Cross-modal Moment Localization in Videos*

Meng Liu
Shandong University
mengliu.sdu@gmail.com

Xiang Wang
National University of Singapore
xiangwang@u.nus.edu

Liqiang Nie
Shandong University
nieliqiang@gmail.com

Qi Tian
[1]Huawei Noah's Ark Lab
[2]University of Texas at San Antonio
tian.qi1@huawei.com

Baoquan Chen
[1]Peking University
[2]Shandong University
baoquan.chen@gmail.com

Tat-Seng Chua
National University of Singapore
dcscts@nus.edu.sg

ABSTRACT

In this paper, we address the temporal moment localization issue, namely, localizing a video moment described by a natural language query in an untrimmed video. This is a general yet challenging vision-language task since it requires not only the localization of moments, but also the multimodal comprehension of textual-temporal information (e.g., "first" and "leaving") that helps to distinguish the desired moment from the others, especially those with the similar visual content. While existing studies treat a given language query as a single unit, we propose to decompose it into two components: the relevant cue related to the desired moment localization and the irrelevant one meaningless to the localization. This allows us to flexibly adapt to arbitrary queries in an end-to-end framework. In our proposed model, a language-temporal attention network is utilized to learn the word attention based on the temporal context information in the video. Therefore, our model can automatically select "what words to listen to" for localizing the desired moment. We evaluate the proposed model on two public benchmark datasets: DiDeMo and Charades-STA. The experimental results verify its superiority over several state-of-the-art methods.

CCS CONCEPTS

• **Information systems → Video search**; **Novelty in information retrieval**;

KEYWORDS

Language-Temporal Attention, Moment Localization, Cross-modal Video Retrieval

ACM Reference Format:
Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240549

*This work was performed when Meng Liu was visiting National University of Singapore as a research intern. Both Liqiang Nie and Baoquan Chen are the corresponding authors.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240549

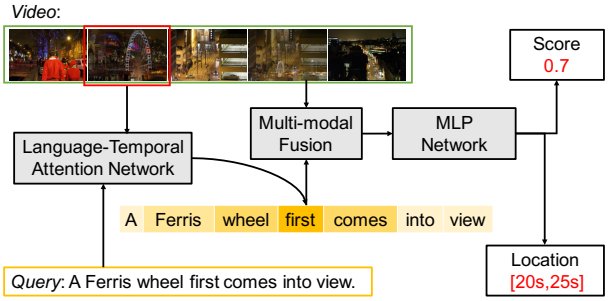


Figure 1: The pipeline of our proposed model.

1 INTRODUCTION

Great progress has been made on video retrieval [5, 26, 31, 39], the task of retrieving videos from a collection to match the given language query. However, moment retrieval remains largely untapped, aiming to find a specific segment (i.e., moment) from a video when given a natural language description. This task, also known as temporal moment localization, has been gaining increasing interests in computer vision. Particularly, given a video and a query like "the child dances over next to the other people", existing solutions commonly use a temporal bounding box (i.e., the start and end time points) to localize the temporal moment corresponding to the query.

In the traditional video retrieval task, the queries are, more often than not, simple keywords expressing the desired action, objects or attributes. In contrast, in the task of temporal moment localization, the given queries are much more complex and sophisticated, which can be arbitrary natural descriptions, like a phrase or a complete sentence. As shown in Figure 1, for example, the sentence "A Ferris wheel first comes into view" is a typical query, emphasizing that a "Ferris wheel" entity appears with a temporal relation "first". A model that only localizes "Ferris wheel" is not satisfactory, since this entity appears twice in the video. Thereby, resolving this query requires both finding a moment that contains a "Ferris wheel" and ensuring that it is the first time of its appearance. Therefore, the key for temporal moment localization is to well comprehend the complex query information and attend to useful words which are the most relevant and significant to localize the desired moment.

We have to mention that, several studies [1, 7] have been proposed to process such complex queries. These prior efforts usually feed the whole description into one offline language processor (e.g., Skip-thoughts [12]) or an online tool (e.g., LSTM [16]) to establish one feature vector for the entire query.

Despite their success, simply treating queries holistically as one feature vector may obfuscate the keywords that have rich temporal and semantic cues. That is, they fail to emphasize the words, such as the "first" in Figure 1, which are significant to localize the desired moment, the first "Ferris wheel" moment, rather than other moments containing the similar visual features. For instance, the second moment also conveying the "Ferris wheel" entity. As we can see, the correlation between textual components and temporal moments has not been fully explored. Therefore, it is crucial to build a language processing model to adaptively select the key textual words from the query based on different video context.

In this work, we aim to bridge the research gap by integrating a language processing module, which can capture the spatial-temporal information better, with the moment localization model. We expect our method to exploit the correlation between the textual and visual features and highlight the useful words for the desired moment. Towards this end, we present a c**R**oss-modal m**O**ment **L**ocalization n**E**twork (ROLE) that jointly learns the query representation and temporal moment localization, as illustrated in Figure 1. First, we design a language-temporal attention module to derive effective query representations, adaptively reweighing each word's features according to the query textual information and moment context information. Such query representation can identify "which words to listen to" and be more robust to the query variations that are irrelevant to moment localization. We then stack a multi-modal processing module to jointly model the query and temporal context features. We ultimately train a multi-layer perception (MLP) network to estimate the relevance score and the location of the desired moment. Extensive experiments on two public datasets have well justified that our model outperforms the state-of-the-art baselines significantly.

The contributions of this work are three-fold:

- We present a novel cross-modal temporal moment localization approach, which is able to adaptively encode complex and significant language query information for localizing desired moments.
- We propose a language-temporal attention network which jointly encodes the textual query, local moment, and its context moment information to comprehend query descriptions. To the best of our knowledge, this is the first query attention mechanism based network for the temporal moment localization task.
- We evaluate our proposed model on two large datasets, DiDeMo and Charades-STA, to demonstrate the performance improvement. As a side contribution, we released the data and the codes[1].

