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
https://github.com/MCG-NJU/SportsHHI

Abstract

Video-based visual relation detection tasks, such as video scene graph generation, play important roles in fine-grained video understanding. However, current video visual relation detection datasets have two main limitations that hinder the progress of research in this area. First, they do not explore complex human-human interactions in multi-person scenarios. Second, the relation types of existing datasets have relatively low-level semantics and can be often recognized by appearance or simple prior information, without the need for detailed spatio-temporal context reasoning. Nevertheless, comprehending high-level interactions between humans is crucial for understanding complex multi-person videos, such as sports and surveillance videos. To address this issue, we propose a new video visual relation detection task: video human-human interaction detection, and build a dataset named SportsHHI for it. SportsHHI contains 34 high-level interaction classes from basketball and volleyball sports. 118,075 human bounding boxes and 50,649 interaction instances are annotated on 11,398 keyframes. To benchmark this, we propose a two-stage baseline method and conduct extensive experiments to reveal the key factors for a successful human-human interaction detector. We hope that SportsHHI can stimulate research on human interaction understanding in videos and promote the development of spatio-temporal context modeling techniques in video visual relation detection.

1. Introduction

Video understanding is a fundamental research field in computer vision, which has wide applications in security monitoring, internet video recommendation, and sports video analysis. Remarkable progress has been made in the action recognition task [3, 4, 13, 22, 23, 39, 40] with the advances of video convolution neural networks [3, 11, 45, 46, 49, 50, 52, 55] and video transformers [1, 2, 10, 29, 44, 51, 58].

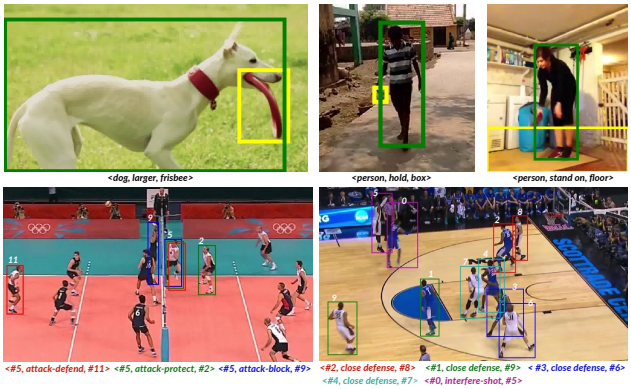

Figure 1. **Comparison between previous video visual relation detection datasets and our SportsHHI.** In the upper row, we show three relation instances from VidVRD and AG datasets. These datasets rarely involve human-human interaction and define semantically simple relations that can be recognized by appearance or prior information. In contrast, the bottom row shows interaction annotations in two sample keyframes of SportsHHI. The bounding boxes and interaction annotation of the same instance are displayed in the same color. SportsHHI provides complex multi-person scenes where various interactions between human pairs occur concurrently. It focuses on high-level interactions that require detailed spatio-temporal context reasoning.

The action recognition task requires tagging each video clip with a single action label. However, real-world applications usually require a more detailed and in-depth understanding of videos, thus more attention has been attracted to fine-grained video understanding tasks such as video action detection [14, 18, 24, 26, 31, 42, 54] and video visual relation detection tasks [7, 19, 20, 35, 36].

Video action detection involves localizing action performers in both space and time, as well as recognizing their action class. While earlier datasets like UCF101-24 [40] and JHMDB [18] primarily focuses on action detection in single-person scenarios, recent large-scale datasets such as AVA [14] and MultiSports [26] have emphasized the importance of modeling human-human interactions for a better recognition of individual actions in multi-person scenes. However, while AVA and MultiSports provide ample multi-

*: Equal contribution. ✉: Corresponding author.

person scenes, neither of them offers explicit human-human interaction definitions and annotations.

Video scene graph generation is a popular video visual relation detection task, which requires the model to localize objects of interest in the video and recognize the relation of each pair of objects. While current datasets like AG [19], VidVRD [35], and VidOR [36] provide explicit relation instance annotations, human-human interactions are barely involved. Moreover, the relation categories defined in these benchmarks primarily include spatial relations, atomic actions, and simple visual comparisons, which are semantically straightforward and often recognizable by appearance information or simple prior knowledge, such as $\langle dog, larger, frisbee \rangle$ and $\langle person, stand\ on, floor \rangle$ in the upper row of Figure 1. This makes current research pay more attention to the recognition of object categories and the utilization of category priors while ignoring reasoning based on spatio-temporal context. However, comprehending high-level interactions between humans is crucial for understanding complex multi-person videos, such as sports and surveillance videos. Spatio-temporal context modeling is an important technology that makes video relation detection different from image relation detection. We argue that there is a need for new datasets that comprehensively explore human-human interactions in complex multi-person scenarios, with high-level relation categories that require detailed spatio-temporal context reasoning.

