Structured Multi-Level Interaction Network for Video Moment Localization via Language Query

Hao Wang[1], Zheng-Jun Zha[1*], Liang Li[2], Dong Liu[1], Jiebo Luo[3]
[1]University of Science and Technology of China,
[2]Institute of Computing Technology, Chinese Academy of Sciences, [3]University of Rochester
whqaz@mail.ustc.edu.cn, {zhazj,dongeliu}@ustc.edu.cn, liang.li@ict.ac.cn, jluo@cs.rochester.edu

Abstract

*We address the problem of localizing a specific moment described by a natural language query. Existing works interact the query with either video frame or moment proposal, and neglect the inherent structure of moment construction for both cross-modal understanding and video content comprehension, which are the two crucial challenges for this task. In this paper, we disentangle the activity moment into boundary and content. Based on the explored moment structure, we propose a novel Structured Multi-level Interaction Network (SMIN) to tackle this problem through multi-levels of cross-modal interaction coupled with content-boundary-moment interaction. In particular, for cross-modal interaction, we interact the sentence-level query with the whole moment while interacting the word-level query with content and boundary, as in a coarse-to-fine manner. For content-boundary-moment interaction, we capture the insightful relations between boundary, content, and the whole moment proposal. Through multi-level interactions, the model obtains robust cross-modal representation for accurate moment localization. Extensive experiments conducted on three benchmarks (*i.e.*, Charades-STA, ActivityNet-Captions, and TACoS) demonstrate the proposed approach outperforms the state-of-the-art methods.*

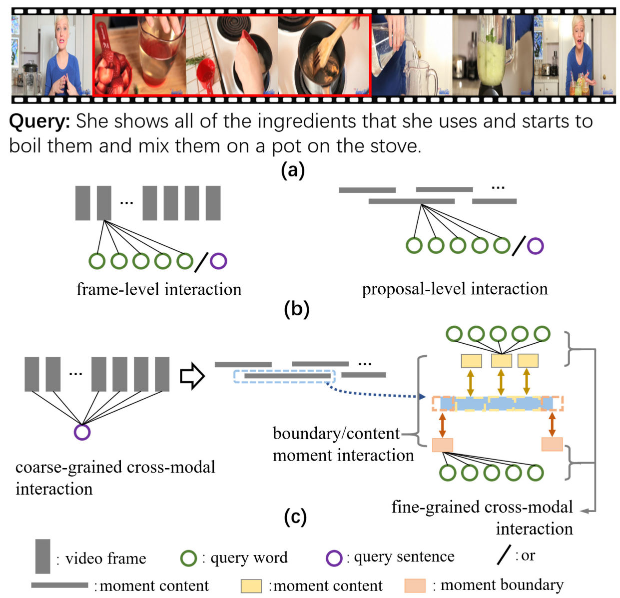

Figure 1. (a): An example of video moment localization queried by a natural language sentence. (b): Existing approaches simply conduct vision-language interaction at frame-level or proposal-level. (c): Our method first fuses the sentence with the video frame as a coarse-grained cross-modal interaction. Then we conduct fine-grained vision-language interaction between words and boundary/content of a moment proposal. Meanwhile, we conduct structured moment interaction to explore the relations between boundary, content, and the moment. (Best viewed in color)

1. Introduction

With the wide popularity of online videos, automatically understanding and analyzing the video content has drawn increasing attention. Recently, due to the limitation of the pre-defined action categories and the flexibility of using a natural language sentence for describing the activity in videos, video moment localization is proposed in the works [1, 7]. Its aim is to localize a temporary segment from an untrimmed video, containing the activity described by the given language query.

For this task, there are two key challenges: (1) cross-modal understanding between the language query and complicated video content, and (2) elaborate video content comprehension for localizing the target moment in videos with complex backgrounds. Existing works locate the moment by interacting the query with either video frame representation [4, 5, 6, 10, 12, 16, 33, 42, 45], or moment proposal representation [1, 7, 9, 19, 20, 43, 44]. For vision-language interaction, these works neglect the inherent structure of

*Corresponding author.

the moment that one is constructed by content and boundary. For video content comprehension, they directly utilize the features of video frame or proposal as the moment representation. These methods miss the discriminative information contained in moment boundary and content and the fine-grained correlation between them to query; therefore, they predict coarse boundary and misaligned moment. In fact, moment content, boundary, and the whole moment have different representation ability to represent a moment, and thus it is non-trivial to (1) perform cross-modal interaction between them and language query respectively for comprehensive cross-modal understanding; (2) conduct the content-boundary-moment interaction (termed as structured moment interaction in this work) for elaborate video content comprehension.

Motivated by the above observations, this paper proposes a novel Structured Multi-level Interaction Network (SMIN) for video moment localization by incorporating multiple levels of vision-language interaction and moment structured interaction into a joint procedure. First, we design the multiple levels of vision-language interaction for detailed vision-language understanding. As illustrated in Figure 1(b)(c), in contrast to previous works that simply use frame-level or proposal-level interaction to fuse video and query, we leverage the inherent moment structure and introduce a coarse-to-fine cross-modal interaction. In detail, the coarse-grained sentence representation is interacted with the video frame in the backbone before proposal generation, while the fine-grained word representation is interacted with boundary and content separately. Second, based on the inherent structure of a moment, we introduce the structured moment interaction by exploiting the structural relationships between content, boundary, and the whole moment. This interaction helps perform the elaborate video comprehension. Finally, we build the content unit, boundary unit, and moment unit for incorporating the multi-level cross-modal interaction and structured moment interaction as a structured multi-level interaction procedure to extract robust moment representation for accurate moment localization.

