataset, the videos are mostly taken in outdoor and indoor scenarios and to understand the nearby scene around a person, the scene labels play a crucial role as show in Fig. 6. Thus, to model the context around a person in a scene, we propose to use two types of contextual information present in videos, namely pose context and scene context. Similar to the appearance network, the context network also takes video frames of dimension $H \times W \times 3$ as input.

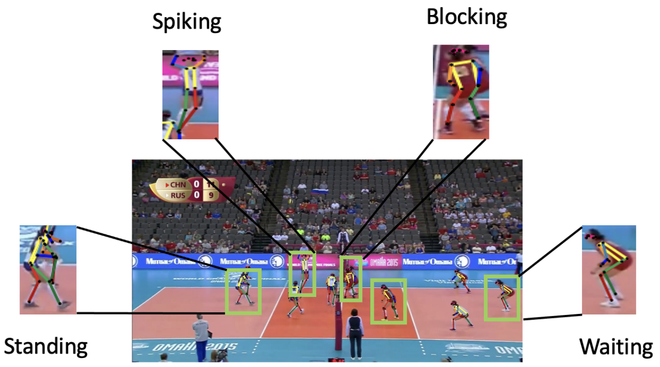


Fig. 4: Each individual action has its own posture. For example, the *spiking* action can be differentiated from *standing* action using pose information. Thus, pose contextual information of a person can provide us crucial cues along with the appearance information for understanding individual actions as well as group activities.

Pose context. Individual actions of players in Volleyball videos depict distinct postures. For example, in Fig. 4 the individual actions such as *standing*, *waiting*, *blocking*, and *spiking* can easily be differentiated by looking at the individual poses. In this paper, we propose to exploit this posture information as context for group activity recognition. To encode pose information of a person, we first pass these video frames through the state-of-the-art HR-Net [34], [38] pose

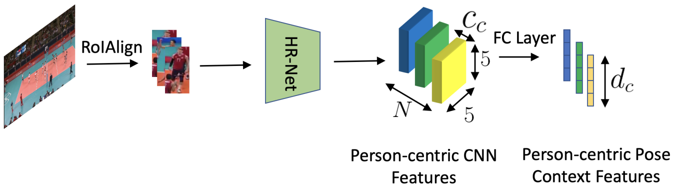
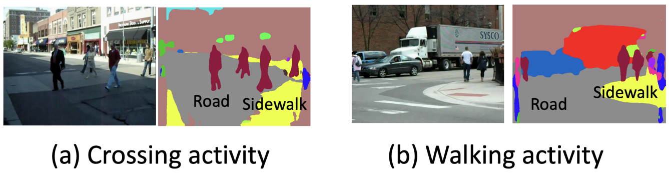

RoIAlign
HR-Net
c_c
5
FC Layer
N
5
d_c
Person-centric CNN Features
Person-centric Pose Context Features

Fig. 5: **The pose context network.** It encodes pose contextual information of every person present in a scene and used as context network in our model. See III-C for more details.

estimation network pretrained on COCO keypoints dataset [44] and extract feature map of dimension $h_c \times w_c \times c_c$ from the penultimate layer of the HR-Net, where h_c, w_c are the height, width and c_c denotes the depth of the context feature map. Then, we use RoIAlign [14] layer to extract fixed size convolutional features of dimension $N \times 5 \times 5 \times c_c$ for each of N person bounding boxes. These convolutional features are then passed through a FC layer to obtain pose contextual representation of dimension $x_c \in \mathbb{R}^{N \times d_c}$, where d_c represents the contextual feature dimension. (see Fig. 5).



(a) Crossing activity

(b) Walking activity

Fig. 6: Scene labels provide important cues about the environment in a video. For example, in the first case (a) a group of people are moving on a road, whereas in the second case (b) people are moving on a sidewalk. This kind of scene contextual information can be exploited for group activity recognition.

Scene context. As shown in Fig. 6, scene labels provide important cues about the environment. In this work, we propose to use the scene labels for group activity recognition. In Fig. 7, we show the context network extracting scene contextual features. For scene contextual features, we pass the video frames through HR-Net [38] segmentation network to obtain pixel-level scene labels of dimension $H \times W$. Specifically, we use HR-Net model pretrained on ADE20K [42] dataset. These scene semantic features are integers (class indices).

