Appendix of *SportsHHI: A Dataset for Human-Human Interaction Detection in Sports Videos*

Tao Wu[1,*] Runyu He[1,*] Gangshan Wu[1] Limin Wang[1,2,✉]
[1]State Key Laboratory for Novel Software Technology, Nanjing University [2]Shanghai AI Lab
{wt,runyu_he}@smail.nju.edu.cn, {gswu, lmwang}@nju.edu.cn

1. More details about SportsHHI dataset

1.1. Full interaction vocabulary

We provide full interaction vocabulary of SportsHHI dataset in Table 1. For comparison, we also provide the full relationship vocabulary of the popular AG dataset [4] for video scene graph generation (VSGG) in Table 2. Our defined interaction classes are of high-level semantics, including technical action, tactical coordination, or confrontation while AG deals with low-level spatial relation or simple atomic action. In video scene graph generation [9, 12, 13], the appearance features of the subject and object can often provide enough cues for relation inference [1, 5, 8, 11, 15, 16]. For example, given that the subject is a person and the object is a clothes, it is highly possible the relation between them is "wearing". However, in human-human interaction detection, the subject and object are both person. Such prior information cannot be used for interaction recognition, and generally, it requires action modeling, relative position encoding and spatiotemporal context modeling.

1.2. Statistics of each sport

We provide statistics of each sport in Table 3. The total number of instances of basketball and volleyball is close. Keyframe interaction instance distribution in basketball is more sparse than in volleyball because basketball videos have longer plain segments of dribbling. Both basketball and volleyball share the characteristics of crowd multi-person scenarios and relatively sparse interaction instance distribution, which requires the methods to distinguish two people without interaction from real interaction instances.

1.3. Statistics of partially invisible instances

For an interaction instance $\langle S, I, O \rangle$, when the subject person or the object person is out of view, we annotate its S or O as "invisible". This occurs due to two main reasons: camera angle switch and fast movement of the people. Table 4 presents statistics of partially invisible interaction instances. In basketball, all partially invisible instances are from the interaction class of *pass-catch*. This is because athletes move quickly during the transition between defense and offense, and sometimes the camera cannot fully capture the "pass-catch" process. In volleyball, partially invisible instances mainly occur in the interaction classes of *serve-*

*: Equal contribution. ✉: Corresponding author.

Basketball	**Volleyball**
jump ball	serve - first pass
pass - catch	co- first pass
drive - defend	first pass - second pass
block - shot	first pass - second attack
interfere - shot	co- attack
pass steal - pass	second pass - attack
dribble steal - dribble	cover attack
dribble - defend	attack - block
dribble - sag	co- block
defend - sag	attack - protect
(with ball) pick-and-roll - defender	co- protect
(with ball) pick-and-roll - teammate	block back - protect
(no ball) pick-and-roll - defender	protect - second pass
(no ball) pick-and-roll - teammate	protect - second attack
pass inbound - catch	attack - defend
close defense	co- defend
	defend - second pass
	defend - second attack

Table 1. **Interaction vocabulary of SportsHHI**

attention	**spatial**	**contact**	
looking at	in front of	carrying	covered vy
not looking at	behind	drinking from	eating
unsure	on the side of	leaning on	holding
	above	have it on the back	lying on
	beneath	not contacting	sitting on
	in	standing on	touching
		twisting	wearing
		wiping	writing on

