A Survey of Temporal Activity Localization via Language in Untrimmed Videos

Yulan Yang
School of Information and Communication Engineering,
Communication University of China
Beijing, China
yyl0217@cuc.edu.cn

Zhaohui Li
School of Information and Communication Engineering,
Communication University of China
Beijing, China
lizhh@cuc.edu.cn

Gangyan Zeng
School of Information and Communication Engineering,
Communication University of China
Beijing, China
zgy1997@cuc.edu.cn

***Abstract*—Video is one of the most informative media which consists of visual, textual and audio contents. As the number of videos on the Internet grows explosively, it is increasingly necessary for machines to understand the semantic information in the videos accurately. Temporally Activity Localization in a video is such a work which needs to localize the video moment that is most semantically similar to a given natural query. This task is quite challenging for that it not only requires to have a deep understanding of the sentences and videos, but also the fine-grained interactions between the two modalities. In this paper, we report a comprehensive survey of existed temporal sentence localization techniques. Firstly, we make a detailed classification and analysis of these methods. Then we discuss the experimental results and performance of existed approaches. Finally, we present some insights for future research direction.**

Keywords-Temporal Activity Localization, Multimodal Interaction, Video Retrieval.

I. INTRODUCTION

The huge video inventory on the network has increased the demand for more intelligent and efficient video retrieval technology. Temporal Activity Localization via Language (TALL) is an emerging multimodal retrieval task, as shown in Figure 1. It aims to localize the start and end time of a target moment in a long video which describes the same activity as the given natural query. This technology greatly reduces human labour and time spent caused by manual annotation on video moment retrieval. Therefore, it is strongly demanded in modern Computer Vision systems, including intelligent surveillance, video editing and producing. Compared to keyword-based and content-based retrieval, TALL requires a deep understanding of the semantics of two modal information and their interaction, which is very challenging, and there are a large number of works devoted to solving these problems. The goal of this paper is to present a detailed and comprehensive review of this topic so that subsequent researchers can have a clearer understanding of the research trend on this task.

In this paper, we first divide the existing methods into three aspects and discuss them in section 2. In section 3, we provide a brief overview of the existing datasets and evaluation metrics. We also discuss and compare the experimental results given by different methods in section 4. Finally, in section 5, we prospect for possible future research and give a conclusion.

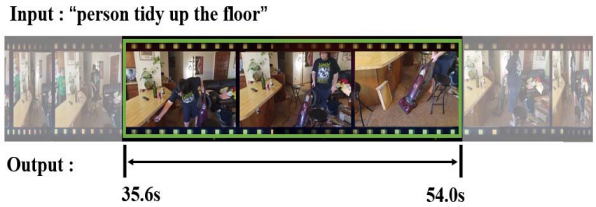

Figure 1. Example of Temporal localization via language.

II. METHODS

Considering different ways of video processing, we divide most of the existed methods of TALL into two main categories. One is called *Proposal-based Methods,* which has relatively higher prediction accuracy but longer processing time and more computing resources. Another is called *Proposal-free Methods*, whose pros and cons are on the opposite. Besides, some researchers claim that most existing methods are trained in a fully-supervised setting, which requires the dense annotations of temporal boundary for each video. However, manually labelling the ground truth (GT) temporal boundaries is time-consuming and expensive, so they turn to the research of weakly-supervised TALL tasks. Also, some researchers have considered the practicability, and they tend to study Video Corpus Moment Retrieval (VCMR) that are closely related to TALL. We will discuss these two works in the *Other Related Research* part.

A. Proposals-based Methods

The basic framework of this these methods is shown in Figure 2. The main idea is first to cut a long video into multiple proposal segments of different scales, and extract features from each proposal and the input query. Then a multimodal processing module is used to perform multimodal interaction and predict their semantic alignment score with the query, as well as time offsets deviated from the GT. Both the earliest works by Gao et al.[5] and Hendricks et al.[11] belong to these methods. Specifically, Gao et.al.[5] introduced a cross-modal regression network CTRL, to temporally model interaction between the sentence and video candidates. They first used a sliding window method to sample candidate video clips of different scales. Then they fused video and language feature by a Multimodal Processing Unit (MPU) which consists of element-wise addition, multiplication and fully connected operations, and sent the fused features to a regression network to generate the alignment score and location offsets. The MCN proposed by Henricks et al. [11] is

978-1-7281-8138-7/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCST50977.2020.00123

slightly different. It directly mapped the features of query and each candidate to a joint space, then ranked these candidates by measuring the distance between visual and text representations.

