Group Activity Recognition via Spatio-Temporal Reasoning of Key Instances

Haoting He
haoting.he@stu.xjtu.edu.cn
Yaochen Li*
yaochenli@mail.xjtu.edu.cn
Yutong Wang
yutongwang@stu.xjtu.edu.cn
Gaojie Li
ligaojie@stu.xjtu.edu.cn
Wei Guo
xgwj3672518@stu.xjtu.edu.cn
Runlin Zou
zourunlin@stu.xjtu.edu.cn

Xi'an Jiaotong University
Xi'an, China

Abstract

The task of group activity recognition is to detect the group behavior performed by a group of people, and detecting the key actors and key frames is particularly important for judging group activity. Therefore, we propose a key instances based spatio-temporal reasoning model. The proposed key instance identification module can identify key roles and key frames from video sequences, and dynamically aggregate the features of related actors through a graph relationship reasoning model. Joint features and RGB features are extracted from the video sequence, and the two are fused through the proposed multi-modal fusion TCT module, which enhances the expressive ability of the original features. In order to infer group activity through spatio-temporal correlation, the improved cross-transformer module is further used to perform spatio-temporal synchronic reasoning on group activity from two dimensions: time and space. Experimental results demonstrate that our proposed method achieves high accuracy on two public general data sets, and outperforms most of state-of-the-art methods.

1 Introduction

The group activity recognition process includes object detection and object tracking, and uses the obtained tracking sequence to identify individual actions, and then uses individual

* Indicates corresponding author.

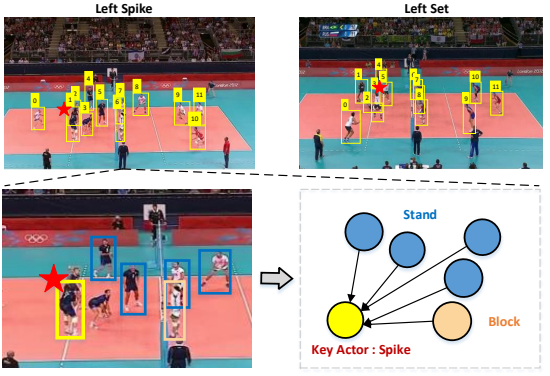

Figure 1: Our method detects the key instances which are denoted by red stars in the video. We build a relation reasoning graph to explicitly model the group activity.

characteristics and interactive information to reason about group activity [6]. The most critical task is how to design an ingenious reasoning module to analyze and obtain the group activity representation of video clips.

Group activity recognition problems often involve only a few key actors who dominate or determine the entire behavioral event [24, 35]. Not all actors contribute to event categories, there are some key actors that contribute more. In this work, we designed a spatio-temporal key instance recognition module to identify key frames in the time dimension and key actors in the spatial dimension. In order to calculate the group relevance score, the global group features need to be obtained. Direct pooling operations on individual features, especially average pooling or max pooling, tend to erase important spatial layout and contextual information in the image. In fact, the object position information and background information in the scene play an important role in identifying group activity [30]. For example, some scene information in the basketball dataset such as the position of the ball, basket, three-point line, and even background information are all important in recognizing group activity. Adding global scene features should be able to reflect the complex structure of the entire scene and introduce object features related to group activity. At the same time, adding global scene features can help improve the generalization ability of the model. By taking into account both global context and local details, the model can better adapt to changes in different scenarios, and can make reasonable judgments and predictions even when faced with unseen scenes.

Fusing joint features and RGB features can achieve complementary advantages [21]. It can not only use joint information to accurately capture the core features of actor actions, but also use RGB image information to supplement scene context and visual details. In this study, we design a TCT multi-modal fusion module to fuse joint features and RGB features. Combining the two features can increase the representation ability of the model, allowing the system to maintain good recognition results when faced with complex situations such as changes in lighting conditions, occlusions, and perspective changes.