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 and 4 detail the temporal moment localization problem and our proposed ROLE model, respectively. We present the experimental results in Section 5, followed by the conclusion and future work in Section 6.

2 RELATED WORK

In this section, we briefly review some studies related to the temporal moment localization. As it is a fairly new task, there are spare literature to refer to. Here we consider three related tasks: grounding referential expressions, temporal action localization, and language grounding in the video.

2.1 Grounding Referential Expressions

The task of grounding referential expressions [17, 21, 43, 44] is to localize an image region described by a given referring expression. And it is usually formulated as a retrieval problem over image regions. Therefore, each image is firstly segmented into a set of region proposals [2, 13, 36, 45], and then different strategies are adopted to score each proposal candidate with respect to the query expression. Finally, the proposal candidate with the highest score is returned as the grounding result.

There are a lot efforts on the computation of matching score between each proposal candidate and the given expression. Mao *et al.* [20] proposed a model jointly considering the local candidate feature and the whole image feature to predict the matching score of each proposal candidate. However, it is insufficient to judge whether a proposal matches the expression. Afterwards, Yu *et al.* [42] found that visual comparison to other objects within an image helps improve performance significantly. Hence, they integrated the contextual feature extracted from other region proposals in the image into the model. However, all the methods aforementioned represent expressions holistically using a Recurrent Neural Network (RNN). Namely, they either predicted a distribution over referential expressions, or encoded expressions into a vector representation. Therefore, they may not well learn the explicit correspondences between the components in the expression and entities in the image. Recently, some researchers have tried to parse the given language expression into textual components instead of treating it as a whole, and align these components with the image regions end-to-end. Hu *et al.* [10] parsed the referential expression into a subject, relation and object with three soft attention maps, and aligned the extracted textual representations with image regions using a modular neural architecture. Similarly, Yu *et al.* [41] decomposed the expression into three modular components via a soft attention, related to the subject appearance, location, and relationship to other objects, respectively.

Although these models are proved to be powerful in their dedicated task, simply extending them to the temporal moment localization task is inappropriate. They may ignore the temporal information of videos, yet it is the most distinctive feature compared to the static images.

2.2 Temporal Action Localization

Temporal action localization is a task that given a long untrimmed video, predicting when a specific action starts and ends [6, 15, 19]. Sun *et al.* [33] addressed the problem of fine-grained action localization from temporally untrimmed web videos via transferring image labels into their model. Later, Shou *et al.* [27] exploited multi-stage 3D ConvNets for temporal action localization in untrimmed long videos in the wild. And Ma *et al.* [19] introduced novel ranking losses within the RNN learning objective to better capture the progression of activities. Meanwhile, Singh *et al.* [30] extended two-stream framework [28] via augmenting full-frame image features with features from a bounding box surrounding the actor. Lately, Gao *et al.* [8] introduced a novel temporal unit regression network

[1] https://acmmm18.wixsite.com/role.

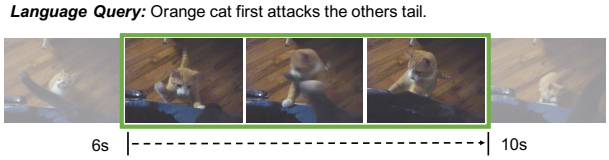


Figure 2: Temporal moment localization via language query in an untrimmed video.

to jointly predict the action proposals and refine the temporal boundaries by temporal coordinate regression. However, these action localization methods are restricted to the pre-defined list of actions. Recently, Gao *et al.* [7] proposed to localize activities by natural language queries. They proposed a cross-modal temporal regression localizer to jointly model the text query and video moments. Hendricks *et al.* [1] designed a moment context network to localize language queries in videos by integrating the local and global video features. Although these two models perform well in their tasks, they always encode the whole query as one single feature, which may be not expressive enough to reveal the information conveyed in the query.

2.3 Language Grounding in the Video

Different from the video retrieval [3, 38] aiming to retrieve a video from a set of video candidates given a natural language query [22, 35, 37], language grounding in the video is a task of spatially grounding objects and actions in a video, or aligning textual phrases to temporal video segments [9, 25, 32]. There are few studies on this issue, and they are severely limited in the natural language vocabulary. Tellex *et al.* [34] proposed a model to retrieve video segments from a home surveillance camera utilizing queries containing a fixed set of spatial prepositions. Yu *et al.* [40] only considered four objects and four verbs to learn representations of words from short video clips paired with sentences. Lin *et al.* [14] proposed a model to locate objects in the video by parsing the descriptions into a semantic graph that is then matched to the visual concepts by solving a linear program. Regneri *et al.* [24] presented a general purpose corpus that aligns high quality videos with multiple natural language descriptions of the actions exhibited in the videos. Bojanowski *et al.* [4] introduced a method automatically providing a time stamp for every sentence, namely aligning a video with its natural language description. Different from the video-text alignment task which gives a video and a set of sentences with temporal ordering, we only input one query into our model. Moreover, methods aligning instructions with videos are restricted to structured videos as they constrain alignment by instruction ordering.

3 TEMPORAL MOMENT LOCALIZATION

In this section, we first formulate the task of temporal moment localization. We then introduce two state-of-the-art models that are the fundamental components in our work.

To formulate the problem, some notations are declared in advance. In particular, we use bold capital letters (e.g., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to denote matrices and vectors, respectively. We employ non-bold letters (e.g., D) to represent scalars, Math calligraphy (e.g., C) as sets, and Greek letters (e.g., λ) as parameters. If not clarified, all vectors are in column form.

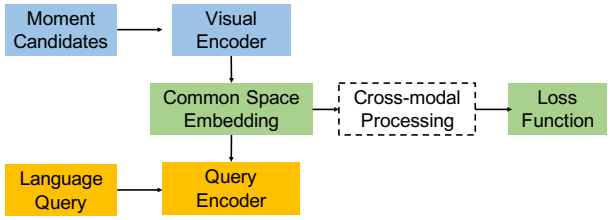


Figure 3: The unified framework of the existing temporal localization methods. The dash line is the unique module of the CTRL model.

