

# Multi-scale 2D Representation Learning for weakly-supervised moment retrieval

Ding Li<sup>1,2</sup>, Rui Wu<sup>3</sup>, Yongqiang Tang<sup>1</sup>, Zhizhong Zhang<sup>1,2</sup>, Wensheng Zhang<sup>1,2</sup>

<sup>1</sup>Institute of Automation, Chinese Academy of Sciences.

<sup>2</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.

<sup>3</sup>Horizon Robotics, China.

(liding2019, tangyongqiang2014, zhangzhizhong2014)@ia.ac.cn, rui.wu@horizon.ai, zhangwenshengia@hotmail.com

**Abstract**—Video moment retrieval aims to search the moment most relevant to a given language query. However, most existing methods in this community often require temporal boundary annotations which are expensive and time-consuming to label. Hence weakly supervised methods have been put forward recently by only using coarse video-level label. Despite effectiveness, these methods usually process moment candidates independently, while ignoring a critical issue that the natural temporal dependencies between candidates in different temporal scales. To cope with this issue, we propose a Multi-scale 2D Representation Learning method for weakly supervised video moment retrieval. Specifically, we first construct a two-dimensional map for each temporal scale to capture the temporal dependencies between candidates. Two dimensions in this map indicate the start and end time points of these candidates. Then, we select top-K candidates from each scale-varied map with a learnable convolutional neural network. With a newly designed Moments Evaluation Module, we obtain the alignment scores of the selected candidates. At last, the similarity between captions and language query is served as supervision for further training the candidates’ selector. Experiments on two benchmark datasets Charades-STA and ActivityNet Captions demonstrate that our approach achieves superior performance to state-of-the-art results.

## I. INTRODUCTION

Video moment retrieval can facilitate a lot of multimedia applications, *e.g.* video surveillance, sport analytics and short-term video recommendation.

Therefore, it has drawn much research interest in recent years[4], [27], [23], [24]. This task aims to search the moment most relevant to the given text query in an untrimmed video. Taking Fig. 1 as an example, given a text query “A person is eating a sandwich.”, we want to know when this event starts and ends in the whole video.

During the past several years, deep learning based approaches have greatly promoted the development of video moment retrieval. Most of these methods use a fully-supervised training manner, which requires accurate annotations of the start and end time points of the corresponding moments for given text queries. However, manually labelling temporal boundary of the moments is time-consuming and of high cost. Besides, the temporal boundaries of moments are usually ambiguous to define, which brings more difficulties for accurate labelling. To remedy the above issues, more recently, intense attention [14], [13], [21], [5] is being paid for developing a weakly-supervised training mechanism, which merely requires video-level description for training data and thus leads to the

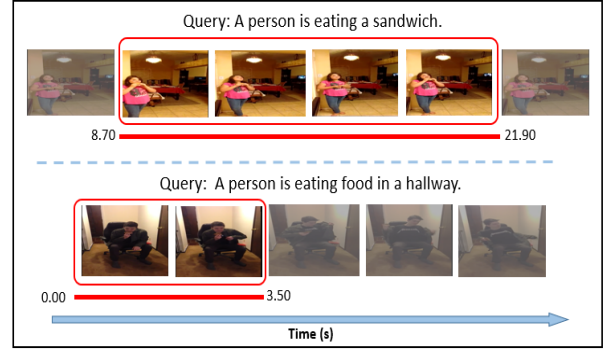


Fig. 1. Video moment retrieval task: Localize the best-matched video moments in an untrimmed video for the given text query.

significant cost saving. Although several works have made in the weakly-supervised setting, there are still two crucial issues requiring to be handled.

First, most existing weakly-supervised methods resort to projecting the text features and video features of moment candidates into some learned unified space, and then calculate the alignment score of candidates with text query, the larger score indicates the higher probability to be the result. However, these methods process each moment candidate individually, thus the relations between video moments are inevitably neglected. Second, the present weakly-supervised moment retrieval methods generally overlook the fact that the variance of temporal scale of video moments is also an important influence factor for moments localization. For example, both text queries shown in Fig.1 are devoted to describing a similar event of person eating sandwich, but the temporal lengths of the corresponding video moments varies greatly.