The segmentation map is then transformed into one-hot encoded map of dimension $H \times W \times N_c$, where N_c represents the number of scene classes. We then pass this one-hot encoded map through three convolutional layers of kernel size k, stride s, and number of output channels c_c to obtain a feature map of dimension $h_c \times w_c \times c_c$ where h_c, w_c, and c_c denote height, width, and number of channels respectively. Note that the receptive field of the feature map, i.e., the amount of surrounding context captured inside a person bounding box, depends on the number of convolution layers, the kernel size, and the amount of stride. Given the N person bounding boxes, we use the RoIAlign [14] layer to extract person-level convolutional features of dimension $N \times 40 \times 40 \times c_c$. This allows us to leverage sufficient information from the region surrounding each detected person and encode it as part of scene context. Then, these features are processed through a global average pooling (GAP) layer. The resulting feature dimension of shape $N \times c_c$ are passed through a FC layer to obtain scene contextual information around a person of dimension $x_c \in \mathbb{R}^{N \times d_c}$.

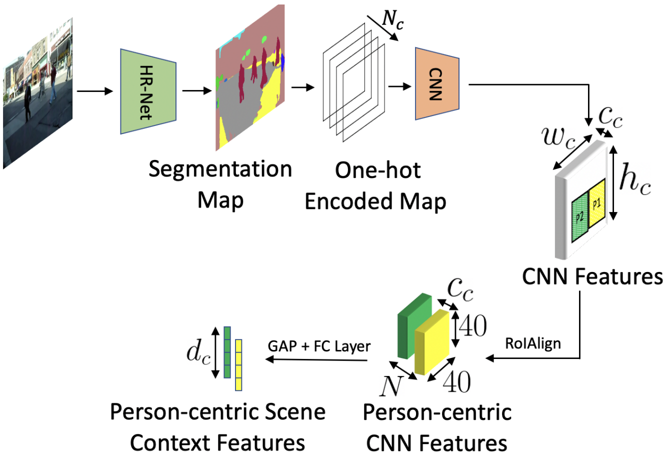


Fig. 7: **The scene context network.** It encodes nearby scene contextual information of each individual present in a scene. We use this as context network in our proposed model. See III-C for more details.

D. Relation Network

To model pair-wise relationship among the people present in a scene, we use GCN [20]. Specifically, we follow the approach proposed by Wu *et al.* [40] to model interactions. We construct a fully connected graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \in \{v_1, v_2, .., v_N\}$ and $\mathcal{E} \in \{e_{1\to1}, e_{1\to2}, .., e_{N\to N}\}$ denote the vertices and edges of the graph respectively. We define each actor as a vertex $v_i = \{(x_i^v, x_i^p)\}$, where $x_i^v = [x_i^a, x_i^c] \in \mathbb{R}^{N \times (d_a + d_c)}$ and x_i^p is the positional features of the i^{th} person respectively. Following [40], we use the distance mask to encode positional information x_i^p. As defined in [40], we also define the edges as follows:

$$e_{i \to j} = h(f_v(x_i^v, x_j^v), f_p(x_i^p, x_j^p)), \tag{1}$$

where $f_v(x_i^v, x_j^v)$ encodes visual relations, $f_p(x_i^p, x_j^p)$ encodes the position relations between two people and h embeds visual and position relations into another vector. This graph $\mathcal{G}$ is then processed with a one layer of GCN and produces a output vector x^r of dimension $N \times d_r$. This vector encodes the relationships that exist among the actors in a scene. For more details of the relation network, readers can refer to the original paper [40]. Finally, the relation feature x_r is fed to two classifiers to predict group activities and individual actions.