In this paper, we propose a new video visual relation detection task: video human-human interaction detection. This task requires detecting human-human interaction instances in video frames. Each human-human interaction instance is formulated as a triplet $\langle S, I, O \rangle$, where S and O denote the bounding boxes of two different people and I denotes the interaction category. For this task, we develop a dataset named SportsHHI, short for Sports Human-Human Interaction. We label interaction instances on the keyframes of basketball and volleyball videos at 5FPS. Two samples of keyframes are shown in the bottom row of Figure 1. SportsHHI has several unique characteristics that distinguish it from other visual relation detection datasets: 1) It is built on basketball and volleyball sports videos, thus containing a large number of complex multiplayer scenes with various interactions between athlete pairs occurring concurrently. 2) The interactions between each pair of humans are transient and change rapidly due to the fast movements of the players. 3) The defined interactions are of high-level semantic, including technical actions, tactical cooperation, and confrontation in sports. Recognizing an interaction instance in SportsHHI usually requires detailed spatio-temporal context reasoning. We hope that SportsHHI can attract more research attention to human-human interaction understanding in complex multi-person videos and promote the development of spatio-temporal context modeling techniques in video visual relation detection.

We test existing action detection and video scene graph generation methods on the SportsHHI dataset and propose a simple and neat baseline method based on these methods. Our baseline method adopts a Faster-RCNN-alike two-stage pipeline [33]: In the first stage, we use an offline human detector to detect person bounding boxes on the keyframe. We exhaust bounding box pairs to generate interaction proposals. In the second stage, we extract features for each proposal and categorize each proposal into an interaction class or background. Our experiments demonstrate that motion features, context information, relative position encoding, and information exchange among proposals are important for human-human interaction detection.

In summary, our contribution is three-fold: 1) We propose a new video relation detection task of human-human interaction detection which aims at exploring the complex interaction between people in multi-person scenarios; 2) We develop the SportsHHI dataset of multi-person sports videos on which high-level human-human interaction is well-defined and finely-labeled; 3) We design a two-stage baseline model and conduct extensive experiments on the SportsHHI dataset to discover the key factors for a successful interaction detector.

2. Related Work

Video action detection. Video action detection aims to locate and recognize action instances on video frames. Earlier datasets like UCF Sports [34], UCF101-24 [40], and J-HMDB [18] typically only include single-person videos. Datasets like DALY [53], AVA [14], AVA-Kinetics [24], and MultiSports [26] contain scenes where multiple people are performing various actions concurrently. AVA sparsely annotates keyframes of movie videos at 1FPS. MultiSports collects videos from sports competitions. Many interaction-related action classes are defined, such as *listen to* in AVA and *second pass* in MultiSports. As they claim, it is important to model the interaction between people to recognize the action category of each person. However, as no interaction annotation is provided, interaction modeling can only be performed implicitly. SportsHHI annotates interactions on keyframes of basketball and volleyball videos from MultiSports at 5FPS. With interaction definitions and annotations provided in SportHHI, human-human interaction detection can be explicitly performed and evaluated.

Video visual relation detection. Video scene graph generation requires localizing object pairs in the video and recognizing the relation of each pair of objects. AG [19] is a large-scale dataset built on Charades [38]. All videos in AG contain only one person, thus no human-human interaction is involved. In VidVRD [35] and VidOR [36], most of the videos contain only one person. Though human-human interaction is involved, the exploration of it is very

limited. Similar to AG, video human-object interaction detection [5, 41] datasets only focus on relations between humans and objects. Besides, the relation categories defined in these datasets only include spatial relations (*e.g. above*), atomic actions (*e.g. lean on*), and simple visual comparisons (*e.g. larger*). The relations are usually semantically simple and can be easily recognized by appearance information. Some trivial relations are also defined like $\langle dog, larger, frisbee \rangle$ in VidVRD. The category of the subject and object can often provide enough cues for relation recognition [7, 20, 28, 32, 43, 47], such as $\langle person, ride, horse \rangle$. Moreover, the relation between two objects changes very slowly in a video. Our SportHHI annotates human-human interaction instances in complex multi-person sports videos. The defined interactions are of high-level semantics and the interaction between two people changes rapidly.

Sports video understanding. Researchers have built many different benchmarks in the sports domain for its challenges in spatio-temporal reasoning and promising application prospects, such as UCF101-24 [40] and MultiSports [26] for video action detection, FineGym [37] and Diving48 [25] for action recognition and temporal action detection, SoccerNet [12] for action spotting, SportsMOT [8] for multi-object tracking, and NBA [56], CAD [6] and Volleyball [17] for group action recognition, etc. Group action recognition (GAR) requires tagging each video clip with a group action label. Implicit modeling of interactions among the players is often performed to improve the accuracy of label predictions. The proposed SportsHHI dataset provides interaction definitions and annotations for explicit human-human interaction exploration.

3. The SportsHHI Dataset

In Sec. 3.1, we will explain several key choices we made while creating the SportsHHI dataset and outline the annotation process. Then in Sec. 3.2, we will present a detailed analysis of the statistics and characteristics of SportsHHI.