The spatio-temporal features of group activity are complexly coupled. Sometimes the spatial apparent features contribute more to group activity, and sometimes the dynamic change features are more critical and accounting for more. To overcome the above challenge, we improve the spatio-temporal cross-transformer [17] to analyze and process the temporal and spatial features of group activity at the same time. In summary, our contributions include:

- We propose a spatio-temporal key instance identification module to identify key frames in the time dimension and key roles in the space dimension. Then graph relationship reasoning module is proposed to model the relationships between individuals.
- We improved the spatio-temporal cross-transformer to enhance the spatio-temporal expression ability of group features. And propose a multi-modal fusion module TCT to effectively fuse multi-modal features.

2 Related work

Individual Action Recognition. In individual action recognition, the majority of the networks utilize the CNN architechtures due to its effectiveness to extract meaningful RGB features. These architechtures can be divided into two categories: 2D CNN and 3D CNN networks. 2D CNNs can avoid huge computation cost. Lin *et al*. [19] proposed the temporal shift module to achieve temporal modeling. Simonyan *et al*. [28] proposed a two-stream CNN architechture for RGB and optical input. The above methods are more efficient compared to 3D CNN but cannot infer complicated temporal relationships. 3D convolutional neural networks can jointly learn spatio-temporal features. Tran *et al*. [32] introduced a 3D Convolutional Neural Network (CNN) derived from VGG architectures, called C3D, which is designed to extract spatio-temporal characteristics from videos. Carreira and Zisserman [3] expanded all 2D convolutional filters within an Inception-V1 model [29] into 3D convolutions. However, 3D CNNs are computationally intensive, which poses challenges for deployment.

Graph Neural Network. Graph neural network (GNN) has emerged as a popular technique due to their ability to extract information from graph-structured data, and are recently widely employed. For instance, graph convolutional network(GCN) [16] extends the traditional convolution operation from images to graph based data. Recurrent graph network [18, 25] employs a recurrent operation to model unseen graph patterns using spatio-temporal structural data. However, operations on graph data only process a local neighborhood, which can not capture long-term relations.

Group Activity Recognition. The current mainstream methods for group activity recognition tasks include RNN [2, 23, 38], attention mechanism [6, 21, 35], GNN [22, 33, 34, 36, 37], and Transformer [8, 9, 17, 31, 39]. These methods are not completely independent. For example, the attention mechanism can be effectively embedded into RNN or GNN. There are also traditional methods for group activity recognition, that is, manual feature methods [13], such as motion boundary histogram (MBH), histogram of gradients (HOG), and using Markov random fields to analyze the relationship between objects. However, hand-generated low-level features cannot represent complex group activities.

Most models proposed in academic papers are basically two-stage models [15].Wu *et al*. [33] models the relationship between people by constructing an actor relationship graph ARG, and continuously optimizes the graph adjacency matrix through graph convolution. SACRF [21] combines the traditional conditional random field method with the graph-based attention mechanism. It includes spatial attention and attempts to divide the entire graph into subgraphs of different sizes. Mahsa *et al*. [6] uses a graph attention network to encode relationship information between nodes. HiGCIN [36] proposes a universal cross inference block (CIB) based on graph structure to utilize the spatiotemporal dependencies hidden be-

