Sports Video Captioning via Attentive Motion Representation and Group Relationship Modeling

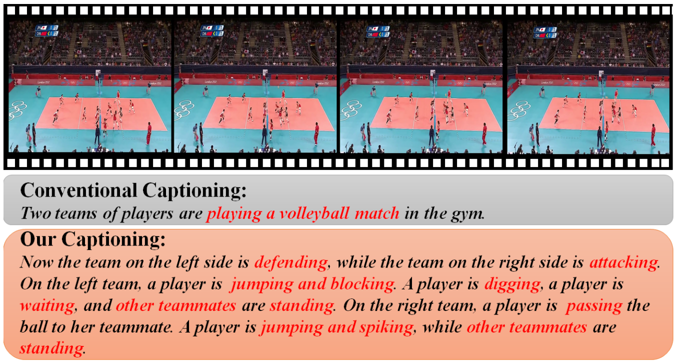
Mengshi Qi, *Member, IEEE,* Yunhong Wang, *Senior Member, IEEE,* Annan Li, *Member, IEEE,* and Jiebo Luo, *Fellow, IEEE*

***Abstract*—Sports video captioning refers to the task of automatically generating a textual description for sports events (*e.g.,* football, basketball or volleyball games). Although a great deal of previous work has shown promising performance in producing a coarse and general description of a video but lack of professional sports knowledge, it is still quite challenging to caption a sports video with multiple fine-grained player's actions and complex group relationship between players. In this study, we present a novel hierarchical recurrent neural network based framework with an attention mechanism for sports video captioning, in which a motion representation module is proposed to capture individual pose attribute and dynamical trajectory cluster information with extra professional sports knowledge, and a group relationship module is employed to design a scene graph for modeling players' interaction by a gated graph convolutional network. Moreover, we introduce a new dataset called *Sports Video Captioning Dataset-Volleyball* for evaluation. The proposed model is evaluated on three widely-adopted public datasets and our collected new dataset, on which the effectiveness of our method is well demonstrated.**

***Index Terms*—Sports Video, Video Captioning, Motion Representation, Group Relationship, RNN.**

I. Introduction

SPORTS video captioning, which aims at elaborately describing events and actions happened in a match with natural language, has captured more attention in computer vision, multimedia and natural language processing communities [1], [2]. In sports videos, a plenty variety of players' actions and interactions occur at the same time, *e.g.,* in a volleyball game (see Figure 1). Automatically generating paragraph to describe more details of sports events has potential huge application value in sports video analysis and sports broadcast. However, the complex variations of the dynamic event and temporal structures make sports video captioning an arduous problem.



Conventional Captioning:

Two teams of players are playing a volleyball match in the gym.

Our Captioning:

Now the team on the left side is defending, while the team on the right side is attacking. On the left team, a player is jumping and blocking. A player is digging, a player is waiting, and other teammates are standing. On the right team, a player is passing the ball to her teammate. A player is jumping and spiking, while other teammates are standing.

Fig. 1. Illustration of sports video captioning task. Conventional captioning always generates a coarse-level textual description for the given video. In contrast, the task of sports video captioning needs to capture more fine-grained individual action details and group relationships between players. The main differences are highlighted in red.

Recently, a great number of researchers strive to this emerging topic. Conventional algorithms for video captioning can be divided into two categories: one is the template-based language model [3]–[5], which generates captions based on the predefined grammar rules, templates of sentences, and correlation between each part of sentence with detected object; and the other is the sequence learning method [6]–[14], which is inspired by Recurrent Neural Nework (RNN), such as Long Short-Term Memory (LSTM) [15] and Gated Recurrent Unit (GRU) [16]. Sequence learning based methods have already achieved the state-of-the-art performance at present for visual captioning. Generally, these approaches are designed based on the encoder-decoder architecture: a encoder is utilized to translate input original video frames to the compact visual features, while a decoder is employed to generate words and sentences by sequence. However, all these previous methods can only generate coarse and general description by a collection of the basic frame-level appearance features, which ignore the motion details of individual action and group activity, resulting in inappropriate for sports video captioning.

In order to fully understand sports events, sports video captioning should take global visual appearance as well as the fine-grained individual motion information into consideration. Theoretically, the action of each individual player is the mainly fine-grained motion information in a sports event, which involves player's articulated movements/pose estimation [17]–[19] and motional trajectory [20], [21]. Capturing and representing these movements accurately and effectively

This work was partly supported by the National Key Research and Development Plan of China (Grant No.2016YFB1001002), the National Natural Science Foundation of China (Grant No. 61573045) and the Foundation for Innovative Research Groups through the National Natural Science Foundation of China (Grant No. 61421003). Jiebo Luo would like to thank the support of New York State through the Goergen Institute for Data Science, NSF (Awards No. 1813709 and No. 1722847), and Futurewei. Mengshi Qi acknowledges the financial support from the China Scholarship Council.

M. Qi, Y. Wang and A. Li are with Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University, Beijing 100191, China. E-mail: {qi_mengshi, yhwang, liannan}@buaa.edu.cn. (Yunhong Wang is the corresponding author).

J. Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627, USA. E-mail: jiebo.luo@gmail.com.

from the untrimmed video would provide more informative cues for captioning. Moreover, in a sports match, especially for ball games, the dynamic motion of a team or group also contains rich information of sports tactics, which is well worth to comprehensively analyzing. Capturing the movement trajectory of team/group can express the change of match situation and tactic strategy, which also is often neglected by traditional methods.

Furthermore, conventional video captioning methods lack sports knowledge, so that their produced text captions cannot describe sports match professionally and accurately. Establishing different dictionaries of sports terms and introducing more specific tactic knowledge into sports video captioning is essential and necessary.

Another difficulty is how to represent the complex group activity happened in sports events. To recognize the group activity under a variety of scenarios, the context information should be taken into consideration, especially the relationship between player and player, between player and objects, and between player and scene. A number of team sports (*e.g.,* soccer and volleyball) contain a variety of mutual interactions (*e.g.,* teammates and opponents) and frequently changeable situations (*e.g.,* attack and defend), thus exactly discovering and understanding group relationships and encoding them for captioning are challenging. As an example, in the "team spiking activity" in Figure 1, one player is spiking and her teammates are waiting or standing to cooperate with her, while their opponents are trying to block and dig.

In addition, attention mechanism [22]–[24] is often introduced to identify the salient visual regions with high objectiveness score and meaningful visual pattern of an image. For the task of visual captioning, the performance can be also improved by attending the spatial salient object and the temporal motion information, and selectively assigning different weights to encode features. As for sports video captioning, the key player's action or movement, such as dunking in basketball and shooting in soccer, invariably play a significant role in a sports event, thus precisely attending to these highlights and retrieving the crucial movement are overwhelmingly critical for sports video captioning.

In this work, to address the above-mentioned issues, we propose a novel hierarchical LSTM-based deep framework for sports video captioning with attentive motion representation and group relationship modeling. In particular, individual pose attribute features and dynamical trajectory cluster information will be fed into a hierarchical encoder-decoder architecture. Generically, the semantic attributes can be leveraged as an extra sports professional knowledge to guide the generation of sports captions, which contains a wealth of sports terminology. Furthermore, we construct a scene graph to model group relationship among players via a gated graph convolutional network, and obtain a plenty of contextual information of sports events. Then, we fuse the motion representation, group relationship and global frame-level features and decode them into natural language utilizing a sequence to sequence architecture with the attention mechanism.

It should be mentioned that this paper is an extension of our conference paper [25]. Compared to the preliminary version, we additionally introduce extra professional sports knowledge to guide the training of the motion representation module, present a group relationship module with the gated graph convolution network and scene graph modeling, and devise a hierarchical bi-directional LSTMs as the encoder-decoder in our proposed framework. Moreover, in the testing process, we conduct extensive experiments on one more public benchmark datasets, *i.e., ActivityNet Dataset* [26], further perform human evaluation for sports video captioning with three criteria, illustrate more qualitative results, and more detailed ablative analysis to demonstrate the effectiveness of each component in our proposed framework in this paper.

The main contribution of this work can be summarized as the following:

- We introduce a novel deep framework for sports video captioning with attentive motion representation and group relationship modeling based on the hierarchical recurrent neural networks.
- A motion representation module is designed to capture player's pose and trajectory information, where we extract semantic attributes from player's skeletons, and cluster trajectory from dynamical movement guided by extra sports professional knowledge.
- A group relationship module is devised to construct a scene graph for modeling the interaction between players by a gated graph convolutional network.
- We annotate a new dataset called *Sports Video Captioning Dataset-Volleyball* that mainly contains volleyball games for evaluation. Meanwhile, extensive experiments on three public benchmarks and our dataset demonstrate the effectiveness and general applicability of our framework. To the best of our knowledge, we are the first to propose such a volleyball video captioning dataset.

The rest of this paper is organized as follows. Related work on video captioning and sports video analysis is briefly discussed in Section II. The proposed framework for sports video captioning is presented in detail in Section III. Then the experimental results are shown and analyzed in Section V. Finally, we draw the conclusion in Section VI.

II. Related Work

A. Video Captioning

Early significant efforts often adopt template-based language methods [3]–[5] that align sentence elements (*e.g.,* subject, verb, object) with detected words from visual contents. Rohrbach *et al.* [4] learned a conditional random fields (CRF) [27] to model the relationships between different components of video contents, and generated sentence descriptions for videos. Xu *et al.* [5] proposed an unified framework to jointly model video and language, by utilizing a compositional language model and a deep neural network.

Very recently, growing sequence learning approaches [6]–[9] have been performed to learn probability distribution in space of video and textual sentence for video captioning. Venugopalan *et al.* [6] proposed an end-to-end sequence-to-sequence model to generate captions for videos, and their model can directly encode the temporal information by LSTM.

Yao *et al.* [7] presented a temporal attention mechanism to automatically select temporal segments for generating video caption. Yu *et al.* [8] proposed a hierarchical recurrent neural network, which consisted of a sentence generator and a paragraph generator for video captioning. Pan *et al.* [9] presented an LSTM-Embedding framework considering the relation between sentence semantic and video content. Furthermore, a great amount of research attended to semantic factors in visual and text content. Pasunuru *et al.* [10] presented a multi-task and knowledge sharing-based method for unsupervised video prediction and language entailment generation. Gan *et al.* [11] developed a semantic compositional network (SCN) to detect semantic concepts or tags from visual contents and employed an LSTM to compose the probability of each tag. Dong *et al.* [12] proposed an interpretive loss via extracting interpretable features from semantically meaningful topics for visual captioning. Pan *et al.* [13] presented a CNN-RNN framework to transfer semantic attribute learned from images and videos. Baraldi *et al.* [28] presented a recurrent video encoding scheme by discovering and leveraging the hierarchical structure of the video. Shen *et al.* [29] trained a model based on weak video-level sentence annotations by using lexical fully convolutional neural networks (Lexical-FCN) [30]. Wang *et al.* [31] added reconstruction sub-network into conventional encoder-decoder framework. Chen *et al.* [32] performed video captioning via picking the informative frames, and Wang *et al.* [33] shared the memory between video features and textual data to guide attention mechanism. Afterward, more modality data are introduced to handle the problem. Hori *et al.* [14] incorporated audio features with image and motion feature for jointly video captioning via multi-model attention mechanism. In addition, reinforcement learning has been introduced into the issue [34].

Krishna *et al.* [26] firstly introduced a dense video captioning model that combined the proposal and the captioning modules to caption each event by a single sentence. Following this work, Wang *et al.* [35] fused the contextual information of past and future attentively with a gating strategy, and Zhou *et al.* [36] adopted a masked transformer framework with self-attention for dense video captioning. Li *et al.* [37] designed a descriptiveness regression-based approach for video proposal localizing and dense captioning jointly.

However, the captions generated by these works are coarse-level and missing lots of fine-grained level or detailed movement occurring in sports videos, consequently they are all inappropriate for sports video captioning. A very closely related work to ours is [38], and their work introduced a sports narrative method by employing player localization, group activity and action information. However, their approach requires pixel-level annotation in each frame and lacks extra professional knowledge of different sports types. While our framework can introduce sports knowledge as the semantic attribute to guide caption generation, and adopt a trajectory clustering approach to capture dynamical motion information. Last but not least, we employ the attention mechanism for further improved performance.

