Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos

Jie Wu
Sun Yat-sen University
wujie10558@gmail.com

Guanbin Li*
Sun Yat-sen University
liguanbin@mail.sysu.edu.cn

Xiaoguang Han
Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong
hanxiaoguang@cuhk.edu.cn

Liang Lin
Sun Yat-sen University
DarkMatter AI
linliang@ieee.org

ABSTRACT

Temporal grounding of natural language in untrimmed videos is a fundamental yet challenging multimedia task facilitating cross-media visual content retrieval. We focus on the weakly supervised setting of this task that merely accesses to coarse video-level language description annotation without temporal boundary, which is more consistent with reality as such weak labels are more readily available in practice. In this paper, we propose a *Boundary Adaptive Refinement* (BAR) framework that resorts to reinforcement learning (RL) to guide the process of progressively refining the temporal boundary. To the best of our knowledge, we offer the first attempt to extend RL to temporal localization task with weak supervision. As it is non-trivial to obtain a straightforward reward function in the absence of pairwise granular boundary-query annotations, a cross-modal alignment evaluator is crafted to measure the alignment degree of segment-query pair to provide tailor-designed rewards. This refinement scheme completely abandons traditional sliding window based solution pattern and contributes to acquiring more efficient, boundary-flexible and content-aware grounding results. Extensive experiments on two public benchmarks Charades-STA and ActivityNet demonstrate that BAR outperforms the state-of-the-art weakly-supervised method and even beats some competitive fully-supervised ones.

KEYWORDS

Temporal grounding of natural language in untrimmed videos, Reinforcement learning, Boundary adaptive refinement

*Corresponding author is Guanbin Li. This work was supported by the State Key Development Program under Grant No. 2016YFB1001004, the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, National Natural Science Foundation of China under Grant No.61976250 and Grant No.61702565, the National High Level Talents Special Support Plan (Ten Thousand Talents Program), the Fundamental Research Funds for the Central Universities under Grant No.18lgpy63, and was also sponsored by CCF-Tencent Open Research Fund.

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

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413862

ACM Reference Format:
Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. 2020. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos. In *Proceedings of ACM (MM'20).* ACM, Seattle, USA, 9 pages. https://doi.org/10.1145/3394171.3413862

1 INTRODUCTION

Temporal grounding of natural language in untrimmed video is a newly-raised and crucial task due to its potential applications in the field of human-robot interaction and cross-media analysis. It aims to locate the temporal segment that is most relevant to the given sentence query in an untrimmed video. Albeit with varying degrees of progress, most of its recent successes [3, 8, 10, 11, 16, 33, 35, 36, 42–44] are involved in a fully supervised setting, i.e., mapping between video interval and the corresponding statement description are available in the training set. It is still arduous to acquire such granular annotations that require a huge amount of manual effort, which becomes a critical bottleneck as this task is pushed toward a larger-scale and more complicated scenario. To alleviate such expensive and unwieldy annotations, [20] proposes to address this task in the weakly supervised setting that learns to infer language-related temporal range from video-level supervision. This weakly supervised paradigm only has access to the video-level language description annotations without their corresponding temporal boundary specification. This is an exceedingly favorable scheme since coarse video-level annotations are more readily available on the internet. In our work, we focus on this weakly supervised paradigm.

Many approaches [3, 8, 10, 16, 20] employ a two-stage "proposal-and-rank" solution pattern to address the task of temporal grounding of natural language. However, these works are indulged in learning more robust cross-modal representations in the rank branch without explicitly considering and modeling boundary-flexible and content-aware proposals. As shown in the left half of Figure 1, "proposal-and-rank" pattern is inherently restrictive as it relies heavily on pre-defined and inflexible sliding windows (e.g., 128 and 256 frames [20]), which results in lacking generalization for videos with considerable variance in length. More rigorously, it raises two additional challenges when it is extended to the weakly supervised setting. First, offset regressive learning [8] for boundary adjustment becomes impractical in the absence of granular annotation. Second, accessing video-query pair during training, the leading model [20] can merely learn cross-modal mappings from the inter-videos,