To summarize, our contributions are as follows:

- We disentangle the inherent structure of moment that one is constructed with boundary and content, and leverage this structure for comprehensive vision-language understanding and elaborate video comprehension.

- We propose a novel Structured Multi-level Interaction Network (SMIN) to incorporate fine-grained cross-modal interaction and detailed structured moment interaction into a joint procedure with the disentangled moment structure.

- We conduct experiments on three popular benchmarks to verify the effectiveness of our approach, which performs superior to the state-of-the-art methods.

2. Related Works

2.1. Temporal Action Detection

Video temporal action detection aims to jointly predict the action label and localize the start and end boundaries of an action proposal in an untrimmed video. It has achieved great progress [2, 3, 8, 15, 23, 28, 27, 34, 35, 47]. Existing approaches can be roughly divided into two-stage approaches [3, 8, 47] consisting of proposal generation and proposal classification and one stage approaches [15, 35] directly detecting action instances without proposal generation. Due to the diverse contents in various videos, predefined action categories cannot completely cover the activities in videos. Therefore, using a language sentence for describing the activity has attracted increasing attention.

2.2. Moment Localization

Moment Localization was proposed by [1, 7], which aims to predict the start and end boundaries of the activity described by a given language query within a video. This task is very challenging since it needs deep vision-language interaction [17, 18, 21, 39, 40] and complete video comprehension [30, 41, 46].

As for vision-language interaction, existing works either used early fusion at frame-level [4, 6, 10, 22, 24, 31, 32, 38, 45] or late fusion at candidate-level [1, 7, 19, 20, 43]. Chen *et al.* [4] incorporated the frame-by-word interactions across video-sentence modalities towards this task. Authors of [6, 22] employed the visual-language attention to encode frame feature with multi-modal context. Zeng *et al.* [38] fused the query with frame features at different temporal scales. Authors of [24] interacted different semantic phrase with video frames. Gao *et al.* [7] combined the information from query sentence and moment proposal and used alignment and regression loss for activity location refinement.

These works neglected the inherent structure of moment for cross-modal interaction, as well as for video comprehension. Hendricks *et al.* [1] concatenated the global video feature to each proposal while Gao *et al.* [7] concatenated the preceding and following clips as context to the central clip. Authors of [42] modeled moment-wise temporal relations via the iterative graph adjustment network. Zhang *et al.* [43] utilized the temporal context from adjacent moments through a two-dimensional temporal map. Wang *et al.* [31] implemented the complementarity of frame-level and candidate-level representations. Authors of [24] took relations between semantic phrases into account through non-local block.

In contrast to existing works, we explore the structure of the moment and disentangle it into content and bound-

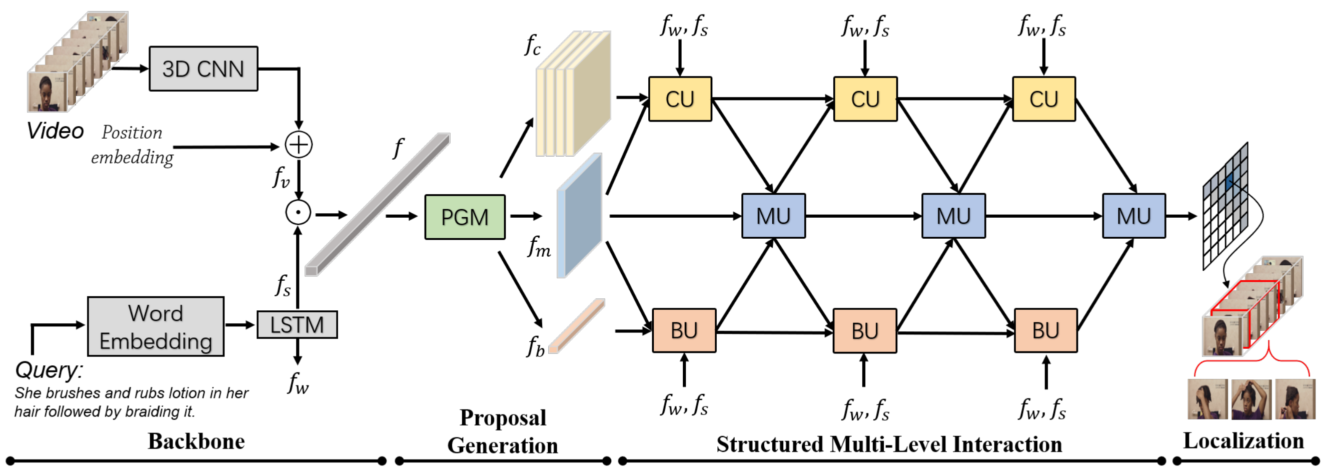

Figure 2. Architecture of the proposed SMIN. We first extract the video feature, word-level and sentence-level query feature and conduct coarse-grained cross-modal interaction. Then we explore the structure of the moment and generate content, boundary, and moment feature in the proposal generation module. Next, we conduct fine-grained cross-modal interaction, together with structured moment interaction in the structured multi-level interaction module. We finally predict the moment most relevant to the query.

ary. Based on the insightful structure, we implement vision-language interaction in a coarse-to-fine manner and model the relationship between content, boundary, and moment as structured moment interaction for video understanding.

3. Proposed Method

3.1. Overview

Given an untrimmed video V and a language query Q, this task aims to retrieve the best matching temporary segment with a start and end timestamps (τ_s, τ_e) referring to the sentence query. Formally, we denote the input video as $V = \{v_i\}_{i=0}^{T_v}$ and language query as $Q = \{q_n\}_{n=0}^{N_q}$, where v_i is the i_{th} frame and w_n is the n_{th} word. T_v and N_q is the length of video and sentence, respectively.

