Group Activity Recognition via Spatio-Temporal Reasoning of Key Instances

Haoting He, Yaochen Li*, Yutong Wang, Gaojie Li, Wei Guo, Runlin Zou

School of Software Engineering, Xi'an Jiaotong University, China

*E-mail: yaochenli@mail.xjtu.edu.cn

Introduction

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. The extensive experimental results well demonstrate the effectiveness of the proposed method.

Methods

The overall architecture of the spatio-temporal reasoning model is shown in Figure 1.

➢ Use ResNet to extract RGB features and Alpha Pose for joint positions. In the first stage, propose the key instance recognition module for identifying key actors and frames, then use GCN for local neighbor aggregation. In the second stage, use TCT fusion module to combine RGB and pose features, followed by spatio-temporal cross transformer to model long-term spatio-temporal relations.

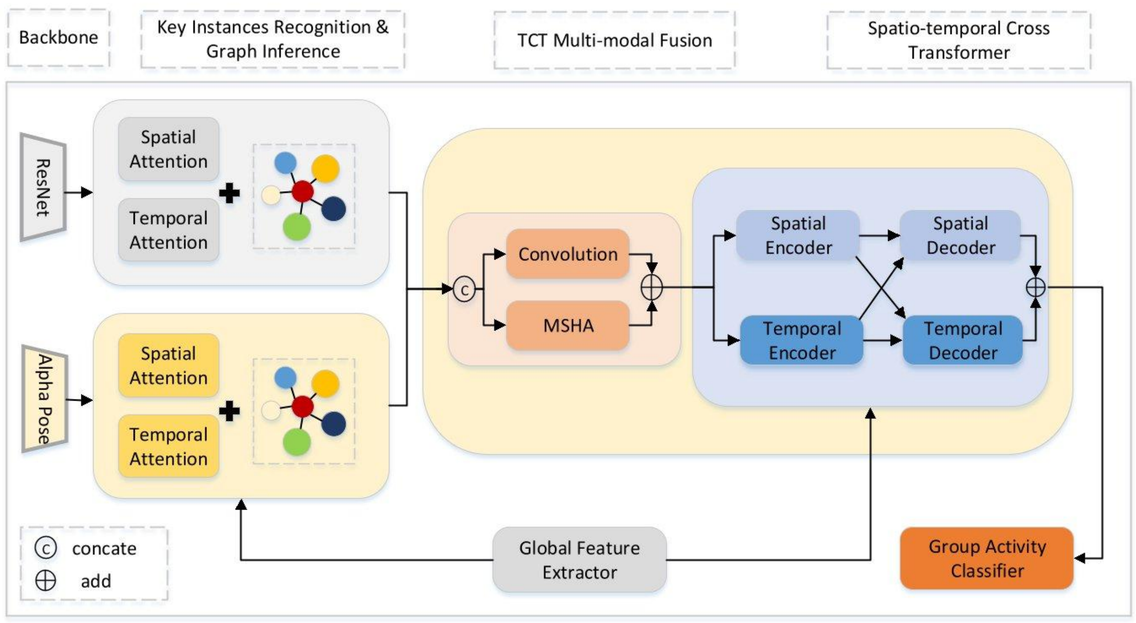

Figure 1. Overall architecture of the spatio-temporal reasoning model.

The overall key instance identification module is shown in Figure 2.

➢ First, use the global feature extractor to extract global scene information, then obtain the scene group features by concatenating it with the pooled features of the individual features extracted from the backbone in the channel dimension. Then, use the spatio-temporal group attention mechanism to identify key roles and frames based on the scene group features. Afterwards, use the graph relationship reasoning module to reason about key roles in the spatial dimension.