B. Sports Video Analysis

In recent years, sports video analysis has became an emerging research topic due to its wide-range audiences and enormous economic potential, which contains a wealth of issues in computer vision and multimedia. With the development of the Internet and mobile devices, a considerable amount of research have been devoted to sports video analysis, which mainly includes player tracking [39]–[41], ball detection [42]–[44], group activity/individual action recognition [45]–[49], understanding specific event during the match [1], [2], [20], [50] and highlight summary [51], *etc*.

For player tracking, Lu *et al.* [39], [40] proposed a Kalman Filter (KF) [52] and conditional random field (CRF) [27] based approach to automatic tracking and labeling players in broadcast sports videos. Considering player's trajectories, Liu *et al.* [41] introduced a Game Context Features (GCF) based model for tracking players in team sports. The model was built to describe the current match state and the distribution of players, and adopted hierarchical trajectories to produce game context features with Random Decision Forest [53]. Morimitsu *et al.* [54] adopted attributed graphs for tracking multiple objects in structured sports videos. While tracking the ball in team sports is intensely hard because of its small size and low-resolution. Wang *et al.* [42], [43] proposed a conditional random field (CRF) [27] and trajectories based method for ball tracking, by exploiting the correlation between the ball and player. Maksai *et al.* [44] performed a Mixed Integer Program considering physical constraints to track the ball, estimate motion of the ball, and different states of the ball's landing. A two-stream deep model for action recognition is introduced via combining RGB and flow frames [47]. Besides, Ibrahim *et al.* [48] devised a hierarchical LSTM model for group activity recognition in the volleyball game. Then, Shu *et al.*. [49] introduced a confidence-energy recurrent network (CERN) with an energy layer to estimate group activity. To handle the high order context modeling problem in group activity recognition, Qi *et al.* [46] presented an attentive semantic RNN model, and Wang *et al.* [55] proposed a recurrent context modeling framework. Event detection is a semantically high-level task, so that it is more complex to address. Xu *et al.* [1] presented an approach for event detection from live sports game based on text and video on the Internet. Zhu *et al.* [20] detected the goal event through extracting tactic information from broadcast soccer videos. Duan *et al.* [2] proposed a mid-level representation between audio-visual processing and semantic analysis for event detection in sports videos. Lin *et al.* [51] performed a context-based method to select highlights and estimate the label of each streaming sports video segment. In addition, in order to classify ego-action sports categories, Kitani *et al.* [50] introduced a fast unsupervised learning method to deal with first-person sports videos.

However, these previous methods are not specially designed for sports video captioning task. In our work, we propose a novel specific framework to describe the details of sports game with attentive motion representation and group relationship modeling.

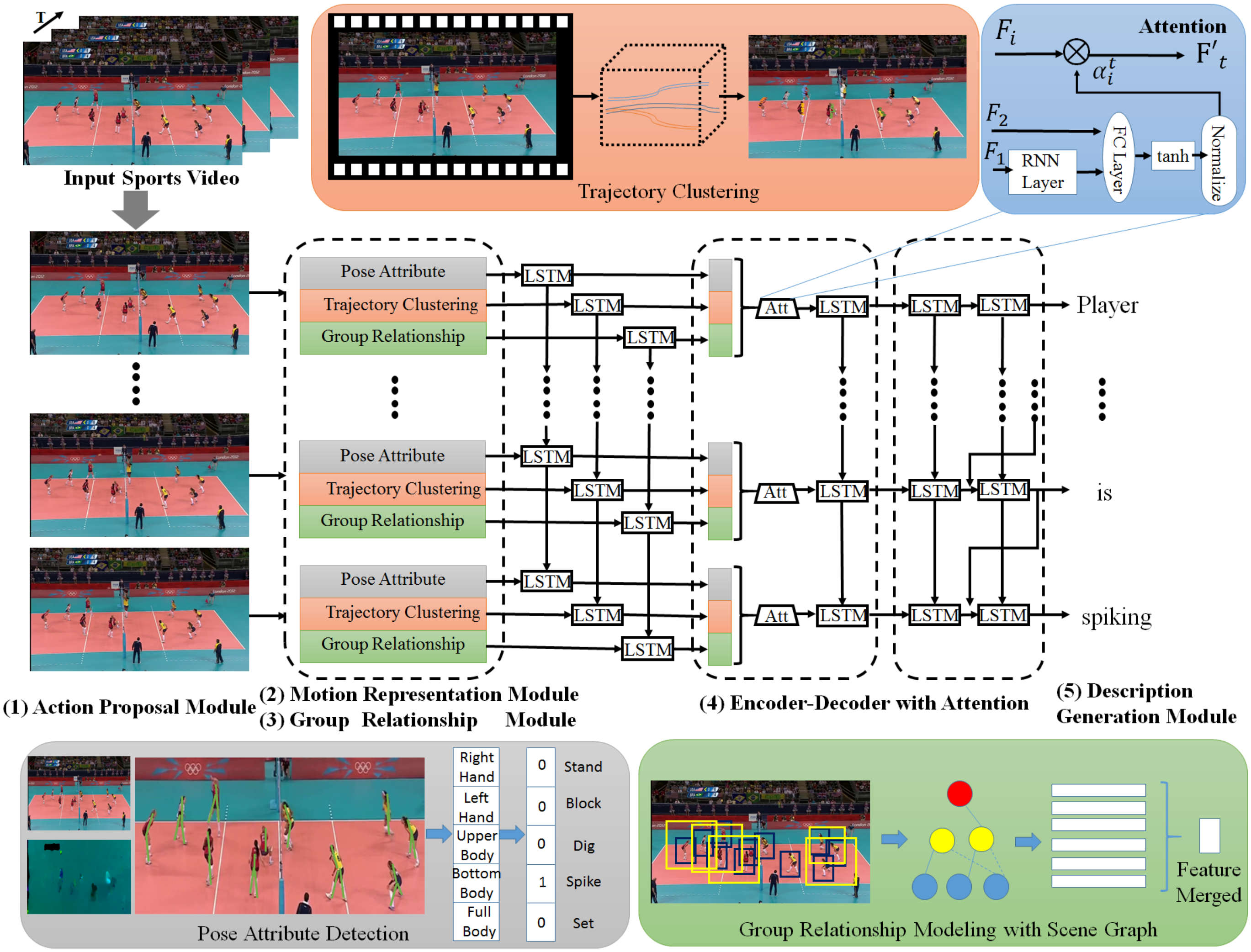


Fig. 2. The overall framework of our proposed sports video captioning model. (1) **Action Proposal Module** segments the whole video into activities of players. (2) **Motion Representation Module** employs detected pose attribute (on the grey background) and trajectory cluster (on the red background) to encode the individual action and the corresponding dynamical movement. (3) **Group Relationship Module** (on the green background) constructs a scene graph to model the interactions among players. (4) Finally, motion and group relationship features are all fused and decoded through an LSTM-based **Encoder-Decoder Architecture** with an attention mechanism (on the blue background). (5) **Description Generation Module** is used to generate textual caption.

III. The Proposed Approach

The framework of the proposed approach for sports video captioning is illustrated in Figure 2. Concretely, our framework includes: (1) action proposal module; (2) motion representation module; (3) group relationship module; (4) encoder-decoder with attention mechanism and (5) description generation module. We adopt a sequence-to-sequence based architecture [6], where the input is a sequence of video frames, and the output is a sequence of words. And the lengths of the input and output are variable. Because of the success of long short-term memory (LSTM) in the visual captioning task, we employ this paradigm in our framework.

A. Action Proposal Module

Given a video, the first task is retrieving and localizing temporal segments that probably contain crucial spatio-temporal group events (*i.e.,* attacking, defending in a sports game) or significant individual actions (*i.e.,* spiking, passing in a volleyball match). In our work, we adopt Deep Action Proposals (DAP) [56] method to generate temporal action proposals. We infer the temporal location and duration of the action proposals from a T-frame video. And each proposal is associated with a confidence score. In practice, the input feature of video frames is extracted from the top layer of a 3D convolutional network (C3D) [57], and then an LSTM network is utilized to encode the sequential information.

B. Motion Representation Module

In order to describe sports video with natural language, more fine-grained details in terms of individual player's action would be of great assistance. Therefore, we design a motion representation module to model player's actions, which consists of a pose attribute detection part and a trajectory clustering part.

1) Pose Attribute Detection: Pose estimation is often used to recognize the individual action in a fine-grained manner.

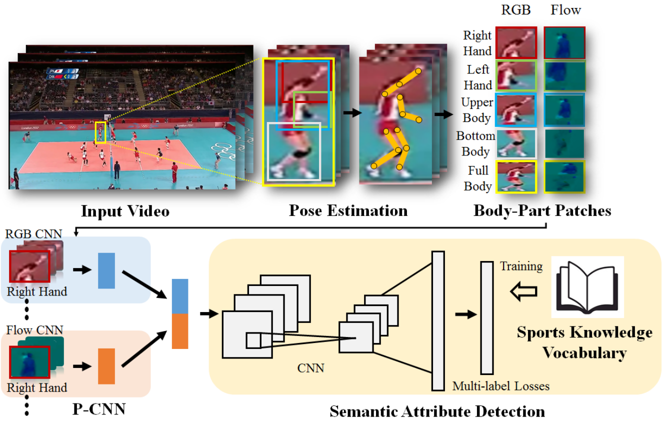


Fig. 3. Pipeline of pose attribute detection in our framework. Firstly, we perform pose estimation to acquire the keypoints with hourglass model [59]. Then, we select five body part patches of RGB and optical flow, *i.e.,* right hand/left hand/upper body/bottom body/full body, as input to P-CNN [18]. P-CNN is able to capture aggregated pose-based deep features combining RGB-CNN and Flow-CNN. Finally, we introduce extra sports knowledge as our attributes vocabulary to train a CNN model for semantic attribute detection.

Given a sequence of video frames, we desire to determine the precise pixel location of critical human body's keypoints, which is awarding to understand individual posture and limb articulation. In our work, we firstly utilize Faster R-CNN [58] to localize all players with the corresponding bounding boxes. Then we have a set of candidate objects with the bounding boxes that represent their location and appearance feature. Based on the predicted probability map, we select several bounding boxes with high confidence score per frame (*e.g.,* we choose 12 bounding boxes on Volleyball Dataset). Afterward, we adopt hourglass model [59] to extract the pose keypoints of each player. Center point of the detected skeleton is utilized to measure the relative offset of each body part, and optical flow values represent the motion of every joint, which manifest the characteristics of the player's movement (*e.g.,* velocity and direction).

As illustrated in Figure 3, we capture the pose-based CNN features with P-CNN [18] from each track of individual player's body parts in the video clip. Based on the position of body joints (*i.e.,* we select five pose parts per player: *right hand, left hand, upper body, bottom body* and *full body*), we crop corresponding RGB and optical flow patches and normalize them to 224 × 224. We adopt VGG-16 network pre-trained on the ImageNet dataset [60] for RGB patches, and motion network in [61] pre-trained on the UCF-101 dataset [62]. The pose-based feature for each player in the video clip can be denoted as F_{pose}, which concatenated all the feature of each pose part.