Figure 2 illustrates the network architecture of the proposed structured multi-level interaction network (SMIN), which consists of four components: (1) backbone for video and query encoding; (2) proposal generation module; (3) structured multi-level interaction module; (4) moment localization module. Specifically, we first extract a sequence of frame-wise video features added with the position embedding of each frame. Meanwhile, we derive word-level and sentence-level features based on the query through a recurrent neural network. Second, we fuse the frame-wise video feature with the word-level query feature as a coarse-grained cross-modal interaction. Next, a proposal generation module generates the moment proposal feature, together with moment content and moment boundary feature. Then, we conduct fine-grained cross-modal interaction and structured moment interaction in a structured multi-level interaction module. We finally localize the most relevant video moment conforming to the query through thoroughly cross-modal interaction and structured moment interaction, based on the robust moment features.

3.2. Video and Query Encoding

For the video, we first extract a sequence of frame-wise video feature by a pre-trained 3D CNN feature extractor. We then uniformly sample T frames of feature over the sequence to obtain a fixed-length of video features sequence $\mathbf{f}_v^{'} = \{\mathbf{f}_{v_i}^{'}\}_{i=0}^{T} \in \mathbb{R}^{T\times d}$, where d denotes the feature dimension. Next, we append the embedding of temporal position to each frame feature, as done in [24]. Thus each frame is aware of its relative position in the video. The position embedding $\mathbf{f}_{pos,i} = \mathbf{W}_{pos}\mathbf{P}_i$, where $\mathbf{W}_{pos} \in \mathbb{R}^{d\times T}$ is a learnable embedding matrix and $\mathbf{P}_i \in \mathbb{R}^{T}$ is the one-hot temporal position vector of each frame. Finally, we obtain the sequence of frame-wise representation: $\mathbf{f}_v = \mathbf{f}_v^{'} + \mathbf{f}_{pos}$.

For the language query, we first extract the embedding vector of each word through the Glove [25] word2vec model. Then, we employ a two-layer bidirectional LSTM network to extract the feature of the query. We compute the sentence-level query feature $\mathbf{f}_s = [\overrightarrow{\mathbf{h}}_{N_q}; \overleftarrow{\mathbf{h}}_1] \in \mathbb{R}^d$ by concatenating the last hidden state of both forward and backward LSTM and calculate a sequence of word-level features $\mathbf{f}_w = \{\mathbf{f}_{w_i}\}_{i=0}^{N_q} \in \mathbb{R}^{N_q\times d}$, where $\mathbf{f}_{w_i} = [\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]$ through the concatenation of hidden states in both directions.

Next, we introduce a coarse-grained cross-modal interaction between the extracted video frame-wise feature and sentence-level query feature, and get the fused features $\mathbf{f} = \{\mathbf{f}_i\}_{i=0}^{T} \in \mathbb{R}^{T\times d}$, where $\mathbf{f}_i = \mathbf{f}_s \odot \mathbf{f}_{v_i}$, and $\odot$ is the Hadamard product operator.

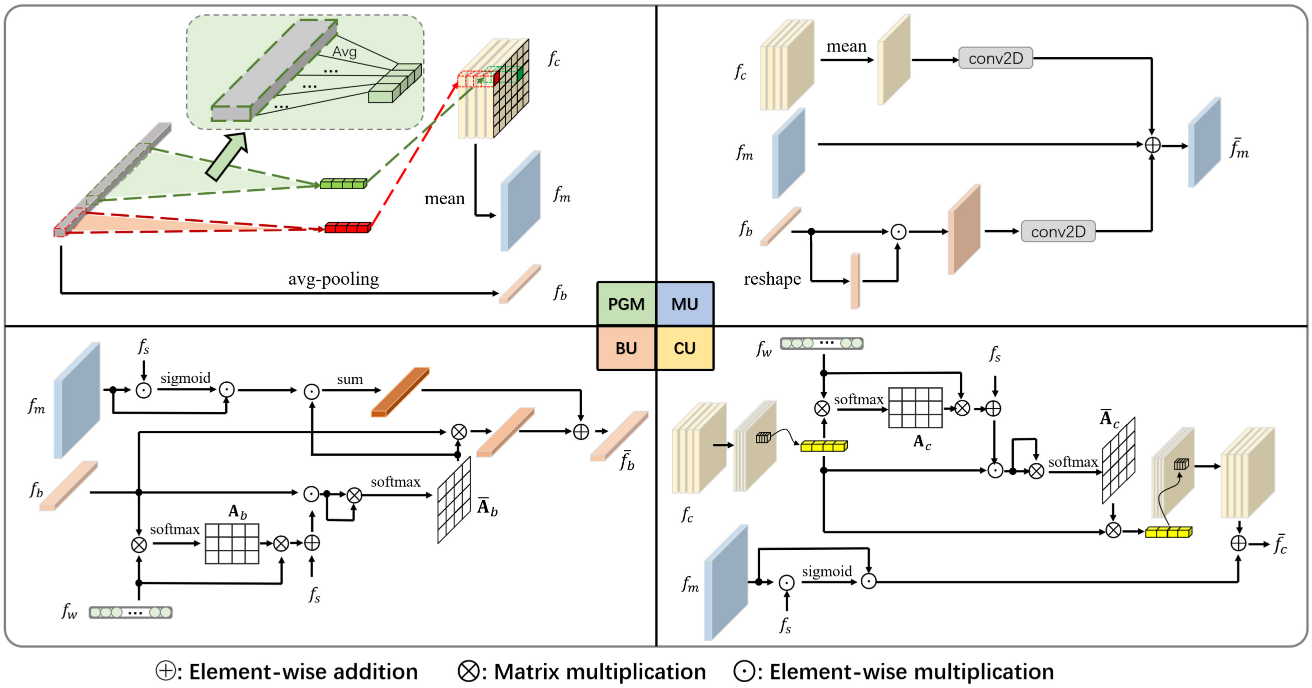

Figure 3. Illustration of the proposal generation module (PGM), boundary unit (BU), content unit (CU), and moment unit (MU).