For the first issue, we spot that this issue has been concerned in several works for fully-supervised temporal action detection. [16] employed self-attention mechanism and update the features by aggregating the information from other candidates with learned weights, but it brings much computational cost. [26] constructed a candidate graph updated by graph neural network(gnn) [18], in which relations between moment candidates are implicitly represented by the edges between candidate nodes. The constructed graph does not characterize

temporal dependencies between nodes explicitly, [27] then proposed the 2D-TAN method, which consists of a single-scale 2D temporal feature map to explicitly represent and capture the temporal dependencies between moment candidates and has achieved promising performance. However, the lack of temporal boundary annotations is not conducive to handle the variance of temporal scale and design loss function in training time, which makes it difficult to directly transfer the proposed framework in these fully-supervised methods into the weakly-supervised task. For the second issue, in fully-supervised setting, *e.g.* 2D-TAN, such challenge is weakened by the strong supervision of massive accurate moment annotations, but we note that the single-scale 2D map applied in [27] achieved limited success when generate more precise moment candidates, which contributes little to the weakly-supervised training.

These considerations motivate us to propose a multi-scale 2D Representation Learning model for Weakly-supervised moment retrieval. The key idea is to construct multiple 2D temporal feature maps with different temporal sampling scale, and then evaluate the alignment scores of moment candidates. The representation learning model resort to the two-stage pipeline, and consists of Multi-scale 2D Temporal Network and Moment Evaluation Module. In the Multi-scale 2D Temporal Network, we construct the Multi-scale 2D Temporal Map and perform convolution over the map to capture the temporal context. In the Moment Evaluation Module, we introduce the video caption module for each input moment candidate, and generate pseudo label for training. We evaluate our proposed method on two popular benchmark datasets for video moment retrieval, *e.g.* Charades-STA and ActivityNet Captions Dataset.

The main contribution of this paper can be concluded as follows:

1. We introduce a novel multi-scale 2D temporal network, which elaborate multi-granularity moment candidates generation and captures temporal dependencies between moment candidates.
2. We propose a moment evaluation module with reconstruction-guided binary cross-entropy loss (RG-BCE loss), which facilitates the weakly-supervised training.
3. Experiment results on the two benchmark datasets (Charades-STA and ActivityNet Captions) verify the effectiveness of our proposed method.

The rest of this paper is organized as follows. In Section II, we briefly review some related works, followed by the introduction of proposed multi-scale representation learning method in Section III. Experimental results and discussions are showed in Section IV. Finally, we conclude this paper in Section V.

## II. RELATED WORKS

In this section, we will mainly focus on the related works of temporal action detection and recent related advances in video moment retrieval via text queries.

Temporal action detection aims at localizing boundaries and classifying category of action instances in untrimmed videos. The two-stage method first generates action instances with temporal boundaries and followed by classifier. These works mainly focus on generating proposals with precise boundaries. [12] adopted three activeness curves to locate flexible proposal boundaries, [26] used graph network to extract features between different proposals, [11] used two feature maps separately for completeness regression and temporal boundary classification. By contract, the one-stage method integrates location and classification into a single step and hence achieves higher efficiency.

Besides, the weakly-supervised temporal action localization only uses video-level action category as label when detecting the temporal boundaries. Autoloc [19] regressed the confidence scores and then generates more accurate proposals. BaS-Net [10] proposed an asymmetrical two-branch weight-sharing architecture to handle the background. However, the temporal action are limited to the pre-defined simple action category, which is not flexible to some video understanding applications.

To overcome aforementioned limitation of temporal action detection, Gao [4] and Hendricks [1] introduced the video moment retrieval via text queries. [4] proposed to jointly model video clips and text queries using multi-modal operations, then alignment scores and location offsets were predicted based on the multi-model representation. [1] proposed to embed both modalities into a common space and minimize the squared distances. [25] followed a two-stage pipeline to retrieve video clips. They first generated query-specific proposals from the videos, then utilized caption reconstruction. In [2], a visual concept based approach was proposed to generate proposals, followed by proposal evaluation and refinement. [24] explored reinforcement learning to find the corresponding segments. [27] introduced the 2d temporal feature map to represent the moment candidates with temporal relations, and achieved better performance.

Inspired by the success of the weakly-supervised temporal action detection, a small number of works are proposed to retrieve best-matching video moment without annotations of temporal boundaries. [3] decomposed the problem of weakly-supervised dense event captioning in videos into a cycle of dual problems: caption generation and moment retrieval, and explores the one-to-one correspondence between the temporal segment and event caption. [14] proposed a weakly-supervised joint visual-semantic embedding framework for moment retrieval, and utilizes the latent alignment for localization during inference. [21] exploited a multi-level co-attention mechanism which comprises of a Frame-By-Word interaction module as well as a novel Word-Conditioned Visual Graph (WCVG), and incorporate the positional embedding in the temporal sequence. [5] designed an alignment branch and a detection branch, and merge the moment-text matching score for the evaluation. [13] constructed a novel semantic completion network for moment candidates evaluation, and exploited the alignment relationship. However, these methods processed the moment candidates individually, which neglects the temporal