However, directly and disorderly pooling all the players' pose-based feature inevitably fails to represent rich semantic information, we desire to further capture more semantic attributes from such raw features. For sports narrative, learning semantic knowledge with respect to specific sports type would be significantly beneficial. Accordingly, we build an attribute vocabulary from the annotated sentences in dataset (*e.g.,* UCF-101 [62] for general action attributes, Volleyball [48] for volleyball action attributes), and we use the top k high-frequency verbs and nouns of them. To be specific, we regard such a vocabulary of pose semantic attribute as sports knowledge. Moreover, we also collect external sports knowledge from public texts on the Internet, *e.g., Wikipedia*, including the words and phrases that humans professionally and commonly utilize to describe the sports events. Then, the pose attribute can be certain object (*e.g.,* hand, leg, head, feet) or individual motion (*e.g.,* spike, dig, pass). When training the pose attribute detection network based on VGG-16, we employ the ground-truth attribute label of each player on the datasets, *e.g.,* individual action label (spiking, passing, standing, etc.) on SVCDV to train the cross-entropy loss function. The input data of the network are pose-based features of each player, and the output data are predicted binary pose attribute vectors. Given n player's pose-based features $\{F^1_{pose}, ..., F^n_{pose}\}$, we take the i-th player's feature F^i_{pose} as input, and employ the last fully connected layer of VGG-16 net to be a k-way attribute classifier (where k refers to the total number of attributes in our built sports knowledge dictionary). Moreover, we formulate the predicted pose attribute vector for the i-th player in the one-hot scheme as $\widehat{y^i} = [\widehat{y^i_1}, \cdots, \widehat{y^i_k}]$, where $\widehat{y^i_k} = 1$ if the player is predicted with the k-th attribute, and $\widehat{y^i_k} = 0$ otherwise. The predicted attribute is determined by the corresponding classification probability scores through the softmax layer of the network. Then, we define the ground-truth attribute annotation of the i-th player as $y^i = [y^i_1, \cdots, y^i_k]$, and the loss function of our proposed attribute detection network can be formulated as the following:

$$L_{att} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}[\widehat{y^i_j}\log(y^i_j) + (1-\widehat{y^i_j})\log(1-y^i_j)]. \quad (1)$$

Particularly, we assign the labels of input data depended on the order of coordinate of players in the frame. By sorting the center point's coordinate dx of each player's bounding box from left to right, we can classify pose attribute with input pose-based feature of each player. After training, we formulate the pose attribute vector by fusing $\widehat{y^i}$ for all the players by order. The fusion order of each player's attribute vector is also performed by the dx coordinates position of the corresponding bounding box. Furthermore, we adopt the attention mechanism in the encoder network (will be described in III-D) to get attentive pose attribute representation F_{pose_att} of n player's poss attribute vector:

$$F_{pose\_att} = \sum_{i=1}^{n}\{\text{Attention Weight}\} \cdot \widehat{y^i}. \quad (2)$$

It is worthy to note that we capture the pose attributes of all the players from a sequence of video clips, and we will produce T copies for all frames of the video based on the encoder-decoder architecture.

2) Trajectory Clustering: The trajectory is good at representing temporal motion in videos, and clustering them into groups can capture the significant dynamical movement information. Inspired by [63], [64], we capture the dense point trajectories $tra = \{tra_1, tra_2, ...tra_M\}$ for a sequence of frames, where M is the number of trajectories. And we set the maximum length of the trajectory as 15 frames. Furthermore, we follow the distance metric in [63] to measure the similarity between trajectory pairs in terms of temporal interval and spatial position. Then we partition all the detected trajectories into groups by computing the affinity matrix between each trajectory pair and utilizing a graph clustering method [63]. Given a video, we can obtain m clusters. We assume the i-th trajectory cluster which contains L trajectories as $tra(i) = \{tra_{i1}, ..., tra_{il}\}$. And defining each trajectory as a position point sequence $tra_{il} = \{(x^1_{il}, y^1_{il}, z^1_{il}), ..., (x^T_{il}, y^T_{il}, z^T_{il})\}$, where $(x^t_{il}, y^t_{il}, z^t_{il})$ is the 3D coordinates of the t-th point in trajectory tra_{il}, and T is the time step of trajectory.

Then we employ convolutional neural networks (CNN) to obtain the trajectory-pooled deep-convolutional representations [65]. We input each frame to the CNN, and obtain a feature map of size $H \times W \times N$, where H, W and N are the number of height, width and channel, respectively. Finally, we achieve an overall feature map $C \in R^{H \times W \times T \times N}$ through concatenating all the feature maps of the video, where T is the length of the video. Then, a trajectory point can be represented with coordinates (x^t, y^t, z^t) (center at $(r \times x^t, r \times y^t, r \times z^t)$ in the feature map, where r denotes the map size ration w.r.t the input size). Thus, the averaged feature of tra_{il} is formulated as the following:

$$F_{tra_{il}} = \frac{1}{T}\sum_{t=1}^{T} C(r \times x^t_{il}, r \times y^t_{il}, r \times z^t_{il}), \tag{3}$$

and the representation of the trajectory cluster is computed via mean pooling of all trajectory features in the same cluster:

$$F_{tra_i} = \frac{1}{L}\sum_{l=1}^{L} F_{tra_{il}}. \tag{4}$$

For a given video, we extract m trajectory clusters and the visual feature of them can be defined as $\{F_{tra_1}, F_{tra_2}, \cdots, F_{tra_m}\}$. Then we adopt the attention mechanism in the encoder network (will be described in III-D) to formulate the overall trajectory feature vector F_{tra}:

$$F_{tra} = \sum_{i=1}^{m}\{\text{Attention Weight}\} \cdot F_{tra_i}. \tag{5}$$

C. Group Relationship Module

Since a wealth of contextual information are contained in the inter-player relationship, only analyzing individual actions is insufficient for understanding sports videos. Therefore, we design a group relationship module by constructing a scene graph [66] to model team-level interaction representations between players, and encode the relation-aware features.

According to aforementioned approach in III-B1, given n detected players through Faster R-CNN [58], $n \times (n-1)$ player pairs can be obtained. We concatenate two types of features of the player pair, *i.e.,* appearance feature and spatial feature, to describe the visual relationship in each frame. The appearance representation for visual relationship is formulated by generating an enclosing bounding box to cover player pair with a small margin. While the spatial representation is formulated by relative positions and sizes of object pairs, which is robust and insensitive to the change of illumination and occlusion. We encode a 5-dimensional vector as spatial representation $x_{spatial} = [\frac{x_{min}}{W_I}, \frac{y_{min}}{H_I}, \frac{x_{max}}{W_I}, \frac{y_{max}}{H_I}, \frac{S_b}{S_I}]$, where $[x_{min}, y_{min}, x_{max}, y_{max}]$ and S_b are bounding box coordinates and area size of detected region b, respectively. W_I, H_I and S_I are width, height and area size of image I, respectively.

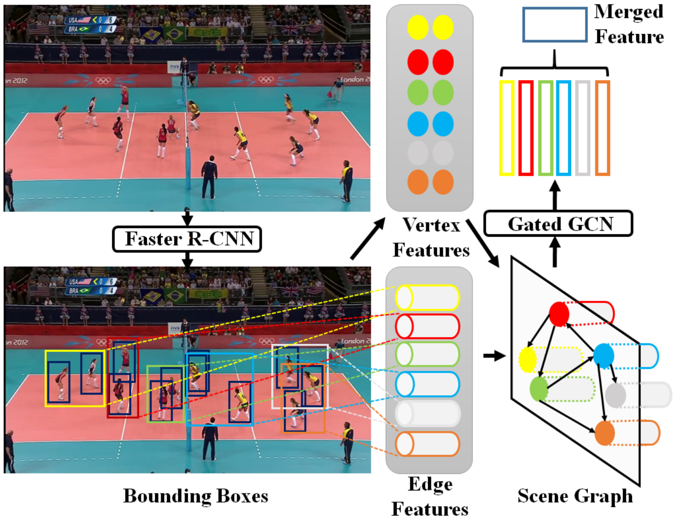


Fig. 4. Illustration of group relationship module in our framework. We firstly employ the Faster R-CNN [58] to obtain a collection of player bounding boxes in a frame. Then, a scene graph is constructed of which the vertex and edge refer to player and their semantic relationship, respectively. Finally, a Graph Convolutional Networks (GCN) is adopted to encode all the edge features with a gate function.

As depicted in Figure 4, we employ a Graph Convolutional Network (GCN) [67] as the encoder of group relationship module, which is often leveraged to refine the representation of each image region and their interaction. Then, all of the encoded region features are fed into an attention-based LSTM network, which assigns different important weights to each region in a frame. Inspired by recent methods on visual relationship detection [68], [69], we predict the semantic relation between players depending on the union bounding box that is able to cover the two players. Specifically, we split all the players into different groups, *e.g.,* choosing two players or three players in a group. As an example, we group K players in a frame into $K \times (K-1)$ player pairs. For the reason that accumulating the representation of all connected edges is not a favorable idea, we incorporate an edge-wise gate function into GCN to determine different weight of each edge.

In particular, we construct a scene graph [46], [69], [70] $G =< V, E >$ by forming all the detected object bounding boxes in each frame to model visual relationship with object pairs, in which V and E are vertex set and edge set, respectively. In the graph G, each node corresponds to a detected object, each edge denotes the interactive relationship. The representation of each edge in the scene graph can be formulated as the following:

$$\begin{aligned} f_{e_{ij}} &= \rho\Big(\sum_{(i,j)\in V} g_{<i,j>}(f_{v_i}, f_{v_j})(U \cdot f_{e_{ij}} + b)\Big), \\ g_{<i,j>}(f_{v_i}, f_{v_j}) &= \sigma(\widetilde{U} \cdot [f_{v_i}, f_{v_j}] + \widetilde{b}), \end{aligned} \tag{6}$$

where $f_{e_{ij}}$, f_{v_i}, and f_{v_j} denote the feature of edge e_{ij} (*i.e.,* relationship region covering node v_i and node v_j), node v_i, and node v_j, respectively. Meanwhile, $g_{<i,j>}$ represents the gate function, σ is the logistic sigmoid function, ρ denotes an activation function (*e.g.,* ReLU), and $[\cdot,\cdot]$ refers to the concatenation operation. U and $\widetilde{U}$ are the transformation matrices of each edge between node v_i and v_j to be trained in GCN, b and $\widetilde{b}$ are the biases. Finally, defining N_E as the total number of edges in a frame, the merged relationship representation F_{rel} of all the edge features can be achieved with attention mechanism (will be described in III-D) as the following:

$$F_{rel} = \sum_{(i,j)\in V}^{N_E} \{\text{Attention Weight}\} \cdot f_{e_{ij}}. \tag{7}$$

D. Encoder-Decoder with Attention Mechanism

We follow the widely-adopted encoder-decoder framework for video captioning. However, the traditional LSTM-based encoder-decoder used in previous video captioning methods is difficult to model long-time dependency and generate long sentences. For the task of sports video captioning, the goal is to generate paragraph description rather than single sentence caption, which stresses on representing wealthy contextual information and relationships between generated sentences. Therefore, as illustrate in Figure 2, we propose a hierarchical Bi-directional LSTMs as encoder-decoder in our proposed framework to address the problem following [8], [26]. In our proposed encoder, we leverage two-layer Bi-directional LSTMs to fuse the motion representation, group relationship feature and frame feature from the given video, resulting in encoding the input video features into a sequence of jointly latent vectors. While we employ another two-layer Bi-directional LSTMs as our decoder, of which the first layer is utilized to generate single word only based on the current state, and the second layer can preserve more contextual information by taking the state of previous generated sentences as input.

In our work, the LSTM based encoder takes attentive motion representation (*i.e.,* pose attribute feature F_{pose_att} and trajectory clustering feature F_{tra}), group relationship feature F_{rel} and frame feature F_{frame} as input, which would be concatenated to formulate total representation F_t in the t-th frame. The updating procedure in the encoder is formulated as

$$\begin{aligned} h_t &= \text{Encoder}(h_{t-1}, F_t), \\ F_t &= [F_{pose\_att}, F_{tra}, F_{rel}, F_{frame}], \end{aligned} \tag{8}$$

where h denotes the hidden state of LSTM in the encoder, F_t refers to the input embedding representations, and $[\cdot]$ denotes concatenation operation.

The decoder takes the encoded representation as input, then sequentially produce the output vector, which denotes the predicted word at each time step. At each time step t, the LSTM updates its hidden state h_t and output y_t based on its previous hidden state h_{t-1} and output y_{t-1} and the encoded embedding F, as the following:

$$[\substack{y_t \\ h_t}] = \text{Decoder}(h_{t-1}, y_{t-1}, F). \tag{9}$$

Next, we will introduce the attention mechanism in the encoder-decoder network.

