Neural Computing and Applications
https://doi.org/10.1007/s00521-021-05807-z

ORIGINAL ARTICLE

# Context-aware network with foreground recalibration for grounding natural language in video

Cheng Chen[1] · Xiaodong Gu[1]

Received: 10 August 2020 / Accepted: 5 February 2021

## Abstract

Grounding natural language in video aims at retrieving a matching moment in a long, untrimmed video described by a referring natural language query. It is a challenging issue due to the dominating influence from noise background in untrimmed video and the complex temporal relationships introduced by the query. Existing methods treat different candidate segments separately in a matching and aligning manner and thus neglect that different target segments require different levels of context information. In this paper, we present the semantic modulation residual module, a novel single-shot feed-forward residual network that explicitly integrates various temporal scale features and introduces less noise to the final moments representation with the guide of query semantic information. To establish more fine-grained interactions between different moments, a global interaction module is embedded in the network. Moreover, the data imbalance issue caused by the sparse annotated moments weakens the effect of binary cross-entropy criterion. Therefore, we design a foreground recalibration mechanism to enhance the intra-class consistency and highlight the positive moments. We evaluate our method on three benchmark datasets i.e., TACoS, Charades-STA and ActivityNet Captions, achieving state-of-the-art performance without any post-processing. In particular, we reach 32.17%, 45.11% and 43.76% under the metric Rank@1, IoU@0.5 on TACoS, Charades-STA and ActivityNet Captions, respectively. Furthermore, ablation studies were performed to show the effectiveness of individual components in our proposed method. We hope that the proposed method can serve as a strong and simple alternative for fine-grained video retrieval.

**Keywords** Text-based moment retrieval · Grounding natural language in video · Semantic modulation based on residual module · Foreground recalibration

## 1 Introduction

Localizing activities in video is a fundamental issue of video understanding in computer vision. Several researches such as temporal proposal generation [1], video summarization [2, 3], video question answering [4], video caption [5, 6] and grounding natural language in video [7] have been proposed for different scenarios. Grounding natural language in video (GNLV) aims at retrieving the start and end points of a specific moment which best matches the language query in a long, untrimmed video. Since classic activities retrieval tasks mainly work on a predefined list of action classes , this formulation establishes a connection channel between language description and visual data for better video understanding.

Most of the existing methods [7–10] mainly adopt a two-step pipeline, moment candidates are first sampled by scanning video sequence with temporal sliding windows, and then, the sampled candidates are ranked by calculating their similarities with the query sentence in a multi-modal common space. As the target segments are of diverse durations and locations, these methods require exhaustively matching the query with large number of overlapping segments, thus resulting in heavy computation. Moreover, since these models independently match different segments with the sentence, they inevitably fail to extract the

✉ Xiaodong Gu
xdgu@fudan.edu.cn

Cheng Chen
chengchen19@fudan.edu.cn

[1] Department of Electronic Engineering, Fudan University, Shanghai 200433, China

Published online: 26 February 2021

inherent temporal dependencies between different moments.

To tackle these limitations in the sample-and-rank framework, some works start to borrow ideas from one-stage object detection [11–13]. Zhang et al. [11] leverage the temporal convolutional network to obtain multi-scale features at different layers, which naturally cover various durations activities. Yuan et al. [12] propose to adjust the feature normalization parameters with reference to the query during the temporal convolution processes. Chen et al. [13] simultaneously assign a set of temporal anchors with multiple scales for each frame and finally localize the moment which corresponds to the sentence.

Although these works have achieved promising performances, it is still challenging to well model the complex temporal relationships introduced by the language sentence and precisely localize the target segments in untrimmed video with noisy background.

On the one hand, the language queries are flexible. As shown in Fig. 1, to retrieve the moment corresponding to the query, "a bunch of people on tubes go by slowly," the model should pay more attention to the former temporal moments in the video. Consider another example shown in Fig. 1 "then we see rapid waters as the tubers go over small falls", a simple word "then" results in complex temporal dependencies. As the visual appearance of frames adjacent to the ground truth is very similar, the model needs to aggregate a long range of context information to capture the temporal relationships between different video moments for precisely localizing the second moment. Therefore, it is critical for the model to adaptively integrate features from different temporal scales to form the final discriminative moment representations.

On the other hand, a real-world video usually contains a large number of frames, but only two frames are annotated. Thus, only a few segments which have large overlap with target moment can be used as positive training samples. As the cross-entropy criterion equally treats each training instance at training time, it is difficult to achieve consistent emphasis of positive moments, i.e., foreground region.

In this paper, a novel context-aware network (CAN) with foreground recalibration is proposed to mitigate the above two issues. CAN consists of a basic multimodal encoder, a semantic modulation residual module, a global interaction module and a foreground recalibration module. Each video clip first meets and interacts with the query sentence in multimodal encoder. In semantic modulation residual module (SMRM), the local clip features are progressively propagated to the top layers through residual connections. The semantic information from the query sentence is explicitly leverage to adaptively modulate the residual connections. By learning knowledge guidance from the language query, the SMRM can integrate the sentence-related contextual information from different temporal convolutional layers more effectively. To further learn the difference between different moment candidates, a self-attention-based global interaction module is designed. The global interaction module (GIM) is performed on the top of SMRM, which already deeply absorbs both semantic and visual information. Thus, the GIM is able to collect more reliable contextual clues for modeling complex temporal dependencies.

To alleviate the dominating influence from noisy background in untrimmed video, we propose to redefine and relabel the foreground region (i.e., positive moments region) in video and embed a foreground recalibration loss at training stage. The foreground recalibration loss is immune to the data-imbalance issue and able to uniformly emphasize the positive moments. Since foreground recalibration loss explicitly exploits the relationship between positive moments region and predicted moments region, it also enables our model to enlarge inter-class differences while preserving intra-class consistency.

Our main contributions are summarized as follows:

(1) A semantic modulation residual module (SMRM) is proposed to efficiently utilize the features from different temporal convolutional layers with the guide of the query semantic. SMRM modulates the temporal convolution operations to better highlight the sentence-related video contents over time.

(2) A global interaction module (GIM) is presented to make the network adaptively extract reliable information for accurately distinguishing different

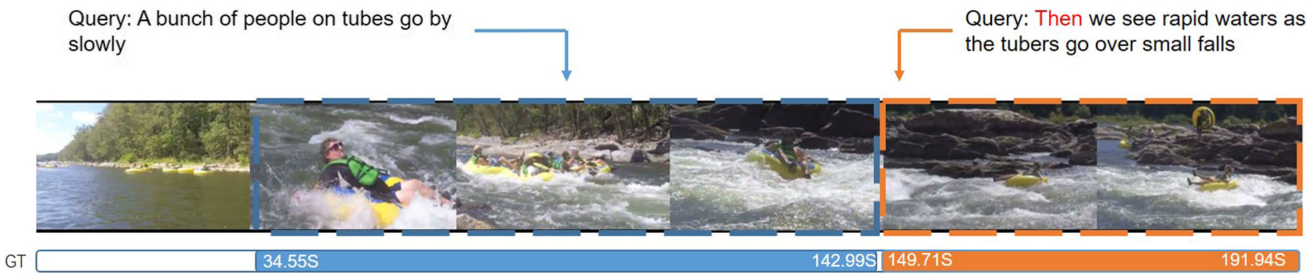
Query: A bunch of people on tubes go by slowly
Query: Then we see rapid waters as the tubers go over small falls
GT
34.55S
142.99S
149.71S
191.94S

**Fig. 1** Example of grounding natural language in video. The target moments are highlighted in blue and orange

moments. GIM calculates the position-aware attention to measure the contributions from different contextual moments

(3) A foreground recalibration mechanism is designed to utilize the relationship between positive moments region and predicted moments region. It is able to better handle data imbalance issue and help the model to focus more on positive moments region without any post-processing. The experimental results verify every component is effective for localizing precise temporal boundaries.

## 2 Related Work

Temporal localization in untrimmed videos includes two major sub-fields: temporal action localization and grounding natural language in video. Here we review some works that are close to ours.

### 2.1 Temporal action localization

Temporal action localization is to predict the temporal boundaries and the label of the activity instances in videos. Earlier works mainly adopt hand-crafted features to train activity classifier [15–17]. Recently there has also been deep learning-based works in this task. For example, Shou et al. [18] develop C3D [19] with a localization loss and achieve state-of-the-art performance. On the other hand, Xu et al. [20] propose an end-to-end network based on Faster-RCNN [21] to generate activity prediction scores. Singh et al. [22] propose a bidirectional RNN-based multi-stream framework to jointly learning features from motion and appearance.

To better utilize the context information in long videos, Zhao et al. [23] propose a structured pyramid pooling strategy to model the visual content structure for producing discriminative representation of the proposal. And Zhang et al. [24] introduce an algorithm to simultaneously predict the temporal boundaries and confidence scores of activity categories in one single shot by imitating the single-shot multi-box detector framework [25].

Although these works have achieved significant performance, they are still limited to work on pre-defined lists of simple actions and fail to handle diverse activities in the real world. Therefore, grounding natural language in video is introduced to tackle this issue.

### 2.2 Grounding natural language in video

Grounding natural language in video is to determine the start and end timestamps of the target video moment which semantically corresponds to the given language query. Early studies about this task are limited to constrained settings [26–28]. Tellex et al. [26] retrieve video moments from surveillance camera by language queries with fixed spatial prepositions. Lin et al. [28] propose to retrieve temporal moments in 21 videos from autonomous driving scenarios.

Recently, larger datasets [7, 8, 14] with videos from more general scenarios and natural language queries are constructed for flexible groundings. In this context, Gao et al. [7] propose to first sample candidate moments by temporal sliding windows and then learn a common embedding space shared by moment features and sentence representations to rank these candidate moments. Hendricks et al. [8] propose to concatenate each local video feature with the global video feature for boosting the moment representations. Ge et al. [9] develop activity classifiers to mine the activity concept and introduce an actionness score to exploit the semantic information from verb object pairs in the language queries. To learn the fine-grained matching relation between video and language, some other methods leverage attention mechanism to deeply integrate sentence information with each video moment. In this context, Chen et al. [13] develop a multi-modal interactor to aggregate frame-by-word interactions between video and query sentence. Yuan et al. [29] propose a co-attention module for better correlating the query sentence with video content. Zhang et al. [11] utilize the graph convolutional network to model the temporal relations among candidate moments. Liu et al. [30] introduce a memory attention mechanism to emphasize the query-related video contents and simultaneously aggregate the context information to learn better representations of candidate moments. However, as described above, these models process each video from a local perspective, while neglect that different language query describes different levels of context information. Besides the crucial context information, the imbalanced distribution of the moments remains to be settled.