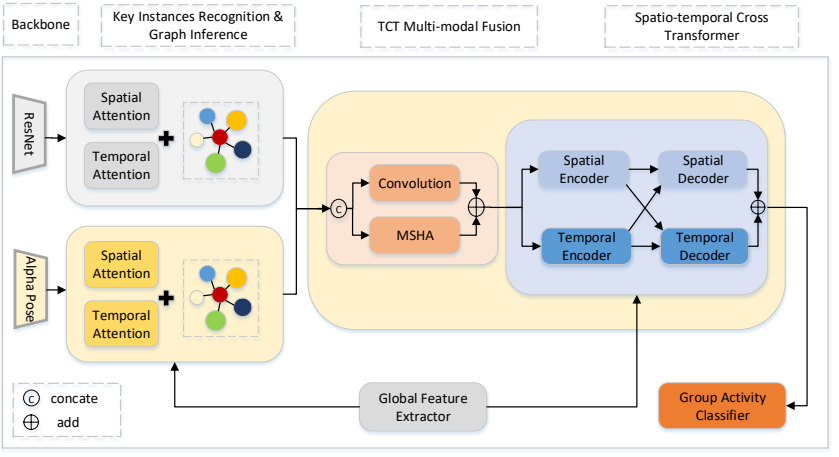

Figure 2: Key instances based spatio-temporal reasoning model. The model includes three submodules: key instance identification, TCT multi-modal fusion and spatio-temporal cross-transformer.

tween feature nodes, greatly reducing the computational complexity. DIN [37] proposes an adaptive dynamic graph reasoning network based on deformable convolution, which can establish adaptive graph structures more effectively. However, their methods can not identify key instances in the group activity and neglect the importance of scene information.

3 Proposed method

In this work, we propose a novel key instances based spatio-temporal reasoning network for group activity recognition. Fig.2 presents an overview of our network. We first use ResNet [10] to extract RGB features from image sequences and use Alpha Pose [7] to extract joint positions of detected individuals. At the first stage, key instance recognition module is proposed to indentify the key actors and key frames, then GCN is used to aggregate information between local neighbors. At the second stage, we use TCT multi-modal fusion module to fuse RGB features and pose features. And then spatio-temporal cross transformer is introduced to model long-term spatio-temporal relations. Due to space constraints, Graph relationship reasoning module and TCT multi-modal fusion module will be supplied as supplementary material. Next we demonstrate how we design the key instances recognition module.

3.1 Spatio-temporal Key Instance Recognition Module

The overall key instance identification module is shown in Fig.3. First, a global feature extractor is used to extract global scene information, and then the scene group features are obtained by concatenating it with the pooled features of the individual feature extracted from the backbone network in the channel dimension. Then, the spatio-temporal group attention mechanism is used to identify key roles and key frames based on the previously obtained group features of the scene. Afterwards, the graph relationship reasoning module is used to reason about the key roles in the spatial dimension. The inference output is residually

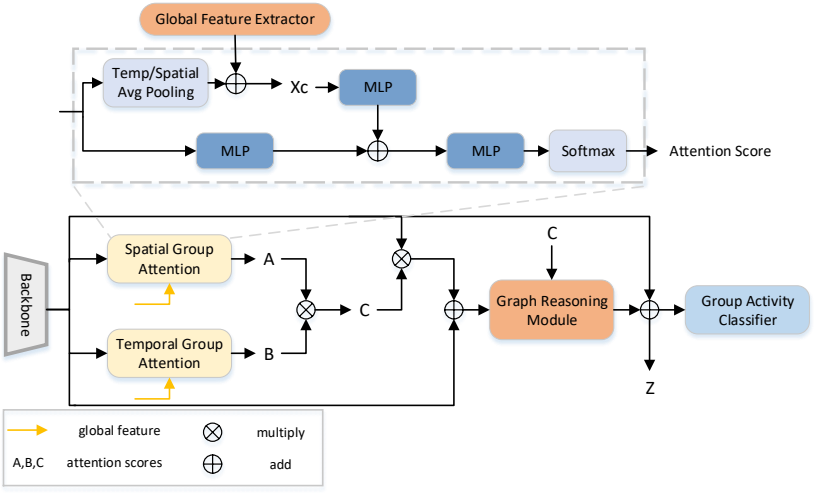

Figure 3: Key instance recognition module. We use temporal and spatial self-attention to obtain group relevance scores.

connected to the individual features extracted from the backbone network to avoid the loss of feature information caused by the graph relationship reasoning module.