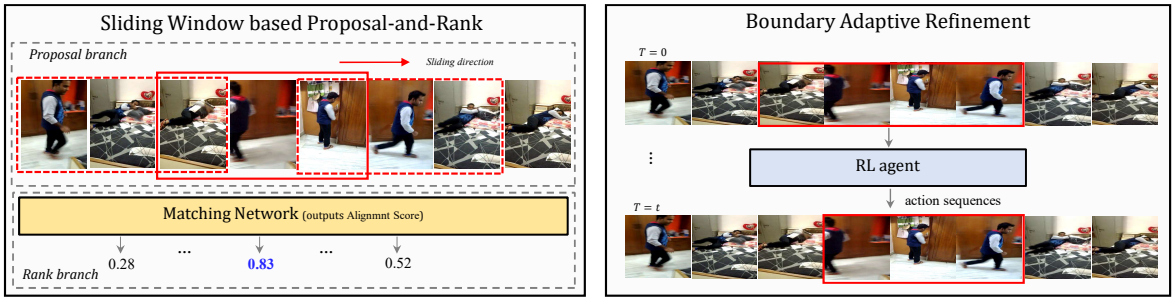


Figure 1: The diagrams of sliding window based proposal-and-rank pattern and the novel boundary adaptive refinement process. The input query in this example is "person goes back to close the door." Traditional pattern is constrained by fixed sliding window templates and has to process extensive candidate segments one by one to localize queries. However, the boundary adaptive refinement manages to flexibly adjust the boundary via a series of actions.

while fails to take into account more subtle and fine-grained semantic concepts within the intra-video. These suboptimal cross-modal mappings generally lead to less accurate boundary prediction.

To better cope with the above issues, as shown in the right half of Figure 1, we formulate the task as a cross-modal matching guided heuristic process, a.k.a, *Boundary Adaptive Refinement (BAR)*. BAR resorts to a tailor-designed reinforcement learning paradigm to adaptively optimize the temporal boundary towards shrinking the cross-modal semantic gap. It is noted that reinforcement learning (RL) has been validated in various tasks of fully supervised video understanding, including video recognition [40] and video referring expression [11]. This work can be regarded as the first attempt to extend RL to weakly supervised temporal localization tasks. Due to the lack of matching supervision for specific video intervals corresponding to the statement,it is non-trivial to design an intensive learning state assessment and reward function which can effectively drive the model to achieve efficient temporal boundary optimization. Our proposed BAR framework hence includes a context-aware feature extractor for encoding the current and contextualized environment state, an adaptive action planner for decision adjustment of direction and interval range, and most typically a cross-modal alignment evaluator for providing an estimate of the alignment score between each segment-query pair in the absence of pairwise supervisory information. This alignment evaluator is crafted to assign a corresponding reward by comparing the alignment score of the consecutive segment-query pair under the guidance of both inter-video and intra-video ranking loss. This modularized component design and heuristic adaptive temporal window adjustment strategy contributes to making the solution pattern more flexible and conforming to the human perception retrieval mechanism; Furthermore, it can be guided and pruned with goal-oriented rewards in a larger search space to extract more accurate temporal window positioning; Moreover, it also attempts to occupy as little time as possible to reach more impressive results.

The contributions of this work are summarized as follows:

• We design a Boundary Adaptive Refinement framework that resorts to reinforcement learning to address the task of weakly supervised temporal grounding of language in video. To the best of our knowledge, we are the first to employ RL to temporal localization task with weak supervision.

• BAR abandons traditional sliding window based proposal-and-rank pattern and employs a novel boundary adaptive refinement process, which contributes to acquiring more efficient, boundary-flexible and content-aware grounding results.

• Experimental results on two benchmark datasets Charades-STA [8] and ActivityNet [12] demonstrate that BAR outperforms the existing state-of-the-art weakly-supervised methods, and even beats some competitive fully-supervised ones.

2 RELATED WORK

Temporal Grounding of Natural Language in Video. Temporal grounding of natural language aims to determine the start and end time of a temporal segment in an untrimmed video that corresponds to a language query. It is a temporal extension of image referring expression comprehension[37–39], and is also a challenging multimedia task which requires cross-modal fusion and fine-grained interactions between the verbal and visual modalities. Many approaches [3, 8, 10, 16, 20] employ a two-stage "proposal-and-rank" manner, which first generates temporal proposals and then selects the one with the highest confidence score. However, these approaches rely on external sliding windows matching and ranking, leading to boundary-inflexible and time consuming. To formulate a computationally efficient framework, Chen *et al.* [2] designed an end-to-end deep neural network that merely performs a single pass to obtain the grounding result. Xu *et al.* [36] proposed a multi-level model to integrate visual-query feature in the earlier stage and further introduced the caption generation as an auxiliary task.