Unlike these existing methods, a novel context-aware network with foreground recalibration is proposed to consider three critical factors for grounding natural language in video, including the multiple temporal scale features from different layers, guiding role of the query semantic and the sufficient emphasis on positive moments region.

## 3 Methodology

In the following, the basic formulation of grounding natural language in video is first introduced. Then, our main framework for grounding natural language in video is proposed, as shown in Fig. 2. Our model CAN consists of

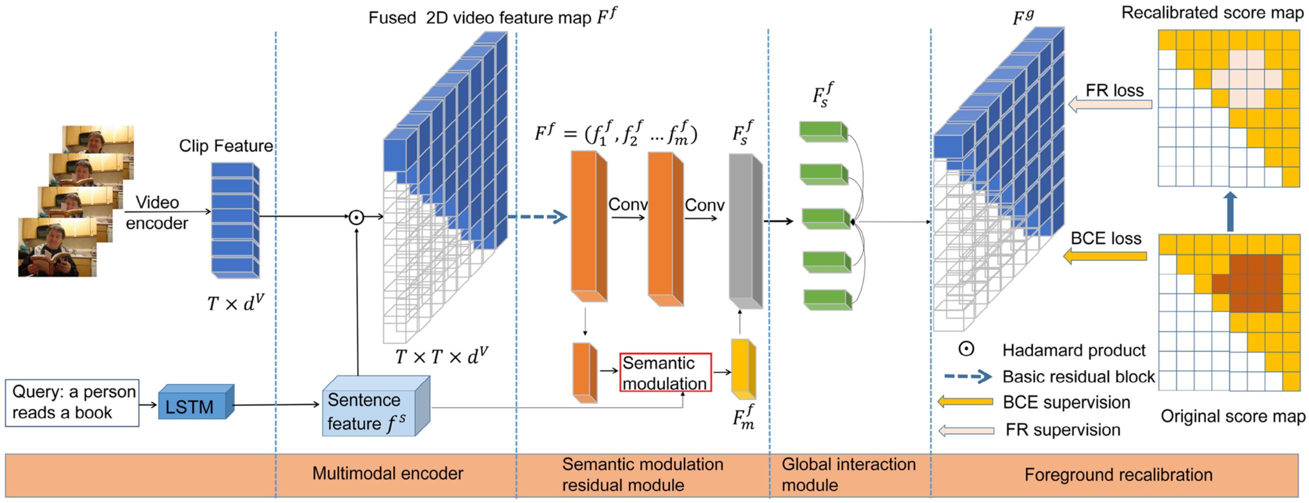

**Fig. 2** The framework of our proposed method. It consists of a multimodal encoder for multimodal representation, a semantic modulation residual module for learning temporal dependencies, a global interaction module for discriminative clues and a foreground recalibration for emphasizing the positive moments

four main components: the multimodal encoder, the semantic modulation residual module, the global interaction module and the foreground recalibration. Please note that all components of our model are deeply coupled and thus can be trained in an end-to-end manner.

## 3.1 Problem formulation

Given a query sentence $S$ and an untrimmed video $V$, our goal is to retrieve the target moment within the video which semantically corresponds to the query. More specifically, the video is a frame sequence, i.e., $V = \{f_i\}_{i=1}^{v}$, where $f_i$ is the $i$th image frame and $v$ is the total number of frames. The retrieved target moment is defined as a temporally continued segment within $V$ which is denoted as $S = (s, e)$, where $s$ and $e$ are the corresponding start and end timestamps, respectively, and the query sentence $Q$ is a word sequence $Q = \{w_i\}_{i=1}^{q}$, where $w_i$ denotes the $i^{th}$ word, and $q$ is the query length.

## 3.2 Multimodal encoder

As shown in Fig. 2, our model consists of four main modules. (1) A basic multimodal encoder that fuses the sentence with video moments features and restructures the fused features to a 2D feature map, and each location of the feature map is specified as a multimodal feature of the corresponding video moment. (2) A semantic modulation residual module that consists of multiple basic residual blocks and semantic modulation residual blocks. In the semantic modulation block, the query sentence attends each location of the multimodal feature map and updates the query-related features of residual connection, so as to better align the sentence and video semantics under various temporal granularities. (3) A global interaction module, where self-attention is adopted to model the moment relations on top of the semantic modulation module. (4) A foreground recalibration module that redefines and relabels the positive moments region and attentively emphasizing the target moments under the supervision of foreground recalibration loss and binary cross-entropy loss. A video is first decomposed into a sequence of clips $V = \{v_i\}_{i=0}^{T}$, where $T$ is the number of clips. For each video clip, its feature $f^V \in R^{d^v}$ is extracted by a pre-trained CNN model, where $d^v$ is the feature dimension. To obtain the features of moment candidates, we max-pool the corresponding clips over a specific time span. Specifically, the moment candidate feature $f_{a,b}^M$ (a and b, respectively, represent the start and end clip) is obtained by $maxpool(f_a^V, f_{a+1}^V, \ldots, f_b^V)$.

For the query sentence, the pre-trained embedding vector for each word is first generated by the GloVe word2vec model [31], and then a three-layer LSTM [32] is employed to encode the language structure of the sentence. The last hidden state of the LSTM is used as sentence feature, i.e., $f^s$ here.

To utilize hierarchical 2D convolutional network for capturing complex temporal dependencies, the same strategy as 2D-TAN [33] is adopted to restructure the whole moment candidates sampled from video clips to a 2D-temporal feature map, denoted as $\mathrm{F}^M \in \mathbb{R}^{T\times T\times d^V}$. The first two dimensions of the 2D temporal feature map represent the indexes of the start and end clips of the moments, respectively, and $d^V$ is the feature dimension. Different locations in the 2D-temporal feature map represent moments of different start and end timestamps. For example, $\mathrm{F}^M[a, b, :]$ indicates the moment starting at A clip and ending at B clip. Noted that, only the upper triangular part of the feature map is valid. Because the start timestamp of the moment should be earlier than the end timestamp. And the lower triangular part of the map is padded with zeros which would not be taken into calculation. In order to correlate the corresponding semantic information between each moment and sentence, we first let 2D-temporal feature map interact with the sentence, which is formulated as:

$$\mathrm{F}^f = \left\| (w^s \cdot f^s) \odot (W^M \cdot \mathrm{F}^M) \right\|_F, \tag{1}$$

where $w^s$ and $W^M$ are learnable parameters, $\|\cdot\|_F$ is Frobenius normalization, and $\odot$ is Hadamard product. With such a cross-modal fusion processing, the fused feature map $\mathrm{F}^f \in \mathbb{R}^{T \times T \times d^f}$ ($d^f$ is the dimension of the fused feature) is able to represent fine-grained interactions between query sentence and video moments. The following context-aware network will gradually compose and correlate such representations, resulting in learning accurate context relationships.

### 3.3 Semantic modulation residual module

As aforementioned, the target moment may have complex temporal relationships with other temporal moments. Therefore, the model needs to extract multiple temporal scale features from the fused representation $\mathrm{F}^f$ and adaptively integrate these features. Recently, several studies [34, 35] about residual learning have revealed that armed with the residual connections the small-scale features from bottom convolutional layers can be transferred to top layers. In the proposed semantic modulation residual module, features encoding multiple temporal scale context relationships are progressively extracted and propagated in consecutive layers, and discriminative features from different range of scales are combined by the residual connections. Moreover, for explicitly formulating the importance of features with different temporal scales, a semantic modulation mechanism is designed to modulate the residual connection by weighting different regions of the video feature maps with referring to the query semantics.

As illustrated in Fig. 2, multiple basic residual blocks and semantic modulation residual blocks are stacked to establish the hierarchical semantic modulation residual module for perceiving more context information. Taking the fused feature map $\mathrm{F}^f$ as input, the basic residual block in this paper is defined as:

$$\mathrm{y} = \mathcal{F}(\mathrm{x}, \{W_i\}) + \mathrm{x}, \tag{2}$$

here x and y denote the input feature map and the output feature map of the block considered. $\mathcal{F}(\mathrm{x}, \{W_i\})$ is performed by two standard temporal convolution operation, $\mathcal{F} + \mathrm{x}$ is the residual connection and element-wise addition.

As illustrated in Fig. 3, semantic modulation module makes the original multi-layer video moment feature maps adaptive to the sentence context through modulating the residual connection. Given the feature map $\mathrm{F}^f$ extracted from a specific convolutional layer, the modulation weights $\beta$ are a function of the sentence context $f^s$ and the current video moment feature $\mathrm{F}^f$. Thus, semantic modulation residual block modulates the residual connection using the modulation weights $\beta$ as:

$$\beta = \phi(f^s, \mathrm{F}^f), \tag{3}$$

$$\mathrm{F}^f_m = f(\mathrm{F}^f, \beta), \tag{4}$$

$$\mathrm{F}^f_s = \mathcal{F}(\mathrm{F}^f, \{W_i\}) + \mathrm{F}^f_m, \tag{5}$$

where $\mathrm{F}^f_m$ is the modulated residual feature, $\mathrm{F}^f_s$ is the feature map yielded by the semantic modulation residual block, $\phi(\cdot)$ is the modulation function that will be detailed in the following, and function $f(\cdot)$ is element-wise multiplication that modulates the video features by modulation weights.

To calculate the modulation weight $\beta$, we first flat the 2D-temporal feature map alone the temporal dimension as $\mathrm{F}^f = (f^f_1, f^f_2 .. f^f_m)$, where $f^f_i$ is the $i^{th}$ moment candidates feature and $m$ is the number of valid candidate moments in the 2D-temporal feature map. $f^f_i$ can be considered as the visual feature of the $i^{th}$ moment in the video.

A direct realization of the semantic modulation function is to calculate the modulation weight $\beta \in d^m$ for each video moment with reference to query sentence. Specifically, given the flatted video feature map $\mathrm{F}^f = (f^f_1, f^f_2 .. f^f_m)$ and the sentence context $f^s$, three fully connected layers are used to generate the weight distributions $\beta$ over the video moments as:

$$\beta = \mathrm{softmax}(W^b \tanh(W_f \mathrm{F}^f + W_s f^s + \mathrm{b})), \tag{6}$$

where $W^b$, $W_f$, $W_s$ and b are learnable parameters. Since the last residual block in the SMRM is modulated, the feature $\mathrm{F}^s$ yielded by the whole semantic modulation residual module is equal to the feature $\mathrm{F}^f_s$ from the last semantic modulation block.