In the group attention module as depicted in Fig.3, this study performs average pooling on individual features $X_{ind} \in \mathbb{R}^{T\times N\times C}$ obtained from the backbone network, followed by element-wise addition with scene tokens acquired from global feature extractor [26] to obtain scene collective feature X_c as shown in Eq.(1). Spatial average pooling is performed to obtain spatial scene group features, while average pooling in both time and space dimensions is conducted to derive video scene group features. The subsequent computation methods for these two distinct types of pooling are similar.

$$X_c = X_{scene} + \text{avg}(X_{ind}) \tag{1}$$

In group activity recognition, some people who have sudden and drastic changes in movement are often more semantically related to group activity. For example, people who suddenly spike the ball determine the overall group activity category. It is hypothesized that if a person's corresponding group attention score is larger, then the individual action will be more relevant to the group activity, and vice versa. First, the DBSCAN clustering algorithm [27] is used to divide the people in the scene into multiple groups, and spatio-temporal group attention is applied to each group to generate the respective group attention weights. Finally, group attention weights are used to enhance key features and suppress features that have a negative impact on group activity analysis. The spatio-temporal group attention module is shown in Fig.3. Assume that there are a total of N individuals in the scene, and these individuals can be divided into G groups through the clustering algorithm. Optimized features are derived through learning group relevance scores α_t^k. These coefficients indicate the degree of correlation between an individual and the group activity:

$$\alpha_t^k = \frac{\exp(e_t^k)}{\sum_{j=S_i}^{E_i}\exp(e_t^j)} \tag{2}$$

$$e_t^k = \text{Relu}(W_{\widehat{x}e}\widehat{x}_t^k + W_{fe}\widehat{x}_{tc} + b_e) \tag{3}$$

where k is the index of actor, $W_{\widehat{x}e}$ and W_{fe} are weight matrices, b_e is the bias vector. $\widehat{x}_{tc}$ is

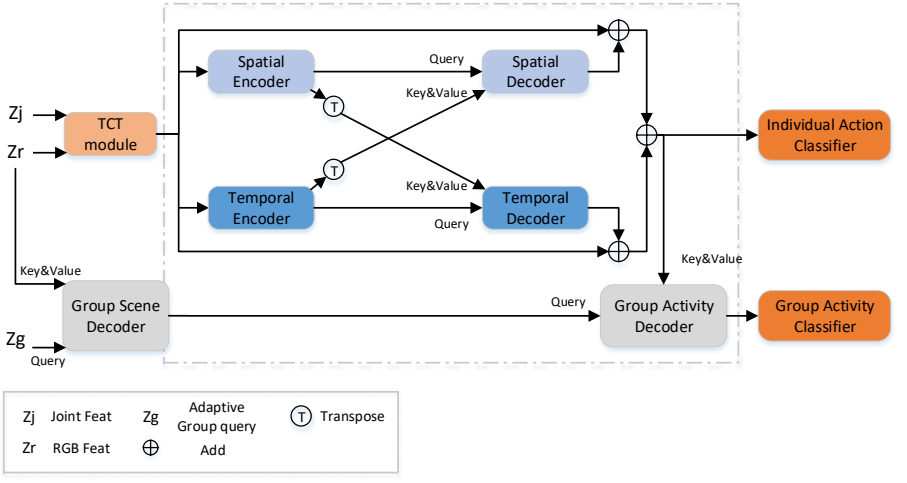

Figure 4: Spatio-temporal cross module. It can infer complex spatio-temporal relationship.

the previously obtained global scene feature of frame t.The group attention scores α_t^k which forme spatio attention matrix $A \in \mathbb{R}^{T\times N}$ enhance features relevant to group behavior while suppressing those unrelated to it. This module identifies key roles crucial for group activity recognition in the spatial dimension, and the principle is analogous when identifying key frames in the temporal dimension, and we obtain temporal attention vector $B \in \mathbb{R}^{T}$.

By combining the association strengths across both temporal and spatial dimensions, we can obtain the spatio-temporal group saliency scores $C = A * B \in \mathbb{R}^{T\times N}$, which are then normalized via softmax:

$$\widetilde{X} = [(1+\gamma^1)\hat{x}^1;(1+\gamma^2)\hat{x}^2;\ldots;(1+\gamma^{T\times N})\hat{x}^{T\times N}] \tag{4}$$

$$\gamma^k = \frac{\exp(C^k)}{\sum_{j=1}^{T\times N}\exp(C^j)} \tag{5}$$

where $k \in \{1,2,\ldots,T\times N\}$.

