Regularized Two-Branch Proposal Networks for Weakly-Supervised Moment Retrieval in Videos

Zhu Zhang*
Zhejiang University
zhangzhu@zju.edu.cn

Zhijie Lin*
Zhejiang University
linzhijie@zju.edu.cn

Zhou Zhao†
Zhejiang University
zhaozhou@zju.edu.cn

Jieming Zhu
Huawei Noah's Ark Lab
jamie.zhu@huawei.com

Xiuqiang He
Huawei Noah's Ark Lab
hexiuqiang1@huawei.com

ABSTRACT

Video moment retrieval aims to localize the target moment in an video according to the given sentence. The weak-supervised setting only provides the video-level sentence annotations during training. Most existing weak-supervised methods apply a MIL-based framework to develop inter-sample confrontment, but ignore the intra-sample confrontment between moments with semantically similar contents. Thus, these methods fail to distinguish the target moment from plausible negative moments. In this paper, we propose a novel Regularized Two-Branch Proposal Network to simultaneously consider the inter-sample and intra-sample confrontments. Concretely, we first devise a language-aware filter to generate an enhanced video stream and a suppressed video stream. We then design the sharable two-branch proposal module to generate positive proposals from the enhanced stream and plausible negative proposals from the suppressed one for sufficient confrontment. Further, we apply the proposal regularization to stabilize the training process and improve model performance. The extensive experiments show the effectiveness of our method. Our code is released at here[1].

CCS CONCEPTS

• **Information systems** → **Video search**; • **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Weakly-Supervised Moment Retrieval; Two-Branch; Regularization

ACM Reference Format:
Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. 2020. Regularized Two-Branch Proposal Networks for Weakly-Supervised Moment Retrieval in Videos. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413967

Figure 1: An example of video moment retrieval.

1 INTRODUCTION

Given a natural language description and an untrimmed video, video moment retrieval [12, 15] aims to automatically locate the temporal boundaries of the target moment semantically matching to the given sentence. As shown in Figure 1, the sentence describes multiple complicated events and corresponds to a temporal moment with complex object interactions. Recently, a large amount of methods [4, 12, 15, 33, 40] have been proposed to this challenging task and achieved satisfactory performance. However, most existing approaches are trained in the fully-supervised setting with the temporal alignment annotation of each sentence. Such manual annotations are very time-consuming and expensive, especially for ambiguous descriptions. But there is a mass of coarse descriptions for videos without temporal annotations on the Internet, such as the captions for videos on YouTube. Hence, in this paper, we develop a weakly-supervised method for video moment retrieval, which only needs the video-level sentence annotations rather than temporal boundary annotations for each sentence during training.

Most existing weakly-supervised moment retrieval works [7, 13, 23] apply a Multiple Instance Learning (MIL) [17] based methods. They regard matched video-sentence pairs as positive samples and unmatched video-sentence pairs as negative samples. Next, they learn the latent visual-textual alignment by inter-sample confrontment and utilize intermediate results to localize the target moment. Concretely, Mithun et al. [23] apply text-guided attention weights across frames to determine the reliant moment. And Gao and Chen et al. [7, 13] measure the semantic consistency between texts and videos and then directly apply segment scores as localization clues. However, these methods mainly focus on the inter-sample confrontment to judge whether the video matches with the given textual descriptions, but ignore the intra-sample confrontment to decide which moment matches the given language

*Both authors contributed equally to this research.
†Zhou Zhao is the corresponding author.
[1]https://github.com/ikuinen/regularized_two-branch_proposal_network

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413967

best. Specifically, as shown in Figure 1, given a matched video-sentence pair, the video generally contains consecutive contents and these are a large amount of plausible negative moments, which have a bit of relevance to the language. It is intractable to distinguish the target moment from these plausible negative moments, especially when the plausible ones have large overlaps with the ground truth. Thus, we need to develop sufficient intra-sample confrontment between moments with similar contents in a video.

Based on above observations, we propose a novel Regularized Two-Branch Proposal Network (RTBPN) to further explore the fine-grained intra-sample confrontment by discovering the plausible negative moment proposals. Concretely, we first devise a language-aware filter to generate an enhanced video stream and a suppressed stream from the original video stream. In the enhanced stream, we highlight the critical frames according to the language information and weaken unnecessary ones. On the contrary, the crucial frames are suppressed in the suppressed stream. Next, we employ a two-branch proposal module to produce moment proposals from each stream, where the enhanced branch generates positive moment proposals and the suppressed branch produces plausible negative moment proposals. By the sufficient confrontment between two branches, we can accurately localize the most relevant moment from plausible ones. But the suppressed branch may produce simple negative proposals rather than plausible ones, leading to ineffective confrontment. To avoid it, we share all parameters between two branches to make them possess the same ability to produce high-quality proposals. Moreover, parameter sharing can reduce network parameters and accelerate model convergence. By the two-branch framework, we can simultaneously develop sufficient inter-sample and intra-sample confrontment to boost the performance of weakly-supervised video moment retrieval.

Next, we consider the concrete design of the language-aware filter and two-branch proposal module. For the language-aware filter, we first project the language features into fixed cluster centers by a trainable generalized Vector of Locally Aggregated Descriptors (VLAD) [1], where each center can be regarded as a language scene, and then calculate the attention scores between scene and frame features as the language-to-frame relevance. Such a scene-based method introduces an intermediately semantic space for texts and videos, beneficial to boost the generalization ability. Next, to avoid producing a trivial score distribution, e.g. all frames are assigned to 1 or 0, we apply a max-min normalization on the distribution. Based on the normalized distribution, we employ a two-branch gate to produce the enhanced and suppressed streams.

As for the two-branch proposal module, two branches have a completely consistent structure and share all parameters. We first develop a conventional cross-modal interaction [4, 40] between language and frame sequences. Next, we apply a 2D moment map [39] to capture relationships between adjacent moments. After it, we need to generate high-quality moment proposals from each branch. Most existing weakly-supervised approaches [7, 13, 23] take all frames or moments as proposals to perform the inter-sample confrontment, which introduces a large amount of ineffective proposals into the training process. Different from them, we devise a center-based proposal method to filter out unnecessary proposals and only retain high-quality ones. Specifically, we first determine the moment with the highest score as the center and then select those moments having high overlaps with the center one. This technique can effectively select a series of correlative moments to make the confrontment between two branches more sufficient.