3.1. Dataset Construction

Selection of the data domain. First, we select to annotate human-human interaction in team sports videos for the following reasons: 1) The field of sports is itself a very promising application area for interaction detection. 2) Team sports videos provide ample multi-person scenes where various interactions between athlete pairs occur concurrently, which are rarely seen in daily-life videos. 3) Team sports videos involve many interactions with high-level semantics, such as tactics coordination or confrontation. 4) Interactions between people in sports game videos can be clearly and comprehensively defined according to game rules and professional athletes' advice. 5) High-quality, diverse sports game videos are easily accessible,

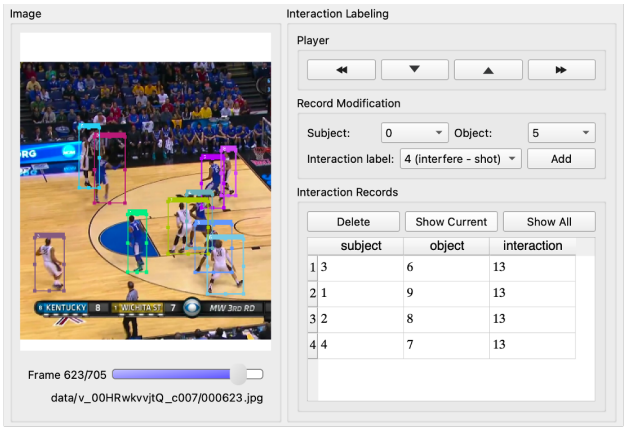

Figure 2. **User interface for interaction annotation.** The person bounding boxes and ids in the keyframe are shown in the left. We can play the video for context information. To add an interaction instance in the current keyframe, the subject person id, object person id, and interaction class should be specified.

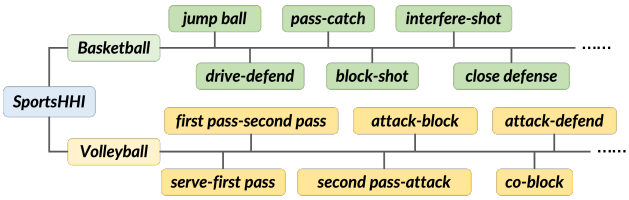

Figure 3. **Interaction classes hierarchy.** There are 34 interaction classes of high-level semantics in total in SportsHHI. 16 for basketball and 18 for volleyball.

while surveillance videos are usually of lower quality and harder to obtain because of privacy. Furthermore, we decide to build the dataset on two popular team sports basketball and volleyball because: 1) The methods can be validated on two different subsets to evaluate their generalization performance on different sports. 2) Football is not included because the instance sparsity and class imbalance (which will be discussed in Sec. 3.2) would be more prominent in football videos.

Interaction classes definition. There are existing benchmarks focusing on atomic relations, which can be recognized by appearance information or prior knowledge. These interaction classes are out of our scope (it does not mean they are solved or not important). Instead, we focus on high-level interactions between athletes, which are semantically complex and require detailed spatio-temporal context reasoning to recognize. With the guidance of professional athletes, we generated the final interaction vocabulary through iterative trial labeling and modification, which is shown in Figure 3. Our defined interaction classes include technical action (*e.g. pass-catch* in basketball), tactical coordination (*e.g. co-block* in volleyball), or confrontation (*e.g. attack-defend* in volleyball).

Interaction instance formulation. Following common practice in AVA and AG datasets, we define interaction in-

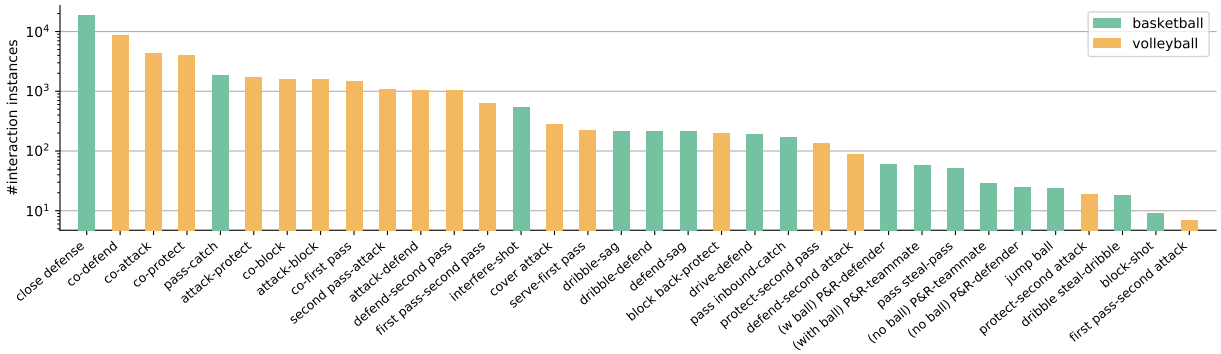

Figure 4. The number of interaction instances of each class sorted by descending order.

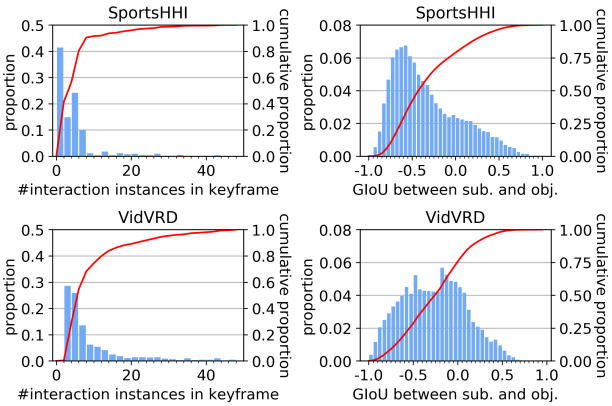

Figure 5. **Statistics comparisons between SportsHHI and VidVRD.** In the left, we compare the distribution of the number of instances in each keyframe. SportsHHI has more keyframes of fewer instances because of the high-level interaction class definition and the property of sports videos. In the right, we compare the distribution of GIoU between the subject and object. The proportion of instances of extremely high and extremely low GIoU between subject and object are both higher than VidVRD,

stances at the frame level, with reference to a long-term spatial-temporal context. Each interaction instance can be formulated as a triplet $\langle S, I, O \rangle$ where S and O denote the bounding boxes of the subject and object person and I denotes the interaction category between them from the interaction vocabulary. When the subject person or the object person is out of view, we annotate S or O as "invisible". This happens infrequently and we will provide statistics about it in the appendix.