Attention Mechanism Conventional methods (*e.g.,* mean pooling operation) always ignore the importance of motion information for video captioning, of which the key player often plays the most remarkable role in the group event. Hence, we adopt a soft attention model to obtain dynamic weighted sum of the pose attribute vector, trajectory cluster representation and group relationship feature (described in Sec III-B1, III-B2 and III-C). Given the encoded embedding F, we denote $F_i \in \{F_1, ..., F_n\}$ as the feature of the i-th players, i-th trajectory clusters, or the i-th relationship edge in the scene graph, where n is the number of players, trajectory clusters or edges. We feed them to a single linear transform layer followed by a softmax function to calculate the attention distribution over F_i, and define $s_i^t = (s_1^t, ..., s_n^t)^T$ as the importance score of the i-th player, i-th trajectory cluster or i-th edge on the frame:

$$s_i^t = U_s \tanh(W_{fs}F_i + W_{hs}h_{t-1}^s + b_s), \tag{10}$$

where U_s, W_{fs}, W_{hs} are the training parameters, and b_s is the bias vector. h_{t-1}^s is the hidden variable from an LSTM unit. Then the attention weight is computed as a normalization of the scores:

$$\alpha_i^t = \frac{\exp(s_i^t)}{\sum_{i=1}^{n} \exp(s_i^t)}. \tag{11}$$

After that, the visual feature at time t is computed by the weighted sum of the frame features, *i.e.,* $F_t^{'}$,

$$F_t^{'} = \sum_{i=1}^{n} \alpha_i^t F_i, \tag{12}$$

where n denotes the number of players, trajectory clusters, or relationship edges, α_i^t is the attention weight of the i-th player, the i-th trajectory cluster, or the i-th relationship edge at time t. With the attention mechanism, the encoder-decoder is able to focus on the salient trajectory movement, key player's pose information and crucial group relationship.

E. Description Generation Module

The goal of sports video captioning in our work is to generate a paragraph including several word sequences to describe a given video. To generate a sentence, the likelihood of generating a word in the n-th sentence is formulated as the following:

$$P(w_t^n | s_{1:n-1}, w_{t-1}^n, F_t, W), \tag{13}$$

where $s_{1:n-1}$ represents all the preceding sentences in the paragraph, w_{t-1}^n means all the previous words in the n-th sentence, F_t are the features that concatenate attentive motion representation and group relationship features in the corresponding frames of the video, and W represents the model parameters to be learned. Furthermore, we define the overall loss function of generating the whole paragraph $s_{1:N}$ as:

$$\mathcal{L}_{cap} = -\sum_{n=1}^{N}\sum_{t=1}^{T_n}\log P(w_t^n|s_{1:n-1}, w_{1:t-1}^n, F_t, W) / \sum_{n=1}^{N} T_n, \tag{14}$$

where N is the number of sentences in the paragraph, T_n is the number of words in the n-th sentence.

IV. Sports Video Captioning Dataset

Sports Video Captioning Dataset-Volleyball (SVCDV) is a new dataset introduced by us that focuses on sports captioning. SVCDV has totally 55 videos with 4,830 short clips collected from *Youtube*, which are mainly high-resolution broadcast Olympic volleyball games. Specifically, the short clips are segmented into different types of group activities, and each short clips has more than 50 frames. It is annotated based on the Volleyball Dataset [48] that is collected to address group activity recognition issue especially. We annotated natural language description of player action and group activity happened in each video, and each sentence with respect to one action or movement. Furthermore, SVCDV has totally 44,436 sentences, of which each video clip has 9.2 sentences on average. Meanwhile, the average sentences per second is 0.366, verbs per sentence is 1.72, and verb ratio is 16.2% in the SVCDV dataset, which are all more than that in other current video captioning datasets (*e.g.,* MSVD [3], [9], MSR-VTT [71], ActivityNet [26]). It demonstrates that SVCDV is very suitable for sports video captioning task. In addition, each player is labeled with a bounding box and one of the nine action labels: *waiting, setting, digging, falling, spiking, blocking, jumping, moving* and *standing*. The whole frame is annotated with one of the eight group activity labels: *right set, right spike, right pass, right winpoint, left winpoint, left pass, left spike* and *left set*. These labels can be utilized for individual pose attribute learning and group relationship modeling. In experiments, we split the dataset into training, validation and testing sets of 65%, 5%, 30%, corresponding to 3,140, 241 and 1,449 video clips, respectively.

V. Experiments

In this section, we conduct extensive experiments in terms of three tasks, *i.e.,* general video captioning, dense-video captioning, and sports video captioning, to fully demonstrate the effectiveness of our approach on four public benchmark datasets. We select **MSVD Dataset** and **MSR-VTT Dataset** for general video captioning, **ActivityNet Captions Dataset** for dense-video captioning, and **Sports Video Captioning Dataset-Volleyball** for sports video captioning. In the following, we firstly introduce the datasets, evaluation metrics and implementation details in brief. Then we describe the comparison methods, present the experimental results and comprehensive analysis as follows.

A. Datasets and Metrics

Microsoft Video Description Dataset (MSVD) [3], [9] contains 1,970 short videos collected from *YouTube*, of which each video describes a single activity in a wide range of topics (*e.g.,* animals, music, actions and sports). In total, the dataset consists of 80,839 sentences with about 40 English descriptions per video clip, and each sentence has about 8 words. Following the same setting in [3], we select 1,200 videos as the training set, 100 for validation and 670 as the testing set.

MSR Video-to-Text Dataset (MSR-VTT) [71] is the largest general video captioning dataset in the size of sentences and vocabulary. It contains 10,000 video clips with 41.2 hours and 200,000 clip-sentence pairs in 20 categories (e.g., news, sports). In average, 20 natural sentences annotated manually for each video clip. The dataset is collected by using a commercial video search engine and covers most of comprehensive categories and diverse contents. Following the same setting in [71], we split it into training, validation and testing sets of 65%, 5%, 30%, corresponding to 6,513, 497 and 2,990 clips, respectively.

ActivityNet Captions Dataset [26] is introduced for dense captioning events and actions in videos. It contains 20k videos with 849 hours and more than 100k sentences collected from ActivityNet [72]. The dataset focuses on long-term event detection and the average length of videos is about 10 minutes, of which each video annotated with 3.65 sentences. Each sentence covers a unique segment of the video to describe multiple events occurred. Each sentence has an average length of 13.48 words, which follows a relatively normal distribution. Following [26], all sentences are pre-processed to be a maximum length of 30 words, and we split it into the training, validation and testing sets of 65%, 5%, 30%.

Metrics: We choose four popular metrics for the evaluation: CIDEr (C) [73], BLEU (B) [74], METEOR (M) [75], and ROUGE-L(R) [76], which are well correlated with human perception. Concretely, CIDEr is used to measure the average cosine similarity between n-grams in the generated description and reference sentences. BLEU is based on the n-gram precision, and we choose 4-gram in our work following previous works [74]. METEOR is computed based on the alignment between generated sentences and reference. ROUGE-L measures the similarity based on the longest common subsequence statistics between a candidate translation and a set of reference translations. We adopt Microsoft COCO evaluation tools[1] to test the performance of video captioning, which has implemented the metrics and evaluation functions.

B. Implementation Details

For video preprocessing, we sample equally-spaced 25 frames in each video, and resize them to 224 × 224 resolution. A VGG-16 network [77] pre-trained on the ImageNet dataset [60] is utilized to extract visual appearance features of frames, and we select a sequence of 4096-dimensional feature vectors produced by the fully connected layer fc7. Moreover,

[1] https://github.com/tylin/coco-caption

we employ the pre-trained C3D [78] network on the Sports-1M dataset [79] to model motion and short-term spatio-temporal activity of videos, and we extract the activation vector from fully-connected layer fc6-1 of C3D network from the input video. Meanwhile, we employ optical flow features captured by the motion network [61] pre-trained on the UCF101 dataset [62] for pose attribute detection.

For text preprocessing, we convert all words to lowercases and split sentences into words and remove punctuation using *wordpunct-tokenizer* method from NLTK toolbox[2]. Consequently, we achieve a vocabulary with 12,593 words from MSVD, 13,065 words from MSR-VTT, 13,560 words from ActivityNet and 7,296 words from SVCDV, where the word with the frequency less than 3 is removed. Furthermore, we utilize the one-hot vector to represent each word in our work.

For pose attribute detection, we elaborate the details regarding how to set the bounding box's size of each pose part. When obtaining a set of body joint positions by human pose detection in each frame, we firstly normalized them with respect to the person size. Hence, the position of each keypoint can be denoted as the relative offsets to the head in the person bounding box, i.e. $< dx, dy, arctan(dy/dx) >$, where "$arctan(dy/dx)$" refers to the orientations. For capturing more contextual information of each body joint, we set a padding parameter "$lside$" to crop more area of bounding box. Then, given a collection of body joint position $P = \{p_1, p_2, \cdots, p_n\}$, where $p_i =< dx_i, dy_i >$ refers to the coordinate of the i-th keypoint, and n denotes the total number of keypoints. To determine the bounding box's size of each body joint, i.e. left hand, right hand, upper body, bottom body and full body, we can crop each bounding box based on the corresponding coordinate of two points, *i.e.,* top-left point and bottom-right point, as the following:
Top-left point

$$\begin{aligned} &< \min(dx_1, dx_2, \cdots, dx_n) - lside, \\ &\max(dy_1, dy_2, \cdots, dy_n) + lside >, \end{aligned} \tag{15}$$

Bottom-right point

$$\begin{aligned} &< \max(dx_1, dx_2, \cdots, dx_n) + lside, \\ &\min(dy_1, dy_2, \cdots, dy_n) - lside > . \end{aligned} \tag{16}$$

Then, the position and size of each body part bounding box can be determined based on top-left point and bottom-right point, and each patch is resized to 224×224 pixels for matching the input layer of pose attribute detection network. In practice, we set "$lside = 20$" in our experiments.

For training our model, we add tag *BOS* and *EOS* to denote the begin and end of each sentence, respectively, which is aimed at making the length of sentences arbitrary. Then we input the *BOS* into the video decoder to start generating video descriptions. For pose attribute representation, we choose 256 and 50 most common words on UCF-101 and SVCDV as general sports action attributes vocabulary and volleyball professional attribute vocabulary, respectively. Then, we train our VGG-16 based attribute prediction model and achieve the final 306-way probabilities vector of attributes. The learning rates for training stage are set to 1×10^{-4}, 1×10^{-4}, 1×10^{-3}, 1×10^{-4} and for MSVD, MSR-VTT, AcitvityNet and SVCDV, respectively. The training batch size is set to 64 for MSVD/MSR-VTT/SVCDV, and 128 for ActivityNet. Meanwhile, we adopt Dropout for regularization with probability 0.5 on the input and output of encoder LSTMs and decoder LSTMs. For LSTMs in our model, the size of hidden states are set to 1,024, and size of embedding representation of video feature and words are set to 512. We select Adam optimizer [80] to update all the parameters in our model. We stop training our model until 200 epochs and the evaluation metric does not improve on the validation set. In the testing, we adopt the beam search strategy with the beam size 5. Our model is implemented using the TensorFlow [81] library with a single NVIDIA GTX 1080Ti GPU.

C. Compared Methods

In order to demonstrate the effectiveness of our proposed approach, we compare our model with following state-of-the-art methods: S2VT [6], LSTM-E [5], TA [7], HRNN [8], HRNE [82], DenseCap [26]. Following the experimental setting in [26], we compare existing video captioning models using ground truth proposals. Specifically, S2VT uses stack LSTMs in both encoder and decoder, and encodes a video using an RNN; LSTM-E utilizes a visual-semantic embedding; TA employs a temporal attention mechanism; HRNN uses a hierarchical decoder to generate captions; DenseCap generates multiple sentences and adopts a winner-take-all scheme to generate the final results[3]. Since not all the papers report full results, we only compare results on the test set. In addition, to examine the importance of different modality features in our proposed framework, we introduce three baseline models to compare with our full model, *i.e.,* "our baseline-1" only with VGG feature [77], "our baseline-2" only with VGG [77] and optical flow feature [61], "our baseline-3" only with VGG [77] and C3D feature [78]. While our full model simultaneously utilize three types of modality feature for video captioning, *i.e.,* VGG feature [77], optical flow feature [61] and C3D representation [78].