Network regularization is widely-used in weakly-supervised tasks [8, 20], which injects extra limitations (i.e. prior knowledge) into the network to stabilize the training process and improve the model performance. Here we design a proposal regularization strategy for our model, consisting of a global term and a gap term. On the one hand, considering most of moments are semantically irrelevant to the language descriptions, we apply a global regularization term to make the average moment score relatively low, which implicitly encourages the scores of irrelevant moments close to 0. On the other hand, we further expect to select the most accurate moment from positive moment proposals, thus we apply another gap regularization term to enlarge the score gaps between those positive moments for better identifying the target one.

Our main contributions can be summarized as follows:

- We design a novel Regularized Two-Branch Proposal Network for weakly-supervised video moment retrieval, which simultaneously considers the inter-sample and intra-sample confrontments by the sharable two-branch framework.
- We devise the language-aware filter to generate the enhanced video stream and the suppressed one, and develop the sharable two-branch proposal module to produce the positive moment proposals and plausible negative ones for sufficient intra-sample confrontment.
- We apply the proposal regularization strategy to stabilize the training process and improve the model performance.
- The extensive experiments on three large-scale datasets show the effectiveness of our proposed RTBPN method.

2 RELATED WORK

2.1 Temporal Action Localization

Temporal action localization aims to detect the temporal boundaries and the categories of action instances in untrimmed videos. The supervised methods [3, 27, 29, 37, 44] mainly adopt the two-stage framework, which first produces a series of temporal action proposals, then predicts the action class and regresses their boundaries. Concretely, Shou et al. [29] design three segment-based 3D ConvNet to accurately localize action instances and Zhao et al. [44] apply a structured temporal pyramid to explore the context structure of actions. Recently, Chao et al. [3] transfer the classical Faster-RCNN framework [26] for action localization and Zeng et al. [37] exploit proposal-proposal relations using graph convolutional networks.

Under the weakly-supervised setting only with video-level action labels, Wang et al. [32] design the classification and selection module to reason about the temporal duration of action instances. Nguyen et al. [24] utilize temporal class activations and class-agnostic attentions to localize the action segments. Further, Shou et al. [28] propose a novel Outer-Inner-Contrastive loss to discover the segment-level supervision for action boundary prediction. To keep the completeness of actions, Liu et al. [20] employ a multi-branch framework where branches are enforced to discover distinctive parts of actions. And Yu et al. [35] explore the temporal action structure and model each action as a multi-phase process.

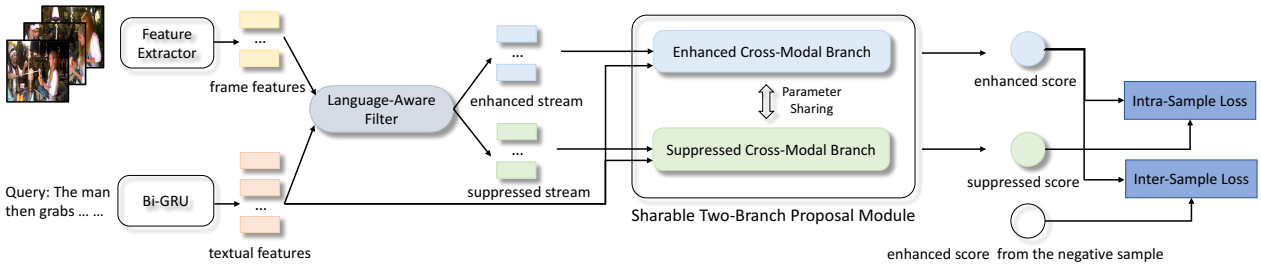


Figure 2: The Overall Architecture of the Regularized Two-Branch Proposal Network.

2.2 Video Moment Retrieval

Video moment retrieval aims to localize the target moment according to the given query in an untrimmed video. Most existing methods employ a top-down framework, which first generates a set of moment proposals and then selects the most relevant one. Early approaches [12, 15, 16, 21, 22] explicitly extract the moment proposals by the sliding windows with various lengths and individually calculate the correlation of each proposal with the query in a multi-modal space. To incorporate long-term video context, researchers [4, 19, 34, 36, 38–40] implicitly produce moment proposals by defining multiple temporal anchors after holistic visual-textual interactions. Concretely, Chen et al. [4] build sufficient frame-by-word interaction and dynamically aggregate the matching clues. Zhang et al. [38] employ an iterative graph adjustment network to learn moment-wise relations in a structured graph. And Zhang et al. [39] design a 2D temporal map to capture the temporal relations between adjacent moments. Different from the top-down formula, the bottom-up framework [5, 6] is designed to directly predict the probabilities of each frame as target boundaries. Further, He and Wang et al. [14, 33] formulate this task as a problem of sequential decision making and apply the reinforcement learning method to progressively regulate the temporal boundaries. Besides temporal moment retrieval, recent works [8, 41, 43] also localize the spatio-temporal tubes from videos according to the give language descriptions. And Zhang et al. [42] try to localize the target moment by the image query instead of the natural language query.

Recently, researchers [7, 10, 13, 18, 23] begin to explore the weakly-supervised moment retrieval only with the video-level sentence annotations. Mithun, Gao and Chen et al. [7, 13, 23] apply a MIL-based framework to learn latent visual-textual alignment by inter-sample confrontment. Mithun et al. [23] determine the reliant moment based on the intermediately text-guided attention weights. Gao et al. [13] devise an alignment module to measure the semantic consistency between texts and videos and apply a detection module to compare moment proposals. And Chen et al. [7] apply a two-stage model to detect the accurate moment in a coarse-to-fine manner. Besides MIL-based methods, Lin et al. [18] propose a semantic completion network to rank proposals by a language reconstruction reward, but ignore the inter-sample confrontments. Unlike previous methods, we design a sharable two-branch framework to simultaneously consider the inter-sample and intra-sample confrontments for weakly-supervised video moment retrieval.

3 THE PROPOSED METHOD

Given a video V and a sentence S, video moment retrieval aims to retrieve the most relevant moment $\hat{l} = (\hat{s}, \hat{e})$ within the video V, where $\hat{s}$ and $\hat{e}$ denote the indices of the start and end frames of the target moment. Due to the weakly-supervised setting, we can only utilize the coarse video-level annotations.