Data preparation. We carefully selected 80 basketball and 80 volleyball videos from the MultiSports [26] dataset to cover various types of games including men's, women's, national team, and club games. The average length of the videos is 603 frames and the frame rate of the videos is 25FPS. All videos have a high resolution of 720P.

Annotation. We follow the sparse annotation strategy of AVA [14] and AG [19] to reduce redundancy and save labor costs, but we annotate keyframes at a higher rate of 5FPS, because in sports videos, the change of interactions is very frequent and fast. In our practice, 5FPS can avoid lots of redundancy and keep up with the changes of interactions.

	#keyframes	#interact.	#inst.	#obj. bbox	#hum. bbox	avg. hum.
AG [19]	234253	25	1.7M	476229	234253	1.00
VidVRD [35]	5834	134	55631	12705	2224	0.38
SportsHHI	11398	34	50649	-	118075	10.36

Table 1. Comparison of statistics between video scene graph generation datasets and SportsHHI. SportsHHI has a comparable scale to VidVRD. The average number of human bounding boxes per frame of SportsHHI is much higher than AG or VidVRD.

We take a two-stage annotation pipeline. The first stage is for **person localization and tracking** and the second stage is for **interaction instance annotation**. The user interface of the software for interaction instance annotation is shown in Figure 2. Though we annotate interaction instances at the frame level, as person id tracking is provided, we can easily generate interaction tubes by linking the same pair of persons with the same interaction class and provide temporal boundaries at a granularity of 5 frames. We will exemplify this in the appendix.

Quality control. For the first stage, all annotations are complemented manually without using detection or tracking models. We double-check each video and manually correct inaccurate bounding boxes and inconsistent person numbers. For the second stage, all the annotators are professional athletes or veteran amateurs. To improve annotation accuracy and completeness, we perform repeated annotation. Each video is distributed to two different annotators. If the two annotators disagree on an annotation record, this record will be checked and decided by the meta-annotator.

3.2. Dataset Statistics and Characteristics

Our SportsHHI provides interaction instance annotations on keyframes of basketball and volleyball videos. As shown in Table 1, SportHHI contains 34 interaction classes, and 11398 keyframes in total are annotated with instances. Current video scene graph datasets deal with general relations between various kinds of objects while our SportsHHI focuses on high-level interaction between humans. It is reasonable that datasets for video scene graph generation have a larger scale than SportsHHI. However, our SportHHI still

has a comparable size to the popular VidVRD dataset for video scene graph generation. Our SportsHHI has more annotated keyframes (11398 versus 5834) and the number of interaction instances is close (55631 versus 50649). One important characteristic of our SportsHHI is the multi-person scenarios. The average number of people per frame in our SportsHHI is much higher than AG and VidVRD. AG only contains one person in each video and there is virtually no multi-person scenario in videos of VidVRD. Human-human interaction is barely involved in these datasets.

We further show some important characteristics of our SportsHHI: 1) As shown in Figure 4, the distribution of the number of interaction instances per class roughly follows Zipf's law. This long-tail distribution requires us to put more attention to classes with fewer instances. 2) As shown in Figure 5 left, compared with VidVRD [35], our SportsHHI has more keyframes with fewer instances. That is because we focus on interaction classes of coordination or confrontation in sports. In long plain segments (*e.g.* segments of dribbling in basketball), few interactions of interest happen, and there are much more interaction instances in video segments with fierce confrontation. The videos are in crowd multi-person scenarios but the interaction instances are relatively sparse, which requires the model to distinguish two people without interaction from real interaction instances. 3) As shown in Figure 5 right, the proportion of instances of low GIoU between subject and object in SportsHHI is significantly higher than in VidVRD, which indicates that the subject and object of many instances are spatially far apart in SportsHHI. The proportion of extremely high GIoU is also higher, which indicates the occlusion is severe in some instances. VidVRD deals with simple relations in daily life videos, where the subject and object usually have a moderate distance.

4. Method

We propose a baseline method for the human-human interaction detection task. As illustrated in Fig. 6, our method adopts a two-stage pipeline. In the first stage, interaction proposals are generated. In the second stage, we construct a representation for each proposal and classify each proposal into an interaction class or background. The following part of this section explains our method in detail.

4.1. Interaction proposal generation

We use an offline human detector for human detection on keyframes. We enumerate all human bounding box pairs as interaction proposals. Formally, for a keyframe at timestamp t, let N denote the number of human boxes in the frame and $B = \{b_1, b_2, \cdots, b_N\}$ denote the boxes. The interaction proposals can be denoted as $P = \{p_1, p_2, \cdots, p_K\}$, where $p_k = (b_{s_k}, b_{o_k})$ and $K = N \times (N-1)$.