D. Result and Analysis

Results on General Video Datasets: To evaluate the generality of our model, we conduct experiments on the MSVD and MSR-VTT, which are both general video captioning datasets covering multiple topics. The results and comparisons can be found in Table I. As can be seen, the proposed method is able to achieve competitive results. On the MSVD dataset, the performance of our method is slightly no more 2% worse than the state-of-the-art methods across all metrics. Meanwhile, the performance of our model can get the second best on MSR-VTT across most of metrics. The results imply that better fine-grained motion representation and inter-object relationship features can effectively enhance the performance of general video captioning. Particularly, it is worth noting that our model

[2] http://www.nltk.org

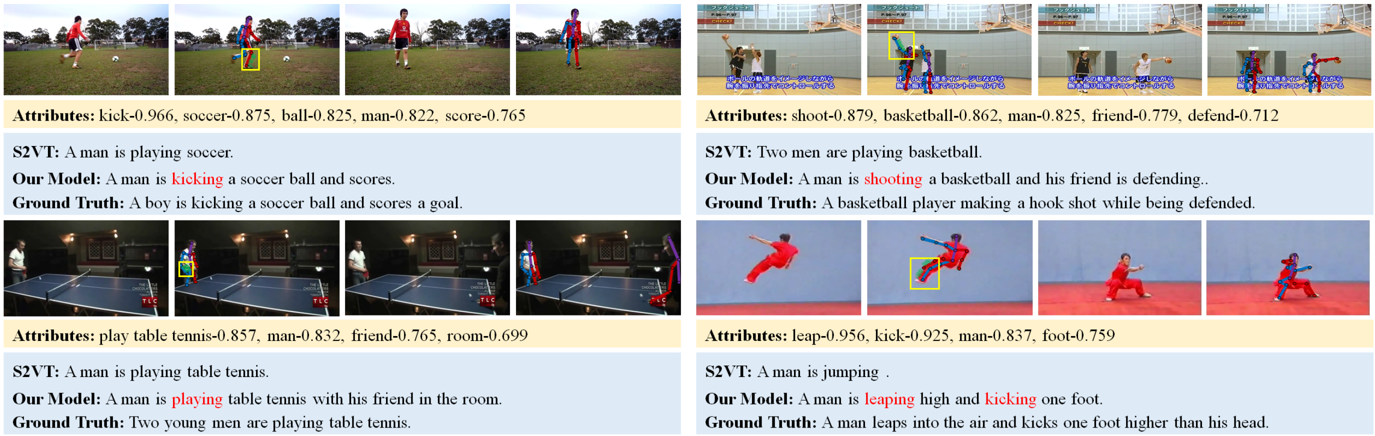
[3] In the experiments, the parameter settings of above-mentioned methods are adopted from corresponding papers.

TABLE I
PERFORMANCE COMPARISONS OF OUR FULL MODEL, BASELINE MODELS AND THE STATE-OF-THE-ART APPROACHES WITH DIFFERENT VIDEO FEATURES AND BACKBONES ON MSVD/MSR-VTT/ACITIVTYNET DATASET. (V) DENOTES VGG16, (G) DENOTES GOOGLENET, (R) DENOTES RESNET-152, (C) DENOTES C3D AND (O) DENOTES OPTICAL FLOW. "OURS W/O KNOWLEDGE" AND "OURS W/O RELATION" DENOTE OUR MODEL WITHOUT EXTRA SPORTS KNOWLEDGE AND GROUP RELATIONSHIP MODULE, RESPECTIVELY. ALL RESULTS ARE CITED FROM CORRESPONDING PAPERS. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD.

Methods	Modality	MSVD						MSR-VTT			ActivityNet					
		B@1	B@2	B@3	B@4	M	C	B@4	M	C	B@1	B@2	B@3	B@4	M	C
basic LSTM	R	80.6	69.3	59.7	49.6	32.7	69.9	-	-	-	-	-	-	-	-	-
S2VT [6]	V	-	-	-	-	29.2	-	-	-	-	-	-	-	-	-	-
S2VT [6]	V+O	-	-	-	-	29.8	-	31.4	25.7	**35.2**	-	-	-	-	-	-
S2VT [6]	C	73.5	59.3	48.2	36.9	29.8	48.6	31.4	25.7	**35.2**	20.4	9.0	4.6	2.6	7.9	21.0
LSTM-E [5]	V	74.9	60.9	50.6	40.2	29.5	-	-	-	-	-	-	-	-	-	-
LSTM-E [5]	C	75.7	62.3	52.0	41.7	29.9	-	-	-	-	-	-	-	-	-	-
LSTM-E [5]	V+C	78.8	66.0	55.4	45.3	31.0	-	-	-	-	-	-	-	-	-	-
TA [7]	R	**81.6**	70.3	**61.6**	**51.3**	**33.3**	**72.0**	-	-	-	-	-	-	-	-	-
TA [7]	V	-	-	-	-	-	-	35.6	25.4	-	-	-	-	-	-	-
TA [7]	C	74.1	58.9	48.2	36.6	29.4	48.1	36.1	25.7	-	-	-	-	-	-	-
TA [7]	G+C	80.0	64.7	52.6	42.2	29.6	51.7	-	-	-	-	-	-	-	-	-
TA [7]	V+C	-	-	-	-	-	-	**36.6**	**25.9**	-	-	-	-	-	-	-
HRNN [8]	V	77.3	64.5	54.6	44.3	31.1	62.1	-	-	-	-	-	-	-	-	-
HRNN [8]	C	79.7	67.9	57.9	47.4	30.3	53.6	-	-	-	19.5	8.8	4.3	2.5	8.0	20.2
HRNN [8]	V+C	81.5	**70.4**	60.4	49.9	32.6	65.8	-	-	20.2	-	-	-	-	-	-
HRNE [82]	G	78.4	66.1	55.1	43.6	32.1	-	-	-	-	-	-	-	-	-	-
HRNE+TA [82]	G	79.2	66.3	55.1	43.8	33.1	-	-	-	-	-	-	-	-	-	-
DenseCap [26]	C	-	-	-	-	-	-	-	-	-	26.5	13.5	7.2	4.0	9.5	**24.6**
Our baseline-1	V	77.6	64.9	55.7	46.5	30.9	63.2	33.9	25.1	29.7	21.2	10.6	5.9	2.7	8.2	21.5
Our baseline-2	V+O	78.2	66.2	56.9	47.3	31.5	65.6	34.6	25.3	30.6	21.9	11.2	6.2	3.1	8.5	22.2
Our baseline-3	V+C	79.9	68.6	58.9	49.1	32.1	69.2	35.7	25.5	32.1	23.6	11.7	6.9	3.9	9.1	22.9
Ours w/o knowledge	V+C+O	80.7	69.5	60.7	50.5	32.7	69.6	36.2	25.6	33.5	25.3	12.9	7.5	4.2	9.7	23.7
Ours w/o relation	V+C+O	80.9	69.6	61.0	50.7	33.1	70.2	36.5	25.7	33.8	25.9	13.2	7.7	4.5	9.8	24.2
Our full model	V+C+O	81.2	69.7	61.3	50.9	33.5	70.3	36.7	25.9	33.9	**26.6**	**13.9**	**8.2**	**4.9**	**9.9**	**24.6**

can be easily integrated with the compared methods for general video captioning. From the table, we find that S2YT performs much worse than other models in the MSVD dataset since it encodes long sequences of video by mean pooling. H-RNN performs slightly better due to its attentive object-level features. In addition, we have seen that utilizing a more powerful or advanced representation can improve the performance, thus some methods with ResNet features perform significantly better than C3D features (*e.g.,* TA with ResNet feature obtains the best performance than that with other features on MSVD dataset). Specifically, it should be pointed out that our method focuses on sports video captioning and also has the ability for general video captioning. Moreover, we illustrate in Figure 5 quite a few qualitative results, including example video clips and corresponding description sentences generated by our proposed model and the baseline method (*i.e.,* S2VT [6] in our experiments). As shown in Figure 5, we can see that our proposed model is able to produce more accurate, reasonable, logic language caption than the baseline. As an example, we can find that our modal can accurately capture the salient motion "kicking a soccer ball" in the top-left panel of the figure, while the S2VT baseline fails to identify the detailed individual action and only produces the sentence "playing the soccer". From the other three examples in the figure, we can draw similar conclusions. Because the videos in these two datasets have a small amount of group relationship, we mainly analyze the effect of pose attribute detection and trajectory clustering introduced in our model. We list the pose attributes detected in the four examples which center on personal actions, and also visualize the trajectory clusters that are assigned higher attention weights. Clearly, our model can extract the significant individual action and salient movement of person accurately in the video clips, and generate more refined and elaborate video description. Meanwhile, the semantic attributes captured by our model can be regarded as a better advanced semantic video features than low-level features in the task of video captioning. For instance, in the down-right panel of the figure, our method produces the caption "leaping high and kicking the foot" by exactly detecting the pose attributes "leap" and "kick" from the video clip, and precisely attend to the most significant trajectory clusters of "leaping" and "kicking" action. These better results demonstrate the effectiveness of motion representation module in our framework.

Results on ActivityNet Dataset: ActivityNet Captioning Dataset focuses on describing human actions in videos. We report results using action proposal module in our proposed model for segmenting the video and testing first three sentences in each video, because almost all the videos in the dataset contain at least three sentences. As noticed in Table I, our approach achieves the best performance in terms of all the metrics, such as BLEU@3, BLEU@4 and METEOR. Moreover, we observe that using attentive player's motion representation and group relationship features in our model achieves superior performance. In contrast, several state-of-the-art methods (*e.g.,* HRNN) encode the whole video features by mean pooling, which lost more motion details of player. Although DenseCap adopts action proposal and attention model, they cannot capture more pose-based motion details. Furthermore, we are aware that integrating optical flow information with RGB video can further improve the accuracies of action recognition and video captioning. And the performance would be further enhanced across all evaluation metrics by combining better video features and player's motion features. In addition, due to the limit of quantitative metrics, we show



Attributes: kick-0.966, soccer-0.875, ball-0.825, man-0.822, score-0.765

S2VT: A man is playing soccer.
Our Model: A man is kicking a soccer ball and scores.
Ground Truth: A boy is kicking a soccer ball and scores a goal.

Attributes: shoot-0.879, basketball-0.862, man-0.825, friend-0.779, defend-0.712

S2VT: Two men are playing basketball.
Our Model: A man is shooting a basketball and his friend is defending..
Ground Truth: A basketball player making a hook shot while being defended.

Attributes: play table tennis-0.857, man-0.832, friend-0.765, room-0.699

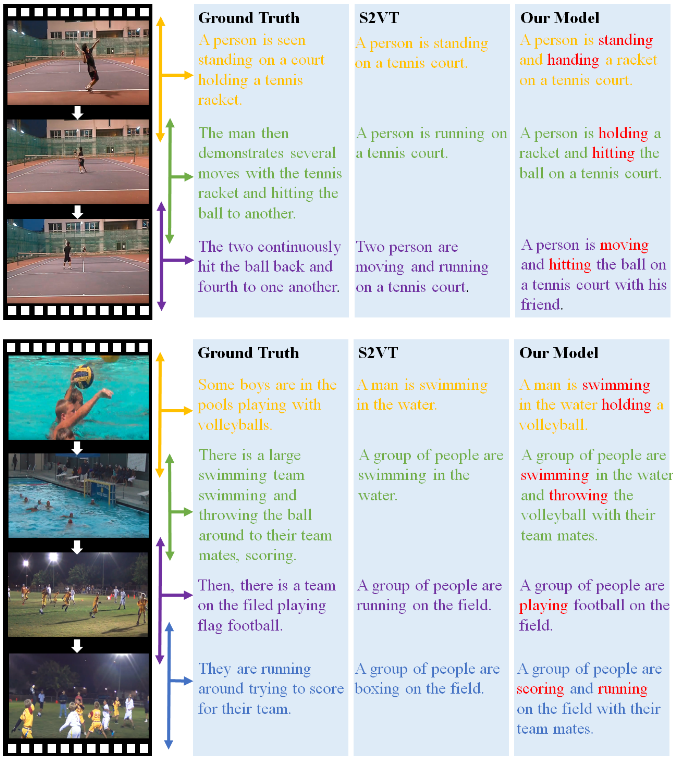
S2VT: A man is playing table tennis.
Our Model: A man is playing table tennis with his friend in the room.
Ground Truth: Two young men are playing table tennis.