Table 2. **Relationship vocabulary of AG**

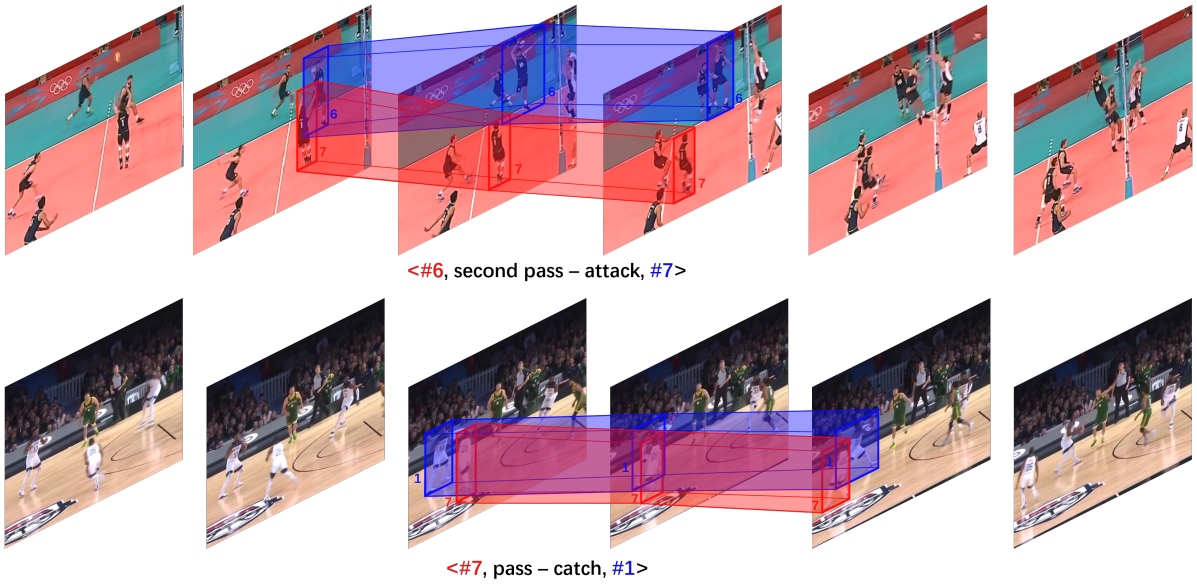


Figure 1. **Tube-level interaction instances.** We generate tube-level interaction instances by linking the same pair of people with the same interaction type across adjacent keyframes. The subjects are displayed in red and the objects in blue.

	#keyframes	#interact.	#inst.	#hum. bbox	avg. hum.
Basketball	6791	16	22455	64549	9.51
Volleyball	4607	18	28194	53526	11.62
SportsHHI	11398	34	50649	118075	10.36

Table 3. **Statistics of SportsHHI**

interact. class	#inv. ins.	#ins.	class %	all %
pass - catch	9	1894	0.48	0.04
others	0	0	0	0

(a) Basketball

interact. class	#inv. ins.	#ins.	class %	all %
serve - first pass	13	225	5.78	0.05
co- first pass	52	1464	3.55	0.18
others	5	-	-	0.02

(b) Volleyball

Table 4. **Statistics of partially invisible instances.** The table displays the count and percentage of partially invisible instances for each interaction class in the given sport. The percentage is calculated based on the total number of instances in that class as well as all instances in the sport.

first pass and *co-first pass*. This is because sometimes the camera switches its focus from the serving player to capturing the entire court during the serve. The percentage of partially invisible instances is low because the high-level semantics we focus on are often also the focus of the view and can attract the camera's attention the most.

1.4. Tube-level interaction instance generation

We annotate interaction instances on the keyframes at 5FPS. Person id tracking is available, making it simple to create tube-level interaction instances by linking the same pair of individuals with the same interaction type across adjacent keyframes. Temporal boundaries can be provided at a granularity of 5 frames. Figure 1 illustrates two tube-level interaction instances generated using this approach. Most existing video visual relation detection benchmarks and methods focus on frame-level instance detection. We follow their lead and define interaction instances at the frame level. However, our SportsHHI can be readily adapted to tube-level video visual relation detection, which could potentially become a new research trend in the future.

