Learning Graph-based Residual Aggregation Network for Group Activity Recognition

Wei Li, **Tianzhao Yang**, **Xiao Wu*** and **Zhaoquan Yuan**
School of Computing and Artificial Intelligence, Southwest Jiaotong University
liwei@swjtu.edu.cn, tianzhao@my.swjtu.edu.cn, wuxiaohk@gmail.com, zqyuan@swjtu.edu.cn

Abstract

Group activity recognition aims to understand the overall behavior performed by a group of people. Recently, some graph-based methods have made progress by learning the relation graphs among multiple persons. However, the differences between an individual and others play an important role in identifying confusable group activities, which have not been elaborately explored by previous methods. In this paper, a novel Graph-based Residual AggregatIon Network (GRAIN) is proposed to model the differences among all persons of the whole group, which is end-to-end trainable. Specifically, a new local residual relation module is explicitly proposed to capture the local spatiotemporal differences of relevant persons, which is further combined with the multi-graph relation networks. Moreover, a weighted aggregation strategy is devised to adaptively select multi-level spatiotemporal features from the appearance-level information to high-level relations. Finally, our model is capable of extracting a comprehensive representation and inferring the group activity in an end-to-end manner. The experimental results on two popular benchmarks for group activity recognition clearly demonstrate the superior performance of our method in comparison with the state-of-the-art methods.

1 Introduction

Group activity recognition (GAR) has many potential applications including video surveillance, social behavior understanding, and sports video analysis. Compared with individual action recognition (e.g., running or jumping), GAR is a more challenging task (e.g., talking together or queuing up) that lies not only in the recognition of individual behaviors, but also in the exploration of the interactions among different persons [Masato *et al.*, 2011]. Therefore, GAR is more easily explained and makes sense in practice.

Given a video sequence, most of existing methods are executed in the following sequential process. First, individual features are extracted based on a backbone network. Second, the spatiotemporal interactions are explored by the graph neural networks [Wu *et al.*, 2019; Yan *et al.*, 2020] or recurrent neural networks (RNNs) [Wang *et al.*, 2017]. Finally, a holistic feature vector is generated to represent the group activity by typically pooling over personal features. However, most of existing methods are limited to the coarse individual-level features, and the fine-grained relations have not been considered as mentioned in [Qi *et al.*, 2020]. Specifically, the differences between an individual and others are ignored, which are vital to suppress the interference of local similar motions.

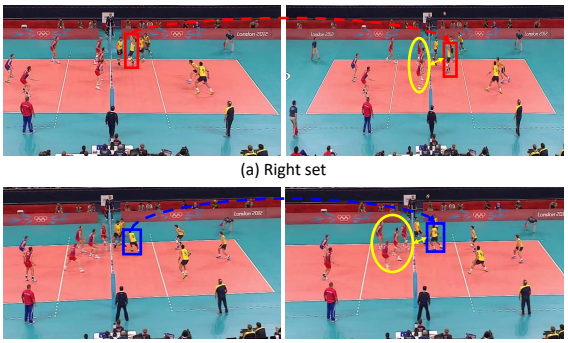

Figure 1: Examples of the "right set" and "right pass" activities in the volleyball game. It is difficult to distinguish them based on the individual behaviors and the spatiotemporal interactions. Intuitively, it is crucial to focus on the differences (yellow double-arrow) between an individual and others (yellow ellipse) for providing additional cues.

To better illustrate this, the group activities of "right set" and "right pass" are taken as examples in Figure 1. A player who is "setting or passing" in the volleyball game adjusts his pose over time to handle the ball. Meanwhile, some of his teammates also move around him. It is difficult to distinguish them based solely on appearance-level features and their interactions, leading to unsatisfactory recognition performance. Fortunately, although the players of both the "right set" and "right pass" have similar motions, their teammates or opponents may form certain latent patterns. For example, the stance and pose among players will maintain consistent changes in a group activity. Therefore, the differences between an individual and others need to be quantitatively

*Corresponding author.

