Spatial Temporal Network for Image and Skeleton Based Group Activity Recognition [⋆]

Xiaolin Zhai[1,2][0000−0003−1890−8378], Zhengxi Hu[1,2][0000−0001−6119−4185], Dingye Yang[1,2][0000−0002−1039−6817], Lei Zhou[1,2][0000−0002−1940−8597], and ✉Jingtai Liu[1,2][0000−0003−2645−5655]

[1] Institute of Robotics and Automatic Information System, College of Artificial Intelligence, Nankai University
[2] Tianjin Key Laboratory of Intelligent Robotics, Nankai University
{ 2120210410,hzx,1711502}@mail.nankai.edu.cn,zhouleinku@gmail.com,
liujt@nankai.edu.cn

Abstract. Group activity recognition aims to infer group activity in multi-person scenes. Previous methods usually model inter-person relations and integrate individuals' features into group representations. However, they neglect intra-person relations contained in the human skeleton. Individual representations can also be inferred by analyzing the evolution of human skeletons. In this paper, we utilize RGB images and human skeletons as the inputs which contain complementary information. Considering different semantic attributes of the two inputs, we design two diverse branches, respectively. For RGB images, we propose Scene Encoded Transformer, Spatial Transformer, and Temporal Transformer to explore inter-person spatial and temporal relations. For skeleton inputs, we capture the intra-person spatial and temporal dynamics by designing Spatial and Temporal GCN. Our main contributions are: i) we propose a spatial-temporal network with two branches for group activity recognition utilizing RGB images and human skeletons. Experiments show that our model achieves 97.1% MCA and 96.1% MPCA on the Collective Activity dataset and 94.0% MCA and 94.4% MPCA on the Volleyball dataset. ii) we extend the two datasets by introducing human skeleton annotations, namely human joint coordinates and confidence, which can also be used in the action recognition task. The code is available at https://github.com/zxll0106/Image_and_Skeleton_Based_Group_Activity_Recognition.

Keywords: Group activity recognition · Video analysis · Scene understanding.

1 Introduction

Group activity recognition [6] is widely used in social behavior understanding, service robots, and autonomous driving cars, therefore playing a vital role in

[⋆] This work is supported in part by National Key Research and Development Project under Grant 2019YFB1310604, in part by National Natural Science Foundation of China under Grant 62173189.

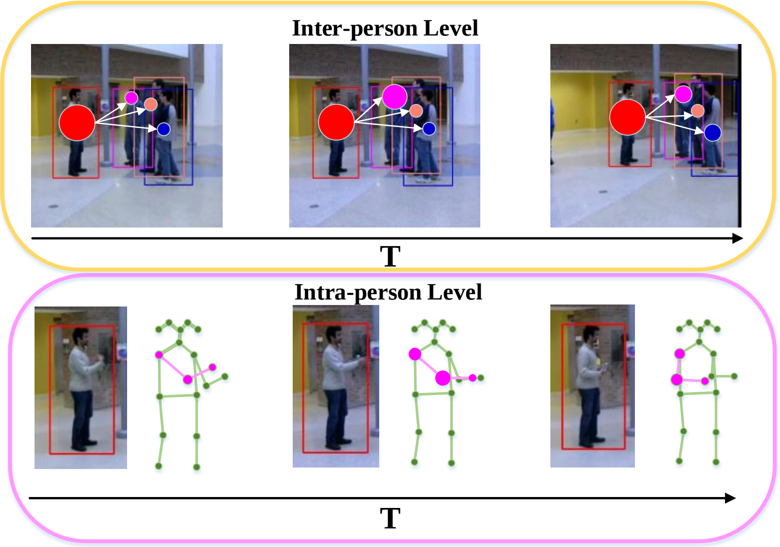


Fig. 1. Intuitive examples of a *talking* video clip in the Collective Activity dataset. On the inter-person level, we capture the spatial relation in the group at the same frame, and temporal dynamics of interaction between consecutive frames. On the intra-person level, the human skeleton reveals the more detailed evolution of the individual action, so we model the complex and diverse spatial-temporal individual features.

video analysis and scene understanding. The goal of group activity recognition is to understand what they are doing in the multi-person scene. In the previous methods, great efforts have been made to verify the effectiveness of modeling inter-person relations. However, intra-person relations in the human skeleton convey fine-grained actions information, but receive less attention in the group activity recognition task.

To model the relationship between individuals, graph neural networks are applied to represent inter-person social relations. Nodes in the graph are individuals and the edges are formed by the relationship between them. Recent attention in group activity recognition has also focused on Transformer [28] which was proposed in the field of natural language processing. Transformer in group activity recognition [37,11,19] is utilized to model the inter-person relation and combine individuals' features. In addition, some methods [11,2,19] utilized optical flow frames which contain different motion features from RGB images. Previous works show impressive improvement, which suggests the effectiveness of modeling inter-person relations. However, they neglect that intra-person information of the human skeleton contains fine-grained motion dynamics.

Human skeletons convey information complementary to RGB images and optical flow frames. Human bodies can be viewed as a whole composed of a trunk and limbs, while human actions can be regarded as motions of joints and bones. And skeletons can be represented by 2D position sequences of human joints. With the development of pose estimation networks, we can obtain more accurate human joint coordinates. From reliable skeleton data, we can model intra-person spatial-temporal dynamics and extract an effective representation.