Weakly Supervised Learning. Weakly-supervised learning is a research setup that aims at optimizing a model without substantial manual labeled information. Many computer vision and multimodal tasks such as salient object detection [14], captioning [6], language grounding [20, 22], referring expression grounding[18] have explored the weakly-supervised setup, since granular annotations are much more source-consuming compared to coarse annotations. Wang *et al.* [31] proposed a weakly supervised collaborative learning framework to resolve the task of weakly supervised object detection, which only requires image-level labels. In the video domain, Duan *et al.* [6] formulated a new task: weakly supervised dense event captioning. The goal of this task is to detect and describe all events of interest contained in a video without dense

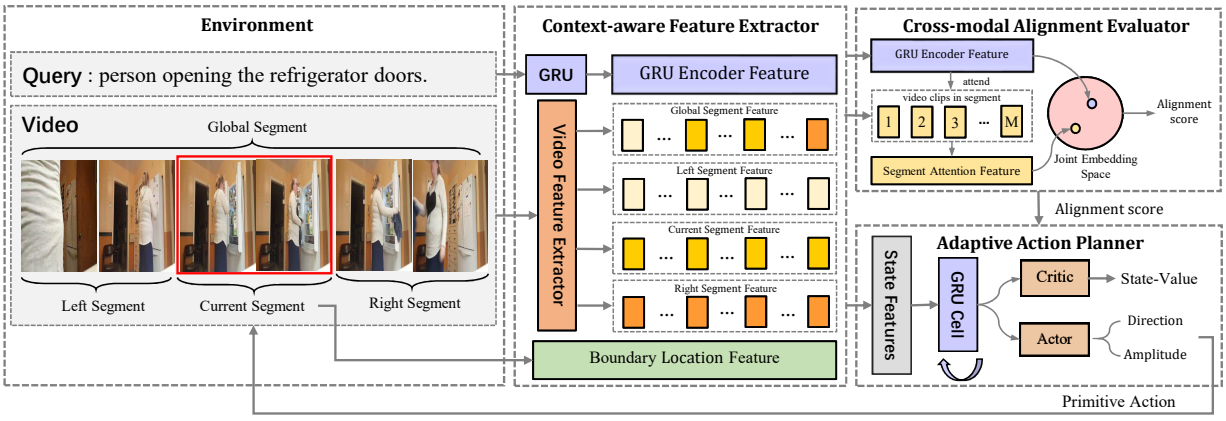


Figure 2: The overall architecture of the Boundary Adaptive Refinement (BAR) framework, which consists of a context-aware feature extractor, an adaptive action planner and a cross-modal alignment evaluator.

segment annotations for model training. The work that closely related to ours is [20]. Mithun *et al.* [20] designed Text-Guided Attention (TGA) mechanism to leverage latent alignment between video frames and sentence descriptions to address the same task as us.

Reinforcement Learning. Reinforcement learning (RL) is originated from the neuroscientific and psychological understandings of how humans learn to optimize their behaviors in an environment. It can be mathematically formulated as a Markov Decision Process (MDP) in a sequential decision-making manner. Recently, RL technique [34] has been utilized to imitate human's thinking pattern to address various tasks, which generally can be formulated as a MDP that executes a series of actions to accomplish the task-specific objective[4, 11, 24, 26, 41]. Ranzato *et al.* [24] used the REINFORCE algorithm to train the captioning model in sequence level by optimizing the non-differentiable metric directly. Yeung *et al.* [40] adopted REINFORCE algorithm to optimize an end-to-end approach for reasoning the temporal bounds of action and its category. He *et al.* [11] resorted to RL based method to address the fully-supervised version of the studied task, which utilized the temporal IoU as reward indicator. Our work offers the first attempt to extend RL to accomplish the proposed task with weak supervision. To estimate an accurate reward function in the absence of pairwise supervisory information, a cross-modal alignment evaluator is crafted to provide tailor-designed rewards.

3 METHODOLOGY

3.1 Problem Formulation

Following the common-used formulation [8], we represent a video V by N clips $\{V_1, V_2, ..., V_N\}$, each clip corresponds to a small chunk of sequential frames. Taking V and a text query T as inputs, the studied task aims to output a video segment $[j, k]$ (j and k indicate the start and end clip indices respectively) that semantically matches the query description. Our work focuses on the weakly supervised setting of this task. Specifically, only a set of V-T pairs are provided but the video segment annotation for each pair is not available. Inspired by the observation that humans usually locate interest events during a long video with a heuristic search strategy, we propose to formulate this task as a Markov Decision Process. A Boundary Adaptive Refinement (BAR) framework is thus designed: starting from an initial segment, the reinforcement learning technique is utilized to refine its temporal boundary progressively. The overall architecture of the proposed BAR framework is depicted in Figure 2. As illustrated, this modular framework employs a context-aware feature extractor to encode the environment state into cross-modal contextual concepts. The cross-modal alignment evaluator is crafted to provide a tailor-designed reward and the termination signal for the iterative refinement process. An adaptive action planner is designed to reason the direction and amplitude of the action from contextualized observation adaptively, instead of shifting a fixed amplitude every step [11]. The details of these modularized components will be described in the following sections.