2. More discussion about the baseline method

2.1. Comparisons with current Video VRD methods

When designing the baseline method, we followed some practices of the current video scene graph generation and video human-object interaction detection methods, such as relative position encoding. However, there are still three major differences between current video visual relation detection (Video VRD) methods [1, 5, 8, 11, 15, 16] and our baseline method: 1) They rely on appearance features extracted by image object detector and overlook motion modeling of the subject and object while we adopt 3D backbone for better modeling of each person's action. The semantic level of relation classes defined by previous datasets is relatively low and the appearance feature is often sufficient

Training	Validation	HHICls					HHIDet				
		mAP	R@150	R@100	R@50	R@20	mAP	R@150	R@100	R@50	R@20
Basketball	Basketball	3.21	95.10	90.28	66.42	25.21	0.99	77.40	67.36	46.92	17.63
Volleyball	Volleyball	15.65	83.52	74.85	58.11	34.06	8.09	65.18	54.04	36.21	20.35
SportsHHI	Basketball	3.52	94.94	91.44	78.16	52.97	1.39	79.80	71.35	53.35	29.22
SportsHHI	Volleyball	15.82	84.23	75.41	59.05	34.00	7.65	65.52	53.61	34.29	18.25
SportsHHI	SportsHHI	10.69	89.25	82.93	68.13	43.72	4.93	72.22	61.92	42.99	23.89

Table 5. **HHICls and HHIDet results.** We show the results of training the model on the basketball or volleyball part of the SportsHHI training set and validating on the corresponding part of the validation set. We further show the results of training the model on the whole training set and validating on the basketball part, volleyball part, and whole validation set. ViT-B backbone is used.

for recognition, such as $\langle dog, larger, frisbee \rangle$. However, action modeling is very important for interaction recognition on SportsHHI. For example, to distinguish between *defend - second pass* and *defend - second attack*, we need a good modeling of the object person's action to distinguish whether he is passing the ball or attacking. When we replace the motion features in interaction representation with appearance features, the performance drops significantly. 2) They are dependent on the accuracy of the image object detector in identifying object categories. For example, STTran [1] adds the category embedding of the subject and object to the relation representation, which provides strong prior information. For instance, knowing that the subject and object are *human* and *horse* respectively, the relation category is very likely to be *ride*. However, SportsHHI does not have such a priori as all subjects and objects are humans. 3) Current Video VRD methods tend to treat different relation instances as independent individuals, while our baseline method exchanges information among interaction instances. In SportsHHI, sometimes, recognizing an interaction requires information from other interaction instances. For example, to recognize an interaction of *co-defend* in volleyball, we need to know there exists an interaction of *attack-defend*.

2.2. Comparisons with action detection methods

Current action detection methods [10, 14, 19, 21] typically adopt the two-stage detection paradigm. Proposal person bounding boxes are generated in the first stage and RoI features are extracted for each proposal for action classification. Our baseline method for human-human interaction detection follows the two-stage pipeline. Unlike action detection methods, the RoI feature is insufficient for classification in the second stage. Experimental results show that context information, relative position encoding, and information exchange among the proposals are necessary for interaction recognition. Some methods model the interaction between people through attention mechanisms to improve the accuracy of action recognition of each individual person. However, the interaction modeling is implicitly performed as no interaction annotation is provided for supervision. The quality of interaction modeling can only be indirectly evaluated through the accuracy of action classification. In contrast, our baseline method deals with explicit human-human interaction modeling and prediction. The performance can be directly evaluated on the SportsHHI dataset.

2.3. Comparisons with action recognition methods

Some methods for action recognition [18] or group action recognition [2, 3, 17, 20] implicitly model the relations among people or objects to improve the action classification accuracy. For example, Wang et al. [18] adopt a GCN to implicitly model relations among all people and objects in the video. Each node in their GCN is the appearance feature of a person or object and the nodes are connected according to similarity and adjacency. By performing graph convolution on the GCN, the information from all people and objects is gradually aggregated, which benefits the classification of the video. The goal of our baseline methods is different. Our method aims to identify whether there is an interaction between each pair of people and recognize the type of interaction. We adopt a Transformer to exchange information among interaction proposal representations to assist the recognition of each single proposal rather than aggregate global information for video-level classification.