3.2 Spatio-temporal Cross Module Based on Multi-modal Fusion

Although the above key instance recognition module can perform relational reasoning to obtain the category of group activity, it cannot model the complex spatio-temporal relationships of group activity. Therefore, we propose a spatio-temporal cross-Transformer model based on the multi-modal fusion module TCT to address the issue of group feature lacking strong spatio-temporal representation capability. The overall processing flow of the spatio-temporal cross Transformer model is shown in Fig.4. The RGB modality and skeleton modality denoted as Z_r and Z_j , are fused using the TCT module, which is demonstrated in the supplementary material. For the fused feature output, temporal and spatial feature encoding is performed separately using encoders along both the temporal and spatial dimensions. After decoding separately in the temporal and spatial dimensions and performing residual connections, an element-wise sum is computed to yield the output Z_k. This output can serve as keys and values for the group activity decoder, through which the final group activity features can be queried. The query for the group activity decoder can be obtained through

the lower branch depicted in Fig.4. We employ adaptive group queries as the query for the group decoder, and utilize the output Z_r of the preceding RGB modality stage as keys and values to obtain group activity features. It is utilized as the final query for the group activity decoder, from which the group activity feature X_G with robust spatiotemporal expressiveness is obtained from the keys and values Z_k.

4 Experiments

4.1 Experimental Setups

Datasets. We conduct experiments on two commonly used group activity datasets: Volleyball dataset and Collective dataset.The volleyball dataset [12] comprises 55 volleyball videos collected from YouTube, divided into 4830 clips, among which 3493 are allocated for training and 1337 for testing. The videos include 8 group action labels: left team passing, right team passing, left team setting, right team setting, left team spiking, right team spiking, left team winpoint and right team winpoint. To obtain ground truth bounding boxes for the unannotated video frames, data provided by [2] is utilized.

The collective dataset [5] consists of 44 videos captured by low-resolution handheld cameras. There are 5 individual actions present: crossing, waiting, queuing, walking, and talking. Group labels for each scene are assigned based on the predominant individual actions of the majority of people.

Implementation Details. This study employs the Adam optimizer, an implementation of stochastic gradient descent, to train the network, with hyperparameters fixed at $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$. For the volleyball dataset, the network is trained for 30 epochs with a batch size of 2. The learning rate is set to 3×10^{-5} and reduced to 1×10^{-5} at the 11th epoch. For the collective dataset, training is conducted over 20 epochs with a batch size of 8 and a learning rate of 1×10^{-4}. The entire model's code is implemented using the PyTorch framework, running on an NVIDIA GeForce RTX 4090 GPU.

4.2 Comparison with State-of-the-art

This study compares our proposed model against previous state-of-the-art (SOTA) approaches. Results on the Volleyball dataset, as shown in Tab.1, reveal that for methods employing ResNet as the backbone network, our model outperforms the best-performing approach by 0.6 percentage points. On the Collective dataset, our model achieves SOTA performance, demonstrating a substantial improvement over the models utilizing ResNet as the backbone.

4.3 Ablation Study

This study proposes a key instance recognition module for identifying critical instances, such as key spatial actors and key temporal frames. An ablation experiment on the identification of key instances is first conducted in both temporal and spatial dimensions, as shown in Tab.2. The use of key frame recognition alone yields relatively poor results, indicating that key frame recognition contributes less to group behavior recognition compared to key actor recognition. However, the simultaneous employment of both key frames and key actors yields better performance than using key actors alone.