Attributes: leap-0.956, kick-0.925, man-0.837, foot-0.759

S2VT: A man is jumping .
Our Model: A man is leaping high and kicking one foot.
Ground Truth: A man leaps into the air and kicks one foot higher than his head.

Fig. 5. Qualitative video captioning examples on the MSVD dataset, including the sample video clips and their corresponding captions produced by our proposed method, S2VT [6] and Ground Truth. Moreover, the detected pose semantic attributes are listed with their corresponding possibility, and the attended trajectory clusters are visualized and marked in the yellow bounding box. The highlights in red denote the important actions and activities in a sports event.

several qualitative examples in Figure 6. As depicted in the figure, it is obvious to see that our proposed framework can detect significant high-level pose semantic attributes to guide the video caption production. Given an example, the top panel of the figure, our model can generate descriptions with more fine-grained actions, *e.g.,* "handing a racket" and "hitting the ball" than the sentences produced by S2VT method. It can be attributed to that the pose attributes "handing", "moving" and "hitting" are precisely detected and directly employed to guide the captioning process in our proposed framework. Similarly, we can see from the down panel of the figure, our model can produce more descriptive and reasonable text sentences with crucial advanced semantic attributes, *e.g.,* "swimming", "holding", "throwing" and "scoring", which enrich the created description with more fine-grained contents. Meanwhile, the attentive trajectory can help our model to capture the critical action in the video, *e.g.,* "holding and throwing the volleyball". In addition, compared with the sentences produced by the baseline method S2VT, our model can generate more comprehensive details to describe the relationship between each person appeared in the videos, such as the terms of "with his friend" and "with their teammates", which can be attributed to the effectiveness of the group relationship module in our framework.

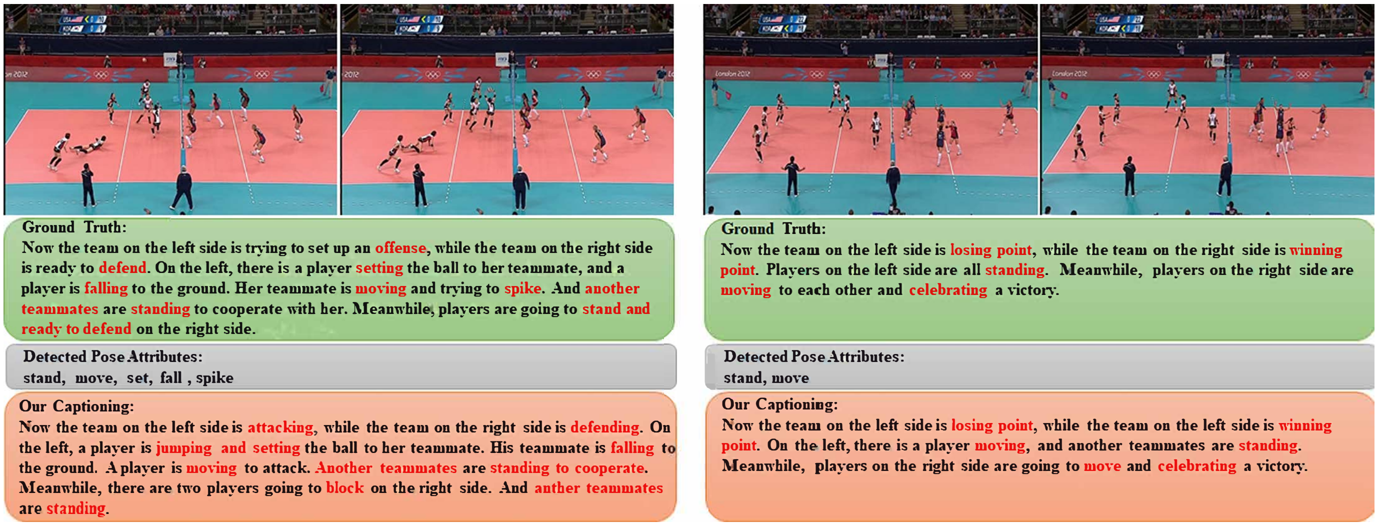


Ground Truth	S2VT	Our Model
A person is seen standing on a court holding a tennis racket.	A person is standing on a tennis court.	A person is standing and handing a racket on a tennis court.
The man then demonstrates several moves with the tennis racket and hitting the ball to another.	A person is running on a tennis court.	A person is holding a racket and hitting the ball on a tennis court.
The two continuously hit the ball back and fourth to one another.	Two person are moving and running on a tennis court.	A person is moving and hitting the ball on a tennis court with his friend.

Ground Truth	S2VT	Our Model
Some boys are in the pools playing with volleyballs.	A man is swimming in the water.	A man is swimming in the water holding a volleyball.
There is a large swimming team swimming and throwing the ball around to their team mates, scoring.	A group of people are swimming in the water.	A group of people are swimming in the water and throwing the volleyball with their team mates.
Then, there is a team on the filed playing flag football.	A group of people are running on the field.	A group of people are playing football on the field.
They are running around trying to score for their team.	A group of people are boxing on the field.	A group of people are scoring and running on the field with their team mates.

Fig. 6. Qualitative dense-video captioning results generated by our proposed method, S2VT [6], and Ground Truth on the ActivityNet dataset. The captions need to temporally localize and describe multiple events occurring simultaneously or overlap in time of a video. The highlights in red denote the important actions and activities in a sports event.

Results on SVCDV Dataset: We evaluate our method on the new SVCDV dataset that mainly contains sports videos. Table II reports the results and comparisons with the state-of-the-art and baseline methods. As it can be noticed in Table II, our approach improves plain techniques and achieves the state-of-the-art performance on SVCDV. We choose S2VT model as a baseline with only global video feature without attentive motion representation and group relationship modeling. The baseline achieves the worst performance that deteriorates our proposed framework by about 5% across all metrics. It obviously manifests that introducing attentive motion representation and group relationship modeling is rewarding to improve the performance of sports video captioning. Although HRNN and DenseCap have the ability to extract context information from the video, more accurate articulate action information is neglected. It strongly suggests that our framework is capable of generating sentences with more fine-grained motion representation and group relationship. Figure 7 illustrates quite a few qualitative captioning results on the test data of SVCDV datasets. As can be seen, our framework can capture more fine-grained action and activity details in the generated text description, and in more accordance with the ground truth. We show quite a few pose attributes detection results for the testing video clips, suggesting pose-based features and



Ground Truth:
Now the team on the left side is trying to set up an offense, while the team on the right side is ready to defend. On the left, there is a player setting the ball to her teammate, and a player is falling to the ground. Her teammate is moving and trying to spike. And another teammates are standing to cooperate with her. Meanwhile, players are going to stand and ready to defend on the right side.

Detected Pose Attributes:
stand, move, set, fall , spike

Our Captioning:
Now the team on the left side is attacking, while the team on the right side is defending. On the left, a player is jumping and setting the ball to her teammate. His teammate is falling to the ground. A player is moving to attack. Another teammates are standing to cooperate. Meanwhile, there are two players going to block on the right side. And anther teammates are standing.

Ground Truth:
Now the team on the left side is losing point, while the team on the right side is winning point. Players on the left side are all standing. Meanwhile, players on the right side are moving to each other and celebrating a victory.

Detected Pose Attributes:
stand, move

Our Captioning:
Now the team on the left side is losing point, while the team on the left side is winning point. On the left, there is a player moving, and another teammates are standing. Meanwhile, players on the right side are going to move and celebrating a victory.

Fig. 7. Qualitative video captioning results on the SVCDV dataset for a specific video clips. The highlights in red denote the important actions and activities in a sports event.

TABLE II
PERFORMANCE COMPARISONS OF OUR METHOD AND THE STATE-OF-THE-ART APPROACHES ON THE SVCDV DATASET AND THE COMPONENTS ANALYSIS OF OUR FRAMEWORK. "W/O" MEANS OUR MODEL WITHOUT SPECIFIC MODULE. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD.

Methods	B@4	R	M	C
S2VT [6]	25.62	45.26	21.55	1.96
HRNN [8]	24.53	44.97	20.96	2.05
DenseCap [26]	26.77	46.78	23.33	2.29
Ours w/o motion	25.71	45.12	21.56	1.88
Ours w/o pose	26.12	45.67	22.31	2.02
Ours w/o trajectory	27.25	46.52	22.79	2.15
Ours w/o relation	27.59	46.76	23.55	2.31
Ours w/o attention	28.38	47.78	23.79	2.36
Ours w/o knowledge	30.69	50.25	25.53	2.72
Ours full model	**31.76**	**51.62**	**26.07**	**2.91**

their corresponding semantic attributes can denote the accurate action and skeleton movement of players, and further improve the captioning performance with more specific and refined action information. Meanwhile, it is beneficial that such semantic attributes introduced into the model can be deemed as extra professional sports knowledge. Moreover, the attended trajectory clusters detected by our model, which highlight the significant motion of the crucial players and demonstrate their necessity in our motion representation module for improved the performance of fine-grained sports video captioning. However, our generated caption fails to describe a few of exactly players' actions or activities in several cases (*e.g.*, 'blocking' is mistaken as 'standing'), due to that several actions in the video share high similarities and occlusions in the video. More training data and advanced action detection model can be beneficial to better distinguishing these actions.

Human Evaluation of Sports Video Captioning: Because the conventional evaluation metrics cannot assess the results completely and accurately, we further conduct a human evaluation of sports video captioning on SVCDV to compare our method with the state-of-the-art approaches. We invite 20 volunteers in our experiments to rank the produced 1,000 description sentences from one to ten scores corresponding from worst to best w.r.t three criteria, *i.e.,* (1) Q1: whether the sentences generated is reasonable and readability? (2) Q2: whether the sentences generated can describe the sports event, crucial player's action, group activity and relationship accurately? (3) Q3: whether the sentences generated can help you to understand the match completely and professionally? As illustrated in Table III, the averaged scores of all volunteers apparently indicate that our model achieves the best performance across all three criteria. Specifically, our model obtains the scores 6.725/6.792/6.936 in terms of Q1/Q2/Q3, respectively, improving over the best state-of-the-art DenseCap [26] by about 0.2, 0.7, and 1.0. The relative more improved results w.r.t Q2/Q3 also demonstrate that our model focuses on generating fine-grained text sentences for a sports match, and has the capability to produce the complete description for significant sports highlights, player's action and group activity. It suggests that to some extent our approach can help people to understand the sports game, especially be utilized to assist the blind person in the future.

TABLE III
HUAMAN EVALUATION OF SPORTS VIDEO CAPTIONING WITH OUR METHOD AND THE STATE-OF-THE-ART APPROACHES ON THE SVCDV DATASET. HIGHER SCORE IS BETTER JUDGING BY HUMAN VOLUNTEERS. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD.

Methods	Q1	Q2	Q3
S2VT [6]	6.272	5.531	5.725
HRNN [8]	5.953	4.976	4.623
DenseCap [26]	6.525	5.767	5.969
Ours model	**6.725**	**6.692**	**6.936**

E. Efficiency Analysis

To analyze the efficiency of our proposed approach, we report the mean execution time of our framework and DenseCap [26] on Table IV in terms of sports video captioning on

TABLE IV
EFFICIENCY COMPARISONS OF OUR METHOD AND THE STATE-OF-THE-ART APPROACH ON THE SVCDV DATASET. “OURS W/ PICKING” REFERS TO OUR MODEL INCORPORATED WITH THE METHOD [32]. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD.