3.1 Problem Formulation

Recently, great progress has been made on activity or object localization in longer and untrimmed videos, aiming to localize the temporal activity moments or objects corresponding to the predefined language vocabulary. However, when we attempt to localize a specific moment, such as described as "orange cat first attacks the others tail", a simple action, object, or attribute keyword is insufficient to uniquely identify such golden moment. A straightforward solution is to query with a natural language phrase. Inspired by this, the task of moment localization in a video with language query is proposed. It aims to find a specific temporal segment (i.e., moment) from a video when given a natural language description [1, 7]. Such moments are affiliated with the start and end time points localizing the video.

These models commonly work in the moment retrieval setting: given a video $\mathcal{V} = \{f_1, f_2, \ldots, f_N\}$, where f_i represents the i-th video frame, and a language query Q with the start time τ_s and end time τ_e, i.e., the start time and end time of the desired moment. Then the video is segmented into a set of moment candidates $C = \{c_1, c_2, \ldots, c_M\}$ via the sliding window strategy [1, 7], and each candidate c_i is assigned with a temporal bounding box $[t_s, t_e]$. Therefore, the models only need to estimate the relevance score of each moment candidate and the query.

Figure 2 shows an example of the temporal moment localization. The video depicts that an orange cat looks at the tail of a black one, jumps up to attack the tail of the black, and then falls down. Here, we give a language query "orange cat first attacks the others tail" and expect the moment localization model to return the start time (6s) and the end time (10s) of the corresponding moment (in the green bounding box).

3.2 Moment Localization Model

To well match the query and the moment candidates, an intuitive way is to map the visual features of the moment candidates and the textual feature of the query into a common space, and then minimize the distance of each positive moment-query pairs. Motivated by this intuition, two state-of-the-art methods are proposed. The first is a joint video-language model, named as Moment Context Network[2] (MCN) [1], in which features of query and videos are encouraged to be close in a shared embedding space. Analogously, a cross-modal

[2]https://github.com/LisaAnne/LocalizingMoments.

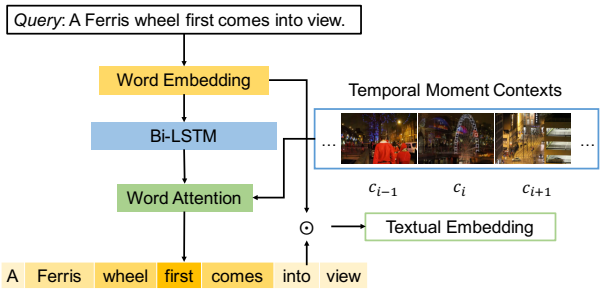


Figure 4: Illustration of the Language-Temporal Attention Network.

localization model, named as Cross-modal Temporal Regression Localizer[3] (CTRL) [7] is proposed, as shown in Figure 3.

These two models differ in several points. 1) Context information. Both models concatenate visual features of the current moment and its contextual moments into a single vector, and then linearly transform it as the final visual feature. For constructing the moment context, MCN takes the entire set of moment candidates, while CTRL adopts the neighbor pre-context and post-context moments. 2) Query encoder. MCN treats the last output of a LSTM network as the query feature; meanwhile, CTRL takes the offline Skip-Thoughts[4] feature to represent the query. And 3) cross-modal processing. CTRL utilizes a multi-modal fusion method to fuse the query and moment features in the common space; whereas, MCN directly calculates the distance between the moment and query features in the common space.

While performing well as compared to the baselines, these models treat the query holistically and overlook the effectiveness of keywords conveying the significant spatial-temporal cues. We note that depending on the distinctiveness of the desired moment, words in one query contribute differently to the estimations. For example, if the target moment is about "a girl in a red coat" among all the people, the words "girl" and "red" should contribute most to the relevance estimation. If the same girl appears more than one time and the given query becomes "the girl in red first appears", the context information "first" should become the crucial contributor to localize the desired moment. Therefore, it is natural and intuitive to integrate an attention mechanism with the moment retrieval modal to read the query word by word and refine its attention based on the temporal context information.

4 OUR MODEL

In this section, we detail our cross-modal moment localization network (ROLE), composing a language-temporal attention network, a multi-modal processing, and a MLP module. In particular, given a moment candidate c_i and a language query Q, we first use the attention network to adaptively reweigh the weight scores of the useful words based on the moment contexts. Thereafter, we leverage the multi-modal processing module to fuse the query and moment representations. A MLP is conducted to calculate the relevance score that measures the compatibility among c_i and Q and the moment location.

[3]https://github.com/jiyanggao/TALL.
[4]https://github.com/ryankiros/skip-thoughts.

4.1 Language-Temporal Attention Network

Our attention network is shown in Figure 4. For a given query Q composed of a sequence of T words $\{w_t\}_{t=1}^{T}$, we project each word w_t into an embedding vector $\mathbf{e}_t$ via Glove [23]. Thereafter, a Bi-directional LSTM is employed to encode the whole query, taking the sequence $\{\mathbf{e}_t\}_{t=1}^{T}$ as input, and outputting a forward hidden state $\mathbf{h}_t^{fw}$ and a backward hidden state $\mathbf{h}_t^{bw}$ at each time t. We then concatenate $\mathbf{h}_t^{fw}$ and $\mathbf{h}_t^{bw}$ into $\mathbf{h}_t$, which contains the information from word w_t and the context words before and after w_t. It can be formulated with the following equations,

$$\begin{cases} \mathbf{e}_t = embedding(w_t), \\ \mathbf{h}_t^{fw} = LSTM^{fw}(\mathbf{e}_t, \mathbf{h}_{t-1}^{fw}), \\ \mathbf{h}_t^{bw} = LSTM^{bw}(\mathbf{e}_t, \mathbf{h}_{t-1}^{bw}), \\ \mathbf{h}_t = [\mathbf{h}_t^{fw}, \mathbf{h}_t^{bw}]. \end{cases} \tag{1}$$