Accordingly, we propose a spatial-temporal network with two branches to capture the intra-person and inter-person relations. For RGB image inputs, we propose Scene Encoded Transformer to model the interaction between individuals and the surrounding scene, and extract informative scene features. As shown in Fig. 1, individuals also take consideration into the reaction of others to determine their behaviors, besides interacting with the scene. We adopt Spatial and Temporal Transformer to infer spatial and temporal inter-person relations, respectively. Moreover, for skeleton inputs, Spatial and Temporal GCN are designed to model intra-person spatial and temporal relations in the human skeleton. In Spatial GCN, spatial information is propagated from one node to another along intra-person connections. Temporal GCN passes temporal dynamics between consecutive frames of the same node. As shown in Fig. 1, exploiting intra-person spatial-temporal relations is important to modeling individuals' features.

The contributions of this work can be summarized as:

(1)We propose a spatial-temporal network with two branches for group activity recognition utilizing RGB images and human skeletons. For RGB image inputs, we propose Scene Encoded Transformer, Spatial Transformer, and Temporal Transformer to model the inter-person relation. For skeleton inputs, we propose Spatial and Temporal GCN to capture intra-person spatial and temporal relations which lack further exploration in the group activity recognition task. Experiment results show that our proposed model achieves 97.1% MCA and 96.1% MPCA on the Collective Activity dataset and 94.0% MCA and 94.4% MPCA on the Volleyball dataset.

(2)We extend the two datasets by introducing human skeleton annotations, namely human joint coordinates and confidences. Extended datasets can be not only used in group activity recognition but also used in skeleton-based action recognition.

2 Related Work

2.1 Group Activity Recognition

Machine learning plays an important role in addressing the issue of group activity recognition. Initially, many approaches designed hand-crafted features and applied probabilistic graphical models [1,4,5,7,12,17]. With the development of deep neural networks, many approaches [16,3,15,24,8] used CNNs to extract the feature map of the video clip. RNNs were also used to infer temporal dynamics of individual actions and group activity [16,8,3,15,24]. [16] utilized the person LSTM layer and the group LSTM layer to extract individual and group features, respectively. [3] designed a unified framework where RNN can reason the probability of individual actions. The semantic graph extracted from text labels and images was applied in stagNet [24] and structural-RNN was used to capture temporal dynamics.

Recent development in graph neural networks has improved the ability of modeling relation between individuals [30,9,23,33,14,38]. ARG[30] constructed actor relation graphs to capture appearance and position relations between actors. [9] utilized I3D as the backbone network to extract spatial-temporal features of a video clip, used self-attention to integrate individuals' features, and used graph attention to model relations between individuals. [23] used the mean-field conditional random fields to infer temporal relations and spatial relations. The social adaptive module designed in [33] has the same structure in the spatial and temporal domain and can infer key instances under the weakly supervised setting. In PRL[14], individuals' relation was represented on the semantic relation graph. Dynamic Inference Network was proposed in [38] to construct person-specific spatial and temporal graphs by designing Dynamic Relation(DR) module and Dynamic Walk(DW) module.

Meanwhile, researchers in group activity recognition [11,19,37,27,40] have shown an increased interest in Transformer [28] which was proposed in the field of natural language processing. Actor-Transformer [11] used RGB frames, optical flow frames, and pose features as input and used Transformer to model group representation. GroupFormer [19] designed a Transformer encoder to capture spatial and temporal features and adopted a Transformer decoder in a cross manner to capture spatial-temporal interactive features. [37] enhanced individuals' representations with the global contextual information and aggregated the relation between individuals using Spatial-Temporal Bi-linear Pooling module.

Previous works paid more attention to exploring the inter-person relation and confirmed its effectiveness in inferring group features. However, they neglected that the human skeleton which contains intra-person relations also conveys different fine-grained information. Different from them, we capture spatial and temporal relations both on intra-person and inter-person levels in parallel by designing two branches.

2.2 Modeling of Interaction

Modeling interaction relationships is an important component in the multi-instance problem, such as action recognition [21,18,34], pedestrian intent estimation [35,36], and trajectory prediction [10,39]. To model the interaction between joint nodes, [21] proposed the MS-G3D network which can disentangle the node representations in different spatial-temporal neighborhoods. Actional links were utilized in [18] to propagate actional information between different joint nodes and structural links were used to expand the respective fields. To exploit the interaction between the target individual and other traffic users, [35] proposed Attentive Relation Network which adopted soft attention to assign the weight of multiple traffic users. AgentFormer [39] utilized individual-aware attention to capture individual-to-itself and individual-to-others relations. Previous works have considered only one of the intra-person and inter-person relations. Different from them, we capture the interaction on the intra-person and inter-peron

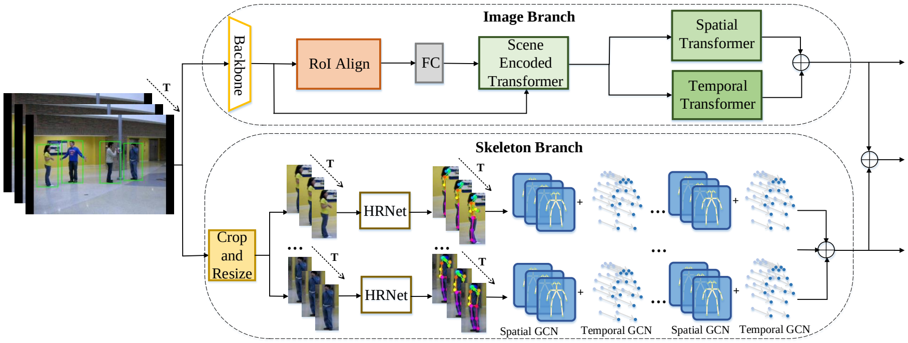


Fig. 2. Overview of the proposed network, which contains the Image Branch and the Skeleton Branch. The Image Branch is proposed to model the inter-person interaction and the Skeleton Branch is designed to extract the intra-person relation.

levels. Considering different semantic attributes of the two levels, we propose the two-branch model to refine different-grained spatial-temporal dynamics.