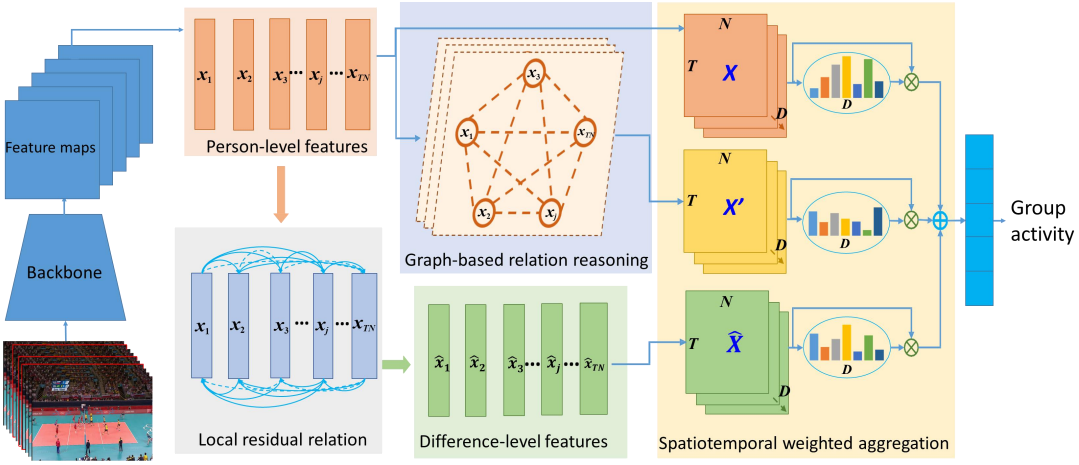

Figure 2: The overview of proposed GRAIN.

modeled as additional information to enhance the features of a whole activity. To this end, a new module is proposed to exploit the local difference relations among the whole group, namely Local Residual Relation Module (LR^{2}M), which is modeled by measuring the sum of residuals within the group. More concretely, LR^{2}M has the ability to capture the latent differences in participants' spatiotemporal features (e.g., the spatial variations and behavior changes over the whole activity), while suppressing the redundancy of temporal actions (e.g., the same motions over time). Furthermore, LR^{2}M can be easily combined into an existing relation graph framework to form the dual-branch relation reasoning.

Empirically, although all persons may be involved in a group activity, only a few motions or interactions are crucial to inferring the class of group activity. For example, the residual or graph-based relations may contain redundant interactions between outlier persons. PRL in [Hu *et al.*, 2020] has validated that it is vital to model the group-relevant and suppress the irrelevant features for GAR. However, most of existing methods treat multi-level features with equal importance, and integrate them by simple summing operators. Thus, it is necessary to find an effective strategy to integrate the multi-level features. Based on dual-branch relation reasoning, a Weighted Aggregation Strategy (WAS) is proposed to learn the weights of multi-level spatiotemporal features. The basic idea is to use adaptive gates to control the information flows from different branches to the holistic feature vector, while suppressing the irrelevant information across multiple persons. Overall, the main contributions can be summarized as follows.

- Going one step further beyond the existing methods, a novel graph-based residual aggregation network (GRAIN) is proposed to encode the latent differences among all persons and learn the refined features for GAR, which is end-to-end trainable. To the best of our knowledge, GRAIN is the first work to model the differences among all the persons in a whole group for GAR.
- To handle the inter-class similarity of different group activities, a new local residual relation module (LR^{2}M) is proposed to model the latent spatiotemporal differences among the whole group.
- To suppress the irrelevant information, a weighted aggregation strategy (WAS) is devised to aggregate multi-level spatiotemporal features.
- Extensive experiments on two publicly available datasets show that GRAIN outperforms the state-of-the-art methods.

2 Related Work

2.1 Action Recognition

Action recognition aims to detect the individual actions from video sequences or still images [Liu *et al.*, 2021]. To capture the temporal dependencies of visual features, RNNs have been used to model the actions in videos. To handle the spatial configurations of articulated skeletons, a two-stream RNN architecture [Wang and Wang, 2017] is proposed to model both temporal dynamics and spatial relations for skeleton-based action recognition. In TEA [Li *et al.*, 2020], both short-range motion excitation and long-range temporal aggregation are designed to capture both short- and long-range temporal evolution. In addition, some graph-based methods have been proposed to model the spatiotemporal information for long-term action recognition [Li *et al.*, 2021a; Zhou *et al.*, 2021]. Different from the focuses of aforementioned methods, our method elegantly explores the different relations in a social group, where multiple persons are involved in modeling their interactions.