To obtain the representation of the given query, a direct way is to average pooling all the word representations $\mathbf{h}_t$. While the above solution seems to be sound and reasonable, the downside is that all the words in the query contribute equally to the query embedding, ignoring their scenario-specific confidence. As mentioned before, solely using such representations may fail to distinguish the desired moment from the moments having the similar visual features. To address the issues, we feed the temporal moment contexts[5] into our attention model, which is capable of assigning the more useful words with higher importance scores. Therefore, given $\mathcal{H} = \{\mathbf{h}_t\}_{t=1}^{T}$, input moment c_i, and its temporal moment contexts c_j ($j \in \{i-n, \ldots, i-1, i+1, \ldots, i+n\}$), n is the neighbor size of moment contexts, we formulate the attention model as follows,

$$\begin{cases} \mathbf{r}_t = f(\mathbf{W}_q \mathbf{h}_t + \sum_{j=i-n}^{i+n} \mathbf{W}_c \mathbf{x}_{c_j} + \mathbf{b}), \\ a_t = \dfrac{\boldsymbol{\beta}^T \mathbf{r}_t}{\sum_{t=1}^{T} \boldsymbol{\beta}^T \mathbf{r}_t}, \end{cases} \tag{2}$$

where $\mathbf{x}_{c_j}$ is the feature vector of each moment candidate, $\boldsymbol{\beta}$ is a trainable vector, $\mathbf{W}_q$ and $\mathbf{W}_c$ respectively represent the embedding matrix for the textual and temporal, $\mathbf{b}$ is a bias vector, and f is the Rectified Linear Unit (ReLU, $max(0; x)$) function.

After establishing the attentive embedding of each word in the query, we can construct the representation for the query as,

$$\mathbf{q} = \sum_{t=1}^{T} a_t \mathbf{e}_t. \tag{3}$$

4.2 Loss Function

Thus far, we have obtained the textual embedding $\mathbf{q}$ and the temporal embeddings of the current input moment and its surrounding moments $\mathbf{x}_{c_j}$. We hence can derive a cross-modal representation for the current query-moment pair by employing the concatenation operator as follows,

$$\mathbf{x}_{c,q} = \mathbf{x}_{c_{i-n}} \oplus \ldots \oplus \mathbf{x}_{c_{i-1}} \oplus \mathbf{x}_{c_i} \oplus \mathbf{x}_{c_{i+1}} \oplus \ldots \oplus \mathbf{x}_{c_{i+n}} \oplus \mathbf{q}, \tag{4}$$

where $\oplus$ denotes the vector concatenation operation. As a result, $\mathbf{x}_{c,q}$ is capable of encoding the information across the textual modality and visual modality.

[5]In this paper, the temporal moment contexts are the n-neighbor moment candidates surrounding the current moment.

Thereafter, we feed $\mathbf{x}_{c,q}$ into a MLP network[6], where the strong representation power of non-linear hidden layers enables complicated interactions among the features of the cross-modal representation. In addition, we can leverage the prediction layer to get the relevance score s_{cq} of the moment-query pair (c, q), as well as the location offset $[t_s - \tau_s, t_e - \tau_e]$ between the current moment and the ground truth,

$$\begin{cases} \mathbf{o}_1 = f_1(\mathbf{W}_1\mathbf{x}_{c,q} + \mathbf{b}_1), \\ \mathbf{o}_2 = f_2(\mathbf{W}_2\mathbf{o}_1 + \mathbf{b}_2), \\ \quad \vdots \\ \mathbf{o}_L = f_L(\mathbf{W}_L\mathbf{o}_{L-1} + \mathbf{b}_L), \end{cases} \tag{5}$$

where $\mathbf{W}_l$, $\mathbf{b}_l$, f_l, and $\mathbf{o}_l$ denote the weight matrix, bias vector, activation function, and the output vector of the l-th hidden layer, respectively. As for the activation function in each hidden layer, we opt for the ReLU unit. Particularly, the output vector $\mathbf{o}_L = [s_{cq}, \delta_s, \delta_e]$ consists of the matching score s_{cq} and the location offsets $\delta_s = t_s - \tau_s$ and $\delta_e = t_e - \tau_e$.

4.2.1 Alignment Loss. Similar to the spirit in [7], we cast the alignment task as a binary classification problem. Given a set of moment candidates C extracted from a video $\mathcal{V}$ and a query Q, we divide the moment-query pairs into two groups: the aligned pairs $\mathcal{P}$ and the misaligned pairs $\mathcal{N}$. We adopt the alignment loss to encourage the former to have positive scores and the latter to have negative scores, as,

$$\begin{aligned} L_{align} = & \sum_{(c,q)\in\mathcal{P}} \lambda_1 \log(1 + \exp(-s_{cq})) \\ & + \sum_{(c,q)\in\mathcal{N}} \lambda_2 \log(1 + \exp(s_{cq})), \end{aligned} \tag{6}$$

where λ_1 and λ_2 are two hyper parameters balancing the weights between the positive and negative pairs.

4.2.2 Location Loss. As the bounding box $[t_s, t_e]$ of the positive moment candidates may not exactly match the ground truth $[\tau_s, \tau_e]$, there is the location offset between the positive candidates and ground truth. We denote the location offset as $[\delta_s^*, \delta_e^*]$, and then the location offset regression is formulated as follows[7],

$$L_{location} = \sum_{(c,q)\in\mathcal{P}} |\delta_s - \delta_s^*| + |\delta_e - \delta_e^*|. \tag{7}$$

As we can see, during the training phase, the offset regression loss is only performed on positive samples. As the testing stage, once we obtain a moment candidate with the highest alignment score, we can add the predicted location with the offset values. As such, the final temporal bounding box will be close to the ground truth.

We devise the optimization framework consisting of the alignment loss and the localization regression loss processes as,

$$L = L_{align} + \lambda L_{location}, \tag{8}$$

where λ is a hyper-parameter to balance the two losses.

