# Dual-AI: Dual-path Actor Interaction Learning for Group Activity Recognition (Supplementary Material)

Mingfei Han*[1], David Junhao Zhang*[2], Yali Wang*[3], Rui Yan[2], Lina Yao[5],
Xiaojun Chang[1,4], Yu Qiao†[3,6]
[1]ReLER, AAII, UTS [2]National University of Singapore
[3]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences [4]RMIT University
[5]University of New South Wales [6]Shanghai AI Laboratory, Shanghai, China
hmf282@gmail.com; xiaojun.chang@uts.edu.au {yl.wang, yu.qiao}@siat.ac.cn

## 1. Implementation Details

For Vollyball and Weak-Volleyball-M, we randomly select $K = 3$ frames with 720×1280 resolution for training and 9 frames for testing, corresponding to 4 frames before the middle frame and 4 frames after. For Collective Activity dataset, we utilize $K = 10$ frames (480×720) of each video clip for training and testing. For NBA dataset, we select $K = 3$ frames (720×1280) around middle frame of each video for training and take 20 frames for testing. For Volleyball and Collective Activity dataset, we use annotated bounding boxes provided by the datasets for training and testing to make fair comparison, *i.e.*, $N = 12$ and $N = 13$ respectively.

**Optimization.** We adopt Adam [6] to learn the network parameters with initial learning rate set to 0.0001. We run 140 epochs in total to obtain the reported results and decay the learning rate by 10 after 60 and 100 epochs. We implement our method based on the released code[1] of [9] and transformer code[2] from Pytorch.

**Weakly Supervised GAR.** We detect actors in the centered frame of each clip with MMDetection Toolbox [1]. Following [12], we further obtain the tracklets by correlation tracker [2] implemented by Dlib [5]. Specificly, we use a Faster-RCNN [8] with ResNet-50 [3] backbone provided by MMDetection, which is pretrained on COCO dataset [7] and further finetuned with person subset of COCO. We use the default configuration provided and sort the detected boxes by the confidence scores. Finally, we select the top $N$ (16 for NBA Dataset and 20 for Weak-Volleyball-M) bounding boxes for actor interaction learning.

**Limited GAR.** We randomly select videos in Volleyball dataset for limited setting. The video ids used for 5%, 10%, 25% and 50% data in our experiments are (1, 38), (1, 23, 38, 54), (1, 6, 10, 15, 18, 23, 32, 38, 42, 48) and (1, 3, 6, 7, 10, 13, 15, 16, 18, 22, 23, 31, 32, 36, 38, 39, 40, 41, 42, 48), respectively. We implement results of other methods with the released code and configurations from [13].

## 2. Dataset

We provide more details of the datasets used for the convenience of result production.

**Volleyball Dataset.** Each clip is annotated with one of 8 group activity classes: *right set, right spike, right pass, right win-point, left set, left spike, left pass and left win-point*. Middle frame of each clip is annotated with 9 individual action labels (*waiting, setting, digging, falling, spiking, blocking, jumping, moving and standing*) and their bounding boxes.

**Collective Activity Dataset.** The individual is annotated with one of the following 6 action categories (*NA, crossing, waiting, queuing, walking and talking*) and their bounding boxes. We follow [10, 11, 13] to merge the *crossing* and *walking* into *moving*. Train-test split follows [13].

**Weak-Volleyball-M.** No standalone distribution is released. This dataset [12] is adapted from Volleyball dataset [4], merging *pass* and *set* into one category and discarding all individual annotations.

**NBA Dataset.** This dataset is available upon request[3], limited by the copyright. Each video clip is annotated with one of the 9 group activities: *2p-succ., 2p-fail.-off., 2p-fail.-def., 2p-layup-succ., 2p-layup-fail.-off., 2p-layup-fail.-def., 3p-succ., 3p-fail.-off., 3p-fail.-def.*. No individual annotations, such as individual action labels and bounding boxes,

* Equal contribution. † Corresponding author.
[1]https://github.com/wjchaoGit/Group-Activity-Recognition
[2]v1.6.0/torch/nn/modules/transformer.py
[3]https://ruiyan1995.github.io/SAM.html