3.3. Proposal Generation

To facilitate structured multi-level interaction, we modify the 2D map method in [43] and generate moment proposal feature, as well as moment content and moment boundary feature in the proposal generation module, as shown in the top left corner of Figure 3.

For a video with T frames, directly enumerating all the possible moment proposal will result in a large number of candidates (i.e., $T \times T$), which is computation-costly. Therefore, we only calculate $L \times L$ moments, where L is much smaller than T. For a specific moment proposal with normalized start time of $\frac{i}{L}$ and normalized end time of $\frac{j+1}{L}$ in the video sequence, where i, j are the indexes range from 0 to $L-1$, we divide all the frame features belong to this moment into C parts and then average the frame features of each part to obtain the moment content feature $\mathbf{f}_c[i,j] \in \mathbb{R}^{C\times d}$. Next, we generate moment proposal feature $\mathbf{f}_m[i,j] \in \mathbb{R}^d$ by averaging C parts of moment content feature. We obtain $\mathbf{f}_c \in \mathbb{R}^{L\times L\times C\times d}$ and $\mathbf{f}_m \in \mathbb{R}^{L\times L\times d}$ by applying this procedure to all the proposals. Directly using the sequence of frame feature as moment boundary feature will cause mismatching in temporal dimension and hinder the following structured moment interaction; therefore, we downsample the frame feature and obtain the moment boundary feature $\mathbf{f}_b \in \mathbb{R}^{L\times d}$. In practice, we find the down-sampling boundary feature improves accuracy since it helps alleviate the imbalance between positive and negative boundary samples as well as the ambiguity of labeling the moment boundary.

3.4. Structured Multi-level Interaction

3.4.1 Boundary Unit

The boundary unit (BU) is shown in the lower-left corner of Figure 3. In this unit, we interact the boundary feature with the word-level query feature to capture fine-grained cross-modal information, as well as the moment proposal feature to build the structural boundary-moment relation.

Boundary-word interaction. We employ a co-attention mechanism to obtain boundary-attended query representation based on the calculated fine-grained attention weights between boundary feature $\mathbf{f}_b$ and word-level query feature $\mathbf{f}_w$, which represent the pair-wise relations between them. The attention weight can be computed as below:

$$\mathbf{A}_b = (\mathbf{f}_b\mathbf{W}_b)(\mathbf{f}_w\mathbf{W}_{bw})^\top \in \mathbb{R}^{L\times N_q}, \tag{1}$$

where $\mathbf{W}_b \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_{bw} \in \mathbb{R}^{d\times d}$ are learnable embedding matrices projecting the two kinds of features into a joint embedding space. Since each row of $\mathbf{A}_b$ represents the similarity between all word feature to a specific boundary feature, we employ the softmax function in each row of $\mathbf{A}_b$ and obtain boundary-attended query representation:

$$\mathbf{f}_{baq} = \text{softmax}(\frac{\mathbf{A}_b}{\sqrt{d}})\mathbf{f}_w \in \mathbb{R}^{L\times d}. \tag{2}$$

The obtained boundary-attended query representation $\mathbf{f}_{baq}$ is then used as a query semantic guidance to adjust the relationship between different boundary features for structured moment interaction.

Boundary-moment interaction. We employ the self-attention mechanism to calculate the similarity map indicating the relationship between different boundary features conditioned on the query semantic guidance of $\mathbf{f}_{baq}$ and $\mathbf{f}_s$. Since $\mathbf{f}_{baq}$ offers fine-grained boundary-attended word-level information and $\mathbf{f}_s$ offers coarse-grained global sentence-level information, we can take full use of query information by combining them as the query semantic guidance. The similarity map $\bar{\mathbf{A}}_b$ can be computed by query-conditioned boundary feature $\mathbf{f}_{bq}$ as below:

$$\mathbf{f}_{bq} = \mathbf{f}_b \odot (\mathbf{f}_{baq} + \mathbf{f}_s) \in \mathbb{R}^{L\times d}, \tag{3}$$

$$\bar{\mathbf{A}}_b = \mathbf{f}_{bq}\mathbf{f}_{bq}^\top \in \mathbb{R}^{L\times L}. \tag{4}$$

Each row of $\bar{\mathbf{A}}_b$ represents the similarity between all boundary feature to a specific boundary feature. We employ the softmax function in each row of $\bar{\mathbf{A}}_b$ and obtain the boundary representation attended by other boundaries. Meanwhile, since each pair of boundaries constitutes a specific moment proposal, this similarity map also indicates the relationship of one boundary to different moments. We integrate moment information to the boundary by summation at the dimension applied with softmax function. This process leads to the following equation:

$$\mathbf{f}_{bb} = \mathrm{softmax}(\frac{\bar{\mathbf{A}}_b}{\sqrt{d}})\mathbf{f}_b \in \mathbb{R}^{L\times d}, \tag{5}$$

$$\mathbf{f}_{bm} = \mathrm{sum}(\mathrm{softmax}(\frac{\bar{\mathbf{A}}_b}{\sqrt{d}}) \odot \mathbf{f}_m) \in \mathbb{R}^{L\times d}. \tag{6}$$

To emphasize query-related information contained in the moment feature, we apply a gate function to the moment feature:

$$\mathbf{g}_m = \sigma(\mathbf{f}_m \odot \mathbf{f}_s), \bar{\mathbf{f}}_m = \mathbf{g}_m \odot \mathbf{f}_m, \tag{7}$$

where σ is sigmoid function and $\mathbf{g}_m \in \mathbb{R}^{L\times L\times d}$ represents the gate value and is dependent on sentence feature $\mathbf{f}_s$ instead of $\mathbf{f}_{bm}$, since $\mathbf{f}_{bm}$ is specified for boundary feature. We replace $\mathbf{f}_m$ in Eq.6 with $\bar{\mathbf{f}}_m$, and finally obtain the updated boundary feature $\bar{\mathbf{f}}_b = \mathbf{f}_{bb} + \mathbf{f}_{bm} + \mathbf{f}_b$.