2.2 Group Activity Recognition

Initial GAR approaches employ the probabilistic graphical models [Amer and Todorovic, 2015] or AND-OR grammar models [Shu *et al.*, 2015] to predict the group activities after extracting the hand-crafted features. The inter-person relations can also provide some important cues [Qi *et al.*, 2020]. ARG [Wu *et al.*, 2019] firstly employs graph convolution networks (GCNs) to learn the appearance and position relations. HiGCIN [Yan *et al.*, 2020] designs a hi-

erarchical graph-based cross inference network to extract multi-level features from the body-region to group-activity for GAR. Inspired by deep reinforcement learning, a progressive relation learning method is proposed to distill the features of group-relevant actions and interactions for building high-level semantic relation graph [Hu *et al.*, 2020]. Furthermore, transformer [Gavrilyuk *et al.*, 2020; Yuan and Ni, 2021; Li *et al.*, 2021b] is employed to bridge the gap between group activity and visual context information. However, the fine-grained relations have still not been elaborately explored by previous methods, such as the differences between an individual and others. These prompt us to develop a novel module to capture the differences, while suppressing the interference of local similar motions.

3 The Proposed Method

3.1 Pipeline

The overview of GRAIN is illustrated in Figure 2. For the backbone network, ResNet18 [He *et al.*, 2016] or VGG16 [Simonyan and Zisserman, 2015] is employed in the experiments to make a fair comparison with previous methods. The stacked appearance-level features can be denoted as $X \in R^{T\times N\times D}$, where T, N, D denote the temporal, spatial and feature dimensions, respectively. First, the residual relation module is built to capture the difference-level features $\hat{X} \in R^{T\times N\times D}$ among all persons of the whole group. Then, the appearance-level features are arranged into the multiple spatiotemporal graphs to encode the interaction-level features among the individuals. After that, the weighted aggregation strategy is introduced to further suppress the irrelevant spatiotemporal features. Finally, following previous methods [Yan *et al.*, 2020; Yuan *et al.*, 2021], the refined features are pooled and fed into a soft-max layer to predict the group activity. The cross-entropy loss function is used to train the model, which can be trained in an end-to-end manner. For the inference of a video clip, the final probability score of group activity is obtained by the output of the soft-max classifier. Details are presented in the following sections.

3.2 Local Residual Relation

Most of existing methods are executed in a similar process, where individual features are fed into a variety of elaborated GCNs for relational reasoning [Wu *et al.*, 2019; Dang *et al.*, 2021]. In fact, when the appearance-level features and their relations are not sufficient, the differences in appearance are crucial for identifying false-positive results, as shown in Figure 1. Although the performance of GAR has been constantly improved, the differences between an individual and others have not been elaborately explored by previous methods. Therefore, how to model the differences is a key component in our framework. The basic idea is to build a new local residual relation module, considering both appearance differences and spatial location constraints of each individual member.

Specifically, the appearance-level features $X = \{x_i\}_{i=1}^{TN}$ is fed into the residual relation module, where $x_i \in R^D$. For convenience, the output of residual relation module $\hat{X}$ is written as $\hat{X} = \{\hat{x}_j\}_{j=1}^{TN}$, where $\hat{x}_j \in R^D$. Then, the appearance

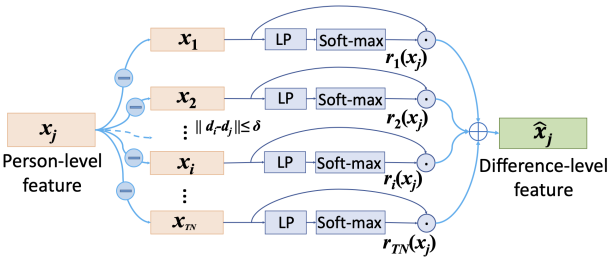

Figure 3: The difference-level features of one person by using the local residual relation module. LP means the linear projection.

difference between x_j and other related persons is computed as follows:

$$\hat{x}_j = \sum_{i=1}^{TN} r_i(x_j)(x_j - x_i) \tag{1}$$

where $r_i(x_j)$ denotes the interaction factor. Intuitively, $r_i(x_j) = 1$ if there exists a residual relation between x_j and x_i, otherwise $r_i(x_j) = 0$. It is easy to see that the appearance difference in Eqn. (1) is not differentiable and cannot be trained directly by back-propagation. Inspired by NetVLAD [Arandjelovič *et al.*, 2018], the soft assignment is employed to indicate other persons associated with x_j. $r_i(x_j)$ is defined as

$$r_i(x_j) = \frac{exp(w_j(x_j - x_i) + b_j)}{\sum_{i=1}^{TN} exp(w_j(x_j - x_i) + b_j)} \tag{2}$$

where w_j and b_j are the learnable weights that project the difference $(x_j - x_i)$ to a scalar for the j-th person.