3 Method

In this section, we outline the pipeline of our method. The overall framework of our method is illustrated in Fig. 2. Our method consists of two branches, namely the Image Branch and the Skeleton Branch. The Image Branch takes RGB frames as input, captures the inter-person interaction, and obtains the spatial-temporal contextual individual and group features. The Skeleton branch takes human skeletons as input, adopts Spatial and Temporal GCN which are suitable for modeling intra-person relations, and obtains complementary individual and group representations. We concatenate the outputs of two branches and pass them to classifiers getting individuals' actions and group activity.

3.1 Image Branch

First, we take T-frame images as the input and obtain individual features from Image Feature Extractor. Then, Scene Encoded Transformer incorporates contextual information in the scene to enhance the individual feature. Considering the continuity and the interactivity of the individual action and group activity, Spatial and Temporal Transformer mine spatial and temporal relations of enhanced individual features. Finally, we concatenate the outputs of Spatial and Temporal Transformer and then apply a pooling layer to get the group feature of the Image Branch.

Image Feature Extractor The input of the Image Branch is T-frame images that are centered on the labeled frame. We utilize the backbone network to extract feature maps $X \in \mathbb{R}^{T\times D\times H\times W}$ of the input frames. Given bounding boxes of N individuals, we apply RoIAlign [13] to extract individuals' features from feature maps. Then a fully-connected layer is adopted to get d-dimensional individuals' features $X_I \in \mathbb{R}^{T\times N\times d}$.

Scene Encoded Transformer Individual actions are influenced by multiple social and environmental elements, in particular the interaction with the scene. Exploring the impact of the scene from a global view lacks further exploration in previous works [30,11,15,38]. To extract relational features of the surrounding scene, we propose Scene Encoded Transformer which captures useful information from the surrounding scene to enhance the individual features. Inspired by Transformer, we note that the attention mechanism plays a critical role in encoding the relative important elements effectively. Taking account into the unique importance of each scene element, we utilize the attention mechanism in Scene Encoded Transformer to calculate the weight based on relations between individuals and the scene. We consider that the scene feature X is composed of $H \times W$ scene elements, and each element contains the visual feature and the position feature. Different scene elements contribute to the target individual unequally, so the attention mechanism is adopted to assign an adaptive weight to each element. To align individuals' features X_I and feature maps X, we project them to d_E dimensional subspaces, obtaining $X'_I \in \mathbb{R}^{T\times N\times d_E}$ and $X' \in \mathbb{R}^{T\times HW\times d_E}$. We view X'_I as the query Q, and view X' as the key K the value V. Then we utilize the similarity between the query Q and the key K to assign the adaptive weight of the value V and compute the weighted sum of value V.

It is difficult to capture order information of inputs because there is no recurrence and no convolution in Transformer [28]. As a result, we encode their position features as follows:

$$\begin{aligned} PE_{(pos,2k)} &= \sin(\frac{pos}{10000^{2k/L}}) \\ PE_{(pos,2k+1)} &= \cos(\frac{pos}{1000^{2k/L}}) \end{aligned} \tag{1}$$

For the ith individuals' features, we encode center coordinates (x_i, y_i) of his bounding box.

$$pos = \begin{cases} x_i, & k \in \{0, 1, \ldots, L/4 - 1\} \\ y_i, & k \in \{L/4, L/4 + 1, \ldots, L/2 - 1\} \end{cases} \tag{2}$$

For the scene feature map, we encode each scene element's coordinate (x, y) of the feature map, where $x \in [0, H], y \in [0, W]$.

$$pos = \begin{cases} x, & k \in \{0, 1, \ldots, L/4 - 1\} \\ y, & k \in \{L/4, d/4 + 1, \ldots, L/2 - 1\} \end{cases} \tag{3}$$

Finally, we obtain the scene encoded individual features $X_E \in \mathbb{R}^{T\times N\times d_E}$ which integrate the scene contextual information to individual features.

Spatial Transformer In order to propagate information between different individuals in the scene, we propose Spatial Transformer to regard individuals as nodes of the graph and aggregate information from surrounding individuals. Our Spatial Transformer captures the target individual's interaction with surrounding individuals and assigns the relative importance weight to them. Then inter-person spatial relations can be integrated into individuals' features. Spatial Transformer views the temporal dimension of scene encoded individuals' features X_E as the batch dimension. We adopt different projection functions to map X_E to Q_S, K_S and V_S, calculate the attention weights of neighbor individuals, and obtain the feature X_S which incorporate spatial information. Positional encoding follows Scene Encoded Transformer.

Temporal Transformer Considering the temporal correlations of individual actions and group activities, we propose Temporal Transformer to encode the evolution of individual features in the temporal domain. Some previous works [9,19] utilized I3D to integrate the temporal feature of the input images, but they ignored individual-level temporal dynamics. Compared to the pooling operation in the temporal domain [30], the attention mechanism in Temporal Transformer can calculate the adaptive importance of individuals in the consecutive frames. Temporal Transformer views the spatial dimension of scene encoded individuals' features X_E as the batch dimension, so we swap the temporal and spatial dimension of X_E reshaping them to $X'_E \in \mathbb{R}^{N \times T \times d_E}$. The self-attention mechanism takes X'_E as inputs and captures contextual temporal information to enrich individuals' features.