3.4.2 Content Unit

The content unit (CU) is shown in the lower right corner of Figure 3. In this unit, we capture fine-grained cross-modal information and explore the structural content-moment relation. Similar to BU, we obtain a similarity map indicating the relationship between different content features of one specific moment via a self-attention mechanism, based on the features after fine-grained cross-modal interaction. We then integrate moment information to the content representation with the guidance of query information.

Content-word interaction. A co-attention mechanism is employed to obtain content-attended query representation. For computational efficiency, we first reduce the channel dimension of content feature $\mathbf{f}_c$ and word feature $\mathbf{f}_w$ from d to d_l. Since the content feature is related to other content features within the same moment, we compute the attention weight between content within a specific moment and word. For the content feature within a specific moment proposal $\hat{\mathbf{f}}_c = \mathbf{f}_c[i,j] \in \mathbb{R}^{C\times d_l}$, we calculate the attention weight as follow:

$$\mathbf{A}_c = (\hat{\mathbf{f}}_c\mathbf{W}_c)(\mathbf{f}_w\mathbf{W}_{cw})^\top \in \mathbb{R}^{C\times N_q}, \tag{8}$$

where $\mathbf{W}_c \in \mathbb{R}^{d_l\times d_l}$ and $\mathbf{W}_{cw} \in \mathbb{R}^{d_l\times d_l}$ are learnable parameters. We then employ the softmax function in each row of $\mathbf{A}_c$ and obtain content-attended query representation:

$$\mathbf{f}_{caq} = \mathrm{softmax}(\frac{\mathbf{A}_c}{\sqrt{d_l}})\mathbf{f}_w \in \mathbb{R}^{L\times d_l}. \tag{9}$$

Content-moment interaction. As in boundary-moment interaction, we combine the content-attended query representation $\mathbf{f}_{caq}$ with channel reduced $\mathbf{f}_s$ as the query semantic guidance. We then obtain the similarity map via a self-attention mechanism based on query-conditioned content feature $\mathbf{f}_{cq}$ as follow:

$$\mathbf{f}_{cq} = \hat{\mathbf{f}}_c \odot (\mathbf{f}_{caq} + \mathbf{f}_s) \in \mathbb{R}^{C\times d_l}, \tag{10}$$

$$\bar{\mathbf{A}}_c = \mathbf{f}_{cq}\mathbf{f}_{cq}^\top \in \mathbb{R}^{C\times C}. \tag{11}$$

Next, a softmax function is employed in each row of $\bar{\mathbf{A}}_c$ and we can obtain the content representation capturing relations to other content within the same moment:

$$\hat{\mathbf{f}}_{cc} = \mathrm{softmax}(\frac{\bar{\mathbf{A}}_c}{\sqrt{d_l}})\hat{\mathbf{f}}_c \in \mathbb{R}^{C\times d_l}, \tag{12}$$

By applying this procedure to all the moment proposals and increasing the channel dimension from d_l to d, we obtain $\mathbf{f}_{cc} \in \mathbb{R}^{L\times L\times C\times d}$. After that, we integrate moment information to content with the similar fashion as in Eq.7 and obtain the updated content feature $\bar{\mathbf{f}}_c = \mathbf{f}_{cc} + \bar{\mathbf{f}}_m + \mathbf{f}_c$.

3.4.3 Moment Unit

The moment unit (MU) is shown in the top right corner of Figure 3. In this unit, we aggregate the boundary feature $\bar{\mathbf{f}}_b$ and content feature $\bar{\mathbf{f}}_c$ from BU and CU into the moment feature. We reshape $\bar{\mathbf{f}}_b$ to $\bar{\mathbf{f}}_b^s \in \mathbb{R}^{L\times 1\times d}$ and $\bar{\mathbf{f}}_b^e \in \mathbb{R}^{1\times L\times d}$, expand them to have the same shape of $\bar{\mathbf{f}}_m$, and fuse them by Hadamard product. We average C parts of $\bar{\mathbf{f}}_c$. This procedure is given by:

$$\bar{\mathbf{f}}_m = \mathrm{Conv2d}(\bar{\mathbf{f}}_b^s \odot \bar{\mathbf{f}}_b^e) + \mathrm{Conv2d}(\mathrm{mean}(\bar{\mathbf{f}}_c)) + \mathbf{f}_m. \tag{13}$$

3.5. Localization

After we obtain the output representation with extensive cross-modal interaction and structured moment interaction, we predict the target moment proposal. We obtain the moment prediction scores p_m by one layer of 2D convolution followed with a sigmoid function from the moment feature $\bar{\mathbf{f}}_m$ of the last layer of MU:

$$p_m = \sigma(\mathrm{Conv2d}(\bar{\mathbf{f}}_m)) \in \mathbb{R}^{L\times L}. \tag{14}$$

We also obtain the start and end boundary prediction scores p_s and p_e by one layer of 1D convolution followed with a sigmoid function from the boundary feature $\bar{\mathbf{f}}_b$ of the last layer of BU:

$$p_s = \sigma(\mathrm{Conv1d}(\bar{\mathbf{f}}_b)), p_e = \sigma(\mathrm{Conv1d}(\bar{\mathbf{f}}_b)) \in \mathbb{R}^{L}. \tag{15}$$

Therefore, the final predicted moment with normalized start time $\frac{i}{L}$ and end time $\frac{j+1}{L}$ can be presented as $(p_m[i,j], p_s[i], p_e[j])$.