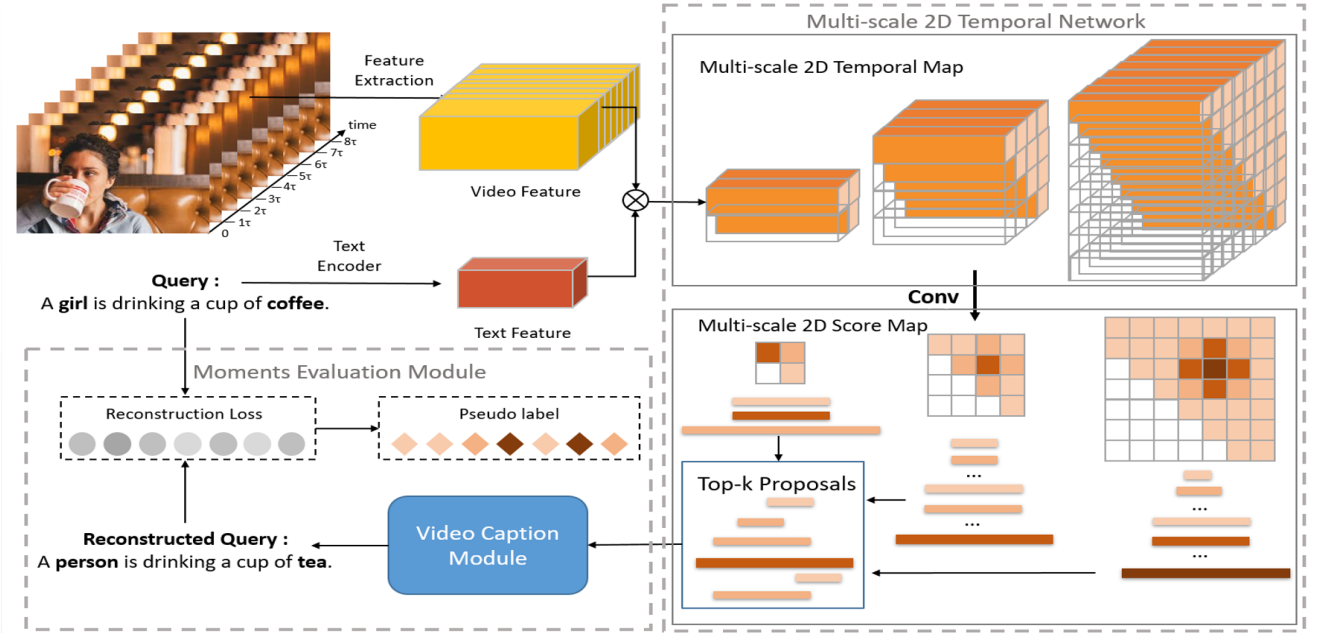


Fig. 2. The Framework of our Multi-scale Representation Learning for weakly-supervised moment retrieval. Taking a video-sentence pair as input, we extract the basic video and text representations by Feature Extractor and Text Encoder. After that, we construct the Multi-scale 2D Temporal Network which consists of 2D feature map and 2D score map. Then, we reconstruct the text query based on the top-k moment candidates in the map, and generate the pseudo labels for training.

context information.

### III. APPROACH

In this section, we first introduce the problem definition of this task, and then present the pipeline for the Multi-scale 2D Representation Learning, including Multi-scale 2D Temporal Network and Moments Evaluation Module. In the proposed pipeline, we utilize a multi-scale 2D temporal map to represent the video segments, and employ 2D Temporal Network on the constructed maps. Then, top-k moment candidates are selected from the set of maps and taken as input of Moments Evaluation Module. Eventually, we can get the final video moments according to the evaluation score.

#### A. Problem Definition

As mentioned before, the goal of this paper is to retrieve video moments of interest in a weakly-supervised setting. Given a video denoted as  $V = \{v_i\}_{i=1}^{N_v}$  and a sentence  $T = \{t_i\}_{i=1}^L$  as text query, we aim to automatically retrieve the most relevant video segment according to the query.  $N_v$  is the number of the frames of the video, and  $L$  is the length of the sentence. Specifically, we can get the best-matched moment  $M = \{\tau_s, \tau_e\}$ , where  $\tau_s, \tau_e$  are the indices of start and end frame respectively. Note that there is no need to have access to the temporal boundary annotations of video moments in the training time.

#### B. Basic Video and Text Representation

This section introduces the basic feature representation of the input text query and untrimmed video.

**Video Representation.** As for the given untrimmed video  $V = \{v_i\}_{i=1}^{N_v}$ , we first split the whole video into several video clips, then each video clip would be used as the input of a pre-trained 3D CNN model. In the procedure of video split, we utilize the multiple fixed intervals to sample frames from the original video, which facilitates the construction of the multi-scale 2D map. Following the setting of [27], spatio-temporal feature are extracted by the pre-trained 3D CNN model, and then passed through a fully-connected layer with  $d^v$  output channels.