With the semantic modulation mechanism, the temporal feature map will further absorb the query semantic information, and the discriminative features from consecutive layers can be better emphasized. Deeper in the module, moment candidates at each location of the feature map are deeper blended with other locations. In addition, the semantic modulation is able to explicitly integrate features

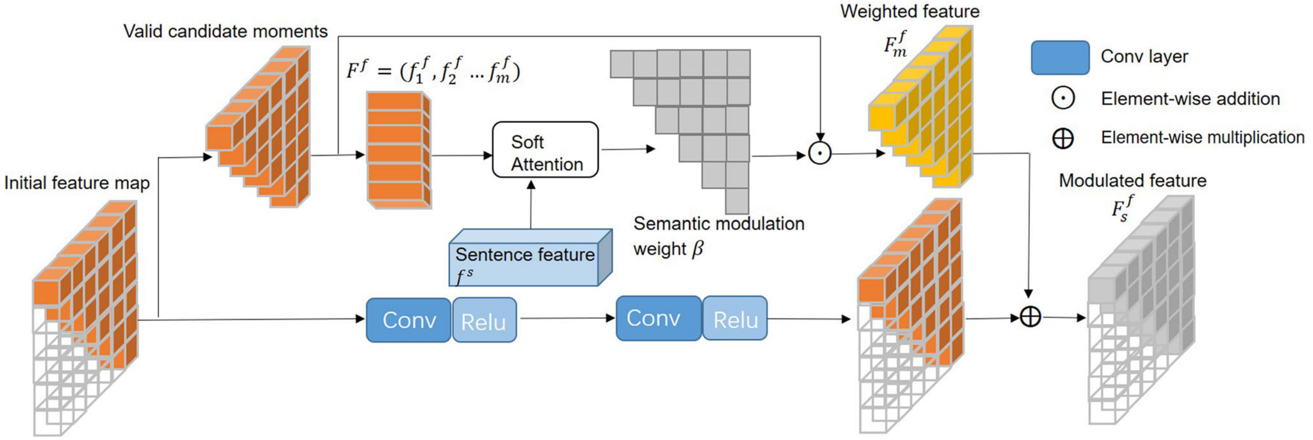

**Fig. 3** Illustration of the semantic modulation residual block. By modulating the residual connection, it adaptively weights multi-scale features with reference to query sentence

from different temporal scales to form the final representations based on the query sentence, which provides rich semantic indications and crucial information for composing the temporal video feature map.

### 3.4 Global interaction module

To better capture the global information for modeling the difference between moment candidates, global interaction is explored by leveraging the self-attention technique [36] on the top of semantic modulation residual module. As our global interaction module is implemented on the layer which already absorbs rich semantic information and discriminative context features, it can further strengthen the representation ability of different locations in the temporal feature map and collect useful grounding clues. Formally, the input to the global interaction module is the video features $\mathrm{F}^s \in \mathbb{R}^{T \times T \times d^f}$ from the semantic modulation residual module. $\mathrm{F}^s$ is first transformed into a temporal sequence $\mathrm{F}^s = (f_1^s, f_2^s \dots f_m^s) \in \mathbb{R}^{d^f \times m}$ by flatting its temporal dimension. As the same of semantic modulation residual module, $m$ is only the number of valid candidate moments. Given the temporal moment sequence, every pair moments are matched to calculate the attention as:

$$\alpha_{j,i} = \frac{exp(s_{ij})}{\sum_{i=1}^{m} exp(s_{ij})}, \tag{7}$$

where $s_{ij} \in (W_q \mathrm{F}^f)^T (W_k \mathrm{F}^f)$, and $\alpha_{ji}$ represents the extent to which the module attends to the $i^{th}$ moment when estimating the correlation of the $jth$ moment and language query. Global information is integrated by the learned attention weight to obtain the output of self-attention layer $\mathrm{F}^g = (f_1^g, f_2^g f_j^g \dots f_m^g) \in \mathbb{R}^{d^v \times m}$, here:

$$f_j^g = \sum_{i=1}^{m} \alpha_{j,i} (W_v f_i^f). \tag{8}$$

In the above formulation, $W_q \in \mathbb{R}^{C \times d^v}$, $W_k \in \mathbb{R}^{C \times d^v}$, and $W_v \in \mathbb{R}^{d^v \times d^v}$ are learnable projection weights. To save memory, $C = \frac{d^v}{8}$ is chosen in all our experiments. Given the weighted moment feature sequence $\mathrm{F}^g$, we restructure it to the 2D feature map $\mathrm{F}^G \in \mathbb{R}^{T \times T \times d^v}$ for further processing.

As shown in Fig. 4, the global interaction module consists of a self-attention layer and two fully connected layers. Two fully connected layers with dropout and rectified linear unit activation are used to further transform the learned feature. Furthermore, two residual connections are also applied to faithfully preserve the temporal dependencies captured by previous semantic modulation residual module.

### 3.5 Foreground recalibration

Given the output feature $\mathrm{F}^g$ of global interaction module, the final prediction is generated as follows:

$$\mathcal{P} = \mathrm{Sigmoid}(\mathbf{W}^p(\mathrm{F}^g)), \tag{9}$$

where $\mathcal{P} \in \mathbb{R}^{T \times T \times 1}$ is the 2D temporal score map. Each $0 < p \in \mathcal{P} < 1$ denotes the matching score between a candidate video moment with the query. $sigmoid(\mathbf{W}^p())$ is implemented by a fully connected layer with Sigmoid function. To supervise the training process, a widely used method is to treat moment candidates within the temporal score map as positive samples if their IoUs (temporal Intersection-over-Union) with ground truth target moment are larger than 0.5, and the overlap prediction loss is realized as a binary cross-entropy (bce) criterion:

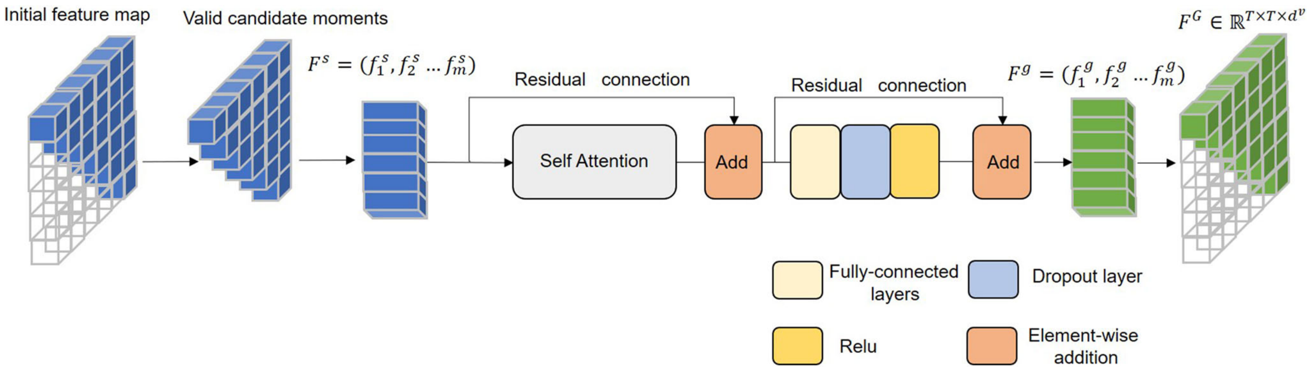

**Fig. 4** Illustration of the global interaction module. It integrates global information by adopting self-attention mechanism, fully connected layer and residual connection

$$L_{bce} = \frac{1}{N} \sum_{i=1}^{N} g_i \log p_i + (1 - g_i) \log(1 - p_i), \tag{10}$$

where $p_i$ and $g_i$ are predicted overlap score of a candidate and ground truth overlap IoU between the candidate and target moment, respectively. N is the total number of valid candidates within the score map.

As shown in Eq. 10, the loss function equally treats each moment candidate and simply accumulates the per moment loss in the whole feature map. Since negative samples often dominate in untrimmed video, the above binary cross-entropy loss cannot effectively prompt the model to emphasize the positive moments region (foreground region) and fails to tackle the data imbalance issue.

To address the data imbalance issue and explicitly highlight the foreground region in the temporal feature map, we propose to reset the positive moments region and define the foreground recalibration loss as:

$$L_{fr} = \frac{\sum((y_i = 1 | g_i \geq \theta) - p_i(y_i = 1 | g_i \geq \theta)) + \sum(p_i - p_i(y_i = 1 | g_i \geq \theta))}{\sum(y_i = 1 | g_i \geq \theta) + \sum p_i}, \tag{11}$$

where $p_i$ is the predicted score of the $i$th moment candidate. $g_i$ is the $i$th candidate's IoU with the ground truth moment. $\theta$ indicates the IoU threshold. As shown in Fig. 5, the labels of the moments with $g_i$ larger than $\theta$ are reset to 1 and regarded as the redefined positive moments to be learned.

As can be seen, instead of using the $g_i$ as supervision label, the loss pushes our model to attend the foreground region where $g_i$ larger than $\theta$ by directly setting the region's labels as 1. In $L_{fr}$, the first term of the molecule represents the difference set between the ground-truth foreground region and intersection of the predicted foreground region and the ground truth, while the second term attends the difference set between the predicted foreground region and the intersection of the predicted foreground region and the ground truth.

When the foreground region predicted by CAN is completely disjoint from the ground-truth foreground set, $L_{fr}$ will give a large penalty to the model. Once the predicted foreground region has no difference with the ground truth, $l_{fr}$ reaches its minimum, i.e., $L_{fr}$=0. Since $L_{fr}$ takes into account the topological relationships among regions, it is able to impose a global constraint on the prediction $\mathcal{P}$, resulting in alleviating the influence from dominating negative examples in training.