3.5.1 Training

We adopt an alignment loss to learn the moment prediction score, which is formulated by:

$$\mathcal{L}_m = -\frac{1}{N_m}\sum_{k=0}^{N_m} y_m^k s_m^k \mathrm{log} p_m^k + (1-y_m^k)(1-s_m^k)\mathrm{log}(1-p_m^k), \tag{16}$$

where p_m^k is the k_{th} output score of p_m representing the k_{th} proposal, y_m^k is the binary label determined by a threshold of 0.5 from IoU score s_m^k between this moment with the ground truth, and N_m is the number of valid moment proposals.

We also adopt a boundary matching loss to learn the boundary prediction score, which is given by:

$$\mathcal{L}_s = -\frac{1}{L}\sum_{k=0}^{L} y_s^k s_s^k \mathrm{log} p_s^k + (1-y_s^k)(1-s_s^k)\mathrm{log}(1-p_s^k), \tag{17}$$

$$\mathcal{L}_e = -\frac{1}{L}\sum_{k=0}^{L} y_e^k s_e^k \mathrm{log} p_e^k + (1-y_e^k)(1-s_e^k)\mathrm{log}(1-p_e^k), \tag{18}$$

where $p_s^k(p_e^k)$ is the k_{th} output score of $p_s(p_e)$, representing the k_{th} boundary. $s_s(s_e) \in \mathbb{R}^L$ is generated by an unnormalized 1D Gaussian $e^{-\frac{x^2}{2\sigma^2}}$ inspired by [14], which gives fewer penalties to the adjacent locations of boundaries. Its center is at $\tau_s(\tau_e)$ and whose σ is set to $(\tau_e - \tau_s)/5$, where τ_s, τ_e are the boundary ground truth. $y_s^k(y_e^k)$ is the binary label determined by $s_s(s_e)$ through a threshold of 0.5. We additionally calculate an auxiliary snippet matching loss:

$$\mathcal{L}_a = -\frac{1}{L}\sum_{k=0}^{L} y_a^k \mathrm{log} p_a^k + (1-y_a^k)\mathrm{log}(1-p_a^k), \tag{19}$$

where $p_a = \sigma(\mathrm{Conv1d}(\bar{\mathbf{f}}_b)) \in \mathbb{R}^L$ and we choose the snippets within the ground truth as positive and others as negative. The snippets close to boundaries are ignored since they may cause ambiguity to determine whether they are within the ground truth.

The total loss function is given by:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_s + \mathcal{L}_e + 0.5\cdot\mathcal{L}_a. \tag{20}$$

3.5.2 Inference

During inference, we use $p_m[i,j]\cdot\sqrt{p_s[i]}\cdot\sqrt{p_e[j]}$ as the final prediction score of the moment with normalized time $(\frac{i}{L}, \frac{j+1}{L})$. We rank all the moment proposals according to their prediction scores followed by a non-maximum suppression (NMS) function.

4. Experiment

4.1. Datasets and Evaluation Metrics

TACoS TACoS [26] consists of 17,344 text-to-clip pairs collected from cooking scenarios. We use the standard split as [7], which has 10146, 4589, and 4083 moment-query pairs for training, validation, and testing, respectively.

Charades-STA Charades-STA [7] was built on Charades [29]. Gao *et al.* [7] generated temporal sentence annotations from the original Charades dataset and result in 12408 and 3720 pairs of sentence-moment for training and testing.

ActivityNet-Captions ActivityNet-Captions [13] consists of 20k videos and 100k descriptions with diverse context, built on ActivityNet v1.3 dataset [11]. Following [43], we use val_1 as validation set and val_2 as testing set. We have 37417, 17505, and 17031 moment-sentence pairs for training, validation, and testing.

Evaluation Metrics Following previous work [7], we adopt the "R@n, IoU=m" metric as the evaluation metric. It is defined as the percentage of at least one proposal in the top "n" predictions that have IoU with ground-truth larger than the thresholds "m".

4.2. Implementation Details

For a fair comparison, we extract the visual features from a pre-trained 3D CNN (i.e, I3D as [24] for Charades-STA, and C3D as [43] for TACoS and ActivityNet-Captions). We uniformly sample T=(64, 128, 128) segments as the input video feature sequence and set the length of 2D feature map L=(16, 32, 64) for Charades-STA, TACoS and ActivityNet-Captions, respectively. For the language query, the pre-trained Glove embedding is employed to each word of the query with a dimension of 300. Each sentence is truncated

Table 1. Performance comparison with other methods on Charades-STA.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
ROLE [20]	12.12	-	40.59	-
CTRL [7]	21.42	7.15	59.11	26.91
ACL [9]	30.48	12.20	64.84	35.13
GDP [6]	39.47	18.49	-	-
CBP [32]	36.80	18.87	70.94	50.19
2D-TAN [43]	39.70	23.31	80.32	51.26
MAN [42]	46.53	22.72	86.23	53.72
DPIN [31]	47.98	26.96	85.53	55.00
DRN [38]	53.09	31.75	89.06	60.05
SCDM [36]	54.44	33.43	74.43	58.08
LGI [24]	59.46	35.48	-	-
SMIN (ours)	**64.06**	**40.75**	**89.49**	**68.09**

Table 2. Performance comparison with other methods on ActivityNet-Captions.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
TGN [4]	27.93	11.86	44.20	24.84
CBP [32]	35.76	17.80	65.89	46.20
CMIN [45]	43.40	23.88	67.95	50.73
LGI [24]	41.51	23.07	-	-
2D-TAN [43]	44.51	26.54	77.13	61.96
DRN [38]	45.45	24.36	77.97	50.30
DPIN [31]	47.27	28.31	77.45	60.03
SMIN (ours)	**48.46**	**30.34**	**81.16**	**62.11**

Table 3. Performance comparison with other methods on TACoS.