3.1 The Overall Architecture Design

We first introduce the overall architecture of our Regularized Two-Branch Proposal Network (RTBPN). As shown in Figure 2, we devise a language-aware filter to generate the enhanced video stream and the suppressed video stream, and next develop the sharable two-branch proposal module to produce the positive moment proposals and plausible negative ones. Finally, we develop the inter-sample and intra-sample losses with proposal regularization terms.

Concretely, we first extract the word features of the sentence by a pre-trained Glove word2vec embedding [25]. We then feed the word features into a Bi-GRU network [9] to learn word semantic representations $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{n_q}$ with contextual information, where n_q is the word number and $\mathbf{q}_i$ is the semantic feature of the i-th word. As for videos, we first extract visual features using a pre-trained feature extractor (e.g. 3D-ConvNet [31]) and then apply a temporal mean pooling to shorten the sequence length. We denote frame features as $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{n_v}$, where n_v is the feature number.

After feature extraction, we devise a language-aware filter to generate the enhanced and suppressed video streams, given by

$$\mathbf{V}^{en}, \mathbf{V}^{sp} = \text{Filter}\,(\mathbf{V}, \mathbf{Q}), \tag{1}$$

where $\mathbf{V}^{en} = \{\mathbf{v}_i^{en}\}_{i=1}^{n_v}$ represents the enhanced video stream and $\mathbf{V}^{sp} = \{\mathbf{v}_i^{sp}\}_{i=1}^{n_v}$ is the suppressed video stream. In the enhance stream, we highlight the critical frame features relevant to the language and weaken unnecessary ones. On the contrary, the significative frames are suppressed in the suppressed stream.

Next, we develop the sharable two-branch proposal module to produce the positive moment proposals and plausible negative ones. The module consists of a enhanced branch and a suppressed branch with the consistent structure and sharable parameters Θ, given by

$$\begin{aligned} \mathrm{P}^{en}, \mathbf{L}^{en}, \mathbf{C}^{en} &= \text{EnhancedBranch}_{\Theta}\,(\mathbf{V}^{en}, \mathbf{Q}), \\ \mathrm{P}^{sp}, \mathbf{L}^{sp}, \mathbf{C}^{sp} &= \text{SuppressedBranch}_{\Theta}\,(\mathbf{V}^{sp}, \mathbf{Q}), \end{aligned} \tag{2}$$

where we feed the enhanced video stream $\mathbf{V}^{en}$ and textual features $\mathbf{Q}$ into the enhanced branch and produce the positive moment

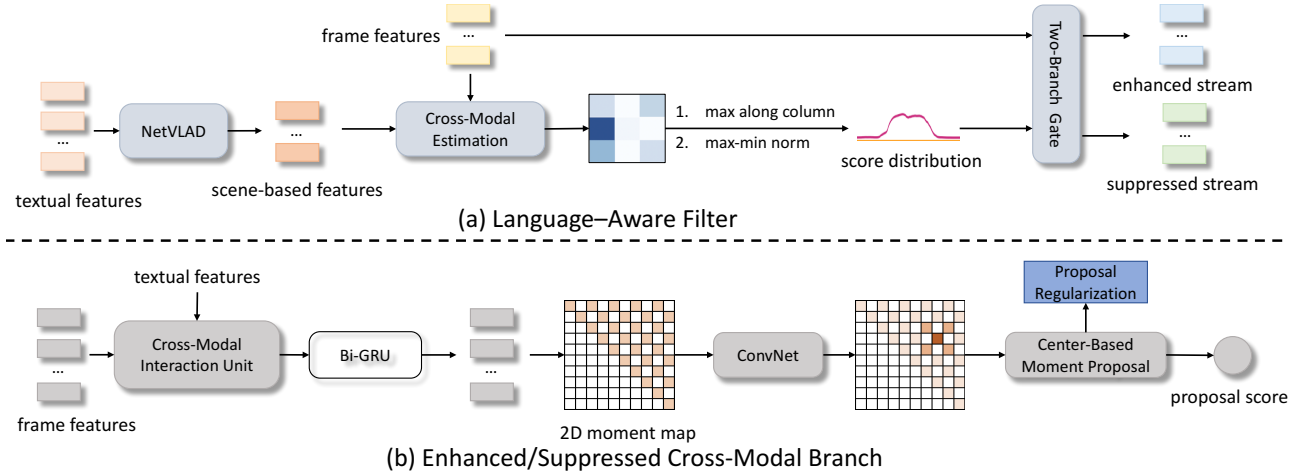


Figure 3: The Concrete Designs of the Language-Aware Filter and Sharable Two-Branch Proposal Module.

proposals $\mathrm{P}^{en} = \{p_i^{en}\}_{i=1}^{T}$, their corresponding temporal boundaries $\mathbf{L}^{en} = \{(s_i^{en}, e_i^{en})\}_{i=1}^{T}$ and proposal scores $\mathbf{C}^{en} = \{c_i^{en}\}_{i=1}^{T}$. The T is the number of moment proposals. Each proposal p_i^{en} corresponds the start and end timestamps (s_i^{en}, e_i^{en}) and the confidence score $c_i^{en} \in (0, 1)$. Likewise, the suppressed branch generates P^{sp}, $\mathbf{L}^{sp}$ and $\mathbf{C}^{sp}$ from the suppressed stream. Next, we can compute the enhanced score $K^{en} = \sum_{i=1}^{T} c_i^{en}$ and suppressed score $K^{sp} = \sum_{i=1}^{T} c_i^{sp}$. The intra-sample loss is given by

$$\mathcal{L}_{intra} = \max(0,\ \Delta_{intra} - K^{en} + K^{sp}), \tag{3}$$

where $\mathcal{L}_{intra}$ is a margin-based triplet loss and Δ is a margin which is set to 0.4. Due to the parameter sharing between two branches, the suppressed branch will select plausible negative proposals. By sufficient intra-sample confrontment, we are able to distinguish the target moment from the intractable negative moments.