As $p_i$ is continuous, the gradient of $L_{fr}$ with reference to $p_i$ can be calculated:

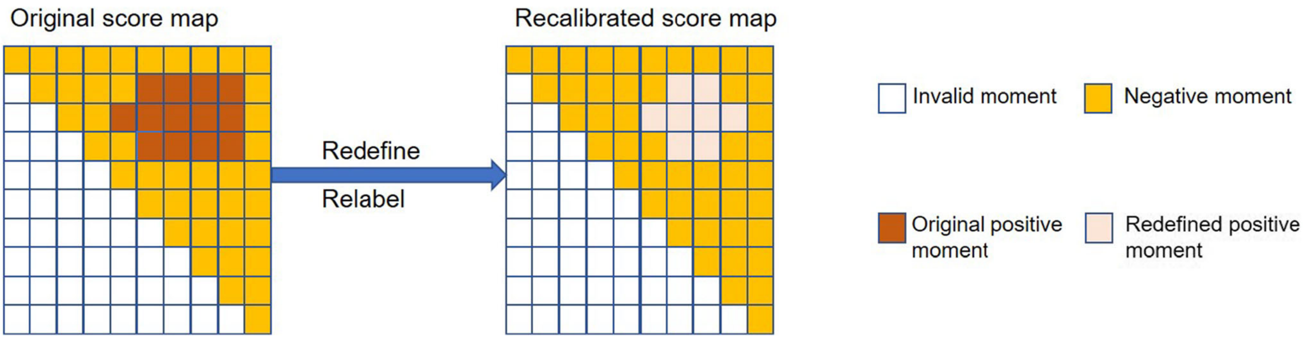

**Fig. 5** Illustration of the semantic modulation residual block. It redefines and relabels the positive moments

$$\frac{\nabla L_{fr}}{\nabla p_i} = \frac{1 - 2(y_i = 1|g_i \geq \theta)}{\sum((y_i = 1|g_i \geq \theta) + \sum p_i)} - \frac{\sum(p_i + (y_i = 1|g_i \geq \theta) - 2p_i(y_i = 1|g_i \geq \theta)}{[\sum(p_i + (y_i = 1|g_i \geq \theta)]^2}. \tag{12}$$

A close look at Eq. 12 reveals that expect $1 - 2(y_i = 1|g_i \geq \theta)$ all terms in the equation are related to the moments in both the prediction $\mathcal{P}$ and the redefined foreground region. As a result, the $L_{fr}$ can produce effective gradient for model learning to reduce the gap between $\mathcal{P}$ and positive moments regions $\mathcal{Y}$, i.e., $\sum_{(} y_i = 1|g_i \geq \theta)$, resulting in highlighting inter-class differences and keeping intra-class consistency.

$L_{fr}$ is considered to enforce the model to consistently highlight the positive moments. On the other hand, to emphasize the difference between different moments, two thresholds $o_{min}$ and $o_{\max}$ are adopted to scale the $g_i$ in Eq. 10 as:

$$g_i^s = \begin{cases} 0 & g_i \leq o_{\min}, \\ \dfrac{g_i - o_{\min}}{o_{\max} - o_{\min}} & o_{\min} < g_i < o_{\max}, \\ 1 & g_i \geq o_{\max}. \end{cases} \tag{13}$$

The scaled overlap prediction loss $L_{bce}^s$ is obtained by replace $g_i$ as $g_i^s$ in Eq. 10. Finally, the total loss function is written as:

$$L = L_{bce}^s + \lambda L_{fr}, \tag{14}$$

where $\lambda$ is used to balance the contributions of the two losses, which is set to 0.2 through cross-validation.

## 4 Experiment

In this section, the effectiveness of our proposed method is evaluated on three public large-scale datasets: TACoS [8], Charades-STA [7], and ActivityNet Captions [14]. We first describe these datasets and our implementation details, and then compare the performances of CAN with the state-of-the-art approaches. Finally, a set of ablation studies are performed to examine the contributions of different components.

### 4.1 Datasets

*TACoS.* It consists of 127 videos selected from the MPII Cooking Composite Activities video corpus [37], which contains different cook activities taken place in the same kitchen scene. The videos in TACoS have long durations, i.e., 7 min. In our experiments, we use the standard split same as [8], i.e., 10,146, 4589 and 4083 moment-query pairs for training, validation and testing, respectively.

*Charades-STA.* The dataset is constructed by [7] on the top of the Charades dataset which is originally designed for action recognition and localization. The same split as [7] is adopted, consisting of 12408 moment-query pairs in training set and 3720 pairs in testing set.

*ActivityNet Captions.* ActivityNet Captions consist of 19209 videos, which are 2 minutes on average. We use the public split in our experiments, which has 37417, 17505 and 17031 moment-query pairs for training, validation and testing, respectively.

The TACoS dataset mainly contains videos depicting human's cooking activities, while Charades-STA and ActivityNet Captions focus on more complicated human activities (i.e., hundreds of activity types in daily life). The difference of Charades-STA and ActivityNet Captions is that Charades-STA mainly contains indoor human activities, while ActivityNet Captions are an open dataset containing both indoor and outdoor activities. We summarize the statistics of these datasets in Table 1, where 'samples' denotes moment query pairs, video duration denotes average length of videos, $N_{\text{vocab}}$ is vocabulary size of lowercase words, $\bar{L}_{\text{query}}$ is average number of words in language query, $\bar{L}_{\text{moment}}$ denotes average length of the target moment in seconds, and $\triangle_{moment}$ is the standard deviation of target moment length in seconds.

### 4.2 Experimental settings

In this section, the evaluation metrics and implementation details of each dataset will be detailed.

#### 4.2.1 Evaluation metrics

For fair comparisons, Rank@n, IoU@m are adopted as evaluation metrics as in previous works [8, 9, 29, 30]. A grounding moment is considered as correct when its IoU with the ground truth moment is larger than a threshold m. More specifically, for each query, we first calculate the temporal Intersection over Union (IoU) between the predicted moments and ground truth. Then for each IoU larger than m, we compute the percentage of top-n results. The metric is on the query level, so the overall performance is the average among all the queries, denoted by:

$$R(n, m) = \frac{1}{N_q} \sum_{i=1}^{N_q} r(n, m, q_i), \tag{15}$$

where $N_q$ represents the total number of queries, and $r(n, m, q_i)$ denotes whether one of the top-n predicted moments of the query $q_i$ has IoU > m.

**Table 1** Statistics of the datasets

| Datasets | TACoS | Charades-STA | ActivityNet Captions |
|---|---|---|---|
| Videos | 127 | 9848 | 19,209 |
| Train/Test/Val samples | 10,146/4083/4589 | 12,408/3729/0 | 37,417/17,031/17,505 |
| Video duration | 7 mins | 30 s | 2 mins |
| Video content type | Cooking Activities | Indoor Activities | Open |
| $N_{\text{vocab}}/\bar{L}_{\text{query}}$ | 2033/10.05 | 1303/7.22 | 12,406/14.78 |
| $\bar{L}_{\text{moment}}/\triangle_{\text{moment}}$ | 5.45 s/7.56 s | 8.22 s/3.59 s | 36.18 s/40.18 s |

### 4.2.2 Implementation details

Our model is implemented on a server equipped with one high-performance NVIDIA 1080Ti GPU, which has 12GB memory. Python with the PyTorch deep learning library is adopted to implement our method. For fair comparisons, we use the same visual encoders as previous methods [7, 33]. Specifically, VGG [38] features for Charades-STA, C3D [19] features for ActivityNet Captions and TACoS. The number of frames within a clip is set to 16 for ActivityNet Captions and TACoS and 4 for Charades-STA. The overlapping between two consecutive clips is set to 0.8 on TACoS, while it is set to 0 on Charades-STA and ActivityNet, i.e., no overlapping. According to the video duration statistics, the time dimensions $T$ of the 2D video feature map $\mathrm{F}^f \in \mathbb{R}^{T\times T\times d^f}$ (i.e., the number of clips) are set as 16 for Charades-STA, 128 for TACoS and 64 for ActivityNet Captions.

Since ActivityNet Captions are the largest dataset, 16 layers convolution networks are used to fit it. For TACoS and Charades-STA, 10 layers convolution networks are adopted for saving memory. Each residual block contains two layers of convolution networks, and thus there are 5 residual blocks for TACoS and Charades-STA, and 8 residual blocks for ActivityNet Captions. The semantic modulation mechanism is performed on the last two residual blocks. The foreground recalibration threshold $\theta$, the scaling thresholds $o_{\text{min}}$ and $o_{\text{max}}$ are related to the evaluation metric Rank@n, IoU@m. The results are reported as $n \in \{1,5\}$ with $m \in \{0.5,0.7\}$ for Charades-STA and ActivityNet Captions datasets, and $n \in \{1,5\}$ with $m \in \{0.1,0.3,0.5\}$ for TACoS dataset. Thus, the scaling thresholds $o_{\text{min}}$ and $o_{max}$ are set to 0.5 and 1.0 for ActivityNet Captions and Charades-STA, and 0.3 and 0.7 for TACoS. The foreground recalibration threshold $\theta$ is set as the max IoU threshold m (i.e., 0.7 for ActivityNet Captions and Charades-STA, and 0.5 for TACoS).

Hidden dimension of the sentence encoder LSTM, dimension of the fused features (i.e., $d^f$) and the filter number of convolution operations used in semantic modulation residual module are all set to 512. According to the matching score $\mathcal{P}$, all the predicted moments are ranked and refined by nonmaximum suppression (NMS) with a threshold of 0.45 during the inference. The whole framework is optimized in an end-to-end way by Adam [39] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 0.0001.

## 4.3 Comparison with state-of-the-arts

The performance of our proposed CAN is compared against the state-of-the-art methods, including:

- Sliding window-based methods: CTRL [7], ACRN [30], SLTF [40], ACL [9], MCF [41], ROLE [10] and MCN [8].
- Reinforcement learning-based methods: SM-RL [42] and TripNet [43].
- Others: DEBUG [44], GDP [45], CBP [46], 2D-TAN [33], MAN [11], TGN [13], CMIN [47], QSPN [48], ABLR [29] and SAP [49].

Tables 2, 3 and 4 report the results on the aforementioned three benchmarks.

Overall, our model CAN achieves the highest accuracy over all evaluation metrics and benchmarks. In most case, CAN ranks the first or second place. It is worth noting that on TACoS dataset and Charades-STA dataset, our CAN significantly surpasses the state-of-the-arts by 6.85% and 4.41% absolute improvements under the metrics Rank@1, IoU@0.5 and Rank@1, IoU@0.7, respectively. Besides, on ActivityNet Captions dataset, CAN also outperforms the state-of-the-art methods CMIN [47] and 2D-TAN [33] with respect to Rank@5, IoU@0.5 and Rank@5, IoU@0.7. It demonstrates the superiority of our proposed method CAN.