3. More experimental results

HHICls and HHIDet results on each sport. In Table 5, we show the results of HHICls and HHIDet on each sport. We first show the result of training and validation on the basketball and volleyball parts of SportHHI respectively. Then we show the results of training on SportsHHI and validation on basketball, volleyball, and both. Using ground-truth human detection results, mAP and Recall of HHICls are higher than HHIDet with a large margin, which indicates the human detection results and the quality of interaction proposals have a significant influence on performance. Overall, our baseline method performs better on volleyball than basketball, because in basketball videos, the interaction patterns are more complex, and the interaction categories are more unbalanced. In the HHICls mode, compared with training the model on the subset of each sport, training on the whole training set of SportsHHI brings validation per-

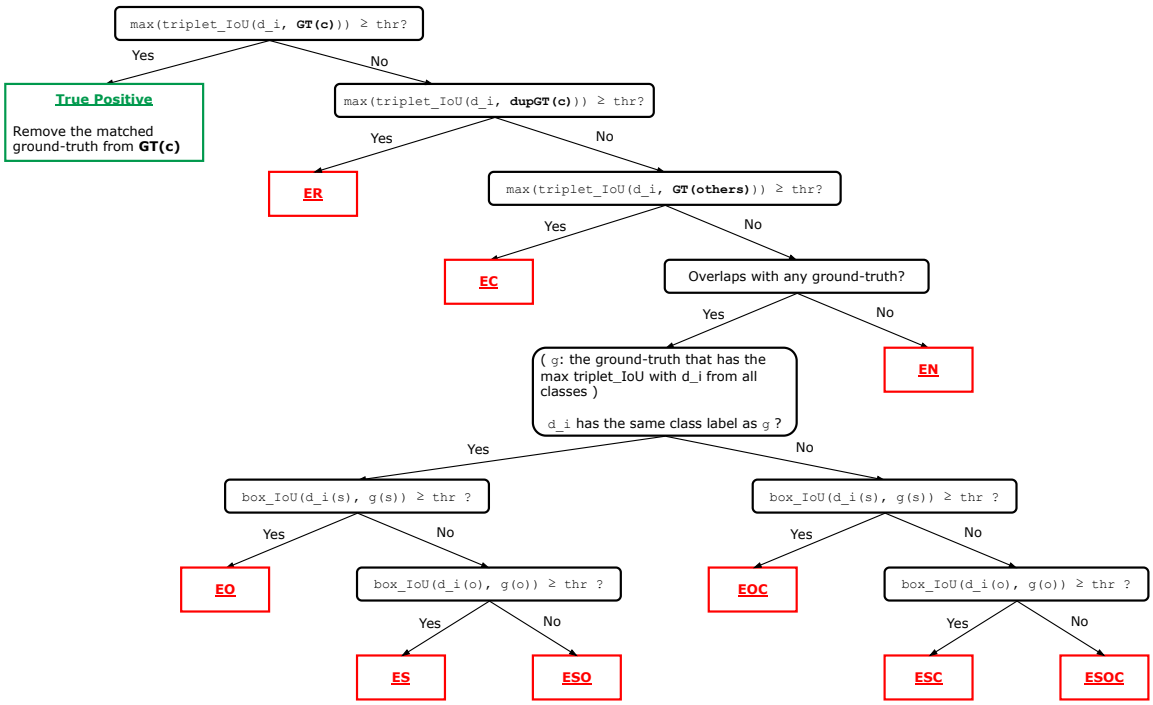


Figure 2. **Error analysis tree.** For each detected triplet `d_i` from a sorted list by descending order of confidence score of class `c`, `d_i(s)` and `d_i(o)` are the subject and object respectively. `g(s)` and `g(o)` are the subject and object for ground truth `g`. `box_IoU` is the traditional IoU score for a pair of boxes. `triplet_IoU` is the minimum of `box_IoU` between subject boxes and object boxes. **`GT(c)`** is the set of ground truths of class `c`. **`dupGT(c)`** is the original copy of **`GT(c)`** and will not change during the error classification process.

formance improvements on both basketball and volleyball. However, in the HHIDet mode, the improvement is not so stable. We speculate that, in the HHICls mode, due to the high quality of interaction proposals, through joint training, the model can learn more general representations. However, in the HHDet mode, low-quality proposals already introduce a lot of noise, and joint training further amplifies the impact of noise.