Positive proposals and negative proposals sampling. We label a proposal p_k as a positive proposal when its subject and object bounding boxes both have an IoU overlap higher than 0.7 with their counterparts in a ground-truth interaction instance. For training, it is possible to input all the interaction proposals to the classification network for loss function calculation, but this will consume large GPU memory and the network optimization will bias towards negative samples as they are dominant. Instead, we use all positive proposals and sample negative proposals by a fixed positive-negative ratio. To increase the proportion of positive proposals and avoid the situation that a ground truth has no corresponding positive proposal, we add all ground-truth human box pairs as positive proposals for training.

4.2. Interaction classification

The presentation of each interaction proposal comprises three parts: the motion features of the subject and object persons, the context features, and the relative position of the subject and object encoding.

Motion features and context features. We sample a video clip centered at the keyframe and employ a 3D video backbone on it for feature map extraction. For an interaction proposal p_k, we apply RoIAlign operation [16] on the feature maps with the bounding boxes b_{s_k} and b_{o_k} and then transform the RoIAligned feature to dimension d_v with a linear layer. The motion features are denoted as $f_k^{ms} \in \mathbb{R}^{d_v}$ and $f_k^{mo} \in \mathbb{R}^{d_v}$. We apply the RoIAlign operation on the features map using the union box of the subject and object for context information. The context feature is also transformed to dimension d_v, denoted as $f_k^c \in \mathbb{R}^{d_v}$

Relative position information encoding. Following [7, 57], given an interaction proposal p_k, we generate a spatial mask for the subject and object person respectively. We resize the union box of b_{s_k} and b_{o_k} to a fixed size of $M \times M$. We then generate the spatial mask of the subject bounding box $m_k^s \in \mathbb{R}^{M \times M}$. For each pixel in m_k^s, if it is inside the subject bounding box b_{s_k}, its value is set to 1, else 0. The spatial mask of the object bounding box $m_k^o \in \mathbb{R}^{M \times M}$ is generated in the same way. The two spatial masks are flattened, concatenated, and transformed to dimension d_p. We use the final vector $f_k^p \in \mathbb{R}^{d_p}$ as relative position encoding.

Proposal representation. We use the concatenation of motions features f_k^{ms} and f_k^{mo}, context features f_k^c, and relative position encoding f_k^p as the representation x_k of interaction proposal p_k. Formally,

$$x_k = \langle f_k^{ms}, f_k^{mo}, f_k^c, f_k^p \rangle , \tag{1}$$

where $\langle , \rangle$ is the concatenation operation.

Information exchange among proposals. We denote the representation of the K interaction proposals as $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_K\}$. Sometimes, recognizing an interaction requires information from other interaction instances. For

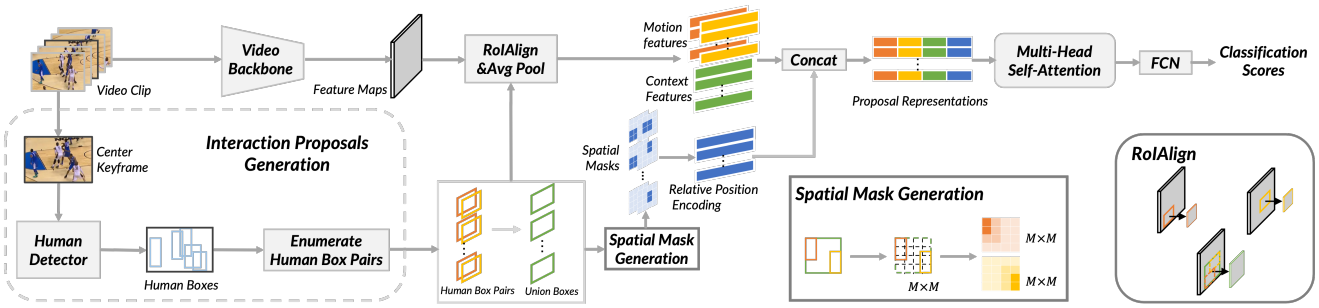

Figure 6. **An overview of our baseline method.** Given a video clip centered at a keyframe, we first use an offline human detector to detect human bounding boxes on the keyframe. All the human bounding box pairs are enumerated as interaction proposals. Then, we construct representations for the proposals. We perform the RoIAlign operation on the video feature maps extracted by a video backbone for motion features and context features. We generate spatial masks for each proposal, which are flattened and transformed into a vector. The vector is used as relative position encoding. The motion features, context features, and relative position encoding are concatenated as proposal representations. Multi-head self-attention is applied to the representations for information exchange. Finally, an FCN classification head is used to predict the classification scores.

example, to recognize an interaction of *co-defend* in volleyball, we need to know there exists an interaction of *attack-defend*. To capture the information from other interaction proposals, we apply a standard multi-head self-attention (MHSA) [48] operation on X. Formally,

$$X' = \text{MHSA}(X). \tag{2}$$

Classification head. Finally, X' is input to a fully connected network (FCN) for classification score prediction. Let S denote the predicted scores. Formally,

$$S = \text{FCN}(X'), \tag{3}$$

where $S \in \mathbb{R}^{K\times(C+1)}$, C denotes the number of interaction classes.

5. Experiments

5.1. Experimental Setup

Dataset. We train and validate our model on the SportsHHI Dataset. The dataset is split into the training and validation set by video. Evaluation on a very small number of examples could be unreliable, so we evaluate the 28 classes that have at least 10 instances in both the training and validation set. Instances with an invisible subject or object are excluded from evaluation. In total, the current version contains 38,527 instances from 8,719 keyframes for training and 12,122 instances from 2,679 keyframes for validation. We report results on the validation set.