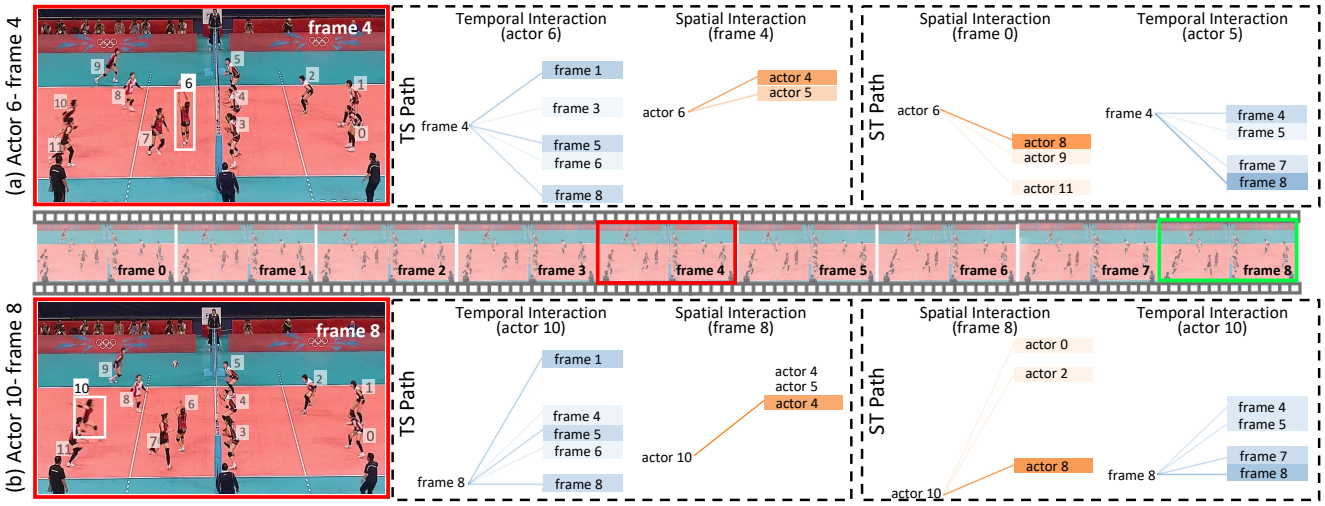


Figure 1. Actor interaction visualization for *l-set* activity with connected lines. Brighter color indicates stronger relation. (a) For actor 6 in frame 4, we visualize the temporal interaction with same actors in different frames for ST and TS paths; similarly, we visualize the spatial interaction with different actors in frame 4. (b) We visualize the actor interaction for actor 10 in frame 8 in the same way.

are provided.

## 3. Visualization

We provide visualization of actor interactions for *l-set*, as shown in Fig. 1. The attention weight between actors is represented by connected lines, and the brightness of the lines represents the scale of the attention weight. Orange and Blue lines correspond to the Spatial and Temporal interaction, respectively.

As shown by spatial interaction in Fig. 1 (a), the player setting the ball (actor 6) is more related with defending players in TS path, who are "jumping" and "blocking" (actor 4 and actor 5). Differently, in ST path, actor 6 has wider connections with accompanying players, who are "moving" (actor 8 and actor 9) and "jumping" (actor 10) cooperatively to tackle the ball falling on different position and prepare for next ball contact. Similarly, as shown by spatial interaction in Fig. 1 (b), the actor 10 is related to defending player (actor 4) in TS path, and related to both accompanying player (actor 8) and defending players (actor 0 and actor 2), showing complementary patterns. As for temporal interaction in both (a) and (b), the anchor actor is more evenly related to other frames (frame 1, frame 5 and frame 8) in TS path, and more related to late frames (frame 7 and frame 8) in ST path, which shows different evolution patterns.

## References


[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[2] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference, Nottingham, September 1-5, 2014*. Bmva Press, 2014. 1

[3] 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. 1

[4] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1980, 2016. 1

[5] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 1

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 1

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1

[9] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recog-

nition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9964–9974, 2019. 1

[10] 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*, pages 1292–1300, 2018. 1

[11] 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*, 2020. 1

[12] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020. 1

[13] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proceedings of the IEEE international conference on computer vision*, 2021. 1