3.2 Context-aware Feature Extractor

The context-aware feature extractor takes a video-query pair (V-T) from the external environment and encodes it into the context-aware cross-modal concepts. Each word in query T is firstly encoded using GloVe [23] embeddings and then fed into the GRU [5] to capture long-range dependencies. The summarized query representation $\mathbf{E}$ is obtained from the last hidden state of the GRU. A pre-trained video feature extractor (C3D [30] or TSN [32]) is used to extract the clip-level feature for each video clip. A video segment is represented as a set of clip features, i.e., $\mathbf{F} = \{\mathbf{F_1}; ...; \mathbf{F_i}; ...; \mathbf{F_M}\} \in \mathbb{R}^{d_k \times M}$. $\mathbf{F_i} \in \mathbb{R}^{d_k}$ denotes the clip-level feature for the video clip V_i and M is the number of clips in the corresponding video segment. At each time step, the updated boundary divides the whole video into three parts: left segment, current segment and right segment. And we collect all clip-level features within the corresponding boundary into a set to obtain the three corresponding segment-level features. Rather than directly taking the current segment's feature as independent inputs [11], this extractor also leverages context-aware contextual cues derived from other segments in the video (i.e., left and right segment feature) for state encoding. Furthermore, the extractor explicitly involves the normalized boundary location $\mathbf{L_{t-1}}$ into the encoded features to provide some notion of

relative position:

$$\mathbf{L_{t-1}} = [\frac{l^s_{t-1}}{N}, \frac{l^e_{t-1}}{N}] \tag{1}$$

where l^s_{t-1} and l^e_{t-1} denote the start and end clip indices of the boundary, respectively. $t = \{1, \cdots, T_{max}\}$ and T_{max} indicates the maximum iterations in the refinement process.

3.3 Cross-modal Alignment Evaluator

The cross-modal alignment evaluator is designed specially to address two critical issues in our RL-based approach. On the one hand, this evaluator is crafted to assign a target-oriented reward to address the difficulty that the adaptive action planner can not directly obtain a reliable reward function without granular boundary annotations. On the other hand, this alignment evaluator manages to determine an accurate stop signal to terminate the refinement process. Given a video segment, the dimension of each clip feature is reduced to the same as the summarized query representation $\mathbf{E}$ via a filter function θ, which consists of a fully-connected layer followed by ReLU [13] and Dropout [28] function. $\mathbf{E}$ is taken to create a temporal attention over all video clips, which manages to emphasize crucial video clips and weaken inessential parts. Concretely, the scaled dot-product attention mechanism [19] is utilized to obtain attention weight a_i and the segment attention feature $\mathbf{A}$:

$$a_i = softmax(\frac{\mathbf{E} \odot \theta(\mathbf{F_i})}{\sqrt{k}}), \quad \mathbf{A} = \sum_{i=1}^{M} a_i \theta(\mathbf{F_i}) \tag{2}$$

where $\odot$ indicates the dot product operation between two vectors. k is the dimension of $\mathbf{E}$. Then the segment attention feature and query representation are mapped to a joint embedding space to compute the alignment score S:

$$S = \text{L2Norm}(\mathbf{A}) \odot \text{L2Norm}(\mathbf{E}) \tag{3}$$

The alignment score can be regarded as a reward estimate to provide reliable reward. Specifically, the evaluator measures the alignment score of the consecutive segment-query pairs, and assigns the corresponding reward r_t:

$$r_t = \text{sign}(S^c_t - S^c_{t-1}) \tag{4}$$

where S^c_t denotes the alignment score of the current segment and sentence query at time step t. This reward function returns +1 or -1. Basically, if the next boundary has a higher alignment score than the current one, the reward r_t of the action a_t moving from the current window to the next one is +1, and -1 otherwise. Such binary rewards reflect more clearly which action can drive the boundary towards the ground-truth and thus facilitate the agent's learning.

3.4 Adaptive Action Planner

The adaptive action planner is designed to infer action sequences to refine the temporal boundary. To get a fixed-length visual representation, we utilize a mean pooling layer over feature set $\mathbf{F}$ of the global, current, left and right segment, obtaining the pooling features $\mathbf{f^g}, \mathbf{f^c_{t-1}}, \mathbf{f^l_{t-1}}, \mathbf{f^r_{t-1}}$ respectively. Then the cross-gated interaction method [7] is further adopted to enhance the effects of the relevant segment-query pairs. Concretely, the current pooling feature $\mathbf{f^c_{t-1}}$ is gated by query representation $\mathbf{E}$, and meanwhile the gate of $\mathbf{E}$ depends on $\mathbf{f^c_{t-1}}$:

$$\tilde{\mathbf{f^c_{t-1}}} = \sigma(\mathbf{W^s E}) \odot \mathbf{f^c_{t-1}}, \quad \tilde{\mathbf{E}} = \sigma(\mathbf{W^v f^c_{t-1}}) \odot \mathbf{E} \tag{5}$$

where $\mathbf{W^s}$ and $\mathbf{W^v}$ are parameter matrices and σ denotes the sigmoid function. These cross-modal features are then concatenated and fed into two cascaded fully-connected layers ϕ to get the state activation representation s_t:

$$s_t = \phi(\tilde{\mathbf{E}}, \tilde{\mathbf{f^c_{t-1}}}, \mathbf{f^g}, \mathbf{f^l_{t-1}}, \mathbf{f^r_{t-1}}, \mathbf{L_{t-1}}). \tag{6}$$