E. Training

The entire network is trained in two-stage manner due to the huge memory requirement of the backbone networks used in the appearance and the context network. We first train the appearance network and the context network independently without relational network for group activity and individual action recognition. Then, we freeze the weights of these networks and train the relation network using the features extracted from them. The output of the relation network is then fed to two classifiers to predict group activities and individual actions. The basemodel and GCN is trained using the crossentropy loss as follows:

$$\mathcal{L} = \mathcal{L}_g(y_g, \hat{y}_g) + \mathcal{L}_a(y_a, \hat{y}_a), \qquad (2)$$

where $\mathcal{L}_g$ and $\mathcal{L}_a$ are crossentropy losses, y_g and y_a are the groundtruth labels, $\hat{y}_g$ and $\hat{y}_a$ are the model predictions for group activities and individual actions respectively.

IV. Experiments

A. Datasets and evaluation

We use two publicly available datasets for experimental analysis: Volleyball [17] and Collective Activity [9].

Volleyball. It contains 4830 clips of length 41 frames (3493 for training and 1337 for testing) in total from 55 volleyball sports videos. The annotations include 8 group activities (i.e., *right set*, *right spike*, *right pass*, *right winpoint*, *left winpoint*, *left pass*, *left spike*, and *left set*), 9 individual actions (i.e., *waiting*, *standing*, *digging*, *blocking*, *falling*, *jumping*, *moving*, *setting*, and *spiking*), and bounding boxes for all players in the middle frame of each of the clips. We use tracklet data provided by [5] to obtain the bounding box annotations for all the frames. Following [18], we also use 5 frames before and 4 frames after the middle frame (10 frames in total) for training and testing our proposed model. We evaluate the performance of our model with the accuracy measure for group activity and individual action recognition.

Collective Activity. The collective activity dataset contains 44 videos of varying length recorded with a hand-held camera. The annotations include 5 group activities (i.e., *crossing*, *waiting*, *queuing*, *walking*, and *talking*), 6 individual actions (i.e., *crossing*, *waiting*, *queuing*, *walking*, *talking*, and *NA*), and bounding box coordinates of every individual for every 10th frame. The group activities are determined by the actions performed by the majority of the people. Following [29], we also train on 32 videos and test on 12 videos. We report the performance of our model in terms of group activity recognition accuracy.

B. Implementation details

We use Inception-v3 [35] and VGG19 [33] as our backbone networks for person-centric convolutional appearance feature extraction. Due to the huge memory requirement of VGG19, we do not use VGG19 for volleyball dataset. We resize each frame to 720×1280 for Volleyball dataset and 480×720 for Collective Activity dataset. We resize the appearance features extracted from CNN backbone to 57×87 and extract person-level features of size 5×5 using RoIAlign [14] layer. The depth of the extracted CNN feature map $c_a = 1056$ for Inception-v3 and $c_a = 512$ for VGG19 backbone. The appearance network outputs $d_a = 1024$ dimensional feature for each of the person in a scene.

For pose context, we use the state-of-the-art pose recognition network called HR-Net [34], [38]. We first crop people from the input images and resize to 256×192 using the RoIAlign [14] layer and pass it through the HR-Net backbone to obtain pose contextual features. The context network encodes each of the pose features into $d_c = 256$.

For scene contextual information, we extract segmentation maps for collective activity dataset using HR-Net pre-trained on ADE20K dataset [42] containing 150 classes. We choose $N_c = 9$ common classes out of these 150, such as *road*, *sidewalk* etc. The convolutional layers in the scene context module has a kernel size $k = 3$, stride $s = 2$, and no. of output channels $c_c = 64$. Then, the context network encodes each of the scene context into $d_c = 512$ dimensional vector.

Due to memory constraints, we train our model in two stages: first we finetune the backbone network for group activity recognition task. We refer to this as basemodel throughout our experiments. Second, we fix the backbone network and train the GCN using the features obtained from the basemodel. Note that we do not finetune the HR-Net segmentation model on the target task.

The model is trained to minimize the joint cross-entropy loss described in III-E using Adam [19] optimizer. On Collective Activity dataset, the learning rate is fixed at $1e-5$ when training the basemodel and $1e-4$ when training the GCN model. We use learning rate $1e-5$ and $2e-4$ for basemodel and GCN respectively on volleyball dataset. For Volleyball dataset, we train the basemodel and GCN for 200 and 150 epochs respectively. We train the basemodel and GCN for 100 and 50 epochs respectively for Collective Activity dataset. To reduce the computation during training, we randomly select 1 and 3 frames to train the basemodel and the GCN respectively. During inference, we use all 10 frames of a clip. We use the implementation provided by Wu *et al.* [40], [46] and build our modules on top it.