[6]In our experiments, the number of layers in MLP is set as two.
[7]Here, we adopted L1-norm function.

Table 1: Statistics of the Charades-STA and DiDeMo datasets.

Dataset	# Videos	# Queries	Domain	Video Source
Charades-STA	6,672	16,128	Homes	Daily Activities
DiDeMo	10,464	40,543	Open	Flickr

5 EXPERIMENTS

We first evaluate the effectiveness of our proposed model on two temporal moment localization datasets: Distinct Describable Moments (DiDeMo) dataset and Charades-STA. We then investigate how the well-designed attention network affects the localization.

5.1 Datasets

DiDeMo [1]: This dataset includes distinct video moments paired with descriptions to uniquely localize the moment in the video. It contains over 10,000 personal videos lasting 25-30 seconds duration with over 40,000 localized text descriptions. In the released dataset, each video is segmented into six five-second moments, and each moment is represented by a 4,096-d VGG [29] feature. For language features, they adopted 300 dimensional dense word embeddings Glove [23] to represent each word.

Charades-STA [7]: The Charades-STA dataset contains 6,672 videos. As the released Charades-STA dataset only contains the video-description file, we downloaded videos from the website[8] and extracted features for each moment candidate[9]. Particularly, we first segmented each video into temporal units with window size of 16 frames, and the window's overlap is 12 frames. We then extracted C3D feature[10] for each temporal unit and constructed the moment candidates with different unit sizes of [4,8,16,32]. The temporal feature of each moment candidate is the mean pooling of the features of corresponding units.

The statistics of the datasets are summarized in Table 1. The reported experimental results in this paper are based on the aforementioned datasets[11].

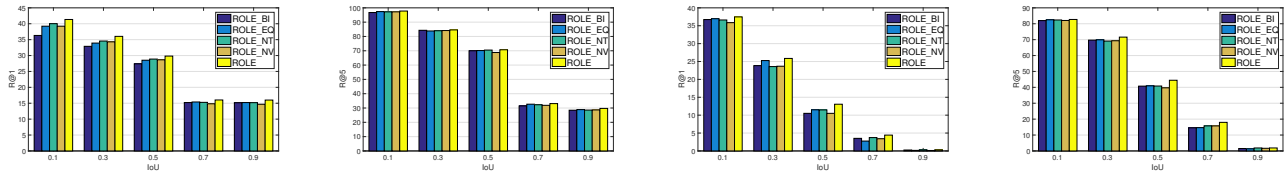
5.2 Experimental Settings

5.2.1 Evaluation Metric. To thoroughly measure our model and the baselines, we adopted "R@n,IoU=m" proposed in [11] as the evaluation metric. Specifically, given a language query, this metric computes the percentage of top-n results having IoU larger than m. We utilized $R(n, m)$ to represent "R@n,IoU=m" in the following description.

5.2.2 Baseline models. We compared our method with the following state-of-the-art baselines:

- **MCN** [1]: This model is designed for localizing the natural language queries in videos by integrating local and global video features. The query feature is extracted by the LSTM model. As it simply assumes that the given queries and video features from the corresponding moments should be close in a common space, the loss function only enforces their

[8]http://allenai.org/plato/charades/.
[9]In our paper, we utilize the videos scaled to 480p as the input videos.
[10]https://github.com/facebook/C3D.
[11]In the following experiments, we set the context moment number n as 1. And the length of context window is set as 128 frames on the Charades-STA dataset and 5 seconds on the DiDeMo dataset.



(a) R@1 vs IoU on DiDeMo (b) R@5 vs IoU on DiDeMo (c) R@1 vs IoU on Charades-STA (d) R@5 vs IoU on Charades-STA

Figure 5: Performance comparison among the variants of our proposed model over the DiDeMo and the Charades-STA datasets. From left to right: (a) the R@1 vs IoU $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the DiDeMo dataset; (b) the R@5 vs IoU $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the DiDeMo dataset; (c) the R@1 vs IoU $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the Charades-STA dataset; and (d) the R@5 vs IoU $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the Charades-STA dataset.

Table 2: Performance comparison between our proposed model and the state-of-the-art baselines on the Charades-STA (p-value*: p-value over $R(1, 0.5)$).

Method	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5	p-value*
MCN	31.26%	13.57%	4.05%	77.60%	50.53%	19.67%	5.30E-09
CTRL	35.05%	21.45%	9.30%	80.43%	65.59%	33.41%	1.80E-02
ROLE	**37.39%**	**25.26%**	**12.12%**	**82.82%**	**70.13%**	**40.59%**	-

Table 3: Performance comparison between our proposed model and the state-of-the-art baselines on the DiDeMo (p-value*: p-value over $R(1, 0.5)$).

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@1 IoU=0.9	R@5 IoU=0.5	R@5 IoU=0.7	R@5 IoU=0.9	p-value*
MCN	23.32%	15.35%	15.31%	41.03%	20.37%	19.77%	3.68E-09
CTRL	26.45%	15.45%	15.38%	68.11%	29.00%	26.12%	5.52E-04
ROLE	**29.40%**	**15.68%**	**15.55%**	**70.72%**	**33.08%**	**29.73%**	-

features to be similar in the shared embedding space. It has been detailed in Section 3.2.

- **CTRL** [7]: This is a cross-modal temporal localization strategy proposed to localize activities by natural language queries. It jointly models the textual query via the Skip-Thoughts and video moments, and outputs similarity score and action boundary. We have described it in Section 3.2.