Moreover, the performance improvements in stricter metrics are more obvious (e.g., Rank@1,IoU@0.5 on TACoS, Rank@1,IoU@0.7 on Charades-STA), and it validates that our proposed method can consistently produce accurate and coherent context information for generating more precise boundaries of predicted video moments.

**Table 2** Performance comparisons on TACoS

| Methods | Rank@1 | | | Rank@5 | | |
|---|---|---|---|---|---|---|
| | IoU(%)@0.1 | IoU(%)@0.3 | IoU(%)@0.5 | IoU(%)@0.1 | IoU(%)@0.3 | IoU(%)@0.5 |
| CTRL [7] | 24.32 | 18.32 | 13.30 | 48.73 | 36.69 | 25.42 |
| MCN [8] | 14.42 | – | 5.58 | 37.35 | – | 10.33 |
| TGN [13] | 41.87 | 21.77 | 18.9 | 53.40 | 39.06 | 31.02 |
| ACRN [30] | 24.22 | 19.52 | 14.62 | 47.42 | 34.97 | 24.88% |
| DEBUG [44] | 41.15 | 23.45 | – | – | – | – |
| ROLE [10] | 20.37 | 15.38 | 9.94 | 45.45 | 31.17 | 20.13 |
| ACL [9] | 31.64 | 24.17 | 20.01 | 57.85 | 42.15 | 30.66 |
| CMIN [47] | 32.48 | 24.64 | 18.05 | 62.13 | 38.46 | 27.02 |
| ABLR [29] | 34.70 | 19.50 | 9.40 | – | – | – |
| 2D-TAN [33] | *47.59* | *37.29* | *25.32* | *70.31* | *57.81* | *45.04* |
| SLTF [40] | 24.67 | 18.07 | 12.36 | 48.78 | 33.20 | 22.86 |
| CBP [46] | – | 27.31 | 24.79 | – | 43.64 | 37.40 |
| MCF [41] | 25.84 | 18.64 | 12.53 | 52.96 | 37.13 | 24.73 |
| SM-RL [42] | 26.51 | 20.25 | 15.95 | 50.01 | 38.47 | 27.84 |
| SAP [49] | 31.15 | – | 18.24 | 53.51 | – | 28.11 |
| **Ours – CAN** | **52**.59 | **42**.84 | **32**.17 | **77**.78 | **65**.71 | **54**.56 |

The best IOU value of each vertical column is in bold

The second-best IOU value of each vertical column is in italics

**Table 3** Performance comparisons on Charades-STA

| Method | Rank@1 | | Rank@5 | |
|---|---|---|---|---|
| | IoU(%)@0.5 | IoU(%)@0.7 | IoU(%)@0.5 | IoU(%)@0.7 |
| CTRL [7] | 23.63 | 8.89 | 58.92 | 29.52 |
| MCN [8] | 17.46 | 8.01 | 48.22 | 26.73 |
| MAN [11] | *41.24* | 20.54 | 83.21 | *51.85* |
| ACRN [30] | 20.26 | 7.64 | 71.99 | 27.79 |
| ROLE [10] | 21.74 | 7.82 | 70.37 | 30.06 |
| ACL [9] | 30.48 | 12.20 | 64.84 | 35.13 |
| SLTF [40] | 23.73 | 9.75 | 73.39 | 32.26 |
| GDP [45] | 39.47 | 18.49 | – | – |
| 2D-TAN [33] | 40.94 | *22.85* | *83.84* | 50.35 |
| SM-RL [42] | 24.36 | 11.17 | 61.25 | 32.08 |
| TripNet [43] | 36.61 | 14.50 | – | – |
| CBP [46] | 36.80 | 18.87 | 70.94 | 50.19 |
| ABLR [29] | 24.36 | 9.01 | – | – |
| SAP [49] | 27.42 | 13.36 | 66.37 | 38.15 |
| QSPN [48] | 35.60 | 15.80 | 79.40 | 45.40 |
| **Ours** – *CAN* | **45**.11 | **27.26** | **86**.10 | **54**.38 |

The best IOU value of each vertical column is in bold

The second-best IOU value of each vertical column is in italics

We also note that the improvement on the ActivityNet Captions is slight, and it is mainly due to the characteristics of the dataset. For example, in ActivityNet Captions, the number of words in sentence query is longer compared to TACoS and Charades-STA. In our method, the entire language query is encoded as a single embedding vector to modulate the convolution processes, and it tends to increase representation ambiguity on those long and complex queries in ActivityNet Captions. As result, the improvement is limited. In the future, we plan to devise fine-grained representation for the query sentence to reduce the query modeling deficiency.

**Table 4** Performance comparisons on Activity Captions

| Method | Rank@1 | | Rank@5 | |
|---|---|---|---|---|
| | IoU(%)@0.5 | IoU(%)@0.7 | IoU(%)@0.5 | IoU(%)@0.7 |
| CTRL [7] | 29.01 | 10.34 | 59.17 | 37.54 |
| MCN [8] | 21.36 | 6.43 | 53.23 | 29.70 |
| CBP [46] | 35.76 | 17.80 | 65.89 | 46.20 |
| 2D-TAN [33] | **44.51** | 26.54 | *77.13* | *61.96* |
| ACRN [30] | 31.67 | 11.25 | 60.34 | 38.57 |
| CMIN [47] | 43.40 | 23.88 | 67.95 | 50.73 |
| TripNet [43] | 32.19 | 13.93 | – | – |
| QSPN [48] | 33.26 | 13.43 | 62.39 | 40.78 |
| ABLR [29] | 36.79 | – | – | – |
| **Ours** – *CAN* | *43.76* | **26**.54 | **77**.71 | **62**.79 |

The best IOU value of each vertical column is in bold

The second-best IOU value of each vertical column is in italics

As aforementioned, the average duration of each video in TACoS is longest among all datasets, and only several moments are used as positive examples. Besides, all the activities in TACoS take place in the same kitchen scenarios with some slightly different cooking objects. Thus, it is hard to learn the fine-grained differences between similar moments with sparse annotations in TACoS. However, our proposed method still outperforms the state-of-the-arts by a large margin under all evaluation metrics.

Furthermore, we compare our approach with temporal convolutional network-based method MAN [11], which utilizes graph convolutional network (GCN) to model relations among different moments. Since it exploits GCN to model context relations for prediction, it outperforms the sliding window methods. From the experimental results, we can see that our CAN performs better than MAN under all evaluation metrics, which further verifies the superiority of our method in integrating context information. In addition, the training and validation processes of CAN on Charades-STA are shown in Fig. 6. According to Fig. 6, it can be concluded that our model converges fast during training and validation.

To evaluate the run-time efficiency of our method, we run experiments with one Nvidia 1080Ti GPU on three datasets. Since the videos in TACoS are relatively long (7 min on average), the experiment on it takes 5 h, 2 h and 20 min for training, validation and testing, respectively. For ActivityNet Captions, it takes about 3 h, 1 h and 15 min for training, validation and testing, respectively. For Charades-STA, it takes about 30 min, 15 min and 80 s for training, validation and testing, respectively. Since Charades-STA contains 3729 testing samples, it can be calculated that our model only takes about 0.02 s to localize one sentence in a given video in it.

## 4.4 Ablation studies

To examine the contributions of each proposed component, ablation studies are performed on Charades-STA to further investigate the effect of semantic modulation residual

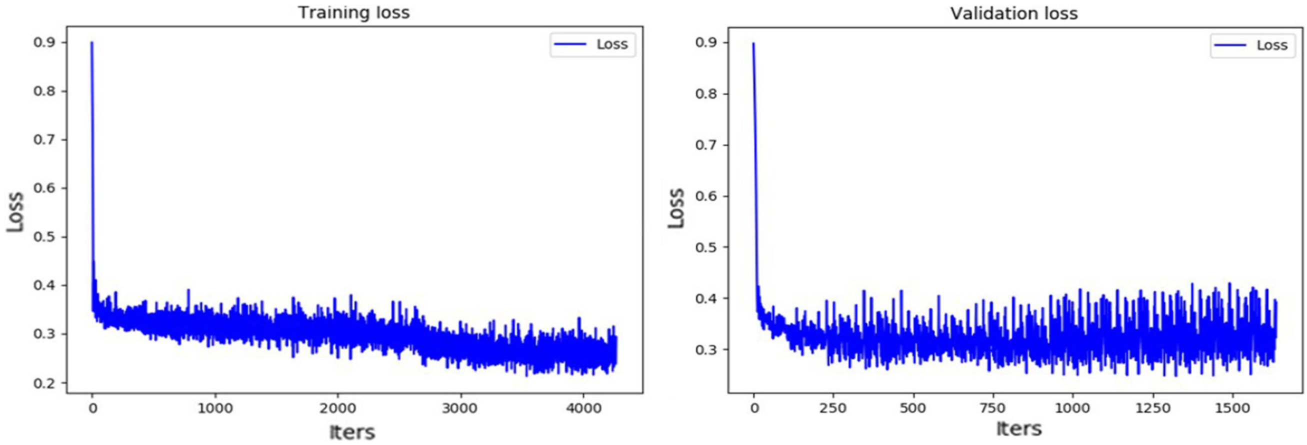

**Fig. 6** Illustration of training and validation losses

module, global interaction module and foreground recalibration mechanism.

#### 4.4.1 Importance of the multimodal encoder

The GNLV task requires to understand both the sentence and video for locating precise temporal regions. As such, current mainstream language grounding methods [8, 9, 11, 12, 29, 30] mainly rely on a set of pre-defined proposals. Thus, we build up the feature map of densely distributed moment proposals by the video clip for further processing. Based on the reconstruct feature map, semantic modulation residual module is devised to correlate the corresponding semantic information from video and language. Besides, the proposal-proposal relations are ignored in previous works; thus, a global interaction module is proposed to explore global context information. Furthermore, as densely sampling proposals inevitably exist data-imbalanced problems, the foreground recalibration is introduced to emphasize the positive moment proposals. In the future, we plan to devise proposal-free method with high localize accuracy for GNLV task. To study the effect of the feature reconstruction alone, the same number of convolution layers is used on top of the reconstruct moment feature map to build a base model, which is denoted as Base-ME in Table 5. On top of the base model, we implement our proposed semantic modulation residual module, global interaction module and foreground recalibration, which are denoted as ME+SMRM, ME+GIM, ME+FR, respectively.