Such contextual features encourage the planner to perform a left-right tradeoff on the video contents and infer a more accurate action. s_t is further fed into a GRU cell to enable the agent to incorporate the memory information about the video segments that have been explored. Then the output state of the GRU is followed by two separate fully-connected layers (i.e., actor and critic) to respectively estimate a policy function $\pi(a_t|s_t)$ and a value approximator $v^\pi(s_t)$. A primitive action $a_t \in \mathcal{A}$ is sampled from the policy function $\pi(a_t|s_t)$ in the training procedure. In our work, the action space $\mathcal{A}$ is composed of four primitive actions: shifting the start/end point backward/forward N/ν clips. ν is an amplitude factor that empirically sets as:

$$\nu = \lfloor 10 \times (1 + 2 \times \tanh(S^c_t - S^g)) \rfloor_+, \tag{7}$$

where $\lfloor\rfloor_+$ denotes the lower bound of a positive integer. S^g and S^c_t denotes the global and current alignment score estimated by the alignment evaluator. tanh is used to constrain the action amplitude to fluctuate around $N/10$ (an empirical number used in [11]). S^g plays as a baseline of the alignment degree to determine ν: when S^c_t is lower, ν becomes smaller and the agent markedly shifts the boundary; when S^c_t becomes higher, ν is larger and the boundary is marginally refined. This adaptive setting enables the agent to determine the action amplitude based on the current observation, which is also in line with human habits.

The state-value $v^\pi(s_t)$ predicted by the critic is the value estimation of the current state. Under the assumption that the critic produces the exact values, the actor is trained based on an unbiased estimation of the gradient.

3.5 Training

Due to its efficiency, the advantage actor-critic (A2C) [29] algorithm is chosen to train our adaptive action planner. Multiple instance learning algorithm with a combined ranking loss $\mathcal{L}_{rank}$ is designed to train the cross-modal alignment evaluator and context-aware feature extractor. The total loss in BAR is summarized as:

$$\mathcal{L} = \mathcal{L}_{A2C} + \eta \mathcal{L}_{rank}, \tag{8}$$

where $\mathcal{L}_{A2C}$ denotes the loss function in the A2C algorithm. η is a trade-off factor between the two losses.

A2C Loss. The adaptive action planner runs T_{max} steps for adjustment during training. Given a trajectory in an episode $\Gamma = \langle s_t, \pi(\cdot|s_t), v^\pi(s_t), a_t, r_t \rangle$, the loss function of the actor $\mathcal{L}_{actor}$ is formulated as:

$$\mathcal{L}_{actor} = -\sum_{t=1}^{T_{max}} [A^\pi(s_t, a_t) \log \pi(a_t|s_t) + \alpha H(\pi(a_t|s_t))], \tag{9}$$

where $A^\pi(s_t, a_t)$ denotes the advantage function and the entropy $H()$ of the policy is introduced into the objective for improving

exploration. $A^{\pi}(s_t, a_t) = Q^{\pi}(s_t, a_t) - v^{\pi}(s_t)$ measures whether or not and how much the action is better than the policy's default behaviour. Temporal-difference (TD) learning is adopted to estimate the Q-value function $Q^{\pi}(s_t, a_t)$ by k-step returns with function approximation:

$$Q^{\pi}(s_t, a_t) = \sum_{l=0}^{k-1} \gamma^l r_{t+l} + \gamma^k v^{\pi}(s_{t+k}) \tag{10}$$

where γ is a constant discount factor. It is noted that the BAR does not suffer from sparse reward issue during training since the reward can be obtained at every step. To optimize the critic, we minimize the mean squared error (MSE) loss $\mathcal{L}_{critic}$ between the Q-value function and the estimated value [21]. And the total A2C loss is a combination of the losses from the actor branch and the critic branch: $\mathcal{L}_{A2C} = \mathcal{L}_{actor} + \mathcal{L}_{critic}$.