Positional encoding in Temporal Transformer uses time order as the position information of different frames.

$$pos = t, t \in [0, T] \tag{4}$$

3.2 Skeleton Branch

We obtain human skeleton data from RGB images using Skeleton Data Extractor. Spatial and Temporal GCN are utilized to model intra-person spatial-temporal features. These features are pooled to form the group feature of the Skeleton Branch.

Skeleton Data Extractor The skeleton data contains intra-person interaction information which is different from inter-person interaction modeled in the Image Branch. It can be represented by 2D coordinates sequences of human joints. We apply HRNet [26] pretrained on the COCO dataset as a pose estimation network. Given the bounding box of each individual, we crop the RGB image and resize it to 192 × 256. HRNet receives cropped images and then predicts human joints coordinates with confidence which are utilized as the input of Spatial and Temporal GCN.

Graph Convolutional Network The input of graph convolutional network (GCN) is graph-structured data which is different from common image-structured data. Layer-wise propagation rules of GCN have a simple structure:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{5}$$

where $\tilde{A} = A + I$, A is the adjacent matrix of the input graph, I is the identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is the degree matrix of $\tilde{A}$, $W^{(l)}$ is the learned weight matrix, $H^{(l)}$ is feature vectors of nodes in the l^{th} layer, and $\sigma(\cdot)$ is the activation function $ReLU(\cdot) = \max(0, \cdot)$.

Spatial GCN Spatial relations of human joints can be represented as a graph, so we utilize Spatial GCN to integrate intra-person spatial relations. We construct the spatial graph $G_t = (V_t, E_t)$ on human joints in the t^{th} frame. $V_t = \left\{v_{ti}|v_{ti}, i = 1, \dots, N^{joint}\right\}$ represents joint nodes, and the attribute of v_{ti} is the coordinates and estimate confidence. The edge set $E_t = \left\{e_{ij}|i, j = 1, \dots, N^{joint}\right\}$ represents the relation between the joint node v_{ti} and v_{tj}. Constructing relation matrix A^S is the key component in GCN. We note that human joints can only move within certain limits and interact with the surrounding joints. Hence, we only consider interaction between the node v_{ti} and its neighborhood set $N(v_{ti}) = \{v_{tj}|d(v_{ti}, v_{tj}) \leq 1, j \neq i\}$. Noting that the motion of joints is driven by the muscles, the movements of muscles are split into two types, namely shortening and lengthening. Muscles shorten pulling the weight towards the body center and muscles lengthen keeping the weight away from the body center. So we divide nodes in the neighborhood set $N(v_{ti})$ into two categories: the proximal set and the distal set.

$$\begin{aligned} \text{Proximal} &= \{v_{tj}|d(v_{tj}, v_{center}) < d(v_{ti}, v_{center}), v_{tj} \in N(v_{ti})\} \\ \text{Distal} &= \{v_{tj}|d(v_{tj}, v_{center}) > d(v_{ti}, v_{center}), v_{tj} \in N(v_{ti})\} \end{aligned} \tag{6}$$

where we regard the neck node as the center v_{center} of the body. The value of A^S_{ij} depends on which set the neighbor node v_{tj} belongs to. v_{ti} in the t^{th} frame are stacked to form the $H^{(0)}$.

Temporal GCN Temporal GCN captures the temporal dynamics of the same joint node in different frames. We construct the temporal graph $G_i = (V_i, E_i)$ on the i^{th} joint node in the different frames, where $V_i = \{v_{ti}|t = 1, 2, \dots, T\}$, $E_i = \{e_{tt'}|t, t' = 1, 2, \dots, T, t' \neq t\}$. We only consider M frames around the t^{th} frame, namely

$$A^T_{tt'} = \begin{cases} 1, & t - \frac{M-1}{2} \leqslant t' \leqslant t + \frac{M-1}{2} \\ 0, & else \end{cases} \tag{7}$$

v_{ti} on the i^{th} joint node are stacked to form $H^{(0)}$.

3.3 Training Loss

To train an end-to-end model, we calculate the loss using the cross-entropy loss:

$$\mathcal{L} = \mathcal{L}_1(y^G, \widehat{y}^G) + \lambda \mathcal{L}_2(y^I, \widehat{y}^I) \tag{8}$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ are the cross-entropy loss, y^G and $\widehat{y}^G$ are the ground truth and the prediction value of the group activity, y^I and $\widehat{y}^I$ are the ground truth and the prediction value of the individual action, and λ is a hyper-parameter balancing the two loss.

4 Experiments

In this section, we conduct experiments on the Collective Activity dataset and the Volleyball dataset. First, we introduce the two datasets which are widely used in group activity recognition. Second, we provide the implementation details of our proposed model. Third, we compare our proposed method with state-of-the-art methods. Finally, we present the ablation study and visualization of our model to verify the effectiveness of our proposed modules.

4.1 Datasets

Collective Activity Dataset The Collective Activity dataset [6] consists of 44 clips composed of frames that range from 194 to 1814. The middle frame of every ten frames contains the bounding box coordinates annotations and individuals' action labels (*NA*, *waiting*, *talking*, *queuing*, *crossing*, and *walking*). Actions of most individuals in the same scene determine the group activity (*waiting*, *talking*, *queuing*, *crossing*, and *walking*). We follow [29,31,32] to merge the label *crossing* with *walking* to the label *moving*. The test set is composed of 1/3 of the video clips and the training set is composed of the rest of the video clips following [24].