5.3 Performance Comparison

We conducted an empirical study to investigate whether our proposed model can achieve better localization performance. The results of all methods on two datasets are presented in Table 2 and 3, respectively. Several observations stand out:

- MCN performs poorly than the other baselines, probably because simply treating the entire moment set as the context feature of each moment candidate can introduce noisy features and lead to negative transfer. Moreover, as it models the relations between the given query and moment features by only enforcing their distance to be close in the common space, the cross-modal relations have not been fully explored. In addition, it utilizes the LSTM network to encode the query, which memorizes all the words and fails to identify the distinctive words.
- CTRL achieves better moment localization results than MCN. Because it not only considers the neighbor moments as contextual information but also contains a cross-modal processing part, which can exploit the interactions across the visual and textual modalities. However, we argued that its expressiveness can be limited by encoding the whole query holistically, and as discussed in Section 1, different words contribute differently in localizing the desired moment within the input video and such coarse-grained feature will bury their varying importance.
- Our proposed ROLE model achieves the best performance, substantially surpassing all the baselines. Particularly, it shows consistent improvements over the aforementioned two baselines. This verifies the importance of integrating temporal moment contexts into the query embedding and selecting distinctive word information from the query.

Note that, in the DiDeMo dataset, since the positive moment-query pairs are well aligned (i.e., there is no location offset between them), we only utilized the alignment loss to train CTRL and ROLE.

In addition, we also conducted significant tests between our model and each of the baselines on the $R(1, 0.5)$ results. We can see that all the p-values are much smaller than 0.05, indicating that the advantage of our model is statistically significant.

5.4 Study of the ROLE

We studied variants of our model to further investigate the effectiveness of the language-temporal attention networks:

- **ROLE_NT**: We eliminated the temporal context information part of our language-temporal attention model. That is, each word attention value is only related to the query and the current moment, without considering its neighbor context.
- **ROLE_NV**: Instead of using the language-temporal attention, we adopted the query attention model which only depends on the embeddings of the query words. Namely, we eliminated all the temporal visual information.
- **ROLE_BI**: We utilized the concatenation of the last output of Bi-LSTM as the query embedding.
- **ROLE_EQ**: We set the weights of the Eqn.(3) as the average value of the words number, i.e., 1/T.

We compared these model variants on the Charades-STA and DiDeMo datasets, respectively. Figure 5 shows the results regarding the component-wise comparison:

- Jointly analyzing the performance of ROLE_NV in Figures 5(a) and 5(b), we found that removing the language-temporal attention hurts the expressiveness adversely and degrades the localization results, especially in term of R@1. This admits that only considering the query information is

insufficient to identify the keywords related to the desired moments and highlight the correlations across modalities.

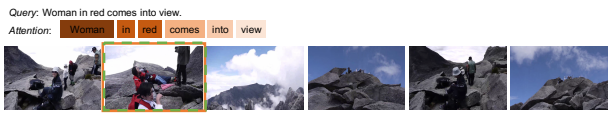
- ROLE_NT performs better than ROLE_NV, indicating that incorporating visual information of the current moment is beneficial to strengthen the comprehension of the given query. And taking advantages of the cross-modal correlations, ROLE_NT is capable of enriching the expressiveness of the model.
- Our proposed ROLE shows consistent improvements over ROLE_NT and ROLE_NV, demonstrating the importance of visual features from the moment context. Discarding the visual features (i.e., ROLE_NV) or only considering the visual feature of the current moment (i.e., ROLE_NT) overlooks the temporal correlations and further limits the expressiveness of query-moment comprehension, and consequently degrades the localization performance. In addition, the improvements over ROLE_BI and ROLE_EQ indicates that not all the words in the query are useful for the localization. Treating the given query holistically as one feature vector (i.e., ROLE_BI) or considering all the words in the query contribute equally (i.e., ROLE_EQ) introduces noisy information to the query embedding, hence influences the the localization performance.

5.5 Attention Visualization

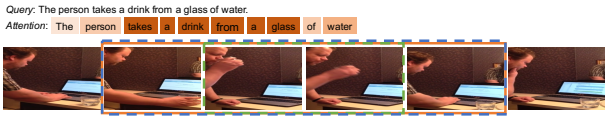
Apart from achieving the superior performance, the key advantage of ROLE over other methods is that its language-temporal attention is able to distinguish the most relevant words to the ground truth moment. Towards this end, we showed some examples, and then visualized their attention values and moment localization results as demonstrated in Figure 6.

Figure 6(a) depicts that people including women and men are enjoying the scenery on the top of a mountain with some sitting on the stone while others standing on the distant highlands. Given a query that "Woman in red comes into view", we expect the retrieved moment should contain a woman wearing a red coat. Intuitively, the distinctive information for localizing the desired moment should be "woman" and "red". We fed the video into our ROLE, as well as the query, and consequently obtained the attention score for each word in the description. Several interesting observations stand out: 1) the word "woman" is marked in the darkest orange, reflecting that this word attracts the most attention; 2) the words "in" and "red" are marked in orange reflecting fewer attentions obtained compared to "woman"; and 3) the remaining three words have the least attention values. These findings are consistent with our expectation, and further demonstrate that our proposed ROLE is capable of adaptively identifying the useful words, according to the temporal moment context. Hence, this verifies the effectiveness of our language-temporal attention network.

Figure 6(b) shows another example, where the video describes a scene: a man is typing on the notebook, drinking water from a glass, and then back to type. When retrieving a moment by "A person takes a drink from a glass of water", we expected our model to distinguish two activities in the video. From the attention result shown in Figure 6(b), we found that the words related to the action "drinking water" attract more attention than other



(a) The moment localization result on the DiDeMo.



(b) The moment localization result on the Charades-STA.

Figure 6: Visualization of the language-temporal attention on DiDeMo and Charades-STA. The Ground truth moments are outlined in the orange box, while the green dash box shows our prediction with the alignment score and the blue dash box shows the localization result with our regression correction. The word attention is presented with different colors, and the darker color states the higher value.