Ranking Loss. In general, the content discrepancy between the inter-videos is higher than that within the intra-video. Hence we resort to multiple instance learning algorithm and first leverage coarse-level semantic concepts from the inter-videos to optimize the framework. Concretely, given the global video feature F^g and its query representation E , it is expected that the alignment score $S(F^g, E)$ (positive pair) is higher than the score $S(F^{g\prime}, E)$ / $S(F^g, E')$ (negative pairs) for any video $F^{g\prime}$ / query $\tilde{E}$ taken from other sample pairs. The inter-video ranking loss [25] is thus defined as:

$$\begin{aligned}\mathcal{L}_{inter} = &\sum_{E'} [\epsilon + S(F^g, E') - S(F^g, E)]_+ \\ &+ \sum_{F^{g\prime}} [\epsilon + S(F^{g\prime}, E) - S(F^g, E)]_+,\end{aligned} \tag{11}$$

where $[x]_+$ denotes a ramp function defined by $\max(0, x)$ and ϵ indicates a margin. $S(F^g, E)$ and S^g are equivalent. The positive and negative pairs are obtained from the same mini-batch.

Inter-videos generally include substantially broad semantic abstractions that are hard to distinguish similar contents in a specific video. To this end, we design the intra-video ranking loss $\mathcal{L}_{intra}$ to capture more subtle concepts in the intra-video to further optimize the network. Expressly, if the score of any one of left, current and right segment-query pairs surpasses the global one during the refinement process, we assume this pair should have higher alignment score than the other two pairs:

$$\begin{aligned}\mathcal{L}_{intra} = &\psi(S_t^c > S^g) \times ([\epsilon + S_t^l - S_t^c]_+ + [\epsilon + S_t^r - S_t^c]_+) \\ &+ \psi(S_t^l > S^g) \times ([\epsilon + S_t^c - S_t^l]_+ + [\epsilon + S_t^r - S_t^l]_+) \\ &+ \psi(S_t^r > S^g) \times ([\epsilon + S_t^c - S_t^r]_+ + [\epsilon + S_t^l - S_t^r]_+),\end{aligned} \tag{12}$$

where S_t^l and S_t^r are the alignment scores of the left segment-query pair and the right segment-query pair at the time step t, respectively. $\psi()$ is a binary indicator function. If the inequality in parentheses holds, $\psi()$ will output 1, otherwise 0. Specifically, when the score of a segment-query pair, say S_t^c, surpasses S_g, the optimization target is to increase the gap between S_t^c and the other two (S_t^l and S_t^r) by increasing S_t^c or decreasing S_t^l and S_t^r . Noted that by lowering S_t^c below S_g might be another option, but this usually becomes increasingly impractical with the progress of inter-video training. In addition, when there exist more than one segment-query pairs of score surpass S_g, the optimization target of $\mathcal{L}_{intra}$ will usually guide the alignment evaluator to suppress the score of the sub-optimal matching pair(s) to be lower than S_g and at the same time drive the action planner to adjust the boundary. Intuitively, $\mathcal{L}_{intra}$ encourages the text query to be closer to a semantically matched video moment than other possible moments from the same video, which contributes to obtaining a content-aware alignment score.

$\mathcal{L}_{intra}$ manages to i) widen the score gap between matched and unmatched segment-query pair to increase the confidence of the alignment evaluation; ii) improve the reward calculation by affecting the alignment evaluator to drive the action planner to achieve better temporal boundary adjustments. To sum up, the combined ranking loss $\mathcal{L}_{rank}$ is defined as:

$$\mathcal{L}_{rank} = \mathcal{L}_{inter} + \lambda \sum_{t=1}^{T_{max}} \mathcal{L}_{intra}, \tag{13}$$

where λ is a weighting parameter to achieve a ranking loss trade-off between the intra-video and the inter-video. In the early stage of this collaborative training scheme, it is very unlikely that the score of a segment-query pair exceeds S_g and $\mathcal{L}_{intra}$ tends to 0, hence $\mathcal{L}_{inter}$ plays a dominant role that learns to transfer the matching between video-query pair to segment-query pair. As the training progresses, $\mathcal{L}_{inter}$ converge gradually and it is more common for the score of segment-query pair to exceed S_g, $\mathcal{L}_{intra}$ begins to play a critical role.

Alternating Update. BAR is trained from scratch and an alternating update strategy is applied to facilitate stable training. Specifically, for each set of $2K$ iterations, we first fix the parameters of the action planner and employ $\mathcal{L}_{rank}$ for model optimization. This setting guarantees a trustworthy initial reward for the action planner. When K iterations are reached, we fix the parameters of the alignment evaluator and feature extractor, and switch $\mathcal{L}_{rank}$ to $\mathcal{L}A2C$ to optimize the action planner for K more iterations. This alternating update mechanism repeats until the model converges.