The experiment results are reported in Table 5 (result on Charades-STA), and it can be observed that all components in our proposed method can bring improvements, which verify the effectiveness of our proposed modules. Besides, our full model outperforms the base model by a large margin (Row 5), and it also validates that our model is able to localize precise temporal boundaries by coupling the proposed modules.

#### 4.4.2 Importance of the semantic modulation residual module

To evaluate the effect of semantic modulation residual module alone, variants of it are studied on Charades-STA: **Base**: The residual connections and semantic modulation mechanism are removed, and the same number of convolutional layers is stacked to generate moment features. **Base + RC**: The residual connection between two consecutive convolutional layers is connected, while the semantic modulation mechanism is not used. **Base + RC + SM**: The sentence representation $f^s$ is utilized to modulate the residual connections of the last two residual blocks.

The experiment results are reported in Table 6; it can be observed that Base+RC already outperforms the base model, which indicates the importance of utilizing residual connections to integrate information from different temporal convolutional layers. Comparing BASE+SC with BASE+SC+SM, it can be found that the performance of our model is further improved with considering fine-grained interactions between video and language. This is because the importance of different temporal scale features can be adaptively measured with semantic modulation, the module can generate more discriminative representation for each candidate moment.

Furthermore, experiments are conducted to investigate whether the performance can be improved by adding more semantic modulation layers. In particular, we denote ×1, ×2, ×3 as the number of residual blocks equipped with semantic modulation, respectively. And the semantic modulation is added to the last block, the last two blocks and last there blocks, respectively. The results are reported in Table 6 (Row 3-5), and it can be observed that adding more semantic modulation layers can improve the performance. The reason is that applying semantic modulation in multiple residual blocks can better modulate the convolution processes. However, too many semantic modulation layers can also lead to over-fitting (Row 5).

**Table 5** Ablation studies for importance of multimodal encoder

| Row | Method | Rank@1 | | Rank@5 | |
|---|---|---|---|---|---|
| | | IoU(%)@0.5 | IoU(%)@0.7 | IoU(%)@0.5 | IoU(%)@0.7 |
| 1 | BASE-ME | 40.54 | 22.90 | 82.90 | 49.62 |
| 2 | ME+SMRM | 43.49 | 24.76 | 85.78 | 51.21 |
| 3 | ME+GIM | 41.96 | 24.41 | 83.25 | 50.65 |
| 4 | ME+FR | 43.76 | 24.95 | 83.12 | 51.02 |
| 5 | Ours-CAN | **45.11** | **27.26** | **86.10** | **54.38** |

The best IOU value of each vertical column is in bold

**Table 6** Ablation studies for importance of semantic modulation residual module

| Row | Method | Rank@1 | | Rank@5 | |
|---|---|---|---|---|---|
| | | IoU(%)@0.5 | IoU(%)@0.7 | IoU(%)@0.5 | IoU(%)@0.7 |
| 1 | BASE | 42.98 | 25.13 | 82.90 | 50.13 |
| 2 | BASE+RC | 43.71 | 26.10 | 85.13 | 53.39 |
| 3 | BASE+RC+SM×1 | 44.65 | 26.64 | 85.40 | 53.23 |
| 4 | BASE+RC+SM×2 | **45.11** | **27.26** | **86.10** | **54.38** |
| 5 | BASE+RC+SM×3 | 44.60 | 26.72 | 85.30 | 53.28 |

The best IOU value of each vertical column is in bold

### 4.4.3 Importance of the global interaction module

Furthermore, we validate our design to augment the moment features with global interaction module (GIM). Our model is re-trained with the following three settings: (1) the global interaction module is removed (Row 1). (2) Instead of performing GIM, the commonly adopted fully connected layer-based global contextual module [7–9] is performed on the top of semantic residual module (Row 2). (3) The proposed GIM is adopted to aggregate context clues for learning discriminative moment features. As shown in Table 7, the proposed self-attention-based approach (Row3) achieves significantly better performance than other baselines, demonstrating the superiority of our proposed GIM over corresponding competitors.

### 4.4.4 Importance of the foreground recalibration

A major component of CAN is foreground recalibration (FR), which redefines and relabels the foreground moments region to alleviate the data imbalance issue and uniformly emphasize positive moments. To examine the contributions of FR, we first remove the foreground recalibration. The results are reported in Table 8 (Row 1), which show a significant decrease in performance under high IoU threshold without foreground recalibration. This comparison validates the effectiveness of FR.

We further vary the foreground region threshold $\theta$ defined in Eq. 11 to study the impact of redefined foreground region scope. The results in terms of different $\theta$ are reported in Table 8 (Row 2–5, Row 9) and Fig. 7 (without loss of generality, Rank@1, IoU@0.7 and Rank@5, IoU@0.7 are selected as evaluation metrics in Fig. 7). It can be observed that the performance increases significantly as the $\theta$ enlarges from 0.4 to 0.5. The underlying reason is that FR can prompt our model to uniformly focus on moments with $g_i \geq 0.4$ when we set $\theta$ as 0.4. However, the lowest IoU threshold of evaluation metrics is 0.5. Thus, the moments with $0.4 \leq g_i < 0.5$ may become noise, resulting in performance decrease.

Moreover, as shown in Table 8 and Fig. 7, the performance gets better as $\theta$ increases from 0.5 to 0.7. This is because that the estimation of the candidates which have high overlap with target moments, i.e., high $g_i$, is extremely crucial for the Non maximum suppression, and when $\theta$ is not large enough, the candidates with small $g_i$ may obtain high predicted score, leading to inferior performance.

In addition, $\theta$ is further increased to 0.8 (Fig. 7 and Table 8). In this case, positive training samples, i.e., moments with $g_i > 0.8$, are extremely rare (average 3.6% of the total training samples), and our FR still achieves satisfying results. This phenomenon further proves that our FR is capable of significantly highlighting the positive moments and promoting the predicted boundaries of video moments to become more precise.

To study the effect of the foreground recalibration loss alone, binary cross-entropy loss is adopted to learn the redefined positive moments which is denoted as CAN-BCE in Table 8. More specifically, the loss function is defined as:

$$L_{BCE} = y_i \log p_i + (1 - y_i) \log(1 - p_i), \qquad (16)$$

where a candidate moment is considered as positive if its IoU (temporal Intersection-over-Union) with ground-truth

**Table 7** Ablation studies for importance of global interaction module

| Row | Method | Rank@1 | | Rank@5 | |
|---|---|---|---|---|---|
| | | IoU(%)@0.5 | IoU(%)@0.7 | IoU(%)@0.5 | IoU(%)@0.7 |
| 1 | CAN-w/o-GIM | 42.02 | 25.40 | 85.30 | 52.23 |
| 2 | CAN-FC | 43.01 | 25.94 | 85.51 | 52.23 |
| 3 | Ours-GIM | **45.11** | **27.26** | **86.10** | **54.38** |

The best IOU value of each vertical column is in bold

**Table 8** Ablation studies for importance of foreground recalibration

| Row | Method | Rank@1 | | Rank@5 | |
|---|---|---|---|---|---|
| | | IoU(%)@0.5 | IoU(%)@0.7 | IoU(%)@0.5 | IoU(%)@0.7 |
| 1 | CAN-w/o-FR | 43.84 | 24.81 | 85.16 | 51.75 |
| 2 | CAN-FR $\theta = 0.4$ | 43.84 | 25.24 | 82.69 | 51.65 |
| 3 | CAN-FR $\theta = 0.5$ | 44.54 | 25.54 | 84.84 | 51.91 |
| 4 | CAN-FR $\theta = 0.6$ | 44.70 | 26.10 | 84.68 | 52.04 |
| 5 | CAN-FR $\theta = 0.8$ | 44.33 | 26.56 | 85.19 | 53.04 |
| 6 | CAN-BCE | 43.84 | 25.62 | 85.78 | 52.47 |
| 7 | CAN-FL | 43.12 | 25.00 | 85.59 | 51.96 |
| 8 | CAN-IL | 44.35 | 26.26 | 85.13 | 53.44 |
| 9 | CAN-FR $\theta = 0.7$ | **45.11** | **27.26** | **86.10** | **54.38** |

The best IOU value of each vertical column is in bold

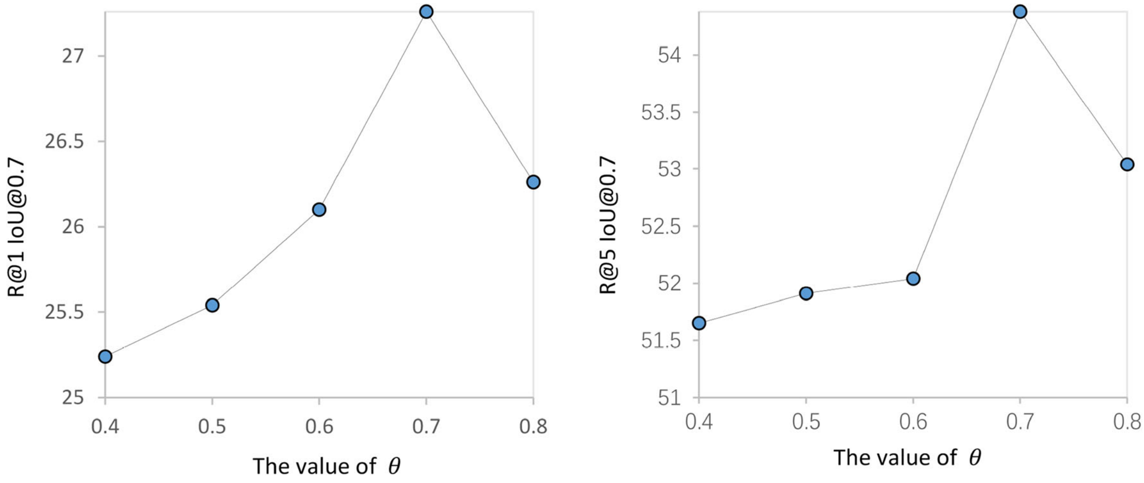

**Fig. 7** Effect of the value of foreground region threshold $\theta$

is larger than 0.7 and its label $y_i$ is set as 1 or $y_i$ is set as 0. The result is shown in Row 6. In this case, $\theta$ is set as 0.7. Comparing with Row 9 which utilizes foreground recalibration loss to learn the same redefined positive moments, it can be observed that roughly using binary cross-entropy to highlight the redefined positive moments will result in worse performance. Besides, we also notice that CAN-BCE performs better than CAN-w/o-FR. This verifies the effectiveness of redefining target moments region.