Method	Year	Backbone	Top1 Group Acc.	
			Volleyball Dataset	Collective Dataset
PCTDM[34]	2018	ResNet18	90.3	92.2
stagNet[23]	2019	VGG16	89.3	89.1
CRM[1]	2019	I3D	92.1	-
ARG[33]	2019	ResNet18	91.1	92.3
PRL[10]	2020	VGG16	91.4	93.8
SACRF[21]	2020	ResNet18	90.7	-
DIN[37]	2021	ResNet18	93.1	95.3
DFWSGAR[12]	2022	ResNet18	90.5	-
GIRN[20]	2022	OpenPose	92.2	-
HiGCIN[36]	2023	ResNet18	91.4	93.0
SPARTAN[1]	2023	ViT-base	92.9	-
Ours	2024	ResNet18	**93.7**	**95.6**

Table 1: Comparison with the state-of-the-art methods on the Volleyball dataset and Collective dataset.

Instance Type	Collective	Volleyball	
	Group Acc.	Group Acc.	Person Acc.
actor	93.2	91.3	80.9
frame	92.6	90.8	79.8
actor+frame	**93.5**	**91.6**	**81.4**

Table 2: Ablation study of key instance type.

Method	Collective	Volleyball	
	Group Acc.	Group Acc.	Person Acc.
avg pool	92.5	90.8	80.7
global feature	93.2	91.4	81.1
avg pool&global feature	**93.5**	**91.6**	**81.4**

Table 3: Ablation study of global group scene feature generation method.

Then we perform an ablation experiment on scene group features. As illustrated in Tab.3, the experiments are categorized into three types: The first approach uses average pooling to extract group features.The second employs a global feature extractor to generate scene tokens, which were subsequently used as scene group features. The third concatenates scene tokens with average pooled individual features along the feature channel dimension, thus constructing the scene group features. The utilization of average pooling produces inferior outcomes. This deficiency stems from the lack of attributes pertinent to group behavior, such as background elements and ball targets. The concatenation of global features with individually pooled individual features emerged as the most efficacious approach, generating highly robust scene group features. These enhanced features were particularly advantageous for harnessing group attention mechanisms to efficiently discern inter-group correlations. Finally, we conduct an ablation experiment on the above-mentioned key instances, graph reasoning, TCT, and cross-trans modules. As shown in Tab.4, incorporating the graph relation reasoning module has the most significant impact on improving model accuracy, contributing an increase of 0.8 percentage points. Moreover, removing this module exerts the greatest detrimental effect on the model, underscoring its crucial role in the process of inferring group activity. Concurrently, eliminating the key instances module incurs the least impact among all components, with a difference of merely 0.2 percentage points compared to the optimal combination. This may be attributed to the implicit analysis of key instances already embed-

Module Type				Volleyball	
Key Instance	Graph Reasoning	TCT	Cross Trans	Group Acc.	Person Acc.
√				91.6	81.6
√	√			92.4	82.5
√	√	√		93.1	82.6
	√	√	√	93.5	82.8
√		√	√	92.7	82.5
√	√		√	93.4	82.7
√	√	√	√	**93.7**	**83.1**

Table 4: Ablation study of various combinations of modules.

ded within the spatio-temporal graph relation reasoning module.