Methods	B@4	R	M	C	time(s)
DenseCap [26]	26.77	46.78	23.33	2.29	**15.6**
Ours full model	**31.76**	51.62	**26.07**	2.91	22.9
Ours w/ PickNet [32]	31.67	**51.69**	25.73	**2.93**	17.3

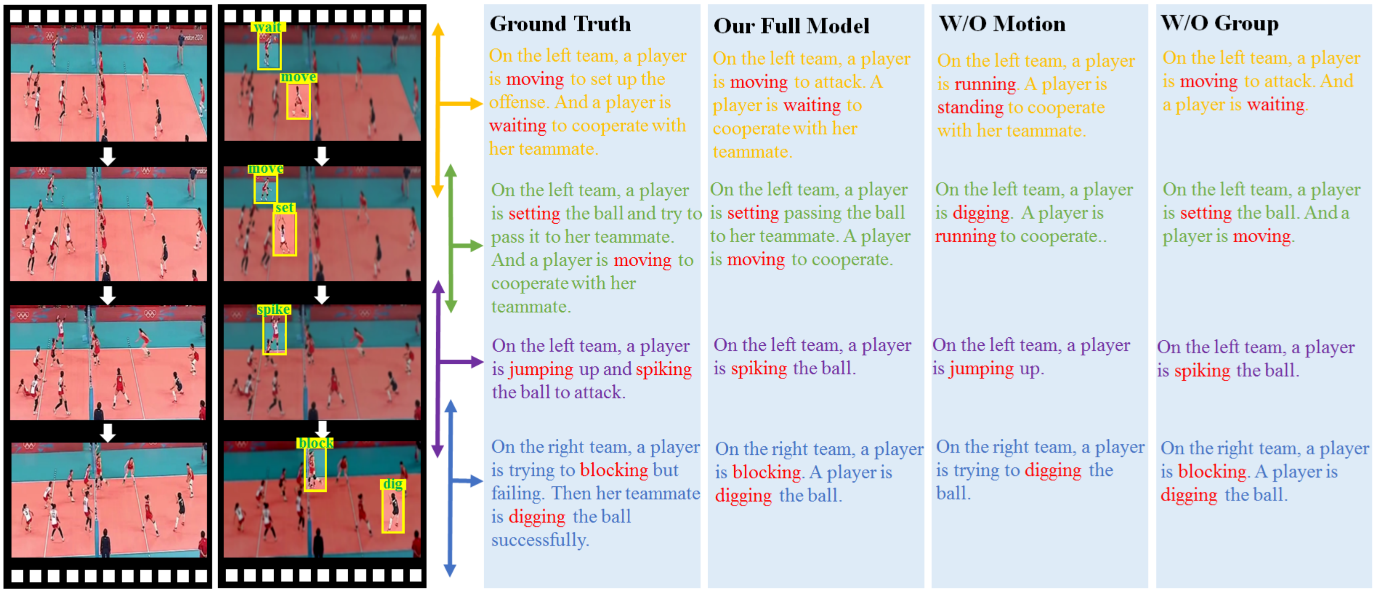
SVCDV. From the table, we can observe that our proposed method need more execution time than DenseCap, because our proposed framework require three types of video representation as input, *i.e.,* pose attribute feature, trajectory clustering feature, and group relationship feature. While DenseCap only need C3D feature of the given video sequence as input. Although our proposed framework achieves better performance than DenseCap by utilizing the well-designed video representations for sports video captioning, extracting these specific features through several sub-nets in our model, *e.g.,* Faster R-CNN, P-CNN, and Hierarchical Encoder-Decoder are exceedingly time consuming. How to get the trade-off balance between execution time and performance of generating captions is an important yet challenging issue. Limited work has devoted to this topic. Therefore, we will leave this as our future work. Furthermore, we examine the efficiency of our model incorporated with the method in [32]. The idea of the method in [32] is that designing a PickNet to select informative frames for video captioning in a reinforcement learning way. We incorporate the PickNet in our framework in the experiment, and report the performance in Table IV. From the table, we can see that our model with PickNet can effectively reduce the execution time from 22.9s to 17.3s, which slightly slow less than 2s compared with DenseCap. Meanwhile, it achieves competitive captioning performance compared with our full model. It denotes that the strategy in [32] by choosing import actions and selecting informative frames from the video is extremely promising and helpful to improve the efficiency of sports video captioning.

F. Ablation Study

To demonstrate the effectiveness of each component in our proposed model (*i.e.,* motion representation module, group relationship module, hierarchical encoder-decoder with attention mechanism, extra sports knowledge and different modality features), we have performed the ablative experiments for analysis.

Motion Representation Module. As can be seen in Table II, we firstly evaluate how much *Motion Representation Module* can help sports video captioning. In our work, we utilize both pose attribute feature and trajectory clustering feature as our motion representation. As a comparison, we only extract the whole video features through C3D model and LSTM (*i.e.,* Ours w/o motion in Table II), which obtains the worst performance. It is indicative of the motion representation module is the most paramount component in our model, and the performance would drop drastically if missing this module (*i.e.,* performances degrade more than 5% across all the metrics compared with the full model), because the raw video feature cannot extract more individual motion details from each frame. Comparing pose attribute feature with trajectory clustering, our method without pose attribute feature achieves worse performance than that without trajectory clustering. Obviously, it proves the pose attribute feature is more crucial than trajectory cluster, because the detected attributes can capture more semantic information to describe individual player motion. As depicted in Figure 8, we show an example of qualitative sports captioning results generated by our proposed full model, our model without attentive motion representation (referred to W/O Motion), our model without group relationship module (referred to W/O Group), and Ground Truth on SVCDV in a dense-captioning manner. Specifically, the description performed by our full model can capture the accurate crucial fine-grained individual action, *e.g.,* setting/spiking/digging/blocking in the Figure 8. While the sentences created by our model W/O Motion only extract a part of individual action and inaccurate action, *e.g.,* missing “blocking” and “spiking” in the third and fourth row, mistaking “setting” by “digging” in the second row. In summary, the proposed motion representation module is very beneficial to generate fine-grained sports video descriptions.

Group Relationship Module. In our framework, we add the GCN based *Group Relationship Module* to model team-level interaction among players, rather than replacing trajectory clustering. In fact, we still adopt the trajectory clustering in Motion Representation Module that is capable of capturing player-level dynamic movement information. While our newly proposed *Group Relationship Module* focuses on encoding the relation-aware feature of sports teams. We perform the ablation study to demonstrate the effectiveness of GCN, namely the proposed *Group Relationship Module* in our framework. In the ablation study, we mainly compare the performance of our full model with our model without Group Relationship Module (“Ours w/o Relation”). Table I and Table II illustrate the quantitative results on MSVD/MSR-VTT/ActivityNet and SVCDV, respectively. As shown in Table I, we can find that our full model incorporated with Group Relationship Module obtains slight improvement (less than 1% across all metrics) compared with our model without Relation. Because the videos of MSVD/MSR-VTT/ActivityNet generally contain single or few people so that the group relationship representation cannot make much difference for captioning performance. In contrast, as shown in Table II, our full model with Group Relationship Module can significantly outperform our model without Relation (“Ours w/o relation”) on SVCDV dataset. For example, our full model achieves a gain of more than 4% w.r.t BLEU@4 and ROUGE-L, and nearly 3% in terms of METEOR. It can be concluded that our proposed GCN based Group Relationship Module can effectively improve the performance of sports video captioning, especially for team sports (such as the volleyball games in SVCDV). Moreover, we can see a deteriorating of performance when we get rid of *Group Relationship Module* in our framework (*i.e.,* performances degrade more than 3% across all the metrics compared with the full model). It is the truth that much context and interactive information exist in sports games, and the scene



Ground Truth	Our Full Model	W/O Motion	W/O Group
On the left team, a player is moving to set up the offense. And a player is waiting to cooperate with her teammate.	On the left team, a player is moving to attack. A player is waiting to cooperate with her teammate.	On the left team, a player is running. A player is standing to cooperate with her teammate.	On the left team, a player is moving to attack. And a player is waiting.
On the left team, a player is setting the ball and try to pass it to her teammate. And a player is moving to cooperate with her teammate.	On the left team, a player is setting passing the ball to her teammate. A player is moving to cooperate.	On the left team, a player is digging. A player is running to cooperate..	On the left team, a player is setting the ball. And a player is moving.
On the left team, a player is jumping up and spiking the ball to attack.	On the left team, a player is spiking the ball.	On the left team, a player is jumping up.	On the left team, a player is spiking the ball.
On the right team, a player is trying to blocking but failing. Then her teammate is digging the ball successfully.	On the right team, a player is blocking. A player is digging the ball.	On the right team, a player is trying to digging the ball.	On the right team, a player is blocking. A player is digging the ball.

Fig. 8. Qualitative sports captioning results generated by our proposed full model, our model without attentive motion representation (denoted as W/O Motion), our model without group relationship module (denoted as W/O Group), and Ground Truth on SVCDV in a dense-captioning manner. The attentive important individual action is marked in the yellow bounding box. The highlights in red denote important actions and activities in a sports event.

graph constructed in our module is able to dynamically discover and model them for elaborate captioning. In addition, Figure 8 illustrates a few of qualitative sports video captioning results. Compared with the generated description of our model without Group Relationship Module (W/O Group), our full model generates more reasonable and semantically informative captions that describe the interaction relationship between players, such as “a player is waiting to cooperate with her teammate” and “passing the ball to her teammate”. As an example, in the second row of the Figure 8, the sentences generated by our model W/O Group only capture the players’ action (setting and moving), missing the cooperation between two teammates, *i.e.,* “a player is trying to pass the ball to her teammate”. As we all know, team-level sports are the most popular and remarkable sports events, such as basketball and soccer. However, most of previous methods directly adopt conventional video captioning approach and neglect the relationship between players, which is arduous to address the task of sports video captioning. For such team sports events, various group-level movements and actions of players can make tactics and the game change frequently and dramatically. Hence, the proposed *Group Relationship Module* models the interaction between players by a scene graph and captures relation-aware representation, which is rewarding to sports video captioning.

Hierarchical Encoder-Decoder Architecture with Attention Mechanism. We perform the experiment to demonstrate the effectiveness of our proposed *Hierarchical Encoder-Decoder Architecture*. Because the proposed framework is mainly focus on generating paragraph description, we choose S2VT [6] that only utilizes one-layer RNN as the compared method on ActivityNet and SVCDV. As shown in Table I and Table II, we can clearly see our full model significantly outperforms S2VT across all the metrics, *e.g.,* our model achieves a gain of nearly 6% w.r.t BELU@4, METEOR and ROUGE-L on SVCDV, suggesting that our proposed hierarchical LSTMs is able to preserve the long-dependency of video streams and generate better descriptions than the single layer RNN model, *i.e.,* S2VT. It also denotes that two-layer LSTM based encoder and decoder can capture more context representation from videos and give coherent paragraph captions. Especially, we can see significant improvements by introducing *attention mechanism* into our encoder-decoder architecture, denoting that attention mechanism is exceedingly useful for sports video captioning. Since the key players invariably play a considerable role for the sports event, as shown in the yellow boxes in Figure 8. Meanwhile, it also reveals that fusing features of video by mean-pooling is not a wise choice.

Importance of Extra Sports Knowledge. To exploit the effectiveness of incorporating extra professional sports knowledge into our framework, we perform the ablative experiments and analysis. Table I and Table II illustrate the quantitative results on MSVD/MSR-VTT/ActivityNet and SVCDV, respectively. From the tables, we can clearly observe that our full model incorporated with extra sports knowledge outperforms our model without professional knowledge and other baseline models. Specifically, our full model with knowledge can obtain slightly better performance on MSVD and MSR-VTT, because these two datasets contain a wide variety of different videos rather than mainly sports videos. While our model with sports knowledge can obviously improve the performance of our model without knowledge (“Ours w/o knowledge”) by about 1% w.r.t BLEU@1 and BLEU@2 on Activity, and more than 1% w.r.t BLEU@4, ROUGE-L and CIDEr on SVCDV. The reason why our model with extra knowledge can improve the performance on such sports and action-centered video datasets is that our pose attribute detection can map human poses to semantic attributes vector, and these attributes learned from sports textual data are beneficial to sports video captioning. Semantic attributes can be regarded as the extra professional sports knowledge collected from the datasets and Internet, and a bridge between visual data and textual data. Such extra knowledge can significantly enrich the vocabulary in the generated description, and can be flexibly applied in different sports events, *e.g.,* basketball and baseball.