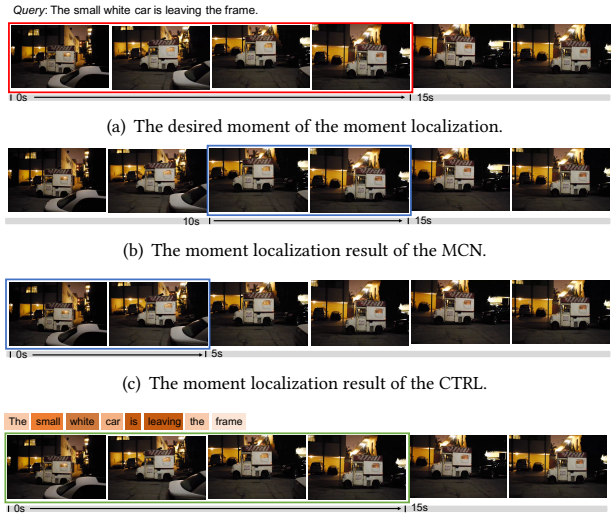
words do. This agrees with our previous analysis. Although the alignment result (i.e., the green dash box) is not that satisfactory, it indeed captures the corresponding action moment. Furthermore, our model can correct the alignment result via the regression part and provide a more accurate result (i.e., the blue dash box). This justifies the effectiveness of our proposed attention mechanism and the regression loss.

5.6 Qualitative Results

To gain the deep insights into our proposed ROLE model, we illustrated several moment localization results via different language queries. In particular, the examples from DiDeMo and Charades-STA are shown in Figure 7 and 8, respectively. In addition, we also displayed the localization results by the baselines.

Figure 7 describes a moving camera scene, where a small white car disappears and another car is still at the original place. We then utilized the aforementioned models to localize the moment corresponding to "The small white car is leaving the frame". Comparing the retrieval results from all the methods, we have the following observations:

- MCN simply returns a moment containing the "white car" from the moment candidates, ignoring the more important distinctive activity "leaving the frame". It is probably because 1) MCN treats the entire candidate set as the global feature to enhance the representation as the current moment. When most moments within the video are related to the another car, the global feature fails to represent the desired scene precisely; and 2) it adopts the last state of LSTM as the textual embedding for the query, which is insufficient to select the keyword like "leaving".
- As Figure 7(c) illustrates, CTRL achieves the unsatisfactory result by returning a moment containing two cars. Compared with MCN, while integrating neighbor moments as the context instead of the whole background, it captures parts of the desired moments since the query is modeled via the offline Skip-Thoughts tool and overlooks the sequential relation "leaving". Synthesizing the above reasons, it only returns the suboptimal moment.



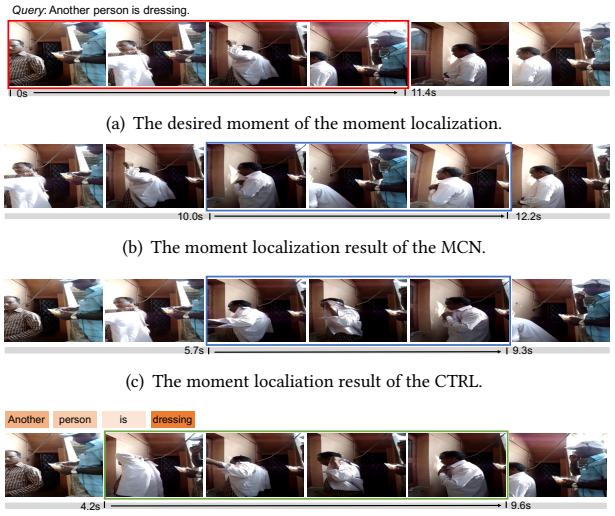
(a) The desired moment of the moment localization.

(b) The moment localization result of the MCN.

(c) The moment localization result of the CTRL.

(d) The moment localization result of the ROLE.

Figure 7: Moment localization results on the DiDeMo dataset. All the above figures are the R@1 results. The gray bar denotes the time line. The red, blue and green bounding boxes denote the ground truth, the result of baseline model and our proposed model, respectively.



(a) The desired moment of the moment localization.

(b) The moment localization result of the MCN.

(c) The moment localiation result of the CTRL.

(d) The moment localization result of the ROLE.

Figure 8: Moment localization results on the Charades-STA dataset. All the above figures are the R@1 results. The gray bar denotes the time line. The red, blue and green bounding boxes denote the ground truth, the result of baseline model and our proposed model, respectively.

- Our proposed ROLE outperforms other baselines as shown in Figure 7(d). The localized moment with the query attention indicates that our model can capture not only the desired entity "small white car" but also the sequential relation "leaving". It again verifies the effectiveness of our proposed temporal moment localization model.

Similarly, as for the example in Figure 8, our model generates more accurate results than those of other models do. As we can see, there are two persons in the video, one is dressing and finishing cloth while the other is having something. Hence the word "dressing" is the most important information to distinguish the desired moment from the others as we expected. Since MCN and CTRL model the language query holistically, the words are considered to have equal contributions to the final prediction. This may introduce noisy information into the textual representation. Moreover, they rarely identify the distinctive words, such as "dressing", from the query. As a result, MCN returns a moment that the person is finishing clothes, while CTRL selects the moment containing the end part of the dressing action. Compared with these two methods, the moment returned by our model has the largest IoU with the ground truth moment. However, it is not equal to the ground truth. The reason may be that we adopted the coarse grain sliding window to generate moment candidates for Charades-STA.

6 CONCLUSION AND FUTURE WORK

In this paper, we present a cross-modal moment localization model to localize a desired moment given the language description. To well model the explicit correspondence between the textual components and temporal moments in the video, we devise a language-temporal attention model to adaptively identify the useful word information based on the temporal context. Moreover, we also perform extensive experiments on two public benchmark datasets to demonstrate the effectiveness of our proposed model. As a byproduct, we have released the data, codes, and parameter settings to facilitate research in the community.

In the future, we plan to deepen or widen our work from the following aspects: 1) We will integrate the spatial information of the corresponding frames into our model. Because there are some descriptions related to the spatial location, such as "left of the table", we should take the spatial relationship among all the objects in a frame into account; 2) We will incorporate reinforce learning into our model to adaptively decide both where to look at next and when to predict. This will not need to generate moment candidates via the multi-scale sliding window segmentation; And 3) we plan to incorporate hashing module [18] into our model to speed up the retrieval process.