However, it should be noted that only a few interactions are crucial to inferring the class of group activity in the real world. Although $\hat{x}_j$ in Eqn. (1) has the ability to capture the residual relation, it may involve some irrelevant persons. As mentioned in ARG [Wu *et al.*, 2019], the relation information in the local scope is more meaningful than the global relation in the modeling of group activities. Based on these facts, a spatial factor δ is introduced to constrain persons who are closer to x_j. The final $\hat{x}_j$ in $\hat{X}$ is formed as

$$\hat{x}_j = \sum_{i=1}^{TN} \frac{exp(w_j(x_j - x_i) + b_j)}{\sum_{i=1}^{TN} exp(w_j(x_j - x_i) + b_j)}(x_j - x_i), \quad s.t.\ \boldsymbol{d}(x_i, x_j) = \|d_i - d_j\| \le \delta \tag{3}$$

where $\boldsymbol{d}(x_i, x_j)$ denotes the Euclidean distance between center points (d_i, d_j) of two persons' bounding boxes. The spatial constraint factor δ acts as a hyper-parameter.

3.3 Dual-branch Relation Reasoning

In order to capture the underlying relation for GAR, both appearance features and position information are used to construct the actor relation graph. It is necessary to model two relations due to their different semantic attributes. Formally, given the appearance features $\{x_i, x_j\}$ and the position coordinates of two persons, a graph of pairwise appearance-level relation can be denoted as $G = \{G_{i,j} \in R^1 | i, j =$

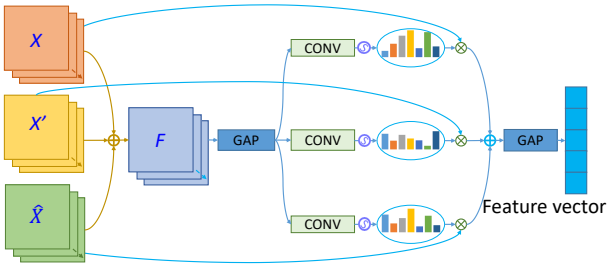

Figure 4: The spatiotemporal weighted aggregation. GAP and CONV mean the global average pooling, the convolutional layers, respectively.

$1, ..., TN\}$, where $G \in R^{TN \times TN}$. The relation value $G_{i,j}$ is expressed as

$$G_{i,j} = \frac{s(d_i, d_j)exp(a(x_i, x_j))}{\sum_{j=1}^{TN} s(d_i, d_j)exp(a(x_i, x_j))} \tag{4}$$

where the embedded dot-product and distance mask in [Wu *et al.*, 2019] are used to model the appearance relation value $a(x_i, x_j)$ and the position relation value $s(d_i, d_j)$, respectively. Intuitively, a soft-max function is used to perform normalization on each graph node. Then, a group of fully-connected graphs $\{G^1, ..., G^g, ..., G^{N_g}\}$ can be built according to Eqn. (4) with different parameters. After the graphs are built, the one-layer of GCNs is used to perform relational reasoning. The output of relation graphs is written as follows:

$$X' = \sum_{g=1}^{N_g} Relu(G^g X W_g) \tag{5}$$

where N_g denotes the number of graphs. $W_g \in R^{d \times d}$ is the learnable graph-specific weight matrix. Furthermore, LR^{2}M can be easily combined with the relation graph module to form the dual-branch relation reasoning.

3.4 Spatiotemporal Weighted Aggregation

Moreover, previous methods tend to directly fuse all features by using the element-wise sum function, which is also suboptimal to suppress the spatiotemporal redundancy with equal importance. Inspired by SKNet [Li *et al.*, 2019], the basic idea of WAS is to use three gates to selectively aggregate the information flows from different branches.

Given a set of $N \times T \times D$-dimensional multi-level features, our goal is to get refined features for GAR. As shown in Figure 4, the features from all branches are first fused by an element-wise summation: $F = \sum_{i=1}^{N_b} X^i$, where N_b denotes the number of branches. Then, the channel-wise statistic $\mathcal{R}$ is computed by simply using global average pooling and a fully convolution layer along the dimension of channels, which is written as follows:

$$\mathcal{R} = W_R(\frac{1}{N \times T} \sum_{i=1}^{N} \sum_{j=1}^{T} F(i, j, :)) \tag{6}$$

where $W_R \in R^{\frac{D}{L} \times D}$ is the learned parameters and L is a dimensional shrinkage coeffcient. Further, the channel-wise weights $\{W^i\}_{i=1}^{N_b}$ can be achieved by fully convolutional layers, followed by the softmax operators:

$$W^i = softmax(w^i \mathcal{R}) \tag{7}$$

where $w^i \in R^{D \times \frac{D}{L}}$ denotes the learnable parameters. The c-th dimension of refined features $\hat{F}$ is computed as

$$\hat{F}_c = \sum_{i=1}^{N_b} W_c^i X_c^i \tag{8}$$

where $W_c^i \in R$ and $X_c^i \in R^{N \times T}$ denote the weight and the specific features of the i-th branch along the c-th dimension of channels, respectively. In our model, $N_b = 3$. Specifically, $\{X^i\}_{i=1}^3$ is replaced by the multi-level features X, X' and $\hat{X}$. Finally, a global pooling operation is performed on $\hat{F}$ to get the group feature vector $\mathcal{F} \in R^D$. The proposed model can be trained in an end-to-end manner, where the cross-entropy loss acts as the loss function.

4 Experiments

4.1 Experiment Settings

Datasets. Two popular benchmarks (*Volleyball Dataset* (VD) [Ibrahim *et al.*, 2016] and *Collective Activity Dataset* (CAD) [Choi *et al.*, 2009]) are used to evaluate our proposed method, while the players' tracklets provided in [Bagautdinov *et al.*, 2017] are employed to perform feature extraction on the whole clips. In CAD, "walking" and "crossing" are merged as a new class of "moving" following [Yan *et al.*, 2020; Yuan *et al.*, 2021].

Evaluation metrics. Two metrics are used to evaluate our method on these datasets, which are *Multi-class Classification Accuracy* (MCA) and *Mean Per Class Accuracy* (MPCA) following previous methods [Yan *et al.*, 2020; Yuan *et al.*, 2021].

Implementation details. Our model is implemented based on Pytorch. Following ARG [Wu *et al.*, 2019], randomly sampling frames from a video clip are selected as the training samples on both two datasets, resulting in $T = 3$ frames as the input to our model. The same backbone network ResNet18 and VGG16 are used to extract 1024-dimensional appearance-level feature vector for each person with ground-truth bounding box. N_g is set to 16. The ADAM optimizer with different learning rates is used to learn the network parameters. The hyper-parameters for ADAM are set as $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. For the training of 40 epochs on VD, the initial learning rate is set to 1×10^{-4} with the dacay rate is 1/3 every 10 epochs. For the training of 30 epochs on CAD, the learning rates are set to 4×10^{-5} and 1×10^{-4} for ResNet18 and VGG16, respectively. The spatial constraint factors δ are set to 0.2, 0.3 of the image width in the training of VD and CAD, empirically. L is set to 16. The batch sizes are set to 2 on both datasets.

4.2 Ablation Studies

Baseline. The person-level features are fed into Actor Relation Graphs (ARG) [Wu *et al.*, 2019], following a global

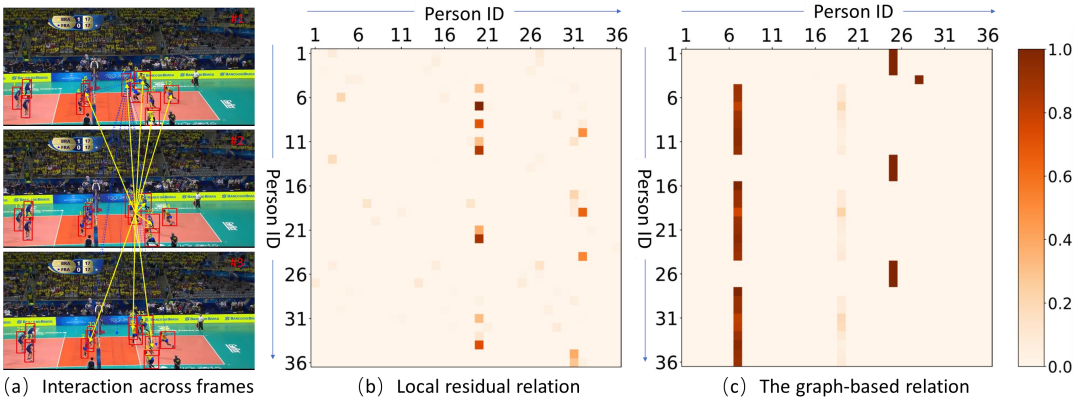

Figure 5: The relation visualizations of a right set activity example. In (a), the yellow solid lines and blue dotted lines show the local residual relation and graph-based relation, respectively. In (b) and (c), the horizontal and vertical axis represent the person's ID and relevant person.