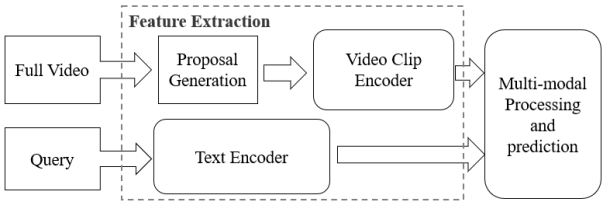

Figure 2. A block diagram of Proposal-based Methods.

As the pioneers of the TALL task, these two works provide a good paradigm for the following work. However, they always encode the whole query as one single feature, which is not expressive enough to reveal the information conveyed in the query, and they always extract the full features in the video clips no matter whether they are related to the query. Additionally, their ways of cross-modal fusion are relatively too simple to capture the complex interactions between two modalities. To improve the localization effectiveness, researchers have refined their works in the improvement of feature extraction and multimodal interaction.

1) The Improvements on Feature Extraction.

The ways of feature extraction used in [5,11] are all rough and straightforward. To further improve the feature extraction module, researchers make advances in the following ways.

a) Attention-based Methods. There are many works [14,15,12,20,30,25] focus on critical feature extraction by attention mechanism. Liu et.al.[14,15] use an attention network in their model ACRN and ROLE, to enhance visual features and keyword features, separately. These approaches work better than[5]. However, they still simply treating videos moments holistically to one global feature vector, and such a coarse-grained feature will result in crucial detail missing and localization inaccuracy. To solve this problem, Jiang et.al.[12] proposed their model SLTA, which extracts some relevant local features by spatial attention and integrates them with global motion features as final visual features. Furthermore, Yuan et.al.[30] proposed a semantic conditioned dynamic modulation (SCDM), which uses soft attention mechanism to extract the key information of the features composed of the sentence and the video, and better correlate sentence-related video contents over time. In this way, more fine-grained and semantically relevant features can be saved.

b) Adding Prior Information. There are still some research works[12,7,34,32] focus on enhancing the learning ability of their models by adding various prior information to the model. For example, Ge et al.[7] extracted the activity concepts from the video and the semantic concepts from queries to improve the temporal grounding accuracy of their model ACL. The semantic concept features are extracted from verb-obj pairs in the query and activity concept features comes from the activity classifiers' prediction scores. Zhang et al.[32] proposed a Moment Alignment Network (MAN), which Utilizes graph-structured relations between different moments to save the temporal structures information in a video and enhance the moment representation explicitly. These prior information provide crucial clues or temporal structure to the model explicitly, avoiding that they could not learn that from the model.

c) Proposals Generation After Interaction. As shown in Figure 2, the proposals generation usually happened before multimodal interaction. It may cause a problem that the localization accuracy will greatly depend on the quality of video proposal clips, which are generated by a dense sliding window method, without considering their relation to query semantics. To tackle this problem, some approaches[1,17,29] have focused on reducing the number of proposals irrelevant to the semantic of the query, by applying multimodal interaction before proposals generation. For example, Li et al. [17] proposed a Bidirectional Single-Stream Temporal Localization (BSSTL) model, which first realizes the fusion of the text with each frame in a full video, then uses an anchor-based method to sample candidate fragments centred on each frame. This way is better than dense sliding window way, but the number of proposals has not decreased too much. Moreover, inspired by the segment proposal network (SPN) used in R-C3D, Xu et al.[29] developed a Query-guided SPN(QSPN) model to get some query-specific segments as proposal candidates. By focusing on generating query-specific clip proposals, filtering unrelated clips, these methods can speed up the whole process.

2) Multimodal Processing.

Early works like Gao et al. [5] proposed a simple MPU that includes element addition, multiplication etc. to fuse features of two modalities, which is insufficient to learn the deep interaction between two modalities. Therefore, a series of works [27,17,29,25] have been presented to learn a deeper interaction between the video and query. For example, instead of using MPU, Wu et al.[27] put forward a Multimodal Circulant Fusion (MCF) module, which uses the circulant matrix of two modalities to explore all possible fine-grained interactions. Li et al.[17] chose to fuse the whole video with the word in a frame-level with various possible interactions like element-wise multiplication, concatenation and so on in BSSTL. Moreover, Xu et al.[29] used a two-layer LSTM to fuse two modalities in QSPN. Specifically, the first layer embeds the word vector of a query, then the second layer takes both the hidden value of the first layer and the visual feature of a visual proposal as inputs, to achieve fine-grained interaction. These improvements on multimodal interaction can boost the fusion and alignment of two modalities, thus improving the accuracy of localization.