Furthermore, ablation studies are performed to compare our proposed foreground recalibration loss with the widely used focal loss [50] in Table 8. The main idea of focal loss is to address class imbalance by down-weighting inliers (easy examples) such that their contribution to the total loss is small. Specifically, focal loss adds a factor $(1 - p_t)^{\gamma}$ to the standard cross-entropy criterion. Setting $\gamma > 0$ reduces the relative loss for well-classified examples. The focal loss is given as:

$$FL = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \tag{17}$$

where $\alpha$ is a balance factor. $\gamma$ and $\alpha$ are empirically set to 2 and 0.25, the result is shown in Row 7. Comparing Row 7 with Row 9, the model with the $L_{fr}$ achieves consistent performance enhancements in terms of all metrics, which proves the effectiveness of our idea of introducing $L_{fr}$ to emphasize the positive moments region.

Previous works [8, 9, 29, 30] treat candidate moments within the temporal feature maps as positive if their IoUs with ground-truth moments are larger than 0.5. And their training objective is realized as a binary cross-entropy loss. In our foreground recalibration loss, we relabel the positive moments region with $g_i \geq \theta$ as 1, which is denoted as $(y_i = 1 | g_i \geq \theta)$ in Eq. 11.

To validate the effectiveness of our relabeling strategy, $g_i$ is directly used as the supervision label by replacing the term $(y_i = 1 | g_i \geq \theta)$ with $g_i$. We compare its performance with our proposed relabel method, as shown in Row 8–9. It can be observed that the performance of our model can be further boosted with our relabeling strategy. This validates that considering the positive moments as a whole further enables the model to consistently focus on the foreground region.

### 4.5 Qualitative results

To qualitatively verify the effectiveness of the CAN method, we illustrate some typical examples of the grounding natural language in video. Figure 8 shows the location prediction results of the CAN. As can be seen, the language queries are very flexible. By intuitive observation, our model can retrieve accurate moment boundaries for the GNLV task.

In addition, we illustrate the retrieval results of the full CAN model and ablation model w/o-FR in Fig. 9. It can be observed that through foreground recalibration processing, our model can further recognize the boundaries of the desired moments. For example, in the first example, the visual appearance of the prediction results yielded by the ablation model is very similar to target moment, and thus the ablation model is bewildered to localize the query. In contrast, as the candidate moments tightly close to the ground truth are emphasized by foreground recalibration processing, the full CAN model is able to achieve higher effectiveness.

Furthermore, as the main component of our CAN, the semantic modulation residual module modulates the residual connections under the query sentence guidance to better attend the query-related video features. Thus, we also visualize the modulation weights (defined in Eq. 6) in Fig. 10 to further understand the modulation process. The orange boxes correspond to the moments with highest modulation weights. It can be observed that the query sentence attentively triggers different video contents with the aim of better aligning their semantics. For example, our model precisely pinpoints the target moment, “Person sits down on a chair” by attending to the action “sits down” in the forth frame. This demonstrates that the semantic modulation strategy successfully focuses on the query-related video contents, which is beneficial to high-quality moment prediction.

As illustrated in Fig. 11, it also shows how our CAN fails to localize the moment “The person closes the door.” In this case, the scenes and the actions of the person around the start and end times are very similar. It is hard to precisely define the start and end timestamps of the desired moment even for humans. Thus, our model inevitably retrieves the wrong start and end points. It is possible that high-quality spatio-temporal features would help to eliminate this type of ambiguity.

## 5 Conclusion

In this paper, an effective context-aware network with foreground recalibration is proposed for grounding natural language in video (GNLV). To adaptively integrate the moment features from different temporal convolutional layers, a semantic modulation residual module is devised to modulate the temporal convolution operations with the guide of query semantics. Meanwhile, a global interaction module is adopted to model complex temporal relationships between different moments for learning discriminative information. Moreover, a foreground recalibration mechanism is designed to alleviate the dominating influence of noisy background moments in untrimmed videos, which is also able to emphasize the positive moments. Extensive experiments on three benchmark datasets show the superiority of our proposed method over the state-of-the-arts.

In the future, we plan to deepen our work from the following aspects: (1) reinforcement learning will be incorporated into our model by controlling an agent to adaptively decide the key part in natural language and video. (2) We plan to devise fine-grained representation for the query sentence to reduce the query modeling deficiency. (3) As our method is anchor based that inevitably

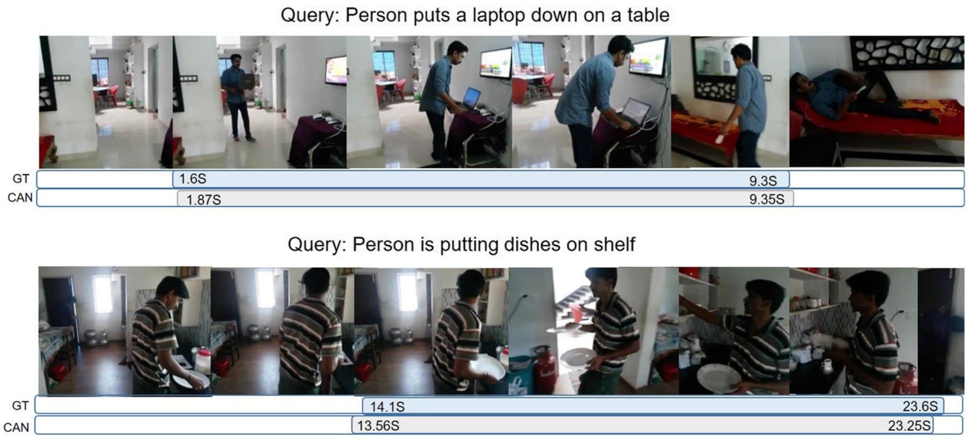
Query: Person puts a laptop down on a table
GT
1.6S
9.3S
CAN
1.87S
9.35S
Query: Person is putting dishes on shelf
GT
14.1S
23.6S
CAN
13.56S
23.25S

**Fig. 8** Qualitative prediction examples of our CAN

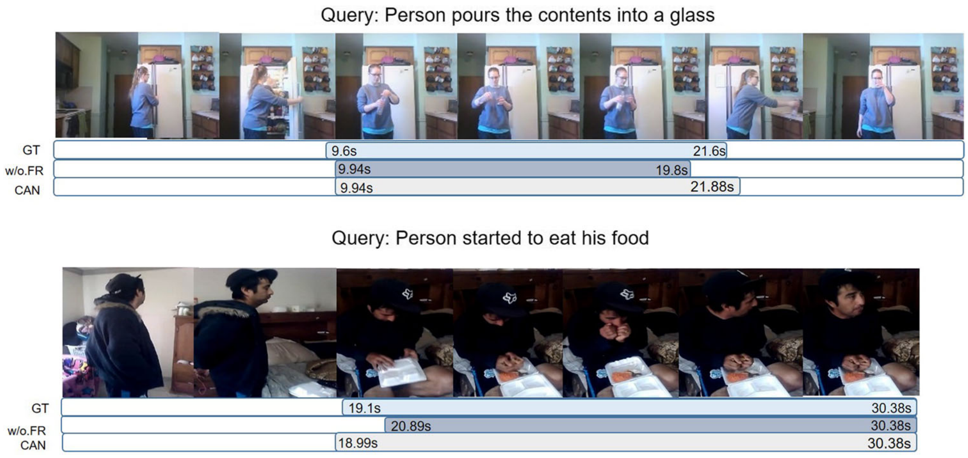

**Fig. 9** Prediction results yielded by the models CAN and w/o-FR

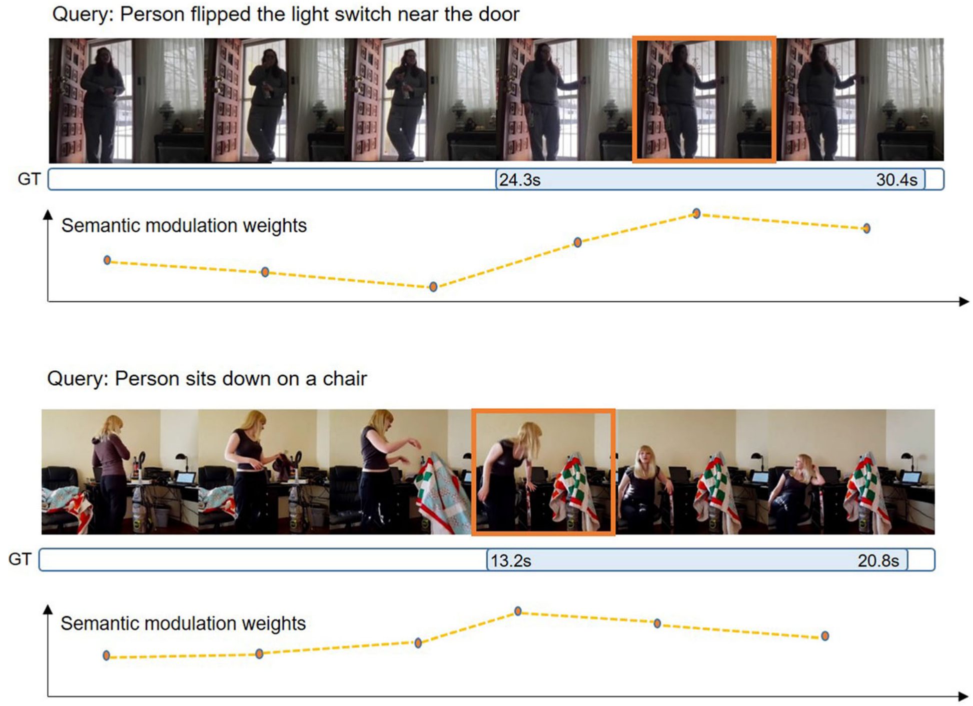

**Fig. 10** The query to video modulation weights in the semantic modulation residual module. Contextual moments corresponding to the highest weights are highlighted in orange boxes (colour figure online)

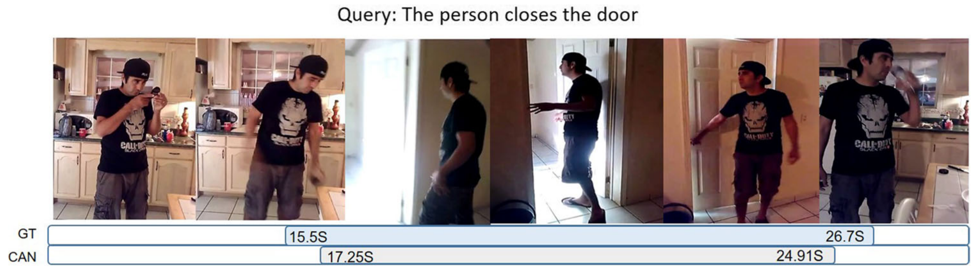

**Fig. 11** Example of failure case of our CAN

have some anchor-related hyperparameters on different datasets, we plan to devise proposal-free detector to reduce design complexity for the GNLV task.