Besides the intra-sample loss, we also develop a inter-sample loss by utilizing the unmatched video-sentence sample, i.e. the negative sample. Specifically, for each video V, we randomly select a sentence from the training set as the unmatched sentence $\overline{S}$ to form a negative sample $(V, \overline{S})$. Likewise, we can randomly choose a video to construct another negative sample $(\overline{V}, S)$. Next, we apply the RTBPN to produce the enhanced scores $\overline{K}_S^{en}$ and $\overline{K}_V^{en}$ for negative samples. The inter-sample loss is given by

$$\mathcal{L}_{inter} = \max(0,\ \Delta_{inter} - K^{en} + \overline{K}_S^{en}) + \max(0,\ \Delta_{inter} - K^{en} + \overline{K}_V^{en}), \tag{4}$$

where the Δ_{inter} is set to 0.6 and $\mathcal{L}_{inter}$ encourages the enhanced scores of positive samples to be larger than negative samples.

3.2 Language-Aware Filter

We next introduce the language-aware filter with the scene-based cross-modal estimation. To calculate the language-relevant score distribution over frames, we first apply a NetVLAD [1] to project the textual features $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{n_q}$ into cluster centers. Concretely, given the trainable center vectors $\mathbf{C} = \{\mathbf{c}_j\}_{j=1}^{n_c}$ where n_c is the number of centers, the NetVLAD accumulates the residuals between language features and center vectors by a soft assignment, given by

$$\alpha_i = \mathrm{softmax}(\mathbf{W}^c \mathbf{q}_i + \mathbf{b}^c),\ \mathbf{u}_j = \textstyle\sum_{i=1}^{n_q} \alpha_{ij}(\mathbf{q}_i - \mathbf{c}_j), \tag{5}$$

where $\mathbf{W}^c$ and $\mathbf{b}^c$ are projection matrix and bias. The softmax operation produces the soft assignment coefficients $\alpha_i \in \mathbb{R}^{n_c}$ corresponding to n_c centers. The $\mathbf{u}_j$ is the accumulated features from $\mathbf{Q}$ for the i-th center. We can regard each center as a language scene and $\mathbf{u}_j$ is the scene-based language feature. We then calculate the cross-modal matching scores between $\{\mathbf{v}_i\}_{i=1}^{n_v}$ and $\{\mathbf{u}_j\}_{j=1}^{n_c}$ by

$$\beta_{ij} = \sigma(\mathbf{w}_a^\top \tanh(\mathbf{W}_1^a \mathbf{v}_i + \mathbf{W}_2^a \mathbf{u}_j + \mathbf{b}^a)), \tag{6}$$

where $\mathbf{W}_1^a$, $\mathbf{W}_2^a$ are projection matrices, $\mathbf{b}^a$ is the bias, $\mathbf{w}_a^\top$ is the row vector and σ is the sigmoid function. The $\beta_{ij} \in (0, 1)$ means the matching score of the i-th frame feature and j-th scene-based language feature. That is, scene-based method introduces an intermediately semantic space for texts and videos.

Considering a frame should be important if it is associated with any language scene, we compute the holistic score for the i-th frame by $\overline{\beta}_i = \max_j\{\beta_{ij}\}$. Then, to avoid producing a trivial score distribution, e.g. all frames are assigned to 1 or 0, we apply a max-min normalization on the distribution by

$$\widetilde{\beta}_i = \frac{\overline{\beta}_i - \min_j\{\overline{\beta}_j\}}{\max_j\{\overline{\beta}_j\} - \min_j\{\overline{\beta}_j\}}. \tag{7}$$

Thus, we obtain the normalized distribution $\{\widetilde{\beta}_i\}_{i=1}^{n_v}$ over frames, where the i-th value means the relevance between the i-th frame and language descriptions. Next, we apply a two-branch gate to produce the enhanced and suppressed streams, denoted by

$$\mathbf{v}_i^{en} = \widetilde{\beta}_i \cdot \mathbf{v}_i,\ \ \mathbf{v}_i^{sp} = (1 - \widetilde{\beta}_i) \cdot \mathbf{v}_i, \tag{8}$$

where the enhance stream $\mathbf{V}^{en} = \{\mathbf{v}_i^{en}\}_{i=1}^{n_v}$ highlights the critical frames and weaken unnecessary ones according to the normalized score, while the suppressed stream $\mathbf{V}^{sp} = \{\mathbf{v}_i^{sp}\}_{i=1}^{n_v}$ is the opposite.

3.3 Sharable Two-Branch Proposal Module

In this section, we introduce the sharable two-branch proposal module, including an enhanced branch and a suppressed branch with a

consistent structure and sharable parameters. The sharing setting can make both branches produce high-quality moment proposals, avoiding the suppressed branch generating too simple negative proposals and leading to the ineffective confrontment. Here we only present the design of the enhanced branch.

Given the enhanced stream $\mathbf{V}^{en} = \{\mathbf{v}_i^{en}\}_{i=1}^{n_v}$ and textual features $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{n_q}$, we first conduct a widely-used cross-modal interaction unit [5, 40] to incorporate textual clues into visual features. Concretely, we perform a frame-to-word attention and aggregate the textual features for each frame, given

$$\begin{aligned} \delta_{ij} &= \mathbf{w}_m^\top \tanh(\mathbf{W}_1^m \mathbf{v}_i^{en} + \mathbf{W}_2^m \mathbf{q}_j + \mathbf{b}^m), \\ \overline{\delta}_{ij} &= \frac{\exp(\delta_{ij})}{\sum_{k=1}^{n_q} \exp(\delta_{ik})}, \ \mathbf{s}_i^{en} = \sum_{j=1}^{n_q} \overline{\delta}_{ij} \mathbf{q}_j, \end{aligned} \tag{9}$$

where $\mathbf{s}_i^{en}$ is the aggregated textual representation relevant to the i-th frame. Then, the cross gate is applied to develop the visual-textual interaction, given by

$$\begin{aligned} &\mathbf{g}_i^v = \sigma(\mathbf{W}^v \mathbf{v}_i^{en} + \mathbf{b}^v), \ \mathbf{g}_i^t = \sigma(\mathbf{W}^t \mathbf{s}_i^{en} + \mathbf{b}^t), \\ &\overline{\mathbf{s}}_i^{en} = \mathbf{s}_i^{en} \odot \mathbf{g}_i^v, \ \overline{\mathbf{v}}_i^{en} = \mathbf{v}_i^{en} \odot \mathbf{g}_i^t, \end{aligned} \tag{10}$$

where $\mathbf{g}_i^v$ is the visual gate, $\mathbf{g}_i^t$ is textual gate and $\odot$ is element-wise multiplication. After it, we concatenate $\overline{\mathbf{v}}_i^{en}$ and $\overline{\mathbf{s}}_i^{en}$ to obtain the language-aware frame feature $\mathbf{m}_i^{en} = [\overline{\mathbf{v}}_i^{en}; \overline{\mathbf{s}}_i^{en}]$.