Method	Backbone	MCA	MPCA
Baseline	ResNet18	89.53	90.13
	VGG16	92.74	92.80
LR^{2}M *w/o SC*	ResNet18	92.22	92.90
	VGG16	92.82	92.95
LR^{2}M	ResNet18	92.74	92.99
	VGG16	93.27	93.26
LR^{2}M *w/DIN*	ResNet18	92.52	92.91
	VGG16	94.02	94.17
LR^{2}M *w/TWA*	ResNet18	92.97	93.37
	VGG16	93.49	93.90
LR^{2}M *w/STWA*	ResNet18	92.00	92.42
	VGG16	93.19	93.30
LR^{2}M *w/WAS*	ResNet18	93.12	93.34
	VGG16	**94.54**	**94.96**

Table 1: Ablation results on VD by using different variants.

pooling layer to get the feature representation of a group activity. Finally, the softmax classifier is employed to predict the activity score [Yuan *et al.*, 2021].

Variants of LR^{2}M and WAS. (1) **LR^{2}M**: the appearance-level features are fed into the LR^{2}M module to capture the local differences among relevant persons (Eqn. 3); (2) **LR^{2}M *w/o SC*** : a standard LR^{2}M module without local spatial constraint is employed to learn the global differences among all persons (Eqn. 1); (3) **LR^{2}M *w/DIN*** : Unlike **LR^{2}M** with ARG, the relation graph follows DIN [Yuan *et al.*, 2021]; (4) **LR^{2}M *w/WAS*** : The proposed weighted aggregation strategy is employed to adaptively select important features along the spatial dimension; (5) **LR^{2}M *w/TWA***: Different from the aggregation direction of **LR^{2}M *w/ WAS***, the frame dimension is used as the aggregation direction to weight all the features; (6) **LR^{2}M *w/STWA*** : this variant involves stacked WAS and TWA, which further investigates the effect of WAS.

As shown in Table 1, it is observed that all models with LR^{2}M outperform the baseline method. MCA and MPCA of **LR^{2}M** yield 92.74% and 92.99% on the ResNet18 backbone, which perform better than **LR^{2}M *w/o SC***. These reflect that explicitly modeling the local differences brings significant performance improvement. Moreover, the space constraint is more stable to model the local differences. The results also show that only a few persons are crucial to inferring the group activity, which are consistent with the conclusion of previous work [Tang *et al.*, 2019]. Combining two existing relation graphs with LR^{2}M can steadily improve the performance. The superiority and robustness may be attributed to the LR^{2}M's ability to learn additional cues. With WAS on the VGG16 backbone, the performance can be further improved. The results indicate that WAS has the capability to weight spatial-temporal features for group activity recognition. It is noted that the **LR^{2}M *w/STWA*** scheme aggregates all features along both spatial and frame dimensions, resulting in the performance dropping dramatically. This observation indicates that the different features weighted along two dimensions will cause confusion for inference process. This may be caused by over-capturing a group of people in a frame while ignoring the information of individuals.

Visualization of LR^{2}M. Figure 5 (a) shows all person-specific interaction across frames. The captured local residual relation and the graph-based relation for a right set activity are shown in Figure 5 (b), (c), which are obtained by parsing two person-specific matrices from the model activation, respectively. Intuitively, the person with ID 7 is setting the ball, who attracts others' attention and builds the interaction between them. Be consistent with existing methods [Yan *et al.*, 2020], the graph-based relation module tends to associate this person with others. Different from this cue, the person with ID 20, who is watching the ball closely, forms another relatively stable relation with relevant people by assigning higher values, as shown in Figure 5 (b). The facts indicate that the relation formed by the local differences is significant in the right set group activity, not just the graph-based relation generated by the person performing setting action.

4.3 Comparison with The State-of-the-arts

For a fair comparison, the RGB video clips are adopted as input for all models and the results with two backbones are reported to demonstrate the effect of our proposed method.