**Text Representation.** The text encoder includes the word embedding and LSTM network [6]. We use GloVe word2vec model to extract the word embedding of each word in the input sentence. For each word  $t_i$  in the input sentence, the respective embedding vector are generated as  $w_i \in R^{d^T}$ ,  $d^T$  is the length of the vector. The embedding vector  $\{w_i\}_{i=1}^L$  are then fed into the three-layer bidirectional LSTM network, and we utilize the last hidden state as the text representation of the sentence. The final text feature are extracted as  $f^T \in R^{L \times d^T}$ , which encodes the input text query.

#### C. Multi-scale 2D Temporal Network

The Multi-scale 2D Temporal Network takes an the basic video and text representation as input, and outputs  $N_s \times K$  segment proposals and corresponding alignment scores respectively.  $N_s$  is the number of scales, and  $K$  is the number of selected segment proposals in each scale.

To get the segment proposals more precisely, we perform multi-scale temporal sampling on the untrimmed video  $V =$

$\{v_i\}_{i=1}^{N_v}$ . Specifically, we first segment it into small video clips. Each video clip consists of  $T$  frames. Then, we repeatedly sample the video clips with  $N_s$  intervals. After  $j$ -th sampling, we get  $N_j$  video clips, and the original video would be converted to  $V = \{S_i^j\}_{i=1, j=1}^{N_j, N_s}$ , where  $S_i^j$  is the sampled clips. Furthermore, we extract the deep 3D CNN feature of each clip as mentioned before, denoted as  $\{f^S\}_{i=1, j=1}^{N_j, N_s}$ . To get a more compact representation, we pass the extracted feature through a fully-connected layer with  $d^v$  output channels. For the  $i$ -th video segment sampled in  $j$ -th scale, the final 3D feature is  $f^S \in R^{d^v}$ , where  $d^v$  is the feature dimension.

The video segments obtained by multi-scale temporal sampling are set as the input of the multi-scale 2D temporal feature map construction, and each grid in the map represents a moment candidate with start and end indexes along the axis. The moment feature of each grid in the 2D map are extracted on the basis of the 3D segment feature  $f^S \in R^{d^v}$ . In the extraction process, we follow the temporal pooling design. For each moment candidate, we perform max-pooling operation on the corresponding segments with the reference of the start and end indexes in the 2D map. For the moment in the  $x$ -th row and  $y$ -th column of the map, we can obtain the feature of this moment candidate:

$$F_{x,y}^j = \begin{cases} \max \text{pool} (f_x^S, f_{x+1}^S, \dots, f_y^S), & \text{if } 0 < x \leq y < N_j \\ 0^S, & \text{else} \end{cases}$$

where this moment starts from time stamp  $x$  and ends in time stamp  $y$ . When  $0 < x \leq y < N_j$ , the value of the moment feature is non-zero. Thus, the 2D map of the  $j$ -th scale is constructed, and the corresponding moment features in the map are extracted by aggregating the video segment features.

We denote the 2D temporal feature map of the  $j$ -th scale as  $F^j \in R^{N_j \times N_j \times d^v}$ .  $N_j$  is the number of sampled segments in the  $j$ -th temporal scale, and also represents the start and end indexes in the  $j$ -th 2D map. Different from the single-scale 2D temporal feature map, we collect all the 2D map constructed in multiple scales, denoted as  $F^M = \{F^j\}_{j=1}^{N_s}$ .

To select the proper candidate moments, we need to sample the possible moments based on the 2D temporal feature map. As introduced in [27], one simple way is to enumerate all the possible consecutive video clips as candidates. While the other way is to sparsely sample the moments, this could efficiently remove the redundant moment candidates, and save the computational cost simultaneously. So we choose the latter sampling strategy, and get the selected moment candidates as proposals.

The 2D Temporal Network mainly consists of the cross-modal fusion and the convolution network over the multi-scale 2D temporal feature map, and update the cross-modal feature by capturing the temporal dependencies on the 2D map.

After the construction of the multi-scale 2D temporal feature map which represent the video moment candidates, the next step is the cross-modal fusion based on the text query features. In order to align the text query with moment candidates in multi-scale 2D map, we first duplicate the

extracted text embedding feature  $f^T$   $N_s$  times, denoted as  $F^T = \{f^T, \dots, f^T\}_{N_s}$ . Then, the text features and video moment features in the map are projected into a common feature space by a fully-connected network. And eventually the cross-modal feature map are fused by the Hadamard product and  $l_2$  normalization. Mathematically, the generation of the cross-modal feature can be formulated as follows:

$$F_{cro} = \|(W^T \cdot F^T \cdot 1^T) \odot (W^M \cdot F^M)\|_F,$$

where  $W^T$  and  $W^M$  represents the parameters of the fully connected layers, which could be learnt in training.  $1^T$  is the transpose of an all-ones vector,  $\odot$  is Hadamard product, and  $\|\cdot\|_F$  denotes Frobenius normalization.