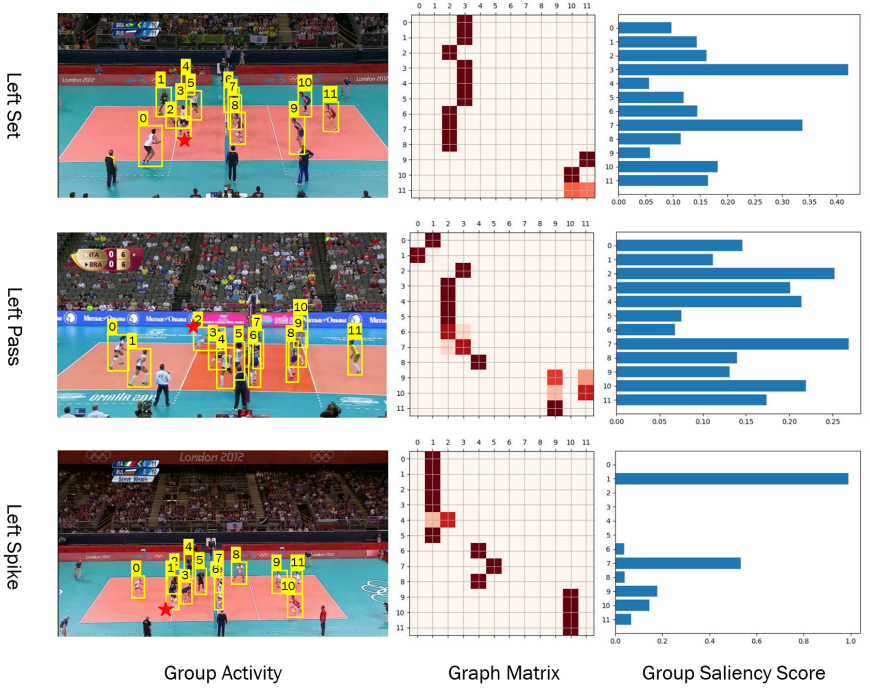

Figure 5: Visualization of key instance in the group activity, graph relation matrix and group saliency score.

4.4 Qualitative Results

Fig.5 visualizes the graph relation matrices from the model's graph reasoning module and the group saliency scores from the key instance recognition module. For each video, we select the graph relation matrix and group saliency scores corresponding to the middle frame for

visualization. It can be observed that columns in the graph relation matrix with the highest number of red blocks correspond to more critical individuals, indicating substantial associations between multiple other roles and the individual represented by that column. The group saliency scores further demonstrate the model's ability to effectively identify key characters within the scene, with these key roles determining the category of the group activity.

5 Conclusion

In summary, we have presented a novel spatio-temporal cross-inference model based on key instances. The key instance recognition module employs temporal group attention and spatial group attention to respectively identify key frames in the temporal dimension and key actors in the spatial dimension. We evaluate on two standard benchmarks (Volleyball and Collective dataset), and conduct thorough ablation studies to demonstrate the effectiveness of our model. Our model has achieved high precision in comparison with other common group activity recognition models on both datasets, meeting the practical requirements for group activity recognition.

Acknowledgement. This work was supported by Key Research and Development Foundationof Shaanxi Province under grant no. 2022GY-080, and Natural Science BasicResearch Plan in Shaanxi Province of China under grant no. 2022J0-631.

References

[1] S. Azar, M. Atigh, A. Nickabadi, and A. Alahi. Convolutional relational machine for group activity recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7884–7893, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00808. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00808.

[2] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3425–3434, 2017. doi: 10.1109/CVPR.2017.365.

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

[4] N. Chappa, P. Nguyen, A. H. Nelson, H. Seo, X. Li, P. Dobbs, and K. Luu. Spartan: Self-supervised spatiotemporal transformers approach to group activity recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5158–5168, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPRW59228.2023.00544. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW59228.2023.00544.

[5] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. pages 1282 – 1289, 11 2009. doi: 10.1109/ICCVW.2009.5457461.

[6] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 177–195, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58545-7.

[7] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, 2022.

[8] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-transformers for group activity recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 836–845, 2020. doi: 10.1109/CVPR42600.2020.00092.

[9] Mingfei Han, David Junhao Zhang, Yali Wang, Rui Yan, Lina Yao, Xiaojun Chang, and Yu Qiao. Dual-ai: Dual-path actor interaction learning for group activity recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2980–2989, 2022. doi: 10.1109/CVPR52688.2022.00300.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 977–986, 2020. doi: 10.1109/CVPR42600.2020.00106.

[12] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1971–1980, 2016. doi: 10.1109/CVPR.2016.217.

[13] Mihir Jain, Hervé Jégou, and Patrick Bouthemy. Better exploiting motion for better action recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2555–2562, 2013. doi: 10.1109/CVPR.2013.330.

[14] D. Kim, J. Lee, M. Cho, and S. Kwak. Detector-free weakly supervised group activity recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20051–20061, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01945. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01945.