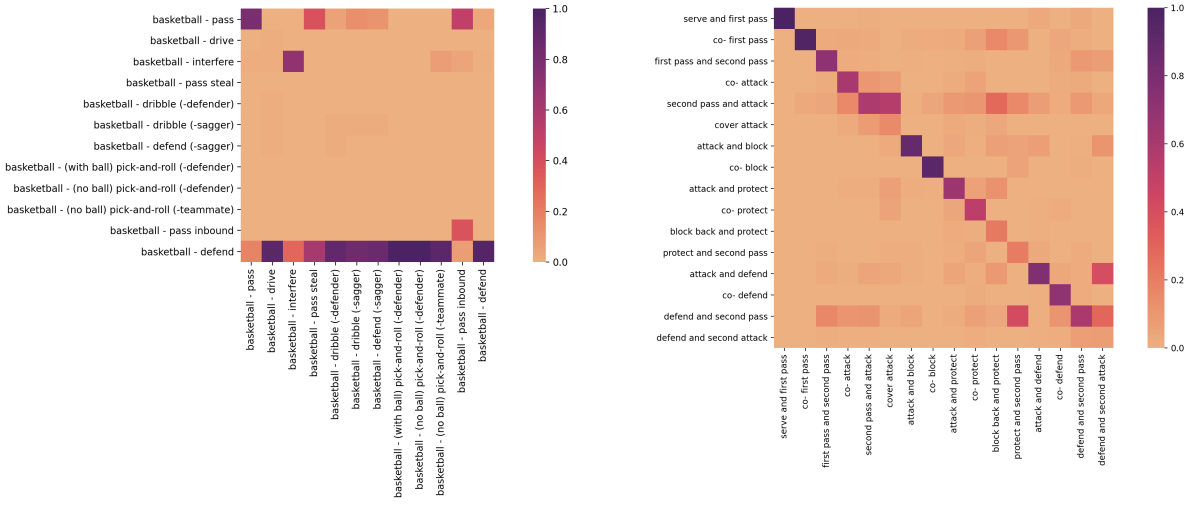
4. More detailed error analysis

4.1. Error analysis tree

Following ACT [6] and MultiSports [7], we analyze error types of the false positives in the predictions to better understand the inherent difficulty in HHIDet. As illustrated in Figure 2, we classify the detection errors into 9 mutually exclusive categories with a decision tree. A more detailed description of each error type is listed below.

- E_R (Errors of repeated detection): a detection result that has a triplet IoU larger than a threshold and the right action class with some ground truth triplet, but the ground truth triplet has already been matched by a detection result with a larger confidence score.
- E_C (Errors of classification): a detection result that has the triplet IoU larger than a threshold with a ground truth, but its interaction class is not the same with the ground truth.
- E_O (Errors of object localization): a detection result that has the same interaction class as a ground truth and the box IoU between the corresponding subject boxes are acceptable, but the box IoU between the object boxes are low.
- E_S(Errors of subject localization): a detection result that has the same interaction class as a ground truth and the box IoU between the corresponding object boxes are acceptable, but the box IoU between the subject boxes are low.
- $E_{S\&O}$ (Errors of subject and object localization): a detection result that has the same interaction class as a ground truth, but neither the object box IoU nor the subject box IoU meets the threshold.
- $E_{O\&C}$ (Errors of object localization and interaction classification): a detection result that has acceptable subject box IoU, but the object box IoU is low and the interaction class is incorrect.
- $E_{S\&C}$ (Errors of subject localization and interaction classification) a detection result that has acceptable object box IoU, but the subject box IoU is low and the interaction class is incorrect.
- $E_{S\&O\&C}$ (Errors of subject localization, object localiza-

Figure 3. **Visualization of typical errors for HHIDet on SportsHHI.** The subject person of a ground-truth or predicted interaction instance is marked in red and the object person is marked in blue. The ground-truth or correctly predicted interaction class labels are displayed in green and wrongly predicted in red.