3.6 Inference

At each time step, BAR executes an action $\hat{a}_t$ via greedy decoding algorithm to adaptively adjust the temporal boundary. And the cross-modal alignment evaluator computes a score S_t^c to provide confidence for alignment degree and termination. Empirically, the final grounding result corresponding to the query usually occupies a reasonable and appropriate video length. Hence to penalize the video segment with abnormal lengths, we propose to update the confidence score with a Gaussian penalty function as follows:

$$P_t = \frac{l_t^e - l_t^s}{N} - \delta, \quad \hat{S_t^c} = S_t^c e^{-\frac{P_t^2}{\tau}} \tag{14}$$

where δ denotes the penalty factor corresponds to abnormal lengths. τ is a modulating factor that as τ increases the effect of the penalty degree is likewise decreased. The segment with the max $\hat{S_t^c}$ during testing is regarded as the final grounding result.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Datasets. We conduct extensive experiments on two benchmark datasets: Charades-STA [8] and ActivityNet [12]. Charades-STA is

Table 1: The performance comparison (in %) of the state-of-the-art methods in fully supervised and weakly supervised setting. "-" indicates that the corresponding values are not available.

			Charades-STA [8]			ActivityNet [12]	
Supervision	Feature	Baseline	tIoU@0.7	tIoU@0.5	tIoU@0.3	tIoU@0.5	tIoU@0.3
Full Supervision	C3D	ROLE [17], ACM MM 2018	-	12.12	25.26	-	-
		MCN [1], ICCV 2017	4.44	13.66	28.99	10.17	22.07
		CTRL [8], ICCV 2017	8.89	23.63	-	14.36	29.10
		ACRN [16], SIGIR 2018	9.65	26.74	47.64	16.53	31.75
		MAC [10], WACV 2019	12.23	29.39	53.34	-	-
		SAP [3], AAAI 2019	13.36	27.42	-	-	-
		QSPN [36], AAAI 2019	15.80	35.60	54.7	27.70	45.30
		ABLR [42], AAAI 2019	-	-	-	36.79	55.67
		SM-RL [33], CVPR 2019	11.17	24.36	-	-	-
		RWM [11], AAAI 2019	13.74	34.12	55.16	34.91	53.00
	TSN	RWM [11], AAAI 2019	17.72	37.23	61.73	37.46	57.29
	I3D	MAN [43], CVPR 2019	22.72	46.53	-	-	-
Weak Supervision	C3D	TGA [20], CVPR 2019	8.84	19.94	32.14	-	-
		WS-DEC [6], NIPS 2018	-	-	-	23.34	41.98
		WSLLN [9], EMNLP 2019	-	-	-	22.70	42.80
		SCN [15], AAAI 2020	9.97	23.58	42.96	29.22	47.23
		BAR (our)	**12.23**	**27.04**	**44.97**	**30.73**	**49.03**
	TSN	BAR (our)	**15.97**	**33.98**	**51.64**	**33.12**	**53.41**

extended from the Charades dataset [27] with generated sentence-clip annotations, which comprises a series of sentence-clip pairs with 12,408 for training and 3,720 for testing. The average length of each video in this dataset is 29.8 seconds and the described clips are 8 seconds long in average. ActivityNet dataset [12] is introduced to validate the robustness of the proposed model with longer and more diverse videos. It contains 37,421 and 17,505 video-sentence pairs for training and testing. The average duration of the videos is 2 minutes and the described temporally annotated clips are 36 seconds long on average.

Evaluation Metrics. We adopt "tIoU@ χ" to evaluate the grounding result. "tIoU @ χ" means the percentage of the queries that have temporal IoU larger than threshold χ.

4.2 Implementation Details

We leverage C3D and the TSN model to encode video representation. The initial boundary is set to $L_0 = [N/4; 3N/4]$. $N/4$ and $3N/4$ denote the start and end clip indices of the boundary respectively. T_{max} is set to 12 and the size of the hidden state in GRU is 1024. The batch size is 12 and the total loss is optimized via the Adam optimizer with the learning rate of 0.001. The margin ϵ in ranking loss is 0.2. The hyper-parameters α and γ is fixed to 0.1 and 0.4, receptively. The factor η and λ are empirically set to 1 and 0.1. The modulating factor τ is set to 0.5 by cross validation. And penalty baseline factor δ is fixed to 0.35 and 1.0 receptively on Charades-STA and ActivityNet. We use $K = 500$ in the alternating update procedure.

4.3 Comparison with the State-of-the-art

We compare the proposed BAR with several state-of-the-art models based on the weakly-supervised and fully-supervised settings in Table 1. On the one hand, BAR significantly outperforms the weakly-supervised method and establishes new state-of-the-art performance on both datasets. Employing the C3D based video feature, BAR boosts the tIoU@0.5 to 27.04% and 30.73%, with an improvement of 3.46%, 1.51% compared with SCN [15] on the two datasets, receptively. Furthermore, it manages to achieve 33.98% (33.12%) in tIoU@0.5 via more powerful TSN feature. It reveals that our approach helps to better obtain accurate video segments. On the other hand, BAR even achieves better or comparable results than some fully-supervised methods. For instance, BAR outperforms QSPN [36] by 3.03% w.r.t tIoU@0.5 on the ActivityNet dataset. This is an inspiring result as it reveals that our model can get impressive results via learning from massive coarse video-level annotations, which is of great benefit to practical application.