Metrics. A prediction is considered as a true positive if and only if its subject and object bounding boxes both have an IoU overlap higher than a preset threshold with their counterparts in a ground-truth interaction instance and the predicted interaction class matches the ground truth. Following VidVRD [35], we set the IoU threshold to 0.5. Following the video scene graph generation task, we use Recall@K (K is the number of predictions) as the evaluation metric. Mean average precision is adopted as an evaluation metric for many detection tasks [27]. However, this metric is discarded by many former visual relation detection benchmarks [19, 30] because of their incomplete annotation. This issue does not exist in SportsHHI. In our experiments, mAP is also reported for the interaction detection task. We argue that mAP is a more difficult and informative metric. Two different modes for model training and evaluation are used: 1)human-human interaction detection (HHIDet) which expects input video frames and predicts human boxes and interaction labels; 2) human-human interaction classification (HHICls) which directly uses gound-truth human bounding boxes and predicts human-human interaction classes.

Implementation details. The human detector we use is a Faster-RCNN [33] detector with a ResNeXt-101-FPN backbone, which is pre-trained on ImageNet [9] and COCO [27] and finetuned on keyframes of SportsHHI. The ratio of positive and negative proposals is set to 1:10. The SlowFast-R50 backbone [11] pre-trained on Kinetics-400 [22] is adopted for video feature extraction. The dimension of features d_v and the dimension of relative position encoding d_p are set to 512 and 256 respectively. We scale the short side of the video frames to 256. We use SGD optimizer with momentum 0.9 and weight decay 1×10^{-5}. The batch size is set to 8. The learning rate is 0.002 for HHICls and 0.004 for HHIDet. The models are trained for 20 epochs.

5.2. Ablation Study

We conduct ablation experiments on SportsHHI to investigate the influence of each design and component of our baseline method. Unless otherwise stated, all ablation experiments are conducted in the HHICls mode, that is, the ground-truth human bounding boxes are used. HHICls mode allows us to study the key factors to predict interactions without the limitations of human detection.

Positive and negative proposal ratio. We investigate the influence of the ratio of positive and negative proposals in Table 3. A moderate proportion of negative samples help the model to distinguish interactions from backgrounds. However, when the positive-negative ratio gets too low, the model optimization will be overwhelmed by negative samples, and thus the performance drops.

Spatial mask generation. We generate two separate spatial masks for the subject and object person. Therefore, the positions of the subject and object can be distinguished. For comparison, we generate a single mask for the two people by setting the values in any box to 1. As shown in Table 4, model performance degrades when using a single mask. As the triplet $\langle S, I, O \rangle$ is ordered in SportsHHI, it is important to distinguish the position of the subject and object.

Interaction proposal representation. We show the importance of motion features f^{mo} and f^{ms}, context feature f^{c} and relative position encoding f^{p} in interaction representation in Table 2. We first show the results of using each feature in lines 1-3 and 8-10. Using only motions features f^{mo} and f^{ms} can achieve acceptable results, which indicates the modeling of the actions of the subject and object person is very important for the recognition of the interaction between them. The performance is much worse when using context feature f^{c} or relative position encoding f^{p} only. Using only the context feature cannot provide specific information of the subject and object and as the subject and object person are often far apart in spatial, the context feature may include much noise. Without visual information, pure prior information of the relative position of the subject and object person is not discriminative enough for interaction recognition. As shown in lines 5-7 and 12-14, when coupling position encoding with motion and context features, the performance improves a lot, which indicates that position information is complementary to visual information. As shown in lines 7 and 14, using all three types of features achieves the best results.

Information exchange among proposals. We ablate the multi-head self-attention module to show the importance of information exchange among proposals. As shown in Table 2, regardless of the representation of the proposal, exchanging information among proposals always brings performance improvement.

5.3. Quantitative Results

Human box detection and interaction proposals generation results. We first show the human detection and interaction results in Table 5. Under IoU threshold of 0.5, both human boxes and interactions have a high recall rate of 98.6 and 98.4. When the threshold improves to 0.7, recall of human boxes drops by 2.6 points and recall of interactions drops by 5.3 points. When improving the IoU threshold to 0.9, both recall rates drop significantly. The results indicate that, despite high recall under a relatively lower IoU threshold, there is still a lot of room for improvement in the quality of human boxes and interaction proposals.

f^{mo}&f^{ms}	f^{c}	f^{p}	Info. Ex.	mAP	R@150	R@100	R@50	R@20
✓	-	-	-	5.00	81.66	74.00	52.73	26.82
-	✓	-	-	2.07	71.75	61.43	38.97	19.12
-	-	✓	-	1.13	50.71	45.24	39.72	30.83
✓	✓	-	-	5.19	81.98	74.81	54.89	28.51
✓	-	✓	-	5.54	83.05	75.45	57.71	31.92
-	✓	✓	-	3.85	76.76	68.57	49.65	27.68
✓	✓	✓	-	5.43	81.71	74.84	58.16	**34.20**
✓	-	-	✓	6.15	84.09	76.59	55.55	28.10
-	✓	-	✓	3.51	73.41	63.11	41.47	20.91
-	-	✓	✓	2.01	58.76	51.60	38.88	26.35
✓	✓	-	✓	6.82	84.01	77.11	58.14	31.86
✓	-	✓	✓	6.94	84.41	76.72	57.44	31.05
-	✓	✓	✓	5.97	80.75	71.82	51.78	31.04
✓	✓	✓	✓	**7.52**	**85.78**	**78.52**	**59.53**	32.76

Table 2. **Ablation study on the interaction representation and information exchanging.** "✓" indicates the corresponding element is enabled while - indicates disabled. "f^{mo}&f^{ms}" denotes motion features of the subject and object person."f^{c}" stands for context features of the union area. "f^{p}" stands for the relative position information encoding. "Info. Ex." stands for the multi-head self-attention operation to exchange information among proposals.