Method	Finetune Backbone	Group Activity ↑
Pose coordinate embedding	✗	25.00%
HR-Net Feature Embedding	✗	86.46%
HR-Net Feature Embedding	✓	**90.95**%

TABLE I: Ablation study on effective representation of pose information for the Volleyball dataset. The results show the benefit of using pose features from HR-Net backbone over direct pose coordinates embedding.

C. Ablation study

We first perform an ablation study on both the volleyball [18] and the collective activity [9] dataset to demonstrate the effectiveness of different components of our model. We use group activity recognition as a metric to evaluate the performance in all our ablation experiments.

How to represent pose context information? We first experiment on the volleyball dataset to investigate the effective way to represent pose information in our model. Specifically, we remove the appearance stream from our model and experiment with only context network. In this work, we explore three types of pose representations. In the first case, we extract the pose coordinates of all joints of dimension 17×2 using HR-Net. We directly embed these coordinates using a fully connected layer to get an embedded vector. In the second case, we extract features from the last convolutional layer of HR-Net model and embed these convolutional features using an affine layer. The third case is similar to the second case except the fact that we finetune the HR-Net backbone on the target task.

Method	Backbone	Group Activity ↑
Appearance only	Inception-v3	91.62%
Pose only	HR-Net	90.95%
Ours, late fusion	Inception-v3 + HR-Net	91.77%
Ours, early fusion	Inception-v3 + HR-Net	**93.04**%

TABLE II: Ablation study on the Volleyball dataset showing the efficacy of contextual information for group activity recognition. In case of early fusion, appearance and pose features are fused together before passing them through the relation network. For late fusion, we fuse pose and appearance features obtained from the relation network before the final classification layers. The results clearly shows the benefit of fusing appearance and pose contextual information

Table I demonstrates the performance of our model on volleyball dataset using these three types of pose representations. Finetuning the HR-Net backbone has proved to be viable representation giving 65.95% of improvement over direct coordinate embedding. We will use this to represent the pose context information in all our subsequent experiments.

How to fuse pose context network with appearance network? Next, we show the efficacy of appearance and context network individually. We also experimentally show the effective way to fuse the appearance and context information on volleyball dataset. First, we remove the context network from our model and train the appearance network with GCN for group activity recognition which is similar to the model proposed in [40]. Table II shows that this achieves 91.62% group activity recognition accuracy. We then remove the appearance network and train with only pose context branch and this leads to a performance of 90.95% accuracy. Then, we fuse the appearance and pose contextual information using late and early fusion techniques. From Table II, it is evident that the early fusion works better resulting in 1.42%, 2.09%, and 1.27% improvement over appearance only, pose only, and late fusion strategy respectively. These experiments suggest that the combination of appearance with pose-level contextual features using early fusion strategy can provide us enriched representations of people present in a scene which can benefit in group activity recognition. We also perform similar experiment on collective dataset. We get 88.50% group activity recognition accuracy with the appearance only network. However, fusing pose contextual information with appearance results in a performance drop of 1.57% (see Table III).

Method	Backbone	Group Activity ↑
Appearance only	Inception-v3	88.50%
	VGG19	88.81%*
Pose only	HR-Net	80.52%
Scene Context only	HR-Net	69.15%
Ours (Appearance + Pose)	Inception-v3 + HR-Net	86.93%
Ours (Appearance + Scene Context)	Inception-v3 + HR-Net	89.93%
	VGG19 + HR-Net	**90.07**%

TABLE III: Ablation study on the Collective dataset showing the efficacy of contextual information for group activity recognition. In this case, scene context is more important than pose context for group activity understanding. *This result reported is from basemodel as adding GCN worsens the group activity recognition performance.