Next, we follow the 2D temporal network [39] to build a 2D moment feature map and capture relationships between adjacent moments. Specifically, the 2D feature map $\mathbf{F} \in \mathbb{R}^{n_v \times n_v \times d_m}$ consists of three dimension: the first two dimensions represent the start and end frame indices of a moment and the third dimension is the feature dimension. The feature of a moment with temporal duration [a, b] is computed by $\mathbf{F}[a, b, :] = \sum_{i=a}^{b} \mathbf{m}_i^{en}$. Note that the location with a > b is invalid and is padded with zeros. And we also follow the sparse sampling setting in [39] to avoid much computational cost. That is, not all moments with a <= b are proposed if the n_v is large. With the 2D maps, we conduct the two-layer 2D convolution with the kernel size K to develop moment relationships between adjacent moments. After it, we obtain the cross-modal features $\{\mathbf{f}_i^{en}\}_{i=1}^{M_{en}}$, where M_{en} is the number of all moments in the 2D map, and compute their proposal scores $\{c_i^{en}\}_{i=1}^{M_{en}}$ by

$$c_i^{en} = \sigma(\mathbf{W}^p \mathbf{f}_i^{en} + \mathbf{b}^p). \tag{11}$$

Next, we employ a center-based proposal method to fiter out unnecessary moments and only retain high-quality ones as the positive moment proposals. Concretely, we first choose the moment with the highest score c_i^{en} as the center moment and rank the rest of moments according to the overlap with the center one. We then select top $T-1$ moments and obtain T positive proposals $\mathrm{P}^{en} = \{p_i^{en}\}_{i=1}^{T}$ with proposal scores $\mathbf{C}^{en} = \{c_i^{en}\}_{i=1}^{T}$. And temporal boundaries (s_i^{en}, e_i^{en}) of each moment are the indices of its location in the 2D map. This method can effectively select a series of correlative moments. Likewise, the suppressed branch has the completely identical structure to generate the plausible negative proposals $\mathrm{P}^{sp} = \{p_i^{sp}\}_{i=1}^{T}$ with proposal scores $\mathbf{C}^{sp} = \{c_i^{sp}\}_{i=1}^{T}$.

3.4 Proposal Regularization

Next, we devise a proposal regularization strategy to inject some prior knowledge into our model, consisting of a global term and a gap term. Due to the parameter sharing between two branches, we only apply the proposal regularization in the enhanced branch.

Specifically, considering most of moments are unaligned to the language descriptions, we first apply a global term to make the average moment score relatively low, given by

$$\mathcal{L}_{global} = \frac{1}{M_{en}} \sum_{i=1}^{M_{en}} c_i^{en}, \tag{12}$$

where M_{en} is the number of all moments in the 2D map. This global term implicitly encourages the scores of unselected moments in the 2D map close to 0, while $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ guarantee positive proposals have high scores.

On the other hand, we further expect to identify the most accurate one as the final localization result from T positive moment proposals, thus it is crucial to enlarge the score gaps between these proposals to make them distinguishable. We perform softmax on positive proposal scores and then employ the gap term $\mathcal{L}_{gap}$ by

$$\overline{c}_i^{en} = \frac{\exp(c_i^{en})}{\sum_{i=1}^{T} \exp(c_i^{en})}, \ \mathcal{L}_{gap} = -\sum_{i=1}^{T} \overline{c}_i^{en} \log(\overline{c}_i^{en}), \tag{13}$$

where T is the number of positive proposals rather than the number M_{en} of all proposals. When the $\mathcal{L}_{gap}$ decreases, the score distribution will become more diverse, i.e. it implicitly encourages to enlarge the score gaps between positive moment proposals.

3.5 Training and Inference

Based on the aforementioned model design, we apply a multi-task loss to train our RTBPN in an end-to-end manner, given by

$$\mathcal{L}_{RTBPN} = \lambda_1 \mathcal{L}_{intra} + \lambda_2 \mathcal{L}_{inter} + \lambda_3 \mathcal{L}_{global} + \lambda_4 \mathcal{L}_{gap}, \tag{14}$$

where λ_* are the hyper-parameters to control the balance of losses.

During inference, we can directly select the moment p_i^{en} with the highest proposal score c_i^{en} from the enhanced branch.

4 EXPERIMENTS

4.1 Datasets

We conduct extensive experiments on three public datasets.

Charades-STA [12]: The dataset is built on the original Charades dataset [30], where Gao et al. apply a semi-automatic way to generate the language descriptions for temporal moments. This dataset contains 9,848 videos of indoor activities and their average duration is 29.8 seconds. The dataset contains 12,408 sentence-moment pairs for training and 3,720 pairs for testing.

ActivityCaption [2]: The dataset contains 19,209 videos with diverse contents and their average duration is about 2 minutes. Following the standard split in [39, 40], there are 37,417, 17,505 and 17,031 sentence-moment pairs used for training, validation and testing, respectively. This is the largest dataset currently.

DiDeMo [15]: The dataset consists of 10,464 videos and the duration of each video is 25-30 seconds. It contains 33,005 sentence-moment pairs for training, 4,180 for validation and 4,021 for testing. Especially, each video in DiDeMo is divided into six five-second clips and the target moment contains one or more consecutive clips. Thus, there are only 21 moment candidates while Charades-STA and ActivityCaption allow arbitrary temporal boundaries.

Table 1: Performance Evaluation Results on Charades-STA ($n \in \{1, 5\}$ and $m \in \{0.3, 0.5, 0.7\}$).