Pos:Neg	mAP	R@150	R@100	R@50	R@20
1:1	6.97	85.92	79.64	**60.96**	32.59
1:5	7.39	**86.25**	77.87	59.86	**32.82**
1:10	**7.52**	85.78	78.52	59.53	32.76
1:15	7.02	85.39	77.87	58.56	31.16
1:20	7.08	85.35	78.43	60.51	32.77
1:25	6.89	85.44	78.31	58.85	31.61

Table 3. **Ablation on the positive and negative proposals ratio.** The best results are obtained at 1:10.

	mAP	R@150	R@100	R@50	R@20
Single Mask	6.77	84.57	77.41	58.27	31.63
Separate Mask	**7.52**	**85.78**	**78.52**	**59.53**	**32.76**

Table 4. **Ablation on spatial mask generation.**"Separate Masks" means generating a spatial mask for the subject and object person respectively while "Single Mask" means using only one spatial mask to represent the pair.

	mAP	AP@0.5	R@0.5	R@0.7	R@0.9
Human Boxes	73.5	92.6	98.6	96.0	58.2
Interactions	-	-	98.4	93.2	37.7

Table 5. **Human detection and interaction proposal generation results.** For human box detection, we report mAP, AP, and Recall. For interaction proposal generation, Recall is reported. We calculate recall under IoU thresholds of 0.5, 0.7, and 0.9.

HHICls and HHIDet results. We provide HHICls and HHIDet results of some existing methods and our baseline in Table 6. STTran [7] and HORT [20] are two well-established video scene graph generation methods. They share two common properties: 1) Using appearance features extracted by image backbone. This is because the semantic level of relation classes defined by previous datasets is relatively low and the appearance feature is often sufficient for recognition, such as $\langle dog, larger, frisbee \rangle$. However,

Method	Backbone	HHICls					HHIDet				
		mAP	R@150	R@100	R@50	R@20	mAP	R@150	R@100	R@50	R@20
STTran [7]	ResNet-101 [15]	3.31	71.29	62.91	42.67	22.14	1.43	51.65	46.54	37.62	20.81
HORT [20]	ResNet-101 [15]	3.75	78.57	67.78	50.33	26.96	1.54	52.47	46.73	36.89	20/93
SlowFast [11]	SlowFast-R50 [11]	5.00	81.66	74.00	52.73	26.82	2.44	62.17	52.77	37.83	22.45
ACARN [31]	SlowFast-R50 [11]	5.44	82.85	75.22	56.53	31.77	2.53	62.25	52.84	37.50	21.85
Our baseline	SlowFast-R50 [11]	7.52	85.78	78.52	59.53	32.76	3.36	64.82	54.62	37.72	21.26
	SlowFast-R101 [11]	8.67	87.35	80.35	60.18	29.70	3.52	65.24	54.76	37.03	20.21
	ViT-B [44]	**10.69**	**89.25**	**82.93**	**68.13**	**43.72**	**4.93**	**72.22**	**61.92**	**42.99**	**23.89**

Table 6. **HHICls and HHIDet results.** We test existing video scene graph generation methods (STTran [7] and HORT [20]) and action detection methods (SlowFast [11] and ACARN [31]) on SportsHHI. Our baseline method achieves the best performance.

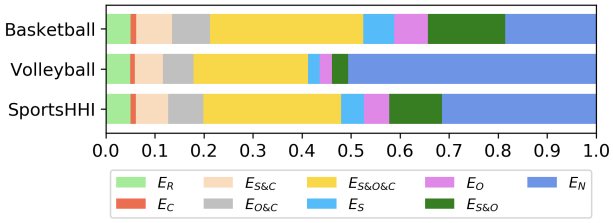

Figure 7. **Error analysis in HHIDet mode.** We show the proportion of each error type among false positives.

action modeling is very important for interaction recognition on SportsHHI. 2) Dependent on the object detection results. They add the category encoding of the objects to the relation representation, which provides strong prior information. For example, 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 prior as all subjects and objects are humans. Since video backbones are used to extract the motion features of each person, the performance of the action detection methods like SlowFast [11] and ACARN [31] is relatively better. Our baseline follows these action detectors to adopt a video backbone for motion features and introduce extra relative position and spatio-temporal context modeling between the subject and object person. Our simple baseline achieves leading performance with the same SlowFast-R50 backbone. When using the stronger VideoMAE [44] pretrained ViT-B backbone, the performance is significantly improved, which shows the importance of spatio-temporal representation.