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
CTRL [7]	18.32	13.30	36.69	25.42
ACRN [19]	19.52	14.62	34.97	24.88
CMIN [45]	24.64	18.05	38.46	27.02
ABLR [37]	19.50	9.40		
ExCL [10]	28.00	13.80	-	-
SCDM [36]	26.11	21.17	40.16	32.18
DRN [38]	-	23.17	-	33.36
2D-TAN [43]	37.29	25.32	57.81	45.04
DPIN [31]	46.74	32.92	62.16	50.26
SMIN (ours)	**48.01**	**35.24**	**65.18**	**53.36**

to a fixed length of (13, 14, 20) for Charades-STA, TACoS, and ActivityNet-Captions. The hidden state size of the bidirectional LSTM is set to 256, and the feature dimension d is set to 512. The number of parts of content features C is set to 4, and d_l in the content unit is set to 128. We stack three layers of boundary unit, content unit, and moment unit. We use an Adam optimizer to train our model, with a learning rate of 0.0005 and a batch size of 64.

4.3. Performance Comparison

We report the result of $n \in \{1, 5\}$ with $m \in \{0.5, 0.7\}$ for Charades-STA, $n \in \{1, 5\}$ with $m \in \{0.5, 0.7\}$ for ActivityNet-Captions and $n \in \{1, 5\}$ with $m \in \{0.3, 0.5\}$ for TACoS, as shown in Table 1, Table 2, and Table 3, respectively. Our method outperforms all competing methods. Specifically, on Charades-STA, SMIN outperforms the previous best method LGI by 4.60% and 5.27% absolute improvement in terms of R@1, IoU=0.5, and R@1, IoU=0.7, respectively. On ActivityNet-Captions, SMIN reaches the highest score with approximately 2% performance improvement concerning R@1, IoU=0.7. On TACoS, SMIN surpasses DPIN with 2.3% and 3.10% performance improvement, regarding R@1, IoU=0.5, and R@5, IoU=0.5, respectively.

Compared to state-of-the-art methods that use frame-level interaction TGN [4], GDP [6], ExCL [10], DEBUG [22], LGI [24], DPIN [31], DRN [38], CMIN [45], CBP [45] or use proposal-level interaction CTRL [7], ACRN [19], 2D-TAN [43], our method performs better. They neglect the inherent structure of moment construction and employ the moment structure for fine-grained cross-modal interaction, leading to relatively lower performances. Besides, we capture the insightful relations between boundary, content and moment through the structured moment interaction. Based on the explored moment structure, our method emphasizes the importance of both fine-grained vision-language understanding and detained video comprehension. Therefore, our method achieves better performance than previous methods.

4.4. Ablation Study

We evaluate the main components of our method on Charades-STA and TACoS in Table 4, where "w/o BU" means without boundary unit (BU), "w/o CU" means without content unit (BU), "w/o BU+CB" means without both BU and CU, and "Full" means the full model. The boundary unit and content unit are two critical components in the structured multi-level interaction module, which conducts fine-grained vision-language interaction and structured moment interaction. From the results in Table 4, the full model outperforms all the compared ablation models on both two datasets, which demonstrates BU and CU are helpful for moment localization since BU contributes to the boundary discrimination while CU benefits the moment alignment.

To evaluate the detailed components in the boundary unit and content unit more deeply, we conduct ablation studies of BU and CU on Charades-STA concerning R@1 in Table 5. "w/o VLI" means without fine-grained vision-language interaction and directly calculate the similarity map by boundary/content feature; "w/o BMI/CMI" means without aggregating moment feature to boundary/content feature; "w/o Gate" means without the gate function when aggregating moment feature to boundary/content feature; and "Full" means the full model. From Table 5, we can

Table 4. Ablation results of main components.

Dataset	Method	R@1 IoU=0.5	R@1 IoU=0.7
Charades-STA	w/o BU	61.99	37.20
	w/o CU	62.22	38.72
	w/o BU+CU	58.65	34.16
	Full	**64.06**	**40.75**
TACoS	w/o BU	46.83	32.99
	w/o CU	44.91	32.89
	w/o BU+CU	43.53	31.91
	Full	**48.01**	**35.24**

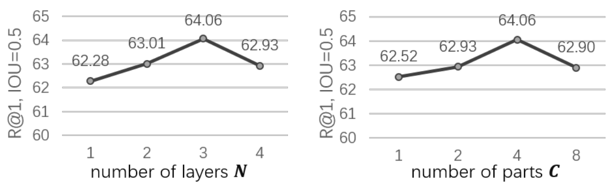

Figure 4. Ablation studies of the hyper-parameters.

learn each detailed component in the boundary/content unit contributes to localizing the target moment. In more detail, by comparing "Full" to "w/o VLI" and "w/o BMI",we observe fine-grained vision-language interaction and structured moment interaction are vital for this task since they leverage detailed information between two modalities and different components of the moment. The result of "w/o Gate" shows the gate function emphasizes query-related information, which benefits this task.