In order to capture the temporal dependencies between moment candidates in the multi-scale 2D feature map, we construct the multi-layer convolution network on the 2D cross-modal feature map  $F_{cro}$ . Through the convolution network, the context information is aggregated from the adjacent moment candidates, and the enlarged receptive field makes it easy to leverage the long-term relations in the video.

#### D. Moments Evaluation Module

We generate the moment candidate alignment scores by the Multi-scale 2D Temporal Network, and select the best-matching moments by this score. However, there is no full annotations of temporal boundaries served as the supervision of the moments evaluation when training in a weakly-supervised manner. Thus, we propose a Moments Evaluation Module and generate the pseudo labels for the moment candidate in the 2D multi-scale score map. In this module, we first select the top- $k$  moment candidates with higher alignment scores from the 2D feature map in each temporal scale, and then perform moments caption based on the selected moments. According to the similarity between the reconstructed text query generated by the moments caption model and the original text query in annotations, we could obtain the pseudo labels as supervision for training.

**Moments Caption.** Observing that the only annotation of this task is the text query, we design the caption module and make the whole model trainable. The framework of Moments Caption has been shown in Fig. 3.

The text embedding vectors  $w_i \in R^{d^T}$  are passed through the first LSTM layer, and the hidden state  $\{h_0^1, h_1^1, \dots, h_L^1\}$  are fused with the cross-modal feature of the selected top- $k$  moment candidates  $F_{cro}$ . Then the fused sequence feature are fed into the second LSTM layer [6], and get the hidden state  $\{h_0^2, h_1^2, \dots, h_L^2\}$ . Finally, we utilize the fully-connected layer and obtain the embedding vector of reconstructed query  $w_i^* \in R^{d^T}$ .

**Moments Evaluation.** As for moments evaluation, the output of the convolution network in the 2D temporal network is passed through a fully-connected network with sigmoid activation function, and obtain the multi-scale 2D score maps, denoted as  $P = \text{sigmoid}(W^F \cdot F_{cro})$ . Each grid in the score map represents the alignment score between the moment candidate and the given text query. Owing to the lack of

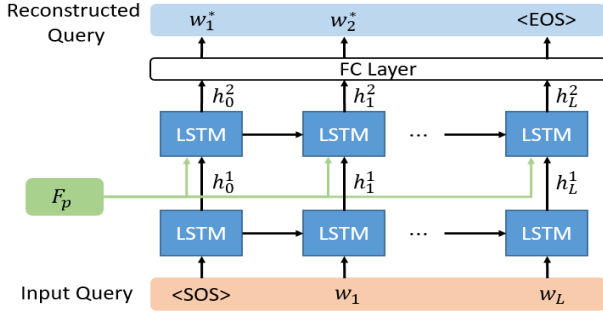


Fig. 3. The Caption Module for reconstruction of the text query. This module consists of a two-layer LSTM network and a fully-connected layer, and generate the caption of the moment candidates based on their cross-modal feature and text query feature.

temporal boundary annotations, we are not able to compute the tIoU between moment candidates and truly-matched moments. Thus, we generate the pseudo labels for further training of the evaluation module, details are illustrated in the loss functions section.

#### E. Loss functions

In this section, we mainly introduce the loss functions for training the framework. The Loss functions in the proposed model consist of a reconstruction loss and a cross-entropy loss. The former is calculated between text query and reconstructed query, and the latter is calculated after the moment candidates evaluation on the 2D multi-scale score map.

Given a video-sentence pair, the reconstruction loss maximizes the normalized log likelihood of the words in the reconstructed query, denoted by:

$$L_{rec} = \frac{-1}{N_s K L} \sum_{k=1}^{N_s} \sum_{l=1}^L \log P(w_l^* | F_{cro}^k, h_{l-1}^2, w_1, \dots, w_{l-1})$$

The reconstruction loss reflects the similarity of the input query and the reconstructed query, the learnt model tends to select moments candidates with lower reconstruction loss. For the  $k$ -th moment candidate in the  $j$ -th scale 2D temporal feature map, the reconstruction loss of the moment is denoted as  $l_k^j$ , and the pseudo label  $y_k^j$  is computed as:

$$l_k^j = \frac{\sum_{l=1}^L \log(w_l^* | F_{cro}^k, h_{l-1}^2, w_1, w_2, \dots, w_{l-1})}{\sum_{k=1}^K \sum_{l=1}^L \log(w_l^* | F_{cro}^k, h_{l-1}^2, w_1, w_2, \dots, w_{l-1})}$$

$$y_k^j = \begin{cases} 0, & l_k^j \geq l_{\max} \\ 1 - l_k^j, & l_{\min} \leq l_k^j < l_{\max} \\ 1, & l_k^j < l_{\min} \end{cases}$$