B. Proposals-free Methods

Although the proposals-based methods are straight forward and achieve some good results, they have two

limitations. One is that it is computationally expensive to compare all the video proposals with the query. Another is the performance highly depends on the proposal generation process, which lacks flexibility. Therefore, later works seek to study the proposals-free methods. The main idea of these methods are directly predicting the start and end time of the target segment without generating candidate segments. According to the different deep learning methods, we divide existing approaches of this type into three categories: methods Combining RNN with Attention, RL-based methods and QA-based methods. We will discuss these works in the following part.

1) Combining RNN with Attention

A full video contains a huge amount of semantic information, but the useful information related to the retrieved query only accounts for a small part. Therefore, to highlight this small valuable part, many research[2,31,21] use recurrent neural networks to encode video and text, then use the attention model to filter the related information from the entire video. For example, Yuan et al.[31] first use two Bi-directional LSTM networks as video and sentence encoder, separately. Then they present a cross-modal co-attention network to generate both query sentence and video attentions, filtering out irrelevant information. Finally, the attention vectors are sent to a novel Attention-Based Location Regression (ABLR) network to regress the temporal boundaries of the query. Cristian et al.[21] adopt a slightly different way, they use a dynamic filter to transfer language information to the visual domain to fuse the sentence and video information, then predict the possibilities of "start and end time" for each frame. It should be noted that Cristian et al.[21] are the first to attend to the subjectivity of annotation and consider this in their model. These methods have many similarities with proposal-based methods, but they require a deeper understanding of the two modalities to achieve high accuracy.

2) Reinforcement Learning-based Methods

With the rise and widespread application of Reinforcement Learning (RL), researchers[10,26,9,28] began to consider applying the idea of RL to solve TALL task. They treated TALL as a sequential decision problem, and solve it by RL-based methods. He et.al.[10] is the first one to use the reinforcement learning idea to solve this task. They used an agent to adjust the predicting temporal boundary according to the learned policy. Specifically, they first initialed a random boundary for the target. Then they learn the current state vector which fuses multimodal features, and use its state vector to determine the next action. The action space consists of 7 different ways to adjust the temporal boundaries. Learning in this way for several times, the model can gradually approach the target segment. Furthermore, Wu et al.[28] think that, to make more reasonable and interpretable decisions in our daily life, people usually infer the deliberative process from coarse-to-fine, so they propose a novel Tree-Structured Policy-based Progressive Reinforcement Learning (TSP-PRL) framework to make coarse-to-fine decisions. The root policy in the tree determines the action categories, including scale variations, left shift and right shift, and the leaf policy determines the step of these actions. Compared with methods based on supervised learning, RL-based methods are more computation resource-saving because they can reach for more flexible boundaries and avoid exhaustive sliding window searches.

3) Question and Answering-based Methods

In addition to treating TALL as a sequential decision problem, some people tried to treat TALL as a Question and Answering (QA) problem by regarding the video as a text paragraph and the target moment as the corresponding answer span. Therefore, some multimodal QA methods [33,2,8] can be used to solve this problem. Ghosh et al.[8] regard each video frame as a word in a passage, and predict the possibility of the frame to be the answer (i.e. the start and end point) by exploiting multimodal interactions between the video and query sentence. However, Ghosh et al.[8] don't adopt the standard span-based QA framework strictly, and they don't consider the differences between traditional text span-based QA and NLVL tasks. Furthermore, Zhang et al.[33] solve TALL task with a standard span-based QA framework VSLNet, which utilizes a Query-Guided Highlighting (QGH) strategy to address the differences between these two tasks.

C. Other Related Research

This section we will not discuss the improvement of the algorithm in the task we mentioned. We mainly give a brief introduction of two new types of research closely related to the TALL.