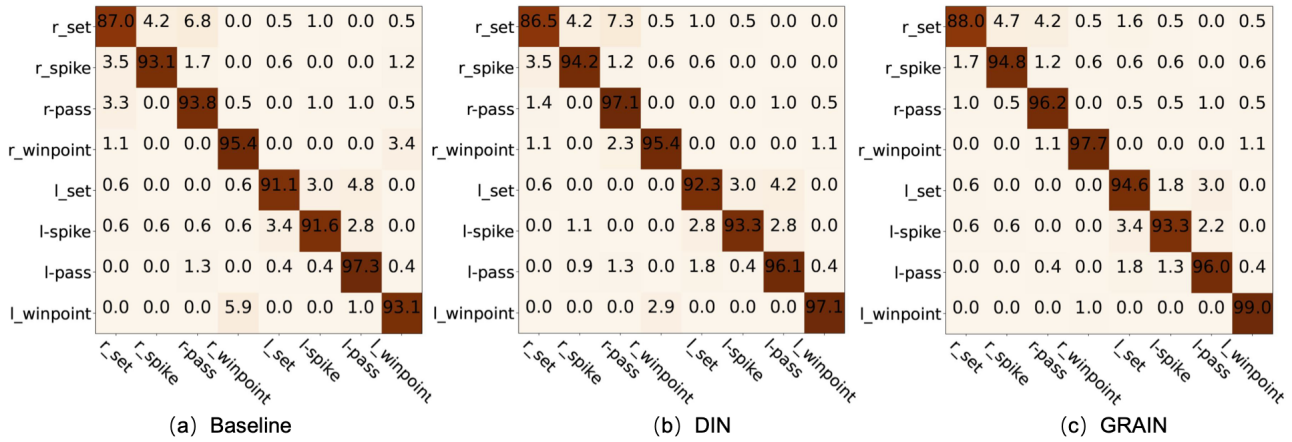

Figure 6: Comparison of the confusion matrices on the Volleyball Dataset. Notably, "l" and "r" are the abbreviations for "Left" and "Right" in the group activity labels. The backbone of three methods is VGG16.

Method	Backbone	MCA	MPCA
ARG [Wu *et al.*, 2019]	Inception-v3	92.5	-
	VGG16	91.9	-
stagNet [Qi *et al.*, 2020]	VGG16	89.3	-
PRL [Hu *et al.*, 2020]	VGG16	91.4	91.8
AT [Gavrilyuk *et al.*, 2020]	HRNet + I3D	94.4	-
HiGCIN [Yan *et al.*, 2020]	ResNet18	91.4	92.0
GLIL [Shu *et al.*, 2020]	Inception-v3	-	93.0
TS [Yuan and Ni, 2021]	Inception-v3	93.3	93.4
	VGG16	94.1	94.4
GFormer [Li *et al.*, 2021b]	Inception-v3	94.1	-
P^{2}CTDM [Yan *et al.*, 2021]	Inception-v3	91.8	92.7
	VGG16	91.5	91.8
DIN [Yuan *et al.*, 2021]	ResNet18	93.1	93.3
	VGG16	93.6	93.8
GRAIN	ResNet18	93.1	93.3
	VGG16	**94.5**	**95.0**

Table 2: Comparison with the state-of-the-art methods on VD.

Results on VD. Our method outperforms all of the aforementioned methods with a considerable margin as shown in Table 2. Although the accuracies of AT [Gavrilyuk *et al.*, 2020], TS [Yuan and Ni, 2021] and GFormer [Li *et al.*, 2021b] are over 94% (MCA), their main improvements are owed to the extra optical flow information or the well-designed transformer module. The results compared with DIN [Yuan *et al.*, 2021] indicate GRAIN is compatible with the spatio-temporal interaction graphs. The confusion matrices on VD are shown in Figure 6.

Results on CAD. Table 3 shows the comparison with different methods on CAD. Among these methods, most of them do not provide MCA and MPCA at the same time. As expected, our method outperforms the state-of-the-art methods. Moreover, GRAIN achieves impressive performance in class "waiting" due to its superiority in capturing the differences between pedestrians along the temporal variations.

Method	Backbone	MCA	MPCA
ARG [Wu *et al.*, 2019]	Inception-v3	91.0	-
	VGG16	90.1	-
stagNet [Qi *et al.*, 2020]	VGG16	89.1	-
PRL [Hu *et al.*, 2020]	VGG16	-	93.8
HiGCIN [Yan *et al.*, 2020]	ResNet18	93.4	93.0
GLIL [Shu *et al.*, 2020]	Inception-v3	-	94.9
TS [Yuan and Ni, 2021]	Inception-v3	-	95.1
	VGG16	-	95.4
GFormer [Li *et al.*, 2021b]	Inception-v3	93.6	-
P^{2}CTDM [Yan *et al.*, 2021]	Inception-v3	-	94.1
	VGG16	-	95.1
DIN [Yuan *et al.*, 2021]	ResNet18	-	95.3
	VGG16	-	95.9
GRAIN	ResNet18	**95.3**	96.1
	VGG16	95.2	**96.5**

Table 3: Comparison with the state-of-the-art methods on CAD.