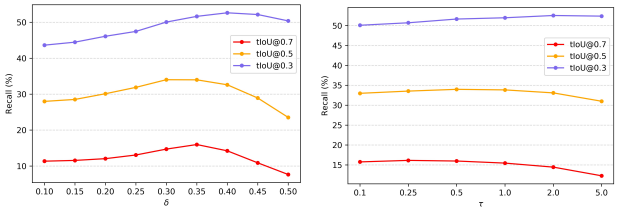
4.4 Ablation Studies

We perform extensive ablation studies and demonstrate the effectiveness of several essential components in BAR. The experiments are conducted on the Charades-STA with the TSN feature. The results are reported in Table 2.

• **Effectiveness of Reinforcement Learning.** More accurate measurement of the factual RL contribution is to directly remove it and use the generated proposals of an off-the-shelf weakly-supervised action localization method [20]. Hence we design a variant (abbreviated as "Ours w/o RL") to follow the above setting. We can observe that removing RL from BAR will lead to a noticeable drop in performance. For example, tIoU@0.5 declines from 33.98% to 25.89%. It reveals that the introduction of RL is fundamental and can bring more flexible and adaptable temporal proposals, this alone is an advantage that cannot be achieved with traditional two-stage frameworks, not to mention its high efficiency.

Table 2: Performance of ablation models.

Metrics	tIoU@0.7	tIoU@0.5	tIoU@0.3
Ours w/o RL	12.37	25.89	45.36
Ours w/ random reward	5.76	8.97	28.82
Initial boundary [N/3; 2N/3]	15.72	33.36	51.33
Initial boundary [N/5; 4N/5]	15.83	33.47	51.20
Ours w/ N/5 amplitude	13.60	31.88	49.65
Ours w/ N/10 amplitude	14.27	32.02	50.25
Ours w/ N/15 amplitude	13.73	31.66	49.29
Ours w/o context	13.62	31.45	49.22
Ours w/o $\mathcal{L}_{intra}$	14.24	30.73	46.82
Ours w/ stop	10.13	24.38	43.22
Ours w/o penalty	13.78	30.97	50.27
Ours	**15.97**	**33.98**	**51.64**



(a) Varying $\delta \in [0.1,0.5]$, τ=0.5 (b) Varying $\tau \in [0.1,5.0]$, δ=0.35

Figure 3: The performance curve of with varying hyper-parameters δ and τ. Best viewed in color.

• **Effectiveness of Tailor-designed Reward.** In order to validate that a target-oriented reward is essential for this task, we design a baseline (abbreviated as "Ours w/ random reward") that samples a random scalar value from the uniform distribution of [-1,1] as the reward for optimization. Table 2 shows that this baseline obtains an exceedingly inferior result, which is approximate to a stochastic one. It indicates that a tailor-designed reward is definitely necessary for the RL setting.

• **Effectiveness of Boundary Initialization.** The initial boundary in this paper is fixed to $L_0 = [N/4; 3N/4]$. To compare different boundary initializations, we design two baselines (denoted as "Initial boundary [N/3; 2N/3]" and "Initial boundary [N/5; 4N/5]") that sets the initial boundary as $[N/3; 2N/3]$ and $[N/5; 4N/5]$, respectively. As reported in Table 2, different boundary initialization is not sensitive to the performance of the algorithm, and all can obtain competitive experimental results, which reflects the robustness of BAR.

• **Effectiveness of Adaptive Setting.** Rather than shifting a fixed distance for each action, BAR can adaptively adjust its action amplitude according to the current state. To demonstrate the superiority of this adaptive setting, we design three variants (named as "Ours w/ N/5 amplitude", "Ours w/ N/10 amplitude" and "Ours w/ N/15 amplitude") that the agent shifts $N/5$, $N/10$ and $N/15$ clips at each step, respectively. As summarized in Table 2, "Ours w/ N/10 amplitude" when set with fixed adjustment strategy. However, our approach with the adaptive setting manages to achieve more impressive performance, which reveals that this adaptive setting is more flexible and effective in our proposed framework.

• **Effectiveness of Context Information.** BAR additionally builds contextualized video representations for action decisions. To investigate the effectiveness of the context information, we design a baseline that removes the context concepts (f^l_{t-1}, f^r_{t-1}) from s_t in Equation 6, abbreviated as "Ours w/o context". From Table 2, we can see