After generating the pseudo labels  $y_k^j$ , we adopt them as supervision of the candidates and design a reconstruction-

guided binary cross-entropy loss (RG-BCE loss):

$$L_{rg-bce} = \frac{1}{N_s K} \sum_{j=1}^{N_s} \sum_{k=1}^K y_k^j \log p_k^j + (1 - y_k^j) \log (1 - p_k^j)$$

In total, the final multi-task loss could be formulated as:

$$L = L_{rg-bce} + \lambda L_{rec}$$

where  $\lambda$  is the hyper-parameter for balancing the reconstruction loss and binary cross-entropy loss.

## IV. EXPERIMENTS

We conduct experiments on two benchmark datasets: Charades-STA and ActivityNet Captions, and evaluate the effectiveness of our multi-scale 2D representation learning for weakly-supervised video moment retrieval. We first introduce the datasets, evaluation metric and implementation details, and then report the experiment results and analysis. Finally, we discuss the impact of the parameter setting in the proposed model.

#### A. Datasets and Evaluation Metric

**Charades-STA.** The Charades dataset [20] is originally proposed in 2016. It contains 9848 videos of daily indoors activities. It is originally designed for action recognition and localization. Gao et al. extend the temporal annotation, labeling the start and end time of moments of the original video dataset with language descriptions and name it as Charades-STA. Charades-STA contains 12408 moment-sentence pairs in training set and 3720 pairs in testing set.

**ActivityNet Captions.** It consists of 19209 videos, whose content are diverse and open. It is originally designed for video captioning task, and recently introduced into the task of moment localization with natural language, since these two tasks are reversible. Following the experimental setting in [27], we use val-1 as validation set and val-2 as testing set, which have 37417, 17505, and 17031 moment-sentence pairs for training, validation, and testing. Currently, this is the largest dataset in this task.

**Evaluation Metric.** We use the evaluation criteria following prior works in literature [14], [13], [21], [5]. We measure rank-based performance R@K (Recall at K) which calculates the percentage of test samples for which the correct result is found in the top-K retrievals to the query sample. We follow [4] for evaluating Charades-STA and ActivityNet Captions dataset, and report results for R@1, R@5 in the condition of IoU=0.3 and IoU=0.5.

#### B. Implementation Details

We utilize a three-layer LSTM for extracting the basic text features, and the feature dimension  $d^v$  and  $d^T$  is 512. We split the whole video into small non-overlapping video clips, and use pre-extracted C3D feature [22], [7], [8] for both Charades-STA and ActivityNet Captions datasets. The number of frames in one clip in Charades-STA is 4, and that in ActivityNet Captions is set to 16. The multi-layer convolution network is 8-layer with kernel size of 5. The dimension of hidden states

in moment caption module is 1024, and the dimension of the Glove embedding [17] is 300. The thresholds  $l_{\max}$  and  $l_{\min}$  in RG-BCE loss are respectively set to 0.7 and 0.1, and we choose the top-10 moment candidates for moments caption. We use Adam [9] with learning rate of  $1 \times 10^{-4}$ , and the batch size 128 for optimization. Non maximum suppression (NMS) [15] with a threshold of 0.5 is applied during the inference.

### C. Quantitative Results and Analysis

In this section, we report the quantitative experiment results and analysis on the two datasets.

**Charades-STA Dataset.** The experiment results on the Charades-STA Dataset are shown in Table 1, and we use the evaluation metric "R@n, IoU=m", where n is {1, 5}, and m is {0.5, 0.7}.

As shown in Table 1, when comparing with other weakly-supervised approaches, our proposed method outperforms the TGA model significantly and achieves better R@1 performance compared with SCN, and the results confirm the effectiveness of context information between moment candidates in the multi-scale 2D representation learning. Through the multi-scale 2D temporal feature map, fine-grained candidates are generated and the context information between candidates is encoded by the temporal network. Although the performance of LoGAN is slightly better than ours, but it constructed a Frame-By-Word interaction and get fine-grained moments representation by co-attention with higher computational cost.