1) Weakly Supervised TALL

The works given above are mostly based on supervised deep learning methods, which inevitably require a large amount of labelled data. Due to the limitations of training data, the practical application of these models will also be limited. Therefore, some research works [6,24,19,3,23,18] proposed a weakly-supervised learning TALL task to alleviate the problem of fine-grained labelling. In these works, you only need to give a description of the whole video to train the entire model, without marking the specific time location. For example, Song et.al.[23] proposed a novel multilevel attention reconstruction model (MARN). This model uses the inter-interaction and intra-interaction between the candidate fragments to learn the language-driven attention weighting map, which can be used in the testing stage Attention weights are learned by a reconstruct loss (visual features to language captioning reconstruction). It can be seen that the solutions to weakly-supervised TALL tend to replace the localization loss of the densely labeled training set with other constraints like reconstruction loss, and the principle of these constraints is that the text and the corresponding video clip can realize the cross-modal transformation due to their similarities in semantics.

2) Video Corpus Moment Retrieval

The methods proposed above all belong to Single Video Moment Retrieval (SVMR), which aims to localize a target moment in a single video corresponding to a given natural query. But in reality, we generally need to find the target segments in a video database, which contains a large number of videos. So [16] and [4] extend SVMR to Video Corpus Moment Retrieval (VCMR), which requires to search for several most relevant video segments from a large video collection instead. It can be seen that VCMR has more practical value than SVMR, and we can pay more attention to it if we consider more of the applicability in TALL.

III. DATASETS AND EVALUATION METRICS

In this section, we will briefly introduce some commonly used datasets, some of which are newly collected and others are modified from other tasks like video captioning. Then we also introduce and analyze commonly used evaluation metrics of TALL.

TABLE I. BASIC INFORMATION ABOUT DATASETS FOR TALL.

Dataset	#Videos/#Clips	#Sentence	Video Source	Domain
TACoS	127/7206	18226	Lab Kitchen	Cooking
Charades-STA	6672/11772	16124	Activity	Indoor Activity
ActivityNet Caption	19209/-	71942	Activity	Open Activity
DiDeMo	10464/20892	40543	Flickr	Open Activity

The widely used datasets includes TACoS [22], Charades-STA [5], ActivityNet Captions [13] and DiDeMo[11], etc. Here is a brief introduction to these datasets in Table Ⅰ.

A. Evaluation Metrics

1) R(n,m). R(n,m) is also called "R@n, IoU=m", which represents the percentage of test samples which satisfied that in top n results, there is at least one segment having higher Intersection over Union (IoU) with GT than m. For each test sample q_i, the score is marked as:

$$r(n, m, q_i) = 1 \quad (1)$$

For all samples in the testing set, the average score is calculated by:

$$R(n, m) = \frac{1}{N}\sum_{i=1}^{N} r(n, m, q_i) \quad (2)$$

R(n,m) is mainly used to evaluate the performance of proposal-based methods. For proposal-free methods, R(1,m) is often used to assess their performance due to there is only one moment predicted by the model.

2) mIoU. mIoU refers to average IoU of top 1 result with GT over all test samples.

IV. COMPARISON OF BENCHMARK DATASETS AND COMMON EVALUATION METRICS

This paper does not conduct a formal experimental evaluation on these TALL methods, and we just present a brief analysis of the experimental results of some typical methods. The results are taken from their papers. We present two sets of results:

(1) We found that most of the methods calculate the R(m,n) and mIoU under the fully supervised conditions, and we summarize these test results on TACoS, Charades-STA, and ActivityNet Caption, separately. These results are shown in Table Ⅱ and Table Ⅲ.

(2) There are a growing number of researches on weakly-supervised TALL, which has great potential in reducing the cost of annotation. We summarize the results in Table Ⅳ.

1) Discussions on Proposal-based Methods

In Table Ⅱ, ACRN[14] and SCDM [30] have added attention mechanism to the process of feature extraction, and we can see that both of them perform better than the basic model CTRL[5], demonstrating the effectiveness of attention mechanisms. In addition, SCDM[30] performs significantly better than ACRN[14], improving 6.4% on average on TACoS. The reason is that in addition to using attention, SCDM[30] also uses hierarchical convolution layers to encode video feature, with the sentence to modulate the temporal convolution process, and the temporal convolution module provide various semantic interactions of different granularities.

TABLE II. RESULTS OF PROPOSAL-BASED METHODS.