7 ACKNOWLEDGMENTS

We would like to thank all reviewers for their comments. This work is supported by Joint NSFC-ISF Research Program (No.61561146397) funded by the National Natural Science Foundation of China and the Israel Science Foundation, the National Basic Research Program of China (973) (No.2015CB352501, No.2015CB352502), National Natural Science Foundation of China (No.61772310, No.61702300, and No.61702302), the One Thousand Talents Plan of China, and the Tencent AI Lab Rhino-Bird Joint Research Program (No.JR201805). This work is also supported by National Science Foundation of China under Grant No.61429201, in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar. In addition, this work is part of NExT research, supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@SG Funding Initiative.

REFERENCES

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing Moments in Video With Natural Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5803–5812.

[2] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. 2014. Multiscale Combinatorial Grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 328–335.

[3] Y Alp Aslandogan and Clement T. Yu. 1999. Techniques and Systems for Image and Video Retrieval. *IEEE Transactions on Knowledge and Data Engineering* 11, 1 (1999), 56–63.

[4] Piotr Bojanowski, Rémi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, and Cordelia Schmid. 2015. Weakly-supervised Alignment of Video with Text. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 4462–4470.

[5] Maaike HT De Boer, Yi-Jie Lu, Hao Zhang, Klamer Schutte, Chong-Wah Ngo, and Wessel Kraaij. 2017. Semantic Reasoning in Zero Example Video Event Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 4 (2017), 1–17.

[6] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. Daps: Deep Action Proposals for Action Understanding. In *Proceedings of the European Conference on Computer Vision*. Springer, 768–784.

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5267–5275.

[8] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. 2017. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 3628–3636.

[9] Haiyun Guo, Jinqiao Wang, Min Xu, Zheng-Jun Zha, and Hanqing Lu. 2015. Learning Multi-view Deep Features for Small Object Retrieval in Surveillance Scenarios. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 859–862.

[10] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling Relationships in Referential Expressions with Compositional Modular Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4418–4427.

[11] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4555–4564.

[12] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought Vectors. In *Proceedings of the Advances in Neural Information Processing Systems*. NIPS, 3294–3302.

[13] Philipp Krähenbühl and Vladlen Koltun. 2014. Geodesic object proposals. In *Proceedings of the European Conference on Computer Vision*. Springer, 725–739.

[14] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2014. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2657–2664.

[15] Tianwei Lin, Xu Zhao, and Zheng Shou. 2017. Single Shot Temporal Action Detection. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 988–996.

[16] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. 2017. Towards micro-video understanding by joint sequential-sparse modeling. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 970–978.

[17] Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-guided Referring Expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7102–7111.

[18] Xin Luo, Liqiang Nie, Xiangnan He, Ye Wu, Zhen-Duo Chen, and Xin-Shun Xu. 2018. Fast Scalable Supervised Hashing.. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 735–744.

[19] Shugao Ma, Leonid Sigal, and Stan Sclaroff. 2016. Learning Activity Progression in LSTMs for Activity Detection and Early Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1942–1950.

[20] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 11–20.

[21] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling Context between Objects for Referring Expression Understanding. In *Proceedings of the European Conference on Computer Vision*. Springer, 792–807.

[22] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkila, and Naokazu Yokoya. 2016. Learning Joint Representations of Videos and Sentences with Web Image Search. In *Proceedings of the European Conference on Computer Vision*. Springer, 651–667.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1532–1543.

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

[25] Remi Ronfard. 2004. Reading Movies: An Integrated DVD Player for Browsing Movies and Their Scripts. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 740–741.

[26] Klaus Schoeffmann and Frank Hopfgartner. 2015. Interactive Video Search. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1321–1322.

[27] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1049–1058.

[28] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the Advances in Neural Information Processing Systems*. NIPS, 568–576.

[29] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014), 1–14.

[30] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. 2016. A Multi-stream Bi-directional Recurrent Neural Network for Fine-grained Action Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1961–1970.

[31] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong. 2011. Multiple Feature Hashing for Real-time Large Scale Near-Duplicate Video Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 423–432.

[32] Young Chol Song, Iftekhar Naim, Abdullah Al Mamun, Kaustubh Kulkarni, Parag Singla, Jiebo Luo, Daniel Gildea, and Henry A Kautz. 2016. Unsupervised Alignment of Actions in Video with Text Descriptions.. In *International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 2025–2031.

[33] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. 2015. Temporal Localization of Fine-Grained Actions in Videos by Domain Transfer from Web Images. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 371–380.

[34] Stefanie Tellex and Deb Roy. 2009. Towards Surveillance Video Search by Natural Language Query. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 1–9.

[35] Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning Language-Visual Embedding for Movie Understanding with Natural Language. *arXiv preprint arXiv:1609.08124* (2016), 1–13.

[36] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective Search for Object Recognition. *International Journal of Computer Vision* 104, 2 (2013), 154–171.

[37] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. 2015. Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 2346–2352.

[38] Rong Yan, Alexander G Hauptmann, and Rong Jin. 2003. Negative Pseudo-relevance Feedback in Content-based Video Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 343–346.

[39] Rong Yan, Jun Yang, and Alexander G Hauptmann. 2004. Learning Query-class Dependent Weights in Automatic Video Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 548–555.

[40] Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded Language Learning from Video Described with Sentences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 53–63.

[41] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. *arXiv preprint arXiv:1801.08186* (2018), 1–14.

[42] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling Context in Referring Expressions. In *Proceedings of the European Conference on Computer Vision*. Springer, 69–85.

[43] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A Joint Speaker Listener-reinforcer Model for Referring Expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7282–7290.

[44] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2017. Parallel Attention: A Unified Framework for Visual Object Discovery through Dialogs and Queries. *arXiv preprint arXiv:1711.06370* (2017), 1–11.

[45] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating Object Proposals from Edges. In *Proceedings of the European Conference on Computer Vision*. Springer, 391–405.