Moreover, our proposed weakly-supervised model outperform the visual-semantic embedding approaches VSA-RNN and VSA-STV by a large margin, and also perform better than some of the fully-supervised method, which indicates our weakly-supervised model could effectively improve the performance without the annotation of the temporal boundaries. Even when comparing with the state-of-the-art fully-supervised method 2D-TAN, the margin of the prediction performance is not so large. Especially, the gap between fully-supervised 2D-TAN and our multi-scale 2D representation learning is not so large, which verifies the rationality of the designed moments evaluation module and RG-BCE loss in our approach.

**ActivityNet Captions Dataset.** The results in Table 2 show the performance comparison with other methods on the ActivityNet Captions Dataset, and we use the evaluation metric "R@n, IoU=m", where n is {1, 5}, and m is {0.3, 0.5}.

Similar to results on Charades-STA, compared with the weakly-supervised methods WS-DEC, WSLN and SCN, our proposed approach has achieved better performance, and even outperform some of the fully-supervised methods. The WS-DEC method designed a iterative process of moments retrieval and caption, which leads to complicated optimization. Compared with WS-DEC, our proposed method has employed the top-k selection on the multi-scale 2D temporal feature map, and has avoided the redundant iteration.

### D. Discussion

In this section, we mainly discuss the impact of the selection of temporal scales and the impact of the loss weight, some of

TABLE I  
PERFORMANCE COMPARISON RESULTS ON CHARADES-STA DATASET.

Method	Training	IoU0.5		IoU0.7	
		R@1	R@5	R@1	R@5
Random	-	8.61	37.57	3.39	14.98
VSA-RNN	Full	10.50	48.43	4.32	20.21
VSA-STV	Full	16.91	53.89	5.81	23.58
CTRL [4]	Full	23.63	58.92	8.89	29.52
2D-TAN [27]	Full	39.70	80.32	23.31	51.26
TGA [14]	Weak	19.94	65.52	8.84	33.51
LoGAN [21]	Weak	<b>34.68</b>	<b>74.30</b>	14.54	<b>39.11</b>
SCN [13]	Weak	23.58	71.80	9.97	38.87
Ours	Weak	<b>30.38</b>	<b>69.60</b>	<b>17.31</b>	<b>34.92</b>

TABLE II  
PERFORMANCE COMPARISON RESULTS ON ACTIVITYNET CAPTIONS DATASET.

Method	Training	IoU0.3		IoU0.5	
		R@1	R@5	R@1	R@5
Random	-	18.64	52.78	7.63	29.49
VSA-RNN	Full	39.28	70.84	23.43	55.52
VSA-STV	Full	41.71	71.05	24.01	56.62
CTRL [4]	Full	47.43	75.32	29.01	59.17
2D-TAN [27]	Full	59.45	85.53	44.51	77.13
WS-DEC [3]	Weak	41.98	-	23.34	-
WSLLN [5]	Weak	42.80	-	22.70	-
SCN [13]	Weak	47.23	71.45	29.22	55.69
Ours	Weak	<b>49.79</b>	<b>72.57</b>	<b>29.68</b>	<b>58.66</b>

the experiment results are listed as follows.

**Impact of Multiple Temporal Scales.** To evaluate the impact of temporal scales of the 2D temporal feature map, we conduct a set of experiments outlined in Table III. When using our proposed multi-scale 2D temporal feature maps ( $N_s = 3$ ), the experiment results are better than that of single-scale 2D temporal feature map ( $N_s = 1$ ,  $N_j = 64$ ), which indicating the effectiveness of the multi-scale 2D feature maps. When setting  $N_j = 64, 24, 4$ , the performance is boosted and better than results of single-scale map by a large margin. The smallest scale in  $N_j$  is ranges in [4, 6, 8], and we get better experiment results when set it as 4, because the smaller value makes the multi-scale temporal maps able to cover more precise moments with longer temporal length.

TABLE III  
EXPERIMENT RESULTS WITH MULTIPLE TEMPORAL SCALES (T-SCALE).

T-Scale	Multi-scale	IoU0.3		IoU0.5	
		R@1	R@5	R@1	R@5
64	✗	44.25	63.66	27.07	51.79
64-24-8	✓	44.52	63.70	25.00	52.05
64-24-6	✓	47.99	66.41	21.09	44.30
64-24-4	✓	<b>49.79</b>	<b>72.57</b>	<b>29.68</b>	<b>58.66</b>





Fig. 4. Qualitative results on Charades-STA dataset. The red line represents the ground truth, and the blue line is the prediction of our method. And we also demonstrate the alignment score and the corresponding reconstruction loss of the predicted video moments.

TABLE IV  
EXPERIMENT RESULTS WITH DIFFERENT LOSS WEIGHT.