[15] Donguk Kim, Sumin Lee, Sangmin Woo, Jinyoung Park, Muhammad Adi Nugroho, and Changick Kim. Multi-modal social group activity recognition in panoramic scene. In *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, 2023. doi: 10.1109/VCIP59821.2023.10402675.

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

[17] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13668–13677, 2021.

[18] Renjie Liao, Yujia Li, Yang Song, Shenlong Wang, Charlie Nash, William L. Hamilton, David Duvenaud, Raquel Urtasun, and Richard Zemel. Efficient graph generation with graph recurrent attention networks. In *NeurIPS*, 2019.

[19] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[20] Mauricio Perez, Jun Liu, and Alex C. Kot. Skeleton-based relational reasoning for group activity analysis. *Pattern Recognition*, 122:108360, February 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.108360. URL http://dx.doi.org/10.1016/j.patcog.2021.108360.

[21] Rizard Renanda Adhi Pramono, Yie Tarng Chen, and Wen Hsien Fang. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 71–90, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8.

[22] 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 Transactions on Circuits and Systems for Video Technology*, 30(2):549–565, 2020. doi: 10.1109/TCSVT.2019.2894161.

[23] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3043–3053, 2016. doi: 10.1109/CVPR.2016.332.

[24] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3043–3053, 2016. doi: 10.1109/CVPR.2016.332.

[25] Luana Ruiz, Fernando Gama, and Alejandro Ribeiro. Gated graph recurrent neural networks. *IEEE Transactions on Signal Processing*, 68:6303–6318, 2020. ISSN 1941-0476. doi: 10.1109/tsp.2020.3033962. URL http://dx.doi.org/10.1109/TSP.2020.3033962.

[26] Michael S. Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[27] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Trans. Database Syst.*, 42(3), jul 2017. ISSN 0362-5915. doi: 10.1145/3068335. URL https://doi.org/10.1145/3068335.

[28] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 568–576, Cambridge, MA, USA, 2014. MIT Press.

[29] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Los Alamitos, CA, USA, jun 2015. IEEE Computer Society. doi: 10.1109/CVPR.2015.7298594. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298594.

[30] Masato Tamura, Rahul Vishwakarma, and Ravigopal Vennelakanti. Hunting group clues with transformers for social group activity recognition. In *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, pages 19–35, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19771-0. doi: 10.1007/978-3-031-19772-7_2. URL https://doi.org/10.1007/978-3-031-19772-7_2.

[31] Masato Tamura, Rahul Vishwakarma, and Ravigopal Vennelakanti. Hunting group clues with transformers for social group activity recognition. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 19–35, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19772-7.

[32] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. doi: 10.1109/ICCV.2015.510.

[33] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9956–9966, 2019. URL https://api.semanticscholar.org/CorpusID:128358806.

[34] Rui Yan, Jinhui Tang, Xiangbo Shu, Zechao Li, and Qi Tian. Participation-contributed temporal dynamic model for group activity recognition. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 1292–1300, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356657. doi: 10.1145/3240508.3240572. URL https://doi.org/10.1145/3240508.3240572.

[35] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. *arXiv preprint arXiv:2007.09470*, 2020.

[36] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: Hierarchical graph-based cross inference network for group activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6955–6968, 2023. doi: 10.1109/TPAMI.2020.3034233.

[37] H. Yuan, D. Ni, and M. Wang. Spatio-temporal dynamic inference network for group activity recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7456–7465, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00738. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00738.

[38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. S. Torr. Conditional random fields as recurrent neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1529–1537, Los Alamitos, CA, USA, dec 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.179. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.179.

[39] Honglu Zhou, Asim Kadav, Aviv Shamsian, Shijie Geng, Farley Lai, Long Zhao, Ting Liu, Mubbasir Kapadia, and Hans Peter Graf. Composer: Compositional reasoning of group activity in videos with keypoint-only modality. *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022)*, 2022.