Volleyball Dataset The Volleyball dataset [15] consists of multiple volleyball clips whose length is 41 frames. The middle frame of each clip contains bounding box coordinates, individual action labels, and group activity labels. Individual action labels contain 9 actions: *setting*, *digging*, *falling*, *jumping*, *blocking*, *moving*, *spiking*, *waiting*, and *standing*. Group activity labels contain 8 activities, namely *right set*, *right pass*, *right spike*, *right winpoint*, *left set*, *left pass*, *left spike*, and *left winpoint*. We split the dataset into the training set composed of 3493 clips and the testing set composed of 1337 clips.

4.2 Implementation details

Referring to the previous methods, we resize the images in the Volleyball dataset to $H \times W = 720 \times 1280$ and images in the Collective Activity dataset to $H \times W = 480 \times 720$. We select $T = 10$ frames in the clips, which contain 5 frames before

Table 1. Comparison with state-of-the-art models on the Collective Activity dataset(CAD) and Volleyball dataset(VD) using Multi-class Classification Accuracy(MCA) and Mean Per Class Accuracy(MPCA) metrics.

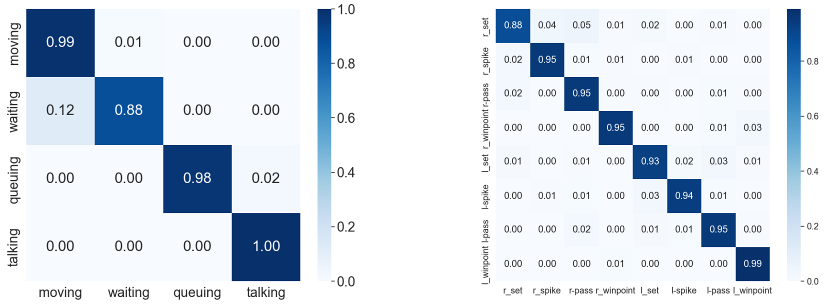
Method	MCA-CAD	MPCA-CAD	MCA-VD	MPCA-VD
HDTM [16]	81.5	-	81.9	-
SBGAR [20]	86.1	-	66.9	-
stagNet [24]	89.1	-	89.3	-
CRM [2]	85.8	94.2	93.0	-
ARG [30]	91.0	-	92.6	-
M. Ehsanpour et al. [9]	89.4	-	93.1	-
PRL [14]	-	93.8	91.4	91.8
Actor-Transformer [11]	91.0	-	93.5	-
DIN [38]	-	95.9	93.6	93.8
Ours-Image	93.7	93.1	92.9	93.5
Ours-Skeleton	95.0	92.8	83.9	83.0
Ours-Image+Skeleton	**97.1**	**96.1**	**94.0**	**94.4**

the middle frames and 4 frames after the middle frames. In the Image Branch, we adopt pretrained VGG16 [25] as the backbone and apply RoIAlign with the crop size $K \times K = 5 \times 5$ on the feature map. After the fully connected layer, the dimension of individuals' features d is 1024. We set scene encoded individual feature dimension $d_E = 1024$. For simplicity, the number of layers in Scene Encoded/Spatial/Temporal Transformer is set as 1. In the Skeleton Branch, we adopt HRNet model $pose\_hrnet\_w32$ which is pretrained on the COCO dataset. Given the bounding boxes of individuals, we crop images and resize them to $H \times W = 256 \times 192$. For the training loss, we set $\lambda = 1$ to balance two tasks. In addition, we set the dropout rate as 0.3 to reduce overfitting.

For training on the Volleyball dataset, we adopt Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. We finetune the backbone network with the learning rate 10^{-5} in 200 epochs and then train the whole model in 30 epochs with the learning rate ranging from 10^{-4} to 10^{-5}. For training on the Collective Activity dataset, we adopt Adam optimizer with the same hyper-parameters. We finetune the backbone network with the same learning rate in 100 epochs and then train the whole model in 30 epochs with the fixed learning rate 10^{-5}. After training the Image Branch and Skeleton Branch separately, we freeze the parameters in the two branches, concatenate the group features extracted by the two branches, and train the classifier layer to obtain the result of the fusion.

4.3 Comparison with the state-of-the-arts

Collective Activity Dataset To verify the effectiveness of our model, we compare our model with the state-of-the-art models on the Collective Activity dataset. The result measured by the Multi-class Classification Accuracy(MCA)



(a) (b)

Fig. 3. (a)The confusion matrix of our proposed model on Collective Activity dataset. (b)The confusion matrix of our proposed model on Volleyball dataset.

and Mean Per Class Accuracy(MPCA) metrics is shown in Table 1. We list the results of the Image Branch, the Skeleton Branch, and the fusion of the two branches. Our Image Branch model and Skeleton Branch model achieves 93.7% MCA and 95.0% MCA respectively, and outperform other methods. Intra-person spatial-temporal dynamics extracted in the Skeleton Branch play a pivotal role in understanding individual and group behaviors. And inter-person relation captured in the Image Branch also models important contextual interaction. The fusion of two branches surpasses all methods by 6.1% MCA and 0.2% MPCA, which demonstrates that the two branches play a complementary role.

To show the effectiveness of our model, we draw the confusion matrix on the Collective Activity dataset shown in Fig. 3(a). It indicates that our model can distinguish well on most classes.