5.4. Error Analysis

Following ACT [21] and MultiSports [26], we analyze error types of the false positives in the predictions to better understand the challenges of human-human interaction detection in SportsHHI. We analyze the predictions of the baseline model with a SlowFast-R50 backbone. We classify false positives in the predictions into 9 mutually exclusive error types. E_R: a prediction targets a ground-truth instance that has already been matched. E_C: a prediction that has an accurate subject and object localization, but wrong interaction class. E_O: a prediction that has accurate subject localization and correct interaction class, but inaccurate object localization. E_S: a prediction that has accurate object localization and correct interaction class, but inaccurate subject localization. $E_{S\&O}$, $E_{O\&C}$, $E_{S\&C}$, $E_{S\&O\&C}$: a prediction that is inaccurate in corresponding aspects while acceptable in other aspects (if any). E_N: a prediction that both the subject and object of it have no overlap with any ground truth.

As shown in Figure 7, E_N is one of the most common error types. This is because SportsHHI is annotated in crowd multi-person scenarios but the interaction instances are relatively sparse due to the characteristic of sports videos and our high-level interaction definition. It is important but difficult to distinguish two people without interaction from real interaction instances. In video action detection [26], localization is relatively accurate and most errors are about classification. However, in interaction detection, the localization of people involved in interaction and the classification of interaction are inseparable. Pure localization errors (E_S and E_O) or pure classification errors (E_C) are relatively rare. The error types that coexisted with localization error and classification error ($E_{O\&C}$, $E_{S\&C}$, $E_{S\&O\&C}$) account for a relatively large proportion.

6. Conclusion

In this paper, we proposed a new video visual relation detection task with a focus on sports human-human interaction understanding. This task deals with the complex interactions between humans in multi-person scenarios. We build a dataset named SportsHHI based on sports videos for this task. Human boxes and interaction instances are exhaustively annotated on keyframes at 5FPS. To benchmark this, we test several existing methods on SportsHHI and propose a baseline method. We conduct extensive experiments on SportsHHI and reveal the importance of motion information, context information, and position information for interaction recognition. We hope our SportHHI dataset and baseline method can inspire future research on human-human interaction understanding in videos.

Acknowledgements. This work is supported by National Key R&D Program of China (No. 2022ZD0160900), National Natural Science Foundation of China (No. 62076119, No. 61921006)), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Luči'c, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. 1

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, pages 813–824, 2021. 1

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1

[4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 1

[5] Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. ST-HOI: A spatial-temporal baseline for human-object interaction detection in videos. In *ICDAR@ICMR*, pages 9–17, 2021. 3

[6] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshops*, pages 1282–1289, 2009. 3

[7] 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, 2021. 1, 3, 5, 7, 8

[8] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *ICCV*, pages 9887–9897, 2023. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[10] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6824–6835, 2021. 1

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 1, 6, 8

[12] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *CVPR Workshops*, pages 1711–1721, 2018. 3

[13] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5843–5851, 2017. 1

[14] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018. 1, 2, 4

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 8

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 5

[17] 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. 3

[18] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013. 1, 2

[19] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10236–10247, 2020. 1, 2, 4, 6

[20] Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. Detecting human-object relationships in videos. In *ICCV*, pages 8086–8096, 2021. 1, 3, 7, 8

[21] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, pages 4415–4423, 2017. 8

[22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 6

[23] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 1

[24] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020. 1, 2

[25] Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: towards action recognition without representation bias. In *ECCV*, pages 520–535, 2018. 3

[26] 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 *ICCV*, pages 13536–13545, 2021. 1, 2, 3, 4, 8

[27] 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 *ECCV*, pages 740–755. Springer, 2014. 6

[28] 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, 2020. 3

[29] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3192–3201. IEEE, 2022. 1

[30] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016. 6

[31] 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 *CVPR*, pages 464–474, 2021. 1, 8

[32] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *MM*, pages 84–93, 2019. 3

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 28, 2015. 2, 6

[34] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 2

[35] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *MM*, 2017. 1, 2, 4, 5, 6

[36] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR*, pages 279–287, 2019. 1, 2

[37] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2613–2622, 2020. 3

[38] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 2

[39] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020. 1

[40] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1, 2, 3

[41] Xu Sun, Yunqing He, Tongwei Ren, and Gangshan Wu. Spatial-temporal human-object interaction detection. In *ICME*, pages 1–6, 2021. 3

[42] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *ECCV*, pages 71–87. Springer, 2020. 1

[43] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *ICCV*, pages 13668–13677, 2021. 3

[44] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 1, 8

[45] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 1

[46] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 1

[47] 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, 2019. 3

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 6

[49] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 1

[50] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. TDN: temporal difference networks for efficient action recognition. In *CVPR*, pages 1895–1904, 2021. 1

[51] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae V2: scaling video masked autoencoders with dual masking. In *CVPR*, pages 14549–14560, 2023. 1

[52] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, pages 413–431, 2018. 1

[53] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Towards weakly-supervised action localization. *CoRR*, abs/1605.05197, 2016. 2

[54] Tao Wu, Mengqi Cao, Ziteng Gao, Gangshan Wu, and Limin Wang. Stmixer: A one-stage sparse action detector. In *CVPR*, pages 14720–14729, 2023. 1

[55] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018. 1

[56] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *ECCV*, pages 208–224, 2020. 3

[57] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 5

[58] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *ICCV*, pages 13577–13587, 2021. 1