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
fully-supervised methods						
VSA-RNN [12]	-	10.50	4.32	-	48.43	20.21
VSA-STV [12]	-	16.91	5.81	-	53.89	23.58
CTRL [12]	-	23.63	8.89	-	58.92	29.52
QSPN [34]	54.70	35.60	15.80	95.60	79.40	45.40
2D-TAN [39]	-	39.81	23.25	-	79.33	52.15
weakly-supervised methods						
TGA [23]	32.14	19.94	8.84	86.58	65.52	33.51
CTF [7]	39.80	27.30	12.90	-	-	-
SCN [18]	42.96	23.58	9.97	95.56	71.80	38.87
RTBPN (our)	**60.04**	**32.36**	**13.24**	**97.48**	**71.85**	**41.18**

4.2 Evaluation Criteria

Following the widely-used setting [12, 15], we apply **R@n,IoU=m** as the criteria for Charades-STA and ActivityCaption and use **Rank@1**, **Rank@5** and **mIoU** as the criteria for DiDeMo. Concretely, we first calculate the IoU between the predicted moments and ground truth, and **R@n,IoU=m** means the percentage of at least one of the top-n moments having the IoU > m. The **mIoU** is the average IoU of the top-1 moment over all testing samples. And for DiDeMo, due to only 21 moment candidates, **Rank@1** or **Rank@5** is the percentage of samples which ground truth moment is ranked as top-1 or among top-5.

4.3 Implementation Details

We next introduce the implementation details of our RTBPN model.

Data Preprocessing. For a fair comparison, we apply the same visual features as previous methods [12, 15, 40], that is, C3D features for Charades-STA and ActivityCaption and VGG16 and optical flow features for DiDeMo. We then shorten the feature sequence using temporal mean pooling with the stride 4 and 8 for Charades-STA and ActivityCaption, respectively. And for DiDeMo, we compute the average feature for each fixed five-second clips as in [15]. As for sentence queries, we extract 300-d word embeddings by the pre-trained Glove embedding [25] for each word token.

Model Setting. In the center-based proposal method, the positive/negative proposal number T is set to 48 for Charades-STA and ActivityCaption and 6 for DiDeMo. During 2D feature map construction, we fill all locations [a, b] if a <= b for DiDeMo. But for Charades-STA, we add another limitation $(b - a) \mod 2 = 1$. And for ActivityCaption, we only fill the location [a, b] if $(b - a) \mod 8 = 0$. The sparse sampling avoids much computational cost. We set the convolution kernel size K to 3, 9 and 3 for Charades-STA, ActivityCaption and DiDeMo, respectively. Besides, the dimension of almost parameter matrices and bias in our model to 256, including the $\mathbf{W}^c, \mathbf{b}^c$ in the NetVLAD, $\mathbf{W}_1^m, \mathbf{W}_2^m$ and $\mathbf{b}^m$ in the frame-to-word attention and so on. We set the dimension of the hidden state of each direction in the Bi-GRU networks to 128. And the dimension of trainable center vectors is 256. During training, we set λ_1, λ_2, λ_3, λ_4 to 0.1, 1, 0.01 and 0.01, respectively. And we use an Adam optimizer [11] with the initial learning rate 0.001 and batch size 64. During inference, we apply the non-maximum suppression (NMS) with a threshold 0.55 while we need to select multiple moments.

Table 2: Performance Evaluation Results on ActivityCaption ($n \in \{1, 5\}$ and $m \in \{0.1, 0.3, 0.5\}$).

Method	R@1			R@5		
	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.1	IoU=0.3	IoU=0.5
fully-supervised methods						
TGN [4]	-	43.81	27.93	-	54.56	44.20
QSPN [34]	-	45.30	27.70	-	75.70	59.20
2D-TAN [39]	-	59.45	44.51	-	85.53	77.13
weakly-supervised methods						
WS-DEC [10]	62.71	41.98	23.34	-	-	-
WSLLN [13]	**75.40**	42.80	22.70	-	-	-
CTF [7]	74.20	44.30	23.60	-	-	-
SCN [18]	71.48	47.23	29.22	90.88	71.45	55.69
RTBPN (our)	73.73	**49.77**	**29.63**	**93.89**	**79.89**	**60.56**

Table 3: Performance Evaluation Results on DiDeMo.

Method	Input	Rank@1	Rank@5	mIoU
fully-supervised methods				
MCN [15]	RGB	13.10	44.82	25.13
TGN [4]	RGB	24.28	71.43	38.62
MCN [15]	Flow	18.35	56.25	31.46
TGN [4]	Flow	27.52	76.94	42.84
MCN [15]	RGB+Flow	28.10	78.21	41.08
TGN [4]	RGB+Flow	28.23	79.26	42.97
weakly-supervised methods				
WSLLN [13]	RGB	19.40	53.10	25.40
RTBPN (our)	RGB	**20.38**	**55.88**	**26.53**
WSLLN [13]	Flow	18.40	54.40	27.40
RTBPN (our)	Flow	**20.52**	**57.72**	**30.54**
TGA [23]	RGB+Flow	12.19	39.74	24.92
RTBPN (our)	RGB+Flow	**20.79**	**60.26**	**29.81**

4.4 Comparison to State-of-the-Art Methods

We compare our RTBPN method with existing state-of-the-art methods, including the supervised and weakly-supervised approaches.

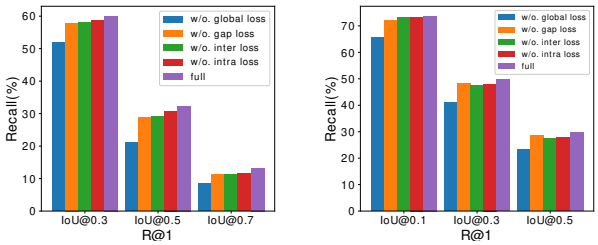
Supervised Method: Early approaches VSA-RNN [12], VSA-STV [12], CTRL [12] and MCN [15] projects the visual features of candidate moments and textual features into a common space for correlation estimation. From a holistic view, TGN [4] develops the frame-by-word interaction by RNN. And QSPN [34] integrates vision and language features early and re-generate descriptions as an auxiliary task. Further, 2D-TAN [39] captures the temporal relations between adjacent moments by the 2D moment map.