**Acknowledgements** This work was supported in part by National Natural Science Foundation of China under Grant 61771145.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. Gao J, Yang Z, Chen K, Sun C, Nevatia R (2017) TURN TAP: temporal unit regression network for temporal action proposals. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, Venice, pp 706–715
2. Peng Y, Ngo Chong-Wah (2006) Clip-based similarity measure for query-dependent clip retrieval and video summarization. IEEE Trans Circuits Syst Video Technol 16:612–627. https://doi.org/10.1109/TCSVT.2006.873157
3. Ji Z, Xiong K, Pang Y, Li X (2020) Video summarization with attention-based encoder-decoder networks. IEEE Trans Circuits Syst Video Technol 30(6):1709–1717. https://doi.org/10.1109/TCSVT.2019.2904996
4. Lei J, Yu L, Bansal M, Berg T-L (2018) Tvqa: localized, compositional video question answering. In: 2018 ACL conference on empirical methods in natural language processing (EMNLP). ACL, Melbourne, Australia, pp 1369–1379
5. Zhang J, Peng Y (2020) Video captioning with object-aware spatio-temporal correlation and aggregation. IEEE Trans Image Process 29:6209–6222. https://doi.org/10.1109/TIP.2020.2988435
6. Zhang J, Peng Y (2019) Object-aware aggregation with bidirectional temporal graph for video captioning. In: 2019 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Long Beach, CA, pp 8327–8336
7. Gao J, Sun C, Yang Z, Nevatia R (2017) TALL: temporal activity localization via language query. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, Venice, IT, pp 5267–5275
8. Hendricks L-A, Wang O, Shechtman E, Sivic J, Darrell T, Russell B (2017) Localizing moments in video with temporal language. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, Venice, IT, pp 5803–5812
9. Ge R, Gao J, Chen K, Nevatia R (2019) MAC: Mining activity concepts for language-based temporal localization. In: 2019 proceedings of the winter conference on applications of computer vision (WACV). IEEE, Waikoloa, HI, USA, pp 245–253
10. Liu M, Wang X, Nie L, Tian Q, Chen B, Chua T-S (2018) Cross-modal moment localization in videos. 2018 ACM international conference on multimedia (MM). ACM, Seoul, S. Korea, pp 843–851
11. Zhang D, Dai X, Wang X, Wang Y, Davis L-S (2019) MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment. In: 2019 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Long Beach, CA, pp 1247–1257
12. Yuan Y, Ma L, Wang J, Liu W, Zhu W (2019) Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: 2019 advance in neural information processing systems (NIPS). MIT Press, Vancouver, CA, pp 534–544
13. Chen J, Chen X, Ma L, Jie Z, Chua T-S (2018) Temporally grounding natural sentence in video. In: 2018 ACL conference on empirical methods in natural language processing (EMNLP). ACL, Brussels, BE, pp 162–171
14. Krishna R, Hata K, Ren F, Li F-F, Niebles J-C (2017) Dense-captioning events in videos. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, Venice, IT, pp 706–715
15. Oneata D, Verbeek J, Schmid C (2013) Action and event recognition with fisher vectors on a compact feature set. In: 2013 IEEE international conference on computer vision (ICCV). IEEE, Sydney, AU, pp 1817–1824
16. Gaidon A, Harchaoui Z, Schmid C (2013) Temporal localization of actions with actoms. IEEE Trans Pattern Anal Mach Intell 35(11):2782–2795. https://doi.org/10.1109/TPAMI.2013.65
17. Tang K, Yao B, Li F-F, Koller D (2013) Combining the right features for complex event recognition. IEEE international conference on computer vision (ICCV). IEEE, Sydney, AU, pp 2696–2703
18. Shou Z, Wang D, Chang S-F (2016) Temporal action localization in untrimmed videos via multi-stage cnns. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Las Vegas, NV, USA, pp 1049–1058
19. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE international conference on computer vision (ICCV). IEEE, Boston, MA, USA, pp 4489–4497
20. Xu H, Das A, Saenko K (2017) R-c3d: region convolutional 3d network for temporal activity detection. 2017 IEEE international conference on computer vision (ICCV). IEEE, Venice, IT, pp 5783–5792
21. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: 2015 advances in neural information processing systems (NIPS). MIT press, Montreal, CA, pp 91–99
22. Singh B, Marks T-K, Jones M, Tuzel O, Shao M (2016) A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Las Vegas, NV, USA, pp 1961–1970
23. Zhao Y, Xiong Y, Wang L, Wu Z, Tang X, Lin D (2017) Temporal action detection with structured segment networks. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, Venice, IT, pp 2933–2942
24. Zhang D, Dai X, Wang X, Wang Y-F (2018) S3d: Single shot multi-span detector via fully 3d convolutional network. In: 2018 British machine vision conference (BMVC). arXiv:1807.08069
25. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg A-C (2016) Ssd: single shot multibox detector. In: 2016 European conference on computer vision(ECCV). Springer, Amsterdam, NL, pp 21–37
26. Tellex S, Roy D (2009) Towards surveillance video search by natural language query. In: 2009 international conference on image and video retrieval. Springer, Santorini Island, GR, pp 1–8
27. Alayrac J-B, Bojanowski P, Agrawal N, Sivic J, Laptev I, Lacoste-Julien S (2016) Unsupervised learning from narrated instruction videos. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Las Vegas, NV, USA, pp 4575–4583
28. Lin D, Fidler S, Kong C, Urtasun R (2014) Visual semantic search: retrieving videos via complex textual queries. In: 2014 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Columbus, Ohio, pp 2657–2664
29. Yuan Y, Mei T, Zhu W (2019) To find where you talk: temporal sentence localization in video with attention based location regression. In: 2019 AAAI conference on artificial intelligence (AAAI). AAAI, Honolulu, Hawaii, USA, pp 9159–9166
30. Liu M, Wang X, Nie L, He X, Chen B, Chua T-S (2018) Attentive moment retrieval in videos. In: 2018 ACM SIGIR conference on research and development in information retrieval (SIGIR). ACM, Ann Arbor Michigan, USA, pp 15–24
31. Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. In: 2014 ACL conference on empirical methods in natural language processing (EMNLP). ACL, Doha, Qatar, pp 1532–1543
32. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput. 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
33. Zhang S, Peng H, Fu J, Luo J (2020) Learning 2D temporal adjacent networks for moment localization with natural language.

In: 2020 AAAI conference on artificial intelligence (AAAI). AAAI, New York, USA, pp 12870–12877

34. Veit A, Wilber M, Belongie S (2016) Residual networks behave like ensembles of relatively shallow networks. In: 2016 advances in neural information processing systems (NIPS). MIT Press, Barcelona, ES, pp 550–558
35. Huang G, Sun Y, Liu Z, Sedra D, Weinberger K (2016) Deep networks with stochastic depth. In: 2016 proceedings of the european conference on computer vision (ECCV). Springer, Amsterdam, NL, pp 646–661
36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: 2017 advances in neural information processing systems (NIPS). MIT Press, California, USA, pp 5998–6008
37. Marcus R, Michaela R, Mykhaylo A, Sikandar A, Manfred P, Bernt S (2012) Script data for attribute-based recognition of composite activities. In: 2012 European conference on computer vision(ECCV). Springer, Firenze, IT, pp 144–157
38. Simonyan K, Zisserman, A (2015) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
39. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
40. Jiang B, Huang X, Yang C, Yuan J (2019) SLTFNet: a spatial and language-temporal tensor fusion network for video moment retrieval. Inf Process Manag. https://doi.org/10.1016/j.ipm.2019.102104
41. Wu A, Han Y (2019) Multi-modal circulant fusion for video-to-language and backward. In: 2019 international joint conference on artificial intelligence (IJCAI). Morgan Kaufmann, Macao, China, pp 1029–1035
42. Wang W, Huang Y, Wang L (2019) Language driven temporal activity localization: a semantic matching reinforcement learning model. In: 2019 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, California, USA pp 334–343
43. Hahn M, Kadav A, Rehg J-M, Graf H-P (2019) Tripping through time: efficient localization of activities in videos. arXiv:1904.09936
44. Lu C, Chen L, Tan C, Li X, Xiao J (2019) Debug: a dense bottom-up grounding approach for natural language video localization. In: 2019 ACL conference on empirical methods in natural language processing(EMNLP). ACL, Hong Kong, pp 5144–5153
45. Chen L, Lu C, Tang S, Xiao J, Zhang D, Tan C, Li X (2020) Rethinking the bottom-up framework for query-based video localization. In: 2020 AAAI conference on artificial intelligence (AAAI). AAAI, New York, USA, pp 10551–10558
46. Wang J, Ma L, Jiang W (2020) Temporally grounding language queries in videos by contextual boundary-aware prediction. In: 2020 AAAI conference on artificial intelligence (AAAI). AAAI, New York, USA, pp 12168–12175
47. Zhang Z, Lin Z, Zhao Z, Xiao Z (2019) Cross-modal interaction networks for query-based moment retrieval in videos. In: 2019 ACM SIGIR conference on research and development in information retrieval(SIGIR). ACM, Paris, FR, pp 655–664
48. Xu H, He K, Plummer B-A, Sigal L, Sclaroff S, Saenko K (2019) Multilevel language and vision integration for text-to-clip retrieval. In: 2019 AAAI conference on artificial intelligence (AAAI). AAAI, Honolulu, Hawaii, USA, pp 9062–9069
49. Chen S, Jiang Y (2019) Semantic proposal for activity localization in videos via sentence query. In: 2019 AAAI conference on artificial intelligence (AAAI). AAAI, Honolulu, Hawaii, USA, pp 8199–8206
50. Lin T, Goyal P, Girshick R, He K, Dollár P (2020) Focal loss for dense object detection. IEEE Trans Pattern Anal and Mach Intell 42:318–327. https://doi.org/10.1109/TPAMI.2018.2858826