Methods	R@1,IoU			R@5,IoU		
	0.7	*0.5*	*0.3*	*0.7*	*0.5*	*0.3*
TACoS						
CTRL[5]	-	13.30	18.32	-	25.42	36.69
ACRN[14]	-	14.63	19.52	-	24.88	34.97
SCDM[30]	-	21.17	26.11	-	32.18	40.16
BSSTL[17]	-	18.73	22.31	-	29.89	40.87
CBP[25]	19.10	24.79	27.31	25.59	37.40	43.64
ACL[7]	-	20.01	24.17	-	30.66	42.15
Charades-STA						
CTRL[5]	8.89	23.63	-	29.52	58.92	-
SCDM[30]	33.43	54.44	-	58.08	74.43	-
QSPN[29]	15.80	35.60	54.7	45.4	79.4	95.6
CBP[25]	18.87	36.80	-	50.19	70.94	-
ACL[7]	12.20	30.48	-	35.13	64.84	-
MAN[32]	22.72	46.53	-	53.72	86.23	-
ActivityNet Caption						
SCDM[30]	19.86	36.75	54.80	41.53	64.99	77.29
QSPN[29]	13.60	27.7	45.3	38.3	59.2	75.7
BSSTL[17]	-	47.68	55.32	-	57.53	70.53
CBP[25]	17.80	35.76	54.30	46.20	65.89	77.63

BSSTL[17], CBP[25] and QSPN[29] all belong to "Proposals Generation After Interaction" methods. It can be seen from Table Ⅱ that all of them have higher accuracy than CTRL[5] and other "Proposals Generation Before Interaction" methods like ACRN[14], which proves that using query information to guide the proposal generation can improve the effectiveness. Additionally, comparing the performance among these models, we can see that CBP[25] achieves better results in both TACoS and Charades-STA datasets, especially in the case of higher IoU, like IoU=0.7. We think it is because CBP[25] adds a boundary submodule to do

binary classification on the boundary of proposals, thus strictly restraining the localization accuracy. In contrast, QSPN[29] performs better than CBP[25] in "R@5" both on Charades-STA and ActivityNet Caption. We argue that the Query-guided SPN used in QSPN[29] make a difference, which helps model select query-specific proposals, so the target moment has more likely to be retrieved in Top 5.

Furthermore, ACL[7] and MAN[32] both add useful prior information. Compared to CTRL[5], ACL[7] only adds visual and semantic concept features to the encoder, but we can see from the Table Ⅱ that ACL[7] largely outperforms CTRL[5] in TACoS and Charades-STA. Additionally, after introducing the temporal graph convolution structure, MAN[32] achieves much higher accuracy than CTRL[5] in Charades-STA. Therefore, we could infer that adding useful prior information is helpful to higher localization accuracy.

Moreover, the ways of multimodal interaction have been improved in BSSTL[17] and QSPN[29], and both of them are better than the CTRL[5], which only uses a simple fusion way. As we can see, in ActivityNet Caption, BSSTL[17] performs better than QSPN[29] in "R@1", but worse in "R@5". It is because BSSTL[17] uses a frame-by-word fine-grained fusion way while QSPN[29] fuses the whole feature of the proposal with each word in the query by LSTM, which is less accurate.

TABLE III. RESULTS OF PROPOSAL-FREE METHODS.

Methods	R@1,IoU				mIoU
	0.7	*0.5*	*0.3*	*0.1*	
TACoS					
TripNet[9]	9.52	19.17	23.95	-	-
VSLNet[33]	20.03	24.27	29.61	-	24.11
Charades-STA					
TripNet[9]	14.50	36.61	51.33	-	-
TSP-PRL[28]	24.73	45.30	-	-	40.93
VSLNet[33]	35.22	54.19	70.46	-	50.02
KL+AL[21]	33.74	52.02	67.53	-	-
ActivityNet Caption					
TripNet[9]	13.93	32.19	48.42	-	-
TSP-PRL[28]	-	38.76	56.08	-	39.21
VSLNet[33]	26.16	43.22	63.16	-	43.19
ABLR[31]	-	36.99	55.67	73.30	36.99
KL+AL[21]	19.26	33.04	51.28	75.25	37.78

2) Discussions on Proposal-free Methods

The original goals of methods presented in Table Ⅲ are reducing the computation cost (this metric is not discussed in this paper). Still, it can be seen that most of their localization accuracy are better than the results of proposal-based methods shown in Table Ⅱ. Among them, TripNet[9] and TRP-PRL[28] are both RL-based methods, and TRP-PRL[28] is superior to TripNet in all datasets, indicating that the coarse-to-fine decision-making method is more helpful to the TALL task. VSLNet [33] uses QA-based methods to solve TALL models, and its results can achieve state of the art, for that QA tasks and TALL tasks are essentially the same problem, and we can learn from each other to solve the problems.