5 Conclusion

In this paper, a graph-based residual aggregation network, termed GRAIN, is proposed to capture the local differences among relevant persons for inferring group activities. Specifically, the local residual relation module is employed to generate the difference-level features. The experiments demonstrate GRAIN outperforms the state-of-the-art methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No: 62001400, 61772436 and 61802053), Sichuan Science and Technology Program (Grant No. 2020YJ0207, 2020YJ0037 and 2021YJ0364), Foundation for Department of Transportation of Henan Province, China (2019J-2-2), Grant of Institute of Applied Physics and Computational Mathematics, Beijing (Grant No. HXO2020-118) and China Postdoctoral Science Foundation (Grant No. 2020M683353 and 2021M702713).

References

[Amer and Todorovic, 2015] Mohamed R Amer and Sinisa Todorovic. Sum product networks for activity recognition. *IEEE TPAMI*, 38(4):800–813, 2015.

[Arandjelovič *et al.*, 2018] Relja Arandjelovič, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE TPAMI*, 40(6):1437–1451, 2018.

[Bagautdinov *et al.*, 2017] 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*, pages 4315–4324, 2017.

[Choi *et al.*, 2009] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *ICCVW*, pages 1282–1289, 2009.

[Dang *et al.*, 2021] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Hierarchical object-oriented spatio-temporal reasoning for video question answering. In *IJCAI*, pages 636–642, 2021.

[Gavrilyuk *et al.*, 2020] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *CVPR*, pages 839–848, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hu *et al.*, 2020] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *CVPR*, pages 980–989, 2020.

[Ibrahim *et al.*, 2016] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, pages 1971–1980, 2016.

[Li *et al.*, 2019] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, pages 510–519, 2019.

[Li *et al.*, 2020] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *CVPR*, pages 909–918, 2020.

[Li *et al.*, 2021a] Dong Li, Zhaofan Qiu, Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Representing videos as discriminative sub-graphs for action recognition. In *CVPR*, pages 3310–3319, 2021.

[Li *et al.*, 2021b] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *ICCV*, pages 13668–13677, 2021.

[Liu *et al.*, 2021] Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *IJCAI*, pages 1743–1749, 2021.

[Masato *et al.*, 2011] Daniele Masato, Timothy J Norman, Wamberto W Vasconcelos, and Katia Sycara. Agent-oriented incremental team and activity recognition. In *IJCAI*, pages 1402–1407, 2011.

[Qi *et al.*, 2020] Mengshi Qi, Yunhong Wang, Jie Qin, Annan Li, Jiebo Luo, and Luc Van Gool. stagnet: an attentive semantic rnn for group activity and individual action recognition. *IEEE TCSVT*, 30(2):549–565, 2020.

[Shu *et al.*, 2015] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Chun Song Zhu. Joint inference of groups, events and human roles in aerial videos. In *CVPR*, pages 4576–4584, 2015.

[Shu *et al.*, 2020] Xiangbo Shu, Liyan Zhang, Yunlian Sun, and Jinhui Tang. Host–parasite: Graph lstm-in-lstm for group activity recognition. *IEEE TNNLS*, 32(2):663–674, 2020.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, pages 1–14, 2015.

[Tang *et al.*, 2019] Jinhui Tang, Xiangbo Shu, Rui Yan, and Liyan Zhang. Coherence constrained graph lstm for group activity recognition. *IEEE TPAMI*, pages 1–12, 2019.

[Wang and Wang, 2017] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *CVPR*, pages 499–508, 2017.

[Wang *et al.*, 2017] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *CVPR*, pages 3048–3056, 2017.

[Wu *et al.*, 2019] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, pages 9964–9974, 2019.

[Yan *et al.*, 2020] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: hierarchical graph-based cross inference network for group activity recognition. *IEEE TPAMI*, pages 1–14, 2020.

[Yan *et al.*, 2021] Rui Yan, Xiangbo Shu, Chengcheng Yuan, Qi Tian, and Jinhui Tang. Position-aware participation-contributed temporal dynamic model for group activity recognition. *IEEE TNNLS*, pages 1–15, 2021.

[Yuan and Ni, 2021] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *AAAI*, pages 3261–3269, 2021.

[Yuan *et al.*, 2021] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *ICCV*, pages 7476–7485, 2021.

[Zhou *et al.*, 2021] Jiaming Zhou, Kun-Yu Lin, Haoxin Li, and Wei-Shi Zheng. Graph-based high-order relation modeling for long-term action recognition. In *CVPR*, pages 8984–8993, 2021.