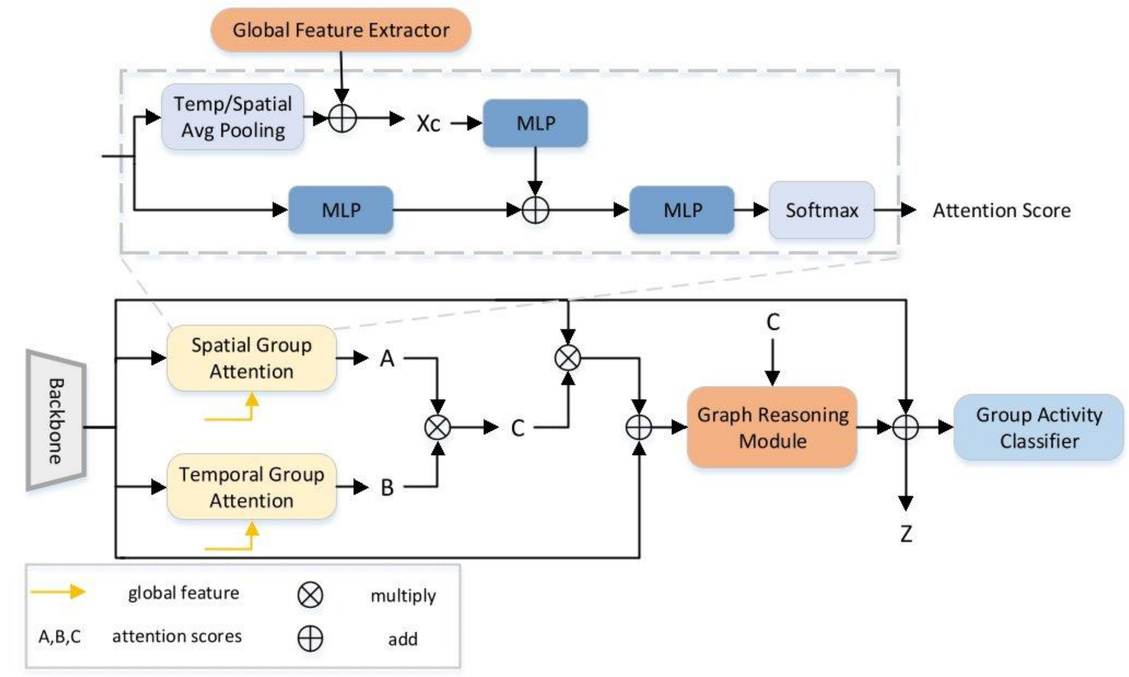

Figure 2. Key instance recognition module.

The overall architecture of the spatio-temporal cross Transformer model is shown in Figure 3.

➢ After decoding in temporal and spatial dimensions and performing residual connections, compute an element-wise sum to yield output Z_K, which serves as keys and values for the group activity decoder to query final features. Adaptive group queries are used as the query for the group decoder, utilizing Z_r from the RGB modality as keys and values to obtain group activity features. This forms the final query for the group activity decoder, deriving the feature X_G with robust spatiotemporal expressiveness.

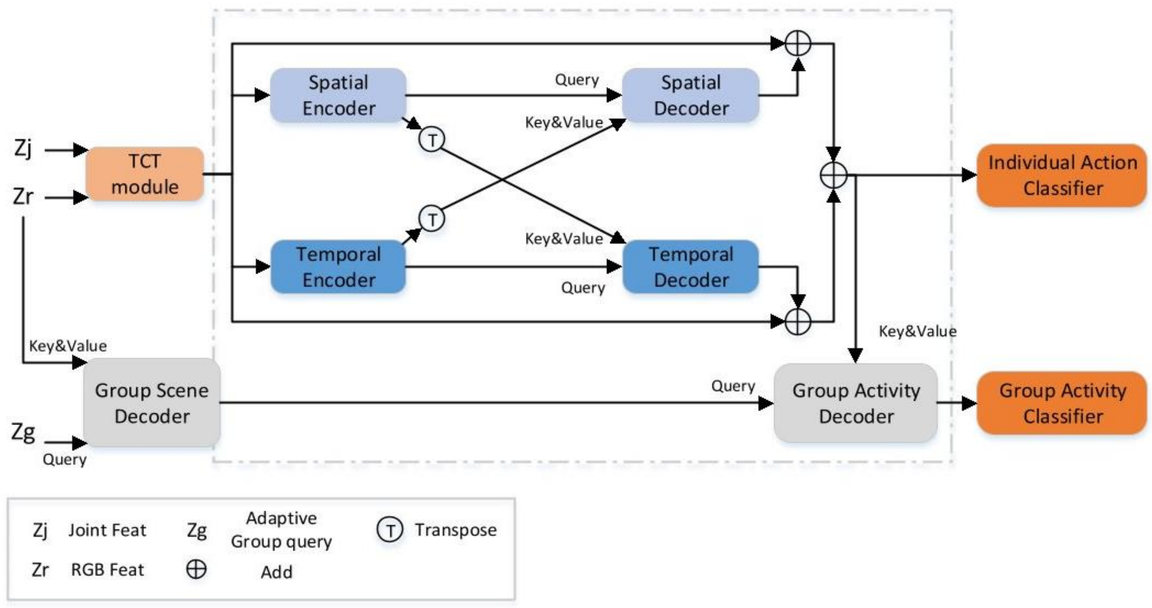

Figure 3. Overall architecture of the spatio-temporal cross Transformer model

Results

This study compares our proposed model against previous state-of-the-art (SOTA) approaches. Results on the Volleyball dataset and Collective dataset, as shown in Table1. Our method achieves the best performance on all tasks.