Efficacy of scene contextual information. Now, we show the experiments with the scene contextual information on collective activity dataset in Table III. We choose a subset of 9 common classes present in collective dataset, namely *building*, *floor*, *road*, *grass*, *car*, *sidewalk*, *path*, *wall*, and *background* to build the segmentation maps. First, we remove the context network from our model and keep appearance only branch. We choose to use Inception-v3 and VGG19 as backbone for the appearance network. This results in 88.50% and 88.81% group activity recognition accuracy for Inception-v3 and VGG19 respectively. Second, we remove the appearance network and keep the context network to train the model, which results in 69.15% accuracy. Then, we fuse the scene contextual information with the appearance features, which leads to an improvement of 1.43% and 1.26% over appearance only network for Inception-v3 and VGG19 backbones respectively.

D. Comparison to the state-of-the-art

Volleyball. Table IV compares our approach to the state-of-the-art methods on Volleyball dataset. As shown in Table IV, our method surpasses all state-of-the-art approaches that use 2D CNN backbone. As mentioned earlier, we use the same relational network proposed in [40] and this makes our appearance only network identical to the model proposed in [40]. Our method outperforms Wu *et al.* [40] by 1.42% for group activity recognition and 1.74% for individual action recognition. Although Azar *et al.* [4] use 3D CNN backbone, our method performs at par with this method. This shows the potential of incorporating contextual cues present in a scene to boost the performance of group activity recognition models.

The most confusion arises between *set*, *spike*, and *pass* activities for Volleyball dataset. For example, the appearance only model struggles to distinguish *right pass* from *right set*. However, our proposed model performs better in distinguishing these two group activities. A similar observation can be made in the case *left set* and *left pass*.

Collective Activity. In Table V, we show and compare the performance of our model to the state-of-the-art. Our method performs better than the basemodel [40] by a margin of 1.26%. The difference between the result reported in Table V

Method	Backbone	Group Activity ↑	Individual Action ↑
Li *et al.* [23]	Inception-v3	66.90%	-
Ibrahim *et al.* [18]	AlexNet	81.90%	-
Shu *et al.* [32]	VGG16	83.30%	-
Biswas *et al.* [6]	AlexNet	83.47%	76.65%
Qi *et al.* [29]	VGG16	89.30%	-
Ibrahim *et al.* [17]	VGG19	89.50%	-
Bagautdinov *et al.* [5]	Inception-v3	90.60%	81.80%
Hu *et al.* [16]	VGG16	91.4%	-
Wu *et al.* [40]*	Inception-v3	91.62%	81.28%
Azar *et al.* [4]	I3D	93.04%	-
Ours (Appearance + Pose Context)	Inception-v3 + HR-Net	**93.04**%	**83.02**%

TABLE IV: Comparison to the state-of-the-art methods on Volleyball dataset with group activity accuracy and individual action accuracy. * This corresponds to the result obtained by retraining the model using the original implementation provided by the authors.

Method	Backbone	Group Activity ↑
Lan *et al.* [22]	-	79.70%
Choi *et al.* [8]	-	80.40%
Deng *et al.* [10]	AlexNet	81.20%
Ibrahim *et al.* [18]	AlexNet	81.50%
Azar *et al.* [4]	I3D	85.75%
Li *et al.* [23]	Inception-v3	86.10%
Shu *et al.* [32]	VGG16	87.20%
Wu *et al.* [40]*	Inception-v3	88.50%
Wu *et al.* [40]*	VGG19	88.81%
Qi *et al.* [29]	VGG16	89.10%
Ours (Appearance + Scene Context)	VGG19	**90.07**%

TABLE V: Comparison to the state-of-the-art methods on Collective dataset with group activity accuracy. * This corresponds to the result obtained by retraining the model using the original implementation provided by the authors.

and [40] is likely due to the non-deterministic behaviour of certain layers in the network and the small size of the dataset as indicated by the authors on their github page [46]. The proposed model also outperforms the previous 2D CNN backbone based approach [29] by 0.97%. Despite the fact that the model proposed by Azar *et al.* [4] uses 3D CNN backbone, our model shows improvement of 4.32% over this method.