We show the effect of the various number of layers N and content parts C on Charades-STA in Figure 4. For the number of layers N, the model achieves the best performance by setting N to 3. Our model is able to leverage comprehensive vision-language interaction and structured moment interaction when we use more layers. Too many layers result in over-smoothing problem and make the performance drop. For the number of content parts C, our model performs best when C is set to 4 since increasing the number of parts in moment content enables our model to capture more detailed information from moment content. Dividing the content into too many parts will largely increase the computation cost and accumulate noise, which harms the performance.

4.5. Qualitative Results

We qualitatively validate the ablation models without fine-grained vision-language interaction and without structured moment interaction in both boundary unit and content unit at the top of Figure 5. We can observe that the ablation models predict coarser boundaries since they lack the crucial detailed interaction for vision-language and moment structure. Besides, we also qualitatively show the ablation models without boundary unit, without content unit,

Table 5. Ablation studies of BU and CU on Charades-STA.

Unit	Method	R@1 IoU=0.5	R@1 IoU=0.7
Boundary Unit	w/o VLI	62.31	38.70
	w/o BMI	62.09	37.87
	w/o Gate	62.69	39.06
	Full	**64.06**	**40.75**
Content Unit	w/o VLI	62.63	39.78
	w/o CMI	62.44	38.95
	w/o Gate	62.55	39.54
	Full	**64.06**	**40.75**

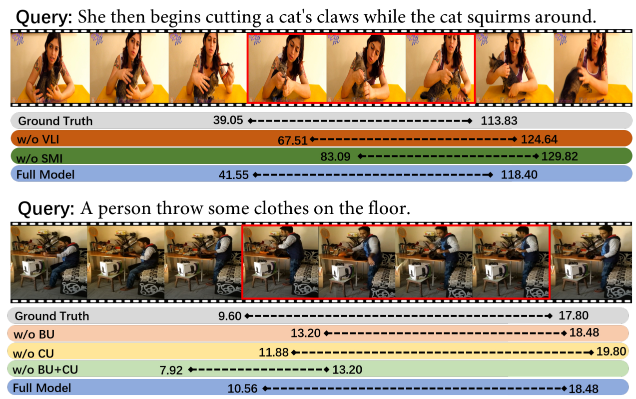

Figure 5. Visualization of the predictions of the SMIN model and the ablation models.

and without both units at the bottom of Figure 5. The result shows that the explored inherent structure of moment construction is crucial for this task since it facilitates both fine-grained vision-language interaction and structured moment interaction between moment and query.

5. Conclusion

In this paper, we propose a new Structured Multi-level Interaction Network (SMIN) for video moment localization by natural language. SMIN leverages the inherent structure of the moment constructed with boundary and content for both vision-language understanding and video comprehension. We design boundary unit, content unit, and moment unit in the structured multi-level interaction module for fine-grained cross-modal interaction between boundary/content and query, and detailed structured moment interaction between boundary, content and moment. Extensive evaluation on three benchmarks has demonstrated the effectiveness of the proposed method.

6. Acknowledgments

This work was supported by the National Key R&D Program of China under Grand 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants U19B2038 and 61771457.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1, 2

[2] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *CVPR*, 2017. 2

[3] Yuwei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*, 2018. 2

[4] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018. 1, 2, 7

[5] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *AAAI*, 2019. 1

[6] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, 2020. 1, 2, 7

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 1, 2, 6, 7

[8] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. In *BMVC*, 2017. 2

[9] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *WACV*, 2019. 1, 7

[10] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*, 2019. 1, 2, 7

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 6

[12] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-modal video moment retrieval with spatial and language-temporal attention. In *ICMR*, 2019. 1

[13] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 6

[14] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 6

[15] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *ACM MM*, 2017. 2

[16] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*, 2018. 1

[17] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for sequence-level image captioning. In *ACM MM*, 2018. 2

[18] Daqing Liu, Hanwang Zhang, Zheng-Jun Zha, and Feng Wu. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, 2019. 2

[19] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, 2018. 1, 2, 7

[20] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *ACM MM*, 2018. 1, 2, 7

[21] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *ICCV*, 2019. 2

[22] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *EMNLP*, 2019. 2, 7

[23] Alberto Montes, Amaia Salvador Aguilera, Santiago Pascual, and Xavier Giro I Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. In *NIPS*, 2016. 2

[24] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 2, 3, 6, 7

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 3

[26] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 6

[27] Zeng Runhao, Gan Chuang, Chen Peihao, Huang Wenbing, Wu Qingyao, and Tan Mingkui. Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 2019. 2

[28] Zheng Shou, Dongang Wang, and Shihfu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 2

[29] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 6

[30] Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. Learning to discretely compose reasoning module networks for video captioning. In *IJCAI*, 2020. 2

[31] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. Dual path interaction network for video moment localization. In *ACM MM*, 2020. 2, 7

[32] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *AAAI*, 2020. 2, 7

[33] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 2019. 1

[34] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 2

[35] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2016. 2

[36] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NIPS*, 2019. 7

[37] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019. 7

[38] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020. 2, 7

[39] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[40] Zheng-Jun Zha, Jiawei Liu, Di Chen, and Feng Wu. Adversarial attribute-text embedding for person search with natural language query. *IEEE Transactions on Multimedia*, 22(7):1836–1846, 2020. 2

[41] Zheng-Jun Zha, Jiawei Liu, Tianhao Yang, and Yongdong Zhang. Spatiotemporal-textual co-attention network for video question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(2):53, 2019. 2

[42] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019. 1, 2, 7

[43] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020. 1, 2, 4, 6, 7

[44] Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *ACM MM*, 2019. 1

[45] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*, 2019. 1, 2, 7

[46] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, 2020. 2

[47] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 2