Volleyball Dataset We further evaluate the effectiveness of our model on the Volleyball dataset and list the results compared with the state-of-the-art models in Table 1. MCA and MPCA metrics are also used to evaluate the results on the Volleyball dataset. Our Image Branch model achieves 92.9% MCA and 93.5% MPCA. It shows that context encoded features and spatial-temporal relations between individuals are pivotal for inferring group features. Our skeleton branch model performs not well on the Volleyball dataset, as we think there are overlapping instances in some images, which will lead to a decline of recognition accuracy. The fusion of two branches achieves 94.0% MCA and 94.4% MPCA and outperforms previous methods, which also demonstrates that inter-person and intra-person relations are complementary.

As shown in Fig. 3(b), the confusion matrix of the Volleyball dataset shows that the accuracy of our model achieves 90% in most classes. Our model rarely confuses the left and right sides because our model integrates the whole scene feature with individuals' features and captures the relation between individuals.

Table 2. Ablation study of the Image Branch on the Volleyball dataset. The base model contains the backbone network, utilizes RoIAlign to extract individuals' features, and applies a classifier layer to obtain the group activity.

Scene Encoded Transformer	Spatial Transformer	Temporal Transformer	MCA	MPCA
			88.0	88.6
✓			92.5	93.1
✓	✓		92.2	92.6
✓		✓	92.5	92.8
✓	✓	✓	**92.9**	**93.5**

Table 3. Ablation study of the Skeleton Branch on the Collective Activity Dataset. **SGCN** denotes Spatial GCN module. **TGCN** denotes Temporal GCN module. **STGCN** denotes a layer composed of Spatial and Temporal GCN.

Components	SGCN	SGCN+TGCN	Layers of STCGN					
	-	-	L=2	L=4	L=6	L=8	L=10	L=12
MCA/MPCA	53.7/42.2	69.9/50.6	78.0/67.9	86.7/77.0	87.2/78.7	88.2/79.0	**95.0/92.8**	93.7/90.0

4.4 Ablation Studies

To evaluate the effectiveness of our proposed modules, we perform ablation study using MCA and MPCA metrics.

Scene Encoded Transformer To study the performance of Scene Encoded Transformer, we append this module to the base model. As shown in Table 2, Scene Encoded Transformer improves 4.5% MCA and 4.5% MPCA. Considering that individuals always interact with the scene, the scene contains latent information related to individuals' actions and group activity. Hence, utilizing relations between individuals and the scene can enhance the group representation.

Spatial and Temporal Transformer In Table 2, Spatial and Temporal Transformer are appended after Scene Encoded Transformer, respectively. Using Spatial or Temporal Transformer alone impacts negatively on the effectiveness of our model, as we consider that spatial and temporal dependency are not considered simultaneously. After the concatenation of Spatial Transformer and Temporal Transformer, our model achieves 92.9% MCA and 93.5% MPCA on the Volleyball dataset, which indicates that fusing the complex spatial-temporal interaction can improve the effectiveness of our model.

Spatial and Temporal GCN We conduct the ablation study on the Skeleton Branch to verify the effectiveness of Spatial and Temporal GCN. As can be seen

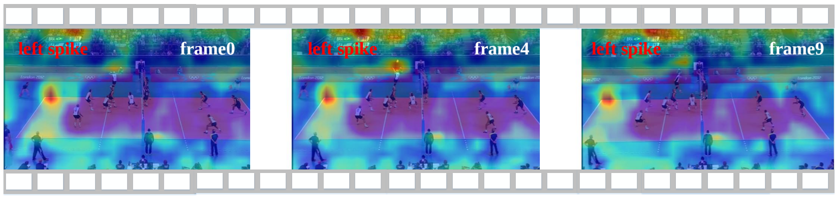


Fig. 4. Visualization results of the attention matrix in Scene Encoded Transformer on the Volleyball dataset.

in Table 3, adding Temporal GCN after Spatial GCN improves the performance from 42.2% MPCA to 50.6% MPCA. It demonstrates that integrating spatial and temporal intra-person relations is fundamental to modeling individuals' representations. In addition, we gradually increase the number of layers, and observe that the performance achieves the best result when the number of layers is set to 10. Stacked layers of Spatial and Temporal GCN can jointly exploit intra-person spatial-temporal relations and integrate them into individuals' features.

4.5 Visualization

Scene Encoded Transformer We visualize the attention map of Scene Encoded Transformer on the Volleyball dataset in Fig. 4. Scene Encoded Transformer emphasizes the location of the volleyball, spectator seats, and referees on the volleyball court. We observe that Scene Encoded Transformer pays more attention to the volleyball at the moment of individual spiking the volleyball. It is clear that key scene features can be captured by Scene Encoded Transformer.

Spatial and Temporal Transformer Visualization results of the *left spike* activity on the Volleyball dataset are shown in Fig. 5. We observe that individual 4 is spiking the volleyball in frame 0. And Spatial Transformer highlights individual 4 and relation with others. This shows that Spatial Transformer pays attention to the key instances and utilizes relations with each other to infer the group activity. In frame 9, individual 6 who is blocking the ball is highlighted by Spatial Transformer. Interaction between individual 6 and individual 4 can also confirm that the group activity is *left spike*. Additionally, Temporal Transformer takes into account the inter-frame influence.

Spatial and Temporal GCN As shown in Fig. 6, we project group features extracted by Spatial and Temporal GCN on the Collective Activity dataset to 2D dimensions using t-SNE [22]. Spatial and Temporal GCN aggregate intra-person spatial and temporal dynamics to the group representations and cluster the representations well. As the layers get deeper, group features are more discriminative. These visualization results verify the effectiveness of our Spatial and Temporal GCN in the Skeleton Branch.