Loss weight	IoU0.3		IoU0.5	
	R@1	R@5	R@1	R@5
$\lambda=0.5$	47.09	74.62	21.63	54.01
$\lambda=1.0$	<b>49.79</b>	72.57	<b>29.68</b>	58.66
$\lambda=2.0$	43.85	<b>79.98</b>	24.68	<b>59.67</b>

**Impact of Loss Weight.** As it shows, the performance of our model is relatively stable when  $\lambda$  is set as 0.5 or 1.0. When  $\lambda = 2.0$ , the R@1 performance drops, while R@5 performance increases. The context information from adjacent clips would benefit the moment caption, so a few video moments with low reconstruction loss do not have the highest tIoU with ground truth, and the R@1 metric drops.

#### E. Qualitative Results

We present some qualitative results on the Charades-STA dataset to illustrate the effectiveness of our method, several examples are shown in Fig. 4. The red line is the ground truth, and the blue line represents the moment prediction. In Fig. 4, case A, E and F are successful cases, and the predictions in cases B, C and D are relatively misaligned compared with the ground truth.

According to the Fig. 4, the successful moments prediction have the higher alignment scores as well as the lower reconstruction losses, which indicates that our moment evaluation module with RG-BCE loss is capable to evaluate the quality of input candidates. The successful samples include moments with long time of duration (e.g. sample A) and those with short time of duration (e.g. sample F), which shows the capability of our proposed method in weakening the negative affects resulting from variation of temporal scales.

Due to the ambiguity of the action moments and existing noise in the scene, our weakly-supervised method has achieved limited success when dealing with these cases. Take sample B as an example, in the untrimmed video, a person opens the refrigerator and then closes it. It is hard to deal with ambiguity of "open" and "close" and localize the best-matching moment without temporal boundary annotations. In sample C, the noise of objects visible in the scene affects the moment evaluation, which leads to the misalignment compared with ground truth.

#### V. CONCLUSION

In this work, we focus on the task of video moment retrieval without manually labelling the start and end time points of moments in training. We address the motivation of considering the various temporal scale of moment candidates as well as the temporal relations between them in weakly-supervised setting, and propose a multi-scale 2D representation learning method, including the multi-scale 2D temporal network and weakly-supervised moments evaluation module with RG-BCE loss. The multi-scale 2D temporal map could generate more precise moment candidates with various temporal scales, and moment-to-text reconstruction facilitate the weakly-supervised training in moments evaluation. The experiment results on the Charades-STA and ActivityNet Captions datasets demonstrated the effectiveness and superiority of our proposed approach.

#### REFERENCES

- [1] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 5803–5812.

- [2] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 8199–8206.
- [3] X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, and J. Huang, "Weakly supervised dense event captioning in videos," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 3059–3069.
- [4] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5267–5275.
- [5] M. Gao, L. Davis, R. Socher, and C. Xiong, "Wslin: Weakly supervised natural language localization networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 1481–1487.
- [6] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.
- [7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675–678.
- [8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725–1732.
- [9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [10] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," *arXiv preprint arXiv:1911.09963*, 2019.
- [11] C. Lin, J. Li, Y. Wang, Y. Tai, D. Luo, Z. Cui, C. Wang, J. Li, F. Huang, and R. Ji, "Fast learning of temporal action proposal via dense boundary generator," *arXiv preprint arXiv:1911.04127*, 2019.
- [12] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.
- [13] Z. Lin, Z. Zhao, Z. Zhang, Q. Wang, and H. Liu, "Weakly-supervised video moment retrieval via semantic completion network," *arXiv preprint arXiv:1911.08199*, 2019.
- [14] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury, "Weakly supervised video moment retrieval from text queries," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 592–11 601.
- [15] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition (ICPR)*, vol. 3. IEEE, 2006, pp. 850–855.
- [16] C. R. Opazo, E. Marresataylor, F. S. Saleh, H. Li, and S. Gould, "Proposal-free temporal moment localization of a natural-language query in video using guided attention," *arXiv preprint arXiv:1908.07236*, 2019.
- [17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [18] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.
- [19] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 154–171.
- [20] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 510–526.
- [21] R. Tan, H. Xu, K. Saenko, and B. A. Plummer, "Logan: Latent graph co-attention network for weakly-supervised video moment retrieval," *arXiv preprint arXiv:1909.13784v2*, 2020.
- [22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015, pp. 4489–4497.
- [23] J. Wang, L. Ma, and W. Jiang, "Temporally grounding language queries in videos by contextual boundary-aware prediction," *arXiv: Computer Vision and Pattern Recognition*, 2019.
- [24] W. Wang, Y. Huang, and L. Wang, "Language-driven temporal activity localization: A semantic matching reinforcement learning model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 334–343.
- [25] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 9062–9069.
- [26] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional networks for temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7094–7103.
- [27] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," *arXiv: Computer Vision and Pattern Recognition*, 2019.