Method	Year	Backbone	Top1 Group Acc.	
			Volleyball Dataset	Collective Dataset
PCTDM	2018	ResNet18	90.3	92.2
stagNet	2019	VGG16	89.3	89.1
CRM	2019	I3D	92.1	-
ARG	2019	ResNet18	91.1	92.3
PRL	2020	VGG16	91.4	93.8
SACRF	2020	ResNet18	90.7	-
DIN	2021	ResNet18	93.1	95.3
DFWSGAR	2022	ResNet18	90.5	-
GIRN	2022	OpenPose	92.2	-
HiGCIN	2023	ResNet18	91.4	93.0
SPARTAN	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.

Figure 4.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.

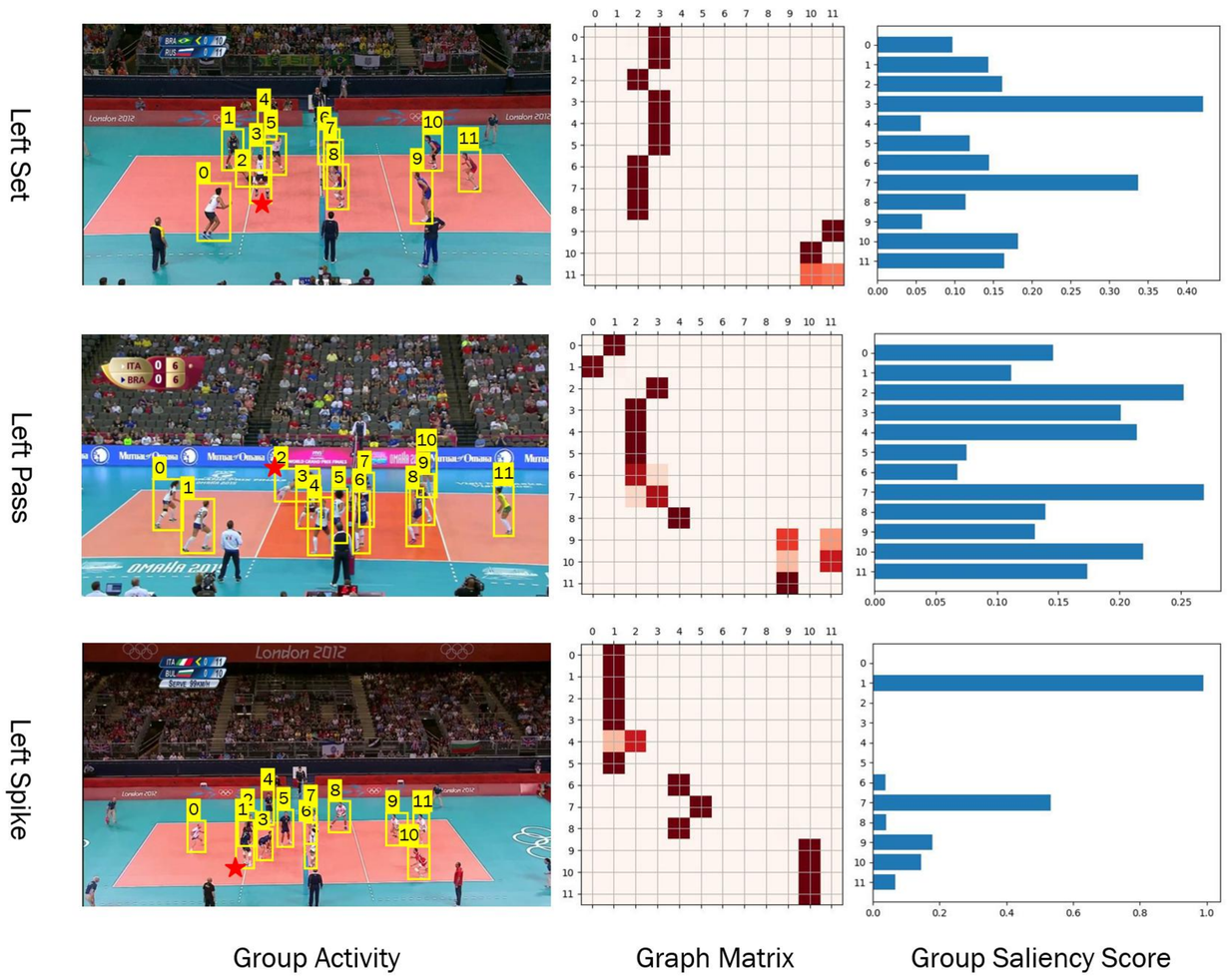

Figure 4. Visualization of key instance in the group activity.

Conclusion

In this work, 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.

BMVC 2024