3) Discussions on different datasets

We also compared the performance of the same method on different datasets. From Table Ⅱ and Table Ⅲ, we can see that most methods perform best on Charades-STA and the worst on TACoS. Take the results of VSLNet[33] and SCDM[30] for example. The accuracy of SCDM[30] on Charades-STA is twice as much as on TACoS, as well as VSLNet[33]. We think it is because there are little changes of a video scene in TACoS datasets, while the events that need to be located are relatively fine-grained. Additionally, although Charades-STA and ActivityNet Captions are both open scene datasets, Charades-STA is limited to indoor activities, and its category and complexity are not as much as the latter, so it is relatively easier to localize, which also can be seen from the results of SCDM[30] and VSLNet[33].

4) Discussions on Weakly Supervised TALL

From Table Ⅳ, we can see that WSLLN[6] performs the best on ActivityNet Caption, and MARN[23] works well on both datasets. Comparing with the supervised learning methods in Table Ⅱ and Ⅲ, it is surprising that the best results of the weakly supervised method WSLLN[6] can be the same as, or even exceed, the best of supervised method VSLNet[33] on ActivityNet. These results show that without the restriction of strong supervision, a large amount of data can still help the model learn the semantic relationship of different modalities implicitly, which also reflects the advantages of big data.

TABLE IV. RESULTS OF WEAKLY SUPERVISED TALL METHODS.

Methods	R@1,IoU			R@5,IoU		
	0.7	*0.5*	*0.3*	*0.7*	*0.5*	*0.3*
Charades-STA						
TGA[19]	8.84	19.94	31.14	33.51	65.52	86.58
CTF[3]	12.90	27.30	39.80	-	-	-
SCN[18]	-	29.22	47.23	-	55.69	71.45
wMAN[24]	13.71	46.53	48.04	37.58	72.17	89.01
MARN[23]	14.81	31.94	48.55	37.40	70.00	90.70
ActivityNet Caption						
SCN[18]	9.97	23.58	42.96	38.87	71.80	95.56
CTF[3]	-	23.60	44.30	-	-	-
MARN[23]	-	29.95	47.01	-	57.49	72.02
WSLLN[6]	22.70	42.80	75.40	-	-	-

V. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this paper, we provide a comprehensive review of the task of localizing moment in a video with natural language. We have given a taxonomy of existing techniques and discussed their pros and cons. We discussed the experimental results of di·erent methods on publicly used evaluation metrics and datasets, subsequently. It can be seen that most of the current work is aimed at improving the accuracy or efficiency of the algorithm. In addition, we can see from some recent research that the research is developing towards more intelligent (e.g., weakly supervised TALL) and practicality (e.g., multi-video TALL). We believe that more studies will focus on these two points in the future. Furthermore,

localizing moments in a video with natural language may also be improved in the following aspects：

(1) Because the video is a kind of multimedia data, the sound and text in the video can be used as relevant information to improve the semantic richness of the video.

(2) Explore the application of zero-shot learning in temporally language grounding to reduce the dependence on supervised deep learning, thereby reducing the cost of manually labelling and improving the scalability of the model.

(3) Using user feedback technology, users can interact with the system during the retrieval process to achieve desired retrieval performance gradually.

REFERENCES

[1] Chen J, Chen X, Ma L, et al. Temporally grounding natural sentence in video[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 162-171.

[2] Chen J, Ma L, Chen X, et al. Localizing natural language in videos[C]. AAAI, 2019.

[3] Chen Z, Ma L, Luo W, et al. Look Closer to Ground Better: Weakly-Supervised Temporal Grounding of Sentence in Video[J]. arXiv preprint arXiv:2001.09308, 2020.

[4] Escorcia V, Soldan M, Sivic J, et al. Temporal Localization of Moments in Video Collections with Natural Language[J]. arXiv preprint arXiv:1907.12763, 2019.

[5] Gao J , Sun C , Yang Z , et al. TALL: Temporal Activity Localization via Language Query[J]. ICCV2017.

[6] Gao M, Davis L S, Socher R, et al. WSLLN: Weakly Supervised Natural Language Localization Networks[J]. arXiv preprint arXiv:1909.00239, 2019.