Weakly-Supervised Method: WS-DEC [10] regards weakly-supervised moment retrieval and dense video captioning as the dual problems. Under the MIL framework, TGA [23] utilizes the text-guided attention weights to detect the target moment, WSLLN [13] simultaneously apply the alignment and detection module to boost the performance, and CTF [7] detects the moment in a two-stage coarse-to-fine manner. Different from MIL-based methods, SCN [18] ranks moment proposals by a language reconstruction reward.

The overall evaluation results on three large-scale datasets are presented in Table 1, Table 2 and Table 3, where we set $n \in \{1, 5\}$, $m \in \{0.3, 0.5, 0.7\}$ for Charades-STA and $n \in \{1, 5\}$, $m \in \{0.1, 0.3, 0.5\}$ for ActivityCaption. The results reveal some interacting points:

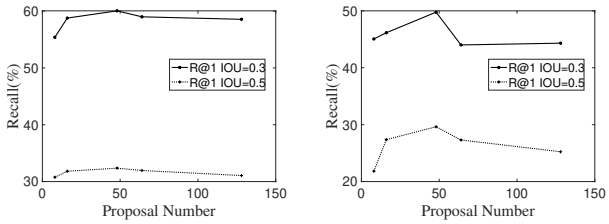
Table 4: Ablation results about the two-branch architecture, filter details and center-based proposal method.

Method	Charades-STA						ActivityCaption					
	R@1			R@5			R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.1	IoU=0.3	IoU=0.5
The Two-Branch Architecture												
w/o. filter	56.43	29.14	11.40	94.86	67.25	37.59	73.54	43.55	26.67	89.79	73.14	57.92
w/o. parameter sharing	32.62	13.87	4.55	80.43	47.06	19.28	**80.47**	48.35	22.92	90.27	75.11	57.03
full model	**60.04**	**32.36**	**13.24**	**97.48**	**71.85**	**41.18**	73.73	**49.77**	**29.63**	**93.89**	**79.89**	**60.56**
The Filter Design												
visual-only scoring	57.85	30.59	12.89	95.78	68.75	40.54	71.82	45.69	27.87	90.52	76.03	58.87
w/o. NetVALD	58.61	31.92	13.14	96.26	70.84	40.70	72.32	45.15	28.08	91.41	77.75	59.91
full model	**60.04**	**32.36**	**13.24**	**97.48**	**71.85**	**41.18**	**73.73**	**49.77**	**29.63**	**93.89**	**79.89**	**60.56**
The Proposal Method												
all-proposal	57.92	30.94	12.16	95.59	68.21	38.84	**82.61**	48.02	21.21	90.37	73.09	55.02
top-k proposal	58.61	31.16	12.63	95.38	69.70	39.55	71.85	47.08	28.25	92.82	77.63	59.89
full model (center-based)	**60.04**	**32.36**	**13.24**	**97.48**	**71.85**	**41.18**	73.73	**49.77**	**29.63**	**93.89**	**79.89**	**60.56**



(a) Charades-STA (b) ActivityCaption

Figure 4: Ablation Results of the Multi-Task Losses.



(a) Charades-STA (b) ActivityCaption

Figure 5: Effect of the Proposal Number on Charades-STA and ActivityCaption Datasets.

- On almost all criteria of three datasets, our RTBPN method achieves the best weakly-supervised performance, especially on Charades-STA. This fact verifies the effectiveness of our two-branch framework with the regularization strategy.
- The reconstruction-based method SCN outperforms MIL-based methods TGA, CTF and WSLLN on Charades-STA and ActivityCaption, but our RTBPN achieves a better performance than SCN, demonstrating our RTBPN with the intra-sample confrontment can effectively discover the plausible negative samples and improve the accuracy.
- On the DiDeMo dataset, our RTBPN outperforms the state-of-the-art baselines using RGB, Flow and two-stream features. This fact suggests our method is robust for diverse features.
- Our RTBPN outperforms the early supervised approaches VSA-RNN, VSA-STV, CTRL and obtains the results comparable to other methods TGN, QSPN and MCN, which indicates even under the weakly-supervised setting, our RTBPN can still develop the sufficient visual-language interacting and retrieve the accurate moment.

4.5 Ablation Study

In this section, we conduct the ablation study for the multi-task loss and the concrete design of our model.

4.5.1 Ablation Study for the Multi-Task Loss. We discard one loss from the multi-task loss at a time to generate an ablation model, including **w/o. intra loss**, **w/o. inter loss** and so on. The ablation results are shown in Figure 4. We can find the full model outperforms all ablation models on two datasets, which demonstrates the intra-sample and inter-sample losses can effectively offer the supervision signals, and the regularized global and gap losses can improve the model performance. The model (w/o. inter loss) and model (w/o. intra loss) have close performance, suggesting intra-sample and inter-sample confrontments are equally important for weakly-supervised moment retrieval. Moreover, the model (w/o. global loss) achieves the worst accuracy, which shows filtering out irrelevant moments is crucial to model training.

4.5.2 Ablation Study for the Model Design. We next verify the effectiveness of our model design, including the two-branch architecture, filter designs and center-based proposal method. Note that the cross-modal interesting unit [40] and 2D temporal map [39] are mature techniques that do not need further ablation.

- **Two-Branch Architecture.** We remove the crucial filter and only retain a branch to perform the conventional MIL-based training without the intra-sample loss as **w/o. filter**. We then keep the entire framework but discard the parameter sharing between two branches as **w/o. parameter sharing**.

- **Filter Design.** We discard the cross-modal estimation and generate the score distribution by only frame features as **w/o. visual-only scoring**. And we remove the NetVALD and directly apply the textual features during the cross-modal estimation as **w/o. NetVALD**.
- **Proposal Method.** During moment proposal generation in two branches, we discard the center-based proposal and sample all candidate moments as **all-proposal**. And we replace the center-based proposal method with a top-k proposal method as **top-k proposal**, where we directly select T moments with the high proposal scores.