(a) Basketball (b) Volleyball

Figure 4. **Confusion matrix of HHICls results on each sport**

tion and interaction classification): a detection result that has low subject box IoU, low object box IoU and the interaction class is incorrect.

- E_N (Errors of not matched): a detection result that has no overlap with any ground truth triplets of any class, indicating there should be no interaction detection results.

4.2. Visualization of error analysis

We provide visualizations of false positives of some error types in Figure 3 to show the challenge of human-human interaction detection on SportsHHI more intuitively.

4.3. Confusion Matrix

We draw the confusion matrix of HHICls predictions in Figure 4. We observe that the model generally performs better on volleyball classes than on basketball classes. This is because the interaction patterns are more complicated and the number of instances of each class is more unevenly distributed in basketball. From Figure 4, we observe the challenges of SportsHHI in the following three aspects.

1. **Handling long-tail distribution.** In basketball data for example, classes like *sag* or *pick-and-roll* have only a small number of instances while *close defense* is very common. The optimization of the interaction classification network will be biased towards the dominant classes. However, the long-tail distribution is natural and inevitable in real-world data and how to catch the rare interactions remains a difficult yet important problem.
2. **Action modeling.** Confusion between volleyball classes *defend - second pass* and *defend - second attack* indicates the importance of the action modeling of each person in SportsHHI. To distinguish between these two classes, we need to accurately identify whether the object person's action is "pass" or "attack". Former video visual relation detection datasets do not emphasize modeling human actions, so the current methods only used appearance features, which is not sufficient for interaction recognition on SportsHHI.
3. **Long-term temporal structure modeling.** As a baseline model, we only leverage spatiotemporal context with a short video clip for neatness and simplicity. However, in order to distinguish between classes like *protect - second pass* and *defend - second pass*, we need longer temporal information to distinguish whether the ball is defended or protected.

References

[1] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *ICCV*, pages 16352–16362. IEEE, 2021. 1, 2, 3

[2] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *CVPR*, pages 4772–4781. IEEE Computer Society, 2016. 3

[3] Mostafa S. Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *ECCV (3)*, pages 742–758. Springer, 2018. 3

[4] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 1

[5] Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. Detecting human-object relationships in videos. In *ICCV*, pages 8086–8096. IEEE, 2021. 1, 2

[6] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, pages 4415–4423. IEEE Computer Society, 2017. 4

[7] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13536–13545, 2021. 4

[8] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *CVPR*, pages 10837–10846. Computer Vision Foundation / IEEE, 2020. 1, 2

[9] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV (1)*, pages 852–869. Springer, 2016. 1

[10] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021. 3

[11] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *ACM Multimedia*, pages 84–93. ACM, 2019. 1, 2

[12] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM International Conference on Multimedia*, Mountain View, CA USA, 2017. 1

[13] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM, 2019. 1

[14] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision*, pages 71–87. Springer, 2020. 3

[15] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *ICCV*, pages 13668–13677. IEEE, 2021. 1, 2

[16] Yao-Hung Hubert Tsai, Santosh Kumar Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi.

Video relationship reasoning using gated spatio-temporal energy graph. In *CVPR*, pages 10424–10433. Computer Vision Foundation / IEEE, 2019. 1, 2

[17] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *CVPR*, pages 7408–7416. IEEE Computer Society, 2017. 3

[18] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV (5)*, pages 413–431. Springer, 2018. 3

[19] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 3

[20] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, pages 9964–9974. Computer Vision Foundation / IEEE, 2019. 3

[21] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware rcnn: A baseline for action detection in videos. In *European Conference on Computer Vision*, pages 440–456. Springer, 2020. 3