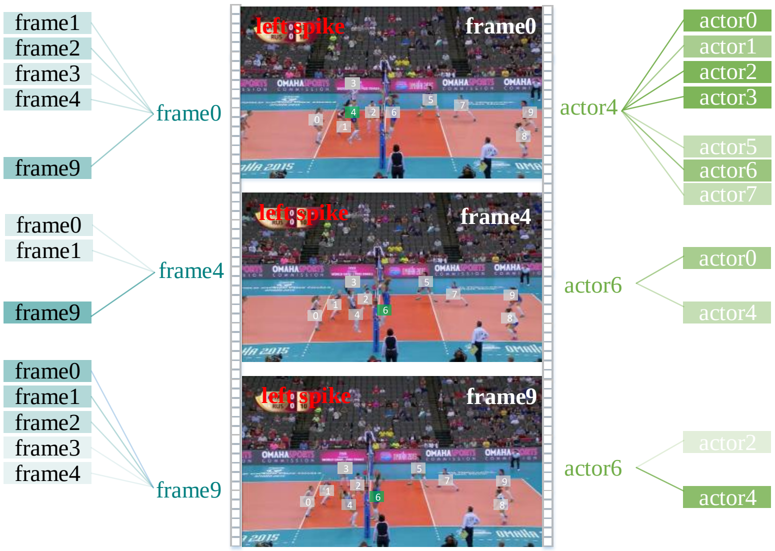


Fig. 5. Visualization results of the attention matrix in Spatial and Temporal Transformer on the Volleyball dataset.

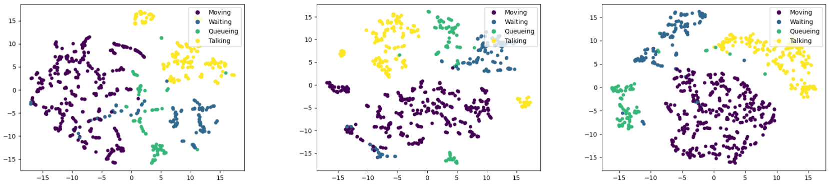


Fig. 6. From left to right, they are t-SNE visualization of 2 layers, 6 layers and 10 layers of Spatial and Temporal GCN on the Collective Activity Dataset.

5 Conclusion

This paper proposes a spatial-temporal network with two branches taking RGB images and human skeletons as input. For RGB image inputs, Scene Encoded Transformer is proposed to incorporate scene features into individuals' features. Spatial and Temporal Transformer are designed to extract spatial and temporal information, respectively. Furthermore, we utilize innovative human skeletons as input and capture spatial and temporal dynamics by utilizing Spatial and Temporal GCN. Experiments demonstrate our proposed model achieves outstanding performance while it can capture spatial and temporal dependencies on the intra-person and inter-person levels.

References

1. Amer, M.R., Lei, P., Todorovic, S.: Hirf: Hierarchical random field for collective activity recognition in videos. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 8694, pp. 572–585. Springer (2014). https://doi.org/10.1007/978-3-319-10599-4_37, https://doi.org/10.1007/978-3-319-10599-4_37
2. Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 7892–7901. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00808, http://openaccess.thecvf.com/content_CVPR_2019/html/Azar_Convolutional_Relational_Machine_for_Group_Activity_Recognition_CVPR_2019_paper.html
3. Bagautdinov, T.M., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 3425–3434. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.365, https://doi.org/10.1109/CVPR.2017.365
4. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 7575, pp. 215–230. Springer (2012). https://doi.org/10.1007/978-3-642-33765-9_16, https://doi.org/10.1007/978-3-642-33765-9_16
5. Choi, W., Savarese, S.: Understanding collective activitiesof people from videos. IEEE Trans. Pattern Anal. Mach. Intell. **36**(6), 1242–1257 (2014). https://doi.org/10.1109/TPAMI.2013.220, https://doi.org/10.1109/TPAMI.2013.220
6. Choi, W., Shahid, K., Savarese, S.: What are they doing? : Collective activity classification using spatio-temporal relationship among people. In: 12th IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2009, Kyoto, Japan, September 27 - October 4, 2009. pp. 1282–1289. IEEE Computer Society (2009). https://doi.org/10.1109/ICCVW.2009.5457461, https://doi.org/10.1109/ICCVW.2009.5457461
7. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011. pp. 3273–3280. IEEE Computer Society (2011). https://doi.org/10.1109/CVPR.2011.5995707, https://doi.org/10.1109/CVPR.2011.5995707
8. Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 4772–4781. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.516, https://doi.org/10.1109/CVPR.2016.516
9. Ehsanpour, M., Abedin, A., Saleh, F.S., Shi, J., Reid, I.D., Rezatofighi, H.: Joint learning of social groups, individuals action and sub-group activities in videos. In:

Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX. Lecture Notes in Computer Science, vol. 12354, pp. 177–195. Springer (2020). https://doi.org/10.1007/978-3-030-58545-7_11, https://doi.org/10.1007/978-3-030-58545-7_11

10. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C.: Vectornet: Encoding HD maps and agent dynamics from vectorized representation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 11522–11530. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01154, https://openaccess.thecvf.com/content_CVPR_2020/html/Gao_VectorNet_Encoding_HD_Maps_and_Agent_Dynamics_From_Vectorized_Representation_CVPR_2020_paper.html
11. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.M.: Actor-transformers for group activity recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 836–845. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00092, https://openaccess.thecvf.com/content_CVPR_2020/html/Gavrilyuk_Actor-Transformers_for_Group_Activity_Recognition_CVPR_2020_paper.html
12. Hajimirsadeghi, H., Yan, W., Vahdat, A., Mori, G.: Visual recognition by counting instances: A multi-instance cardinality potential kernel. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 2596–2605. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7298875, https://doi.org/10.1109/CVPR.2015.7298875
13. 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)
14. Hu, G., Cui, B., He, Y., Yu, S.: Progressive relation learning for group activity recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 977–986. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00106, https://openaccess.thecvf.com/content_CVPR_2020/html/Hu_Progressive_Relation_Learning_for_Group_Activity_Recognition_CVPR_2020_paper.html
15. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III. Lecture Notes in Computer Science, vol. 11207, pp. 742–758. Springer (2018). https://doi.org/10.1007/978-3-030-01219-9_44, https://doi.org/10.1007/978-3-030-01219-9_44
16. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 1971–1980. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.217, https://doi.org/10.1109/CVPR.2016.217
17. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012. pp. 1354–1361. IEEE Computer Society (2012). https://doi.org/10.1109/CVPR.2012.6247821, https://doi.org/10.1109/CVPR.2012.6247821

18. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 3595–3603. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00371, http://openaccess.thecvf.com/content_CVPR_2019/html/Li_Actional-Structural_Graph_Convolutional_Networks_for_Skeleton-Based_Action_Recognition_CVPR_2019_paper.html
19. Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: Groupformer: Group activity recognition with clustered spatial-temporal transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13668–13677 (2021)
20. Li, X., Chuah, M.C.: SBGAR: semantics based group activity recognition. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2895–2904. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.313, https://doi.org/10.1109/ICCV.2017.313
21. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 143–152 (2020)
22. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
23. Pramono, R.R.A., Chen, Y., Fang, W.: Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I. Lecture Notes in Computer Science, vol. 12346, pp. 71–90. Springer (2020). https://doi.org/10.1007/978-3-030-58452-8_5, https://doi.org/10.1007/978-3-030-58452-8_5
24. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Gool, L.V.: stagnet: An attentive semantic RNN for group activity recognition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X. Lecture Notes in Computer Science, vol. 11214, pp. 104–120. Springer (2018). https://doi.org/10.1007/978-3-030-01249-6_7, https://doi.org/10.1007/978-3-030-01249-6_7
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 5693–5703. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00584, http://openaccess.thecvf.com/content_CVPR_2019/html/Sun_Deep_High-Resolution_Representation_Learning_for_Human_Pose_Estimation_CVPR_2019_paper.html
27. Tamura, M., Vishwakarma, R., Vennelakanti, R.: Hunting group clues with transformers for social group activity recognition. CoRR **abs/2207.05254** (2022). https://doi.org/10.48550/arXiv.2207.05254, https://doi.org/10.48550/arXiv.2207.05254

28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
29. Wang, M., Ni, B., Yang, X.: Recurrent modeling of interaction context for collective activity recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 7408–7416. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.783, https://doi.org/10.1109/CVPR.2017.783
30. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 9964–9974. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.01020, http://openaccess.thecvf.com/content_CVPR_2019/html/Wu_Learning_Actor_Relation_Graphs_for_Group_Activity_Recognition_CVPR_2019_paper.html
31. Yan, R., Tang, J., Shu, X., Li, Z., Tian, Q.: Participation-contributed temporal dynamic model for group activity recognition. In: Boll, S., Lee, K.M., Luo, J., Zhu, W., Byun, H., Chen, C.W., Lienhart, R., Mei, T. (eds.) 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018. pp. 1292–1300. ACM (2018). https://doi.org/10.1145/3240508.3240572, https://doi.org/10.1145/3240508.3240572
32. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Higcin: Hierarchical graph-based cross inference network for group activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2020). https://doi.org/10.1109/TPAMI.2020.3034233
33. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Social adaptive module for weakly-supervised group activity recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII. Lecture Notes in Computer Science, vol. 12353, pp. 208–224. Springer (2020). https://doi.org/10.1007/978-3-030-58598-3_13, https://doi.org/10.1007/978-3-030-58598-3_13
34. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 7444–7452. AAAI Press (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17135
35. Yao, Y., Atkins, E., Roberson, M.J., Vasudevan, R., Du, X.: Coupling intent and action for pedestrian crossing behavior prediction. arXiv preprint arXiv:2105.04133 (2021)
36. Yau, T., Malekmohammadi, S., Rasouli, A., Lakner, P., Rohani, M., Luo, J.: Graph-sim: A graph-based spatiotemporal interaction modelling for pedestrian action prediction. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 8580–8586. IEEE (2021)
37. Yuan, H., Ni, D.: Learning visual context for group activity recognition. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI

2021, Virtual Event, February 2-9, 2021. pp. 3261–3269. AAAI Press (2021), https://ojs.aaai.org/index.php/AAAI/article/view/16437

38. Yuan, H., Ni, D., Wang, M.: Spatio-temporal dynamic inference network for group activity recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7476–7485 (2021)
39. Yuan, Y., Weng, X., Ou, Y., Kitani, K.: Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9793–9803. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00967, https://doi.org/10.1109/ICCV48922.2021.00967
40. 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. Proceedings of the 17th European Conference on Computer Vision (ECCV 2022) (2022)