Two activities *crossing* and *walking* lead to the most confusion in the predictions made by the appearance only network. Incorporating scene context reduce this confusion. However, our model needs improvement to discriminate between *waiting* and *walking* activities. This may be due to the imperfect segmentation map generated by the HR-Net backbone.

Recently, Gavrilyuk *et al.* [13] proposed a self-attention based approach achieving 94.4% and 92.8% group activity recognition accuracy on Volleyball and Collective Activity datasets respectively. Although [13] perform somewhat better than our method on these datasets, it would be unfair to compare our model with this method as their main improvement comes from the use of much larger 3D CNN backbone (I3D).

V. Conclusion

In summary, we proposed a framework for group activity recognition. Our model combines two complementary sources of information: appearance and context. Specifically, we leveraged two types of contextual cues, namely pose for Volleyball and scene labels for Collective Activity datasets. This joint representation of each person present in a scene is then used for relational reasoning. The effectiveness of our approach is validated on these two datasets, showing notable improvements over comparable models.

We have evaluated the influence of pose contextual information on Volleyball videos, but its utility for other types of sports such as soccer, cricket, is an interesting avenue to consider. Future work can also focus on exploring pose as well as other types of contextual cues for such sports videos. Crowded scenes provide an additional challenge for our method incorporating scene contextual information, and addressing this is another future research direction.

Acknowledgements

This work was supported in part by the ANR AVENUE project (grant ANR-18-CE23-0011). Avijit Dasgupta is supported by a Google India PhD Fellowship.

References

[1] Bogdan Alexe, Nicolas Heess, Yee W Teh, and Vittorio Ferrari. Searching for objects driven by context. In *NeurIPS*, 2012.

[2] Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. HiRF: Hierarchical random field for collective activity recognition in videos. In *ECCV*, 2014.

[3] Mohamed R Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *ECCV*, 2012.

[4] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *CVPR*, 2019.

[5] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *CVPR*, 2017.

[6] Sovan Biswas and Juergen Gall. Structural recurrent neural network (SRNN) for group activity analysis. In *WACV*, 2018.

[7] Zhe Chen, Shaoli Huang, and Dacheng Tao. Context refinement for object detection. In *ECCV*, 2018.

[8] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, 2012.

[9] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *ICCV workshop*, 2009.

[10] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *CVPR*, 2016.

[11] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018.

[12] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *CVPR*, 2009.

[13] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *CVPR*, 2020.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[15] Geremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.

[16] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *CVPR*, 2020.

[17] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *ECCV*, 2018.

[18] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[21] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012.

[22] Tian Lan, Yang Wang, Weilong Yang, Stephen N Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *TPAMI*, 2011.

[23] Xin Li and Mooi Choo Chuah. Sbgar: Semantics based group activity recognition. In *ICCV*, 2017.

[24] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *CVPR*, 2019.

[25] Matteo Lisotto, Pasquale Coscia, and Lamberto Ballan. Social and scene-aware trajectory prediction in crowded spaces. In *ICCV workshop*, 2019.

[26] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *CVPR*, 2018.

[27] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[28] Jaap Munneke, Valentina Brentari, and Marius Peelen. The influence of scene context on object recognition is independent of attentional focus. *Frontiers in psychology*, 2013.

[29] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *ECCV*, 2018.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[31] Abhinav Shrivastava and Abhinav Gupta. Contextual priming and feedback for faster R-CNN. In *ECCV*, 2016.

[32] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Cern: confidence-energy recurrent network for group activity recognition. In *CVPR*, 2017.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[36] Antonio Torralba, Kevin P Murphy, William T Freeman, Mark A Rubin, et al. Context-based vision system for place and object recognition. In *ICCV*, 2003.

[37] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020.

[38] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

[39] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *CVPR*, 2017.

[40] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, 2019.

[41] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[42] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *IJCV*, 2018.

[43] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Context-reinforced semantic segmentation. In *CVPR*, 2019.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[45] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, Monique Thonnat. VPN: Learning Video-Pose Embedding for Activities of Daily Living. In *ECCV*, 2020.

[46] https://github.com/wjchaoGit/Group-Activity-Recognition