[7] Ge R, Gao J, Chen K, et al. MAC: Mining Activity Concepts for Language-based Temporal Localization[C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019: 245-253.

[8] Ghosh S, Agarwal A, Parekh Z, et al. Excl: Extractive clip localization using natural language descriptions[J]. arXiv preprint arXiv:1904.02755, 2019.

[9] Hahn M, Kadav A, Rehg J M, et al. Tripping through time: Efficient Localization of Activities in Videos[J]. arXiv preprint arXiv:1904.09936, 2019.

[10] He D, Zhao X, Huang J, et al. Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos[J]. arXiv preprint arXiv:1901.06829, 2019.

[11] Hendricks L A , Wang O , Shechtman E , et al. Localizing Moments in Video with Natural Language[J]. ICCV2017:5804-5813.

[12] Jiang B, Huang X, Yang C, et al. Cross-Modal Video Moment Retrieval with Spatial and Language-Temporal Attention[C]//Proceedings of the 2019 on International Conference on Multimedia Retrieval. ACM, 2019: 217-225.

[13] Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Niebles, J. C.2017. Dense-captioning events in videos. In ICCV, 706–715.

[14] Liu M, Wang X, Nie L, et al. Attentive moment retrieval in videos[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 2018: 15-24.

[15] Liu M, Wang X, Nie L, et al. Cross-modal moment localization in videos[C]//2018 ACM Multimedia Conference on Multimedia Conference. ACM, 2018: 843-851.

[16] Lei, Jie, et al. "TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval." arXiv preprint arXiv:2001.09099 (2020).

[17] Li C, Zhao Y, Peng S, et al. Bidirectional Single-Stream Temporal Sentence Query Localization in Untrimmed Videos[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 270-274.

[18] Lin Z, Zhao Z, Zhang Z, et al. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network[J]. arXiv preprint arXiv:1911.08199, 2019.

[19] Mithun N C, Paul S, Roy-Chowdhury A K. Weakly supervised video moment retrieval from text queries[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 11592-11601.

[20] Ning K, Zhu L, Cai M, et al. Attentive Sequence to Sequence Translation for Localizing Clips of Interest by Natural Language Descriptions[J]. arXiv preprint arXiv:1808.08803, 2018.

[21] Opazo C R, Marrese-Taylor E, Saleh F S, et al. Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention[J]. arXiv preprint arXiv:1908.07236, 2019.

[22] Rohrbach, M.; Amin, S.; Andriluka, M.; and Schiele, B.2012a. A database for fine grained activity detection of cooking activities. In CVPR, 1194–1201.

[23] Song Y , Wang J , Ma L , et al. Weakly-Supervised Multi-Level Attentional Reconstruction Network for Grounding Textual Queries in Videos[J]. 2020.

[24] Tan R, Xu H, Saenko K, et al. wman: Weakly-supervised moment alignment network for text-based video segment retrieval[J]. arXiv preprint arXiv:1909.13784, 2019.

[25] Wang J , Ma L , Jiang W . Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction[J]. AAAI2020.

[26] Wang W, Huang Y, Wang L. Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 334-343.

[27] Wu A, Han Y. Multi-modal Circulant Fusion for Video-to-Language and Backward[C]//IJCAI. 2018, 3(4): 8.

[28] Wu J, Li G, Liu S, et al. Tree-Structured Policy based Progressive Reinforcement Learning for Temporally Language Grounding in Video[J]. arXiv preprint arXiv:2001.06680, 2020.

[29] Xu H, He K, Plummer B A, et al. Multilevel language and vision integration for text-to-clip retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 9062-9069.

[30] Yuan Y, Ma L, Wang J, et al. Semantic conditioned dynamic modulation for temporal sentence grounding in videos[C]//Advances in Neural Information Processing Systems. 2019: 536-546.

[31] Yuan Y, Mei T, Zhu W. To find where you talk: Temporal sentence localization in video with attention based location regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 9159-9166.

[32] Zhang D, Dai X, Wang X, et al. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1247-1257.

[33] Zhang H, Sun A, Jing W, et al. Span-based Localizing Network for Natural Language Video Localization[J]. arXiv preprint arXiv:2004.13931, 2020.

[34] Zhang S, Su J, Luo J. Exploiting Temporal Relationships in Video Moment Localization with Natural Language[C]//Proceedings of the 27th ACM International Conference on Multimedia. ACM, 2019: 1230-1238.