Effect of Different Modality Features. Moreover, we also conduct further experiments to examine the effect of different modality features, i.e. VGG, optical flow and C3D as the input features of our framework. The experimental results can be found in the Table I. Note that in our proposed framework, we adopt VGG-16 pre-trained on the ImageNet as our backbone to extract visual appearance features, which is mainly used in pose attribute detection and trajectory clustering in Motion Representation Module and Group Relationship Module. We adopt VGG-16 for fair comparisons with the state-of-the-art methods since most of them employ VGG-16. Furthermore, we employ optical flow features captured by the motion network [61] pre-trained on the UCF101 dataset [62], which is utilized in pose attribute detection to enhance the temporal representation of human pose. In addition, we leverage the C3D model [78] pre-trained on the Sports-1M dataset [79] to extract the video temporal representation, which is used in Action Proposal Module as preprocessing and Encoder-Decoder. Especially, we totally adopt four types of features, i.e. pose attribute feature F_{pose_att} (optical flow and VGG), trajectory clustering features F_{tra} (VGG), group relationship feature F_{rel} (VGG) and frame feature F_{frame} (VGG and C3D) as input of Encoder-Decoder in our full model. Therefore, the VGG feature is the main representation used in each module in our proposed framework, and optical flow and C3D feature are optional representations utilized in certain modules for improved performance. Depended on the above discussion, we introduce three baseline models to compare with our full model: "our baseline-1" only with VGG feature, "our baseline-2" only with VGG and optical flow feature, "our baseline-3" only with VGG and C3D feature. As shown in Table I, we can apparently find that our full model with three types of modality feature (VGG+C3D+Optical Flow) can achieve the best performance compared with other baseline models, suggesting more modality features can be encoded as rich informative representations and hence resulting in better video captioning. Compared to our baseline-1, our baseline-2 can improve the performance by about $1\%\sim2\%$ in terms of all metrics on MSVD/MSR-VTT/ActivityNet, demonstrating the optical flow representation is beneficial to capturing more consistently temporal information of human pose and leading to better captioning results. Meanwhile, the performance of our baseline-3 is clearly superior to our baseline-1 and baseline-2, which should be attributed to the effectiveness and importance of C3D features used in Action Proposal Module and Encoder-Decoder. For example, our baseline-3 outperforms our baseline-1 about $2\%\sim6\%$ in terms of all metrics on MSVD. It denotes that C3D representation can preserve more temporally sequential information from a clip of video. Above all, it is indispensable to simultaneously employ visual appearance features of each frame, optical flow and 3D convolutional representations of video clips for the task of video captioning.

VI. Conclusion

In this study, we propose a novel deep framework for sports video captioning based on jointly capturing attentive motion representation and group relationship feature. Through extracting human pose attribute, trajectory clustering, and group relationship representation, our model is capable of describing more fine-grained information corresponding to dynamic movement of players/teams and various interactions in a sports game. We have evaluated our model on three widely-adopted public datasets and a newly introduced *Sports Video Captioning Dataset-Volleyball*. The experimental results have demonstrated the effectiveness of our framework that achieves competitive or superior performance compared with the current state-of-the-art models.

References

[1] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, "Live sports event detection based on broadcast video and web-casting text," in *Proc. MM*. ACM, 2006.

[2] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu, "A mid-level representation framework for semantic sports video analysis," in *Proc. MM*. ACM, 2003.

[3] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. ICCV*. IEEE, 2013.

[4] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *Proc. ICCV*. IEEE, 2013.

[5] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework." in *Proc. AAAI*, 2015.

[6] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. ICCV*. IEEE, 2015.

[7] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proc. ICCV*. IEEE, 2015.

[8] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. CVPR*. IEEE, 2016.

[9] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proc. CVPR*. IEEE, 2016.

[10] R. Pasunuru and M. Bansal, "Multi-task video captioning with video and entailment generation," in *Proc. ACL*. Association for Computational Linguistics, 2017.

[11] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proc. CVPR*. IEEE, 2017.

[12] Y. Dong, H. Su, J. Zhu, and B. Zhang, "Improving interpretability of deep neural networks with semantic information," in *Proc. CVPR*. IEEE, 2017.

[13] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proc. CVPR*. IEEE, 2017.

[14] C. Hori, T. Hori, T.-Y. Lee, K. Sumi, J. R. Hershey, and T. K. Marks, "Attention-based multimodal fusion for video description," in *Proc. ICCV*. IEEE, 2017.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[17] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*. Springer, 2016.

[18] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proc. ICCV*. IEEE, 2015.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. CVPR*. IEEE, 2017.

[20] G. Zhu, Q. Huang, C. Xu, Y. Rui, S. Jiang, W. Gao, and H. Yao, "Trajectory based event tactics analysis in broadcast sports video," in *Proc. MM*. ACM, 2007.

[21] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*. IEEE, 2011.

[22] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015.

[23] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. CVPR*. IEEE, 2016.

[24] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. CVPR*. IEEE, 2017.

[25] M. Qi, Y. Wang, A. Li, and J. Luo, "Sports video captioning by attentive motion representation based hierarchical recurrent neural networks," in *1st International Workshop on Multimedia Content Analysis in Sports (MMSports' 18)*. ACM, October 2018.

[26] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. ICCV*. IEEE, 2017.

[27] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[28] L. Baraldi, C. Grana, and R. Cucchiara, "Hierarchical boundary-aware neural encoder for video captioning," in *Proc. CVPR*. IEEE, 2017.

[29] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue, "Weakly supervised dense video captioning," in *Proc. CVPR*. IEEE, 2017.

[30] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*. IEEE, 2015.

[31] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proc. CVPR*. IEEE, 2018.

[32] Y. Chen, S. Wang, W. Zhang, and Q. Huang, "Less is more: Picking informative frames for video captioning," in *Proc. ECCV*. Springer, 2018.

[33] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, "M3: Multimodal memory modelling for video captioning," in *Proc. CVPR*. IEEE, 2018.

[34] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," in *CVPR*. IEEE, 2018.

[35] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proc. CVPR*. IEEE, 2018.

[36] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. CVPR*. IEEE, 2018.

[37] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Jointly localizing and describing events for dense video captioning," in *Proc. CVPR*. IEEE, 2018.

[38] H. Yu, S. Cheng, B. Ni, M. Wang, J. Zhang, and X. Yang, "Fine-grained video captioning for sports narrative," in *Proc. CVPR*. IEEE, 2018.

[39] W.-L. Lu, J.-A. Ting, K. P. Murphy, and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," in *Proc. CVPR*. IEEE, 2011.

[40] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1704–1716, 2013.

[41] J. Liu, P. Carr, R. T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *Proc. CVPR*. IEEE, 2013.

[42] X. Wang, V. Ablavsky, H. B. Shitrit, and P. Fua, "Take your eyes off the ball: Improving ball-tracking by focusing on team play," *Computer Vision and Image Understanding*, vol. 119, pp. 102–115, 2014.

[43] X. Wang, E. Tretken, F. Fleuret, and P. Fua, "Tracking interacting objects using intertwined flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2312–2326, 2016.

[44] A. Maksai, X. Wang, and P. Fua, "What players do with the ball: A physically constrained interaction modeling," in *Proc. CVPR*. IEEE, 2016.

[45] M. S. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in *Proc. ECCV*. Springer, 2018.

[46] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool, "stagnet: An attentive semantic rnn for group activity recognition," in *Proc. ECCV*. Springer, September 2018.

[47] D. Annane, J. C. Chevrolet, S. Chevret, and J. C. Rapha?l, "Two-stream convolutional networks for action recognition in videos," in *Proc. NeurIPS*, 2014.

[48] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proc. CVPR*. IEEE, 2016.

[49] T. Shu, S. Todorovic, and S. C. Zhu, "Cern: Confidence-energy recurrent network for group activity recognition," in *Proc. CVPR*. IEEE, 2017.

[50] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *Proc. CVPR*. IEEE, 2011.

[51] Y.-L. Lin, V. I. Morariu, and W. Hsu, "Summarizing while recording: Context-based highlight detection for egocentric videos," in *Proc. CVPR Workshop*. IEEE, 2015.

[52] R. E. Kalman *et al.*, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[53] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[54] H. Morimitsu, I. Bloch, and R. M. Cesar-Jr, "Exploring structure for long-term tracking of multiple objects in sports videos," *Computer Vision and Image Understanding*, vol. 159, pp. 89–104, 2017.

[55] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *Proc. CVPR*. IEEE, 2017.

[56] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Daps: Deep action proposals for action understanding," in *Proc. ECCV*. Springer, 2016.

[57] K. Tang, B. Yao, L. Fei-Fei, and D. Koller, "Combining the right features for complex event recognition," in *Proc. ICCV*. IEEE, 2013.

[58] R. Girshick, "Fast r-cnn," in *Proc. ICCV*. IEEE, 2015.

[59] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. NeurIPS*, 2017.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012.

[61] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. CVPR*. IEEE, 2015.

[62] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[63] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering discriminative action parts from mid-level video representations," in *Proc. CVPR*. IEEE, 2012.

[64] X. Wu, G. Li, Q. Cao, Q. Ji, and L. Lin, "Interpretable video captioning via trajectory structured localization," in *Proc. CVPR*. IEEE, 2018.

[65] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. CVPR*. IEEE, 2015.

[66] M. Qi, W. Li, Z. Yang, Y. Wang, and J. Luo, "Attentive relational networks for mapping images to scene graphs," in *Proc. CVPR*. IEEE, 2019.

[67] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017.

[68] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. CVPR*. IEEE, 2017.

[69] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proc. CVPR*. IEEE, 2017.

[70] M. Qi, Y. Wang, and A. Li, "Online cross-modal scene retrieval by binary representation and semantic graph," in *Proc. MM*. ACM, 2017.

[71] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proc. CVPR*. IEEE, 2016.

[72] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proc. CVPR*. IEEE, 2015.

[73] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. CVPR*. IEEE, 2015.

[74] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. ACL*. Association for Computational Linguistics, 2002.

[75] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. Statistical Machine Translation Workshop*, 2014.

[76] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proc. ACL*. Association for Computational Linguistics, 2004.

[77] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[78] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. ICCV*. IEEE, 2015.

[79] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. CVPR*. IEEE, 2014.

[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[81] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning." in *Proc. OSDI*, 2016.

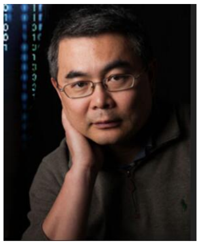
[82] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *Proc. CVPR*. IEEE, 2016.

Mengshi Qi received the B.S. and M.S. degree in computer science from Beijing University of Posts and Telecommunications and Beihang University, Beijing, China, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University. His current research interests include machine learning, computer vision, scene understanding and multimedia retrieval.

Yunhong Wang (M'98-SM'15) received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 1989, and the M.S. and Ph.D. degrees from Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively, all in electronics engineering.

She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, where she is also the Director of Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media. Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing. She is a Fellow of the IAPR.

Annan Li received the B.S. and M.S. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2003 and 2006, and the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. He worked in Singapore as a scientist with Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR) and as a postdoctoral research fellow at National University of Singapore, respectively. He currently works at the School of Computer Science and Engineering, Beihang University. His research interests include computer vision, pattern recognition, and statistical learning. He is a member of IEEE.



Jiebo Luo (S'93-M'96-SM'99-F'09) joined the Department of Computer Science at the University of Rochester in 2011, after a prolific career of over 15 years with Kodak Research. He has authored over 400 technical papers and holds over 90 U.S. patents. His research interests include computer vision, machine learning, data mining, social media, and biomedical informatics. He has served as the Program Chair of ACM Multimedia 2010, IEEE CVPR 2012, ACM ICMR 2016, and IEEE ICIP 2017, and on the Editorial Boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON BIG DATA, Pattern Recognition, Machine Vision and Applications, and ACM Transactions on Intelligent Systems and Technology. He is a Fellow of the IEEE, ACM, AAAI, SPIE and IAPR.