The ablation results on ActivityCaption and Charades-STA datasets are reported in Table 4 and we can find some interesting points:

- The model (w/o. filter) and model (w/o. parameter sharing) have severe performance degradation than the full model. This fact demonstrates that the two-branch architecture with the language-aware filter can develop the intra-sample confrontment and boost the model performance, and the parameter sharing is crucial to make two branches generate high-quality proposals for sufficient confrontment.
- The full model achieves better results than model (visual-only scoring) and model (w/o. NetVALD). It suggests that the cross-modal estimation with language information can generate a more reasonable score distribution than visual-only scoring. And the NetVALD can further enhance the cross-modal estimation by introducing an intermediately semantic space for texts and videos.
- As for the proposal method, the model with the center-based strategy outperforms the model (all-proposal) and model (top-k proposal), which proves our center-based proposal method can discover a series of correlative moments for MIL-based intra-sample and inter-sample training.
- Actually, some ablation models, e.g. model (visual-only scoring) and model (top-k proposal), still yield better performance than state-of-the-art baselines, validating our RTBPN network is robust and does not depend on a key component.

4.6 Hyper-Parameters Analysis

In our RTBPN model, the number of selected positive/negative proposal number T is an important hyper-parameter. Therefore, we further explore its effect by varying the proposal number. Specifically, we set T to 8, 16, 48, 64, 128 on ActivityCaption and Charades-STA datasets and report the experiment results in Figure 5, where we select "R@1, IoU=0.3" and "R@1, IoU=0.5" as evaluation criteria. We note that the model achieves the best performance on both datasets when the number is set to 48. Because too many proposals will introduce irrelevant moments in the model training and affect the model performance. And too few proposals may miss the crucial moments and fail to develop sufficient confrontment, leading to poor performance. Moreover, the trends of the effect of proposal number T on two datasets are similar, which demonstrates this hyper-parameter is insensitive to different datasets.

4.7 Qualitative Analysis

To qualitatively validate the effectiveness of our RTBPN method, we display two typical examples on ActivityCaption and Charades-STA

Figure 6: Qualitative Examples on the Charades-STA and ActivityCaption datasets

in Figure 6, where we show the score distribution from the language-aware filter, the retrieval results from the enhanced branch and suppressed branch and the result of the SCN baseline.

By intuitive comparison, we find that our RTBPN method can retrieve a more accurate moment from the enhanced branch than SCN, qualitatively verifying the effectiveness of our method. And we can observe that the filter gives higher scores to the language-relevant frames than unnecessary ones. Based on the reasonable score distribution, the enhanced branch can localize the precise moment while the suppressed branch can only retrieve the relevant but not accurate moment as the plausible negative proposal.

5 CONCLUSION

In this paper, we propose a novel regularized two-branch proposal network for weakly-supervised video moment retrieval. We devise a language-aware filter to generate the enhanced and suppressed video streams, and then design the sharable two-branch proposal module to generate positive proposals from the enhanced stream and plausible negative proposals from the suppressed one. Further, we design the proposal regularization to improve the model performance. The extensive experiments show the effectiveness of our RTBPN method.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China under Grant No. 2018AAA0100603, Zhejiang Natural Science Foundation LR19F020006 and the National Natural Science Foundation of China under Grant No.61836002, No.U1611461 and No.61751209. This research is supported by the Fundamental Research Funds for the Central Universities 2020QNA5024.

REFERENCES

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5297–5307.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[3] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the Faster R-CNN Architecture for Temporal Action Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1130–1139.

[4] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally Grounding Natural Sentence in Video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 162–171.

[5] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing Natural Language in Videos. In *Proceedings of the American Association for Artificial Intelligence*.

[6] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the Bottom-Up Framework for Query-based Video Localization. In *Proceedings of the American Association for Artificial Intelligence*.

[7] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. 2020. Look Closer to Ground Better: Weakly-Supervised Temporal Grounding of Sentence in Video. *arXiv preprint arXiv:2001.09308* (2020).

[8] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. 2019. Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video. In *Proceedings of the Conference on the Association for Computational Linguistics*.

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in Neural Information Processing Systems*.

[10] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems*. 3059–3069.

[11] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12, Jul (2011), 2121–2159.

[12] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5277–5285.

[13] Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. 2019. WSLLN: Weakly Supervised Natural Language Localization Networks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2019).

[14] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the American Association for Artificial Intelligence*, Vol. 33. 8393–8400.

[15] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*. 5803–5812.

[16] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing Moments in Video with Temporal Language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1380–1390.

[17] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[18] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network. In *Proceedings of the American Association for Artificial Intelligence*.

[19] Zhijie Lin, Zhou Zhao, Zhu Zhang, Zijian Zhang, and Deng Cai. 2020. Moment Retrieval via Cross-Modal Interaction Networks With Query Reconstruction. *IEEE Transactions on Image Processing* 29 (2020), 3750–3762.

[20] Daochang Liu, Tingting Jiang, and Yizhou Wang. 2019. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1298–1307.

[21] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 15–24.

[22] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 843–851.

[23] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11592–11601.

[24] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. 2018. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6752–6761.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 91–99.

[27] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. 2017. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1417–1426.

[28] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. 2018. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision*. 154–171.

[29] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1049–1058.

[30] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *Proceedings of the European Conference on Computer Vision*.

[31] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489–4497.

[32] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. 2017. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[33] Weining Wang, Yan Huang, and Liang Wang. 2019. Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 334–343.

[34] Huijuan Xu, Kun He, L Sigal, S Sclaroff, and K Saenko. 2019. Multilevel Language and Vision Integration for Text-to-Clip Retrieval. In *Proceedings of the American Association for Artificial Intelligence*, Vol. 2. 7.

[35] Tan Yu, Zhou Ren, Yuncheng Li, Enxu Yan, Ning Xu, and Junsong Yuan. 2019. Temporal structure mining for weakly supervised action detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 5522–5531.

[36] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *Advances in Neural Information Processing Systems*. 534–544.

[37] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2019. Graph Convolutional Networks for Temporal Action Localization. In *Proceedings of the IEEE International Conference on Computer Vision*.

[38] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1247–1257.

[39] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *Proceedings of the American Association for Artificial Intelligence*.

[40] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 655–664.

[41] Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Jing Yuan. 2020. Object-Aware Multi-Branch Relation Networks for Spatio-Temporal Video Grounding. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 1069–1075.

[42] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jingkuan Song, and Deng Cai. 2019. Localizing Unseen Activities in Video via Image Query. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

[43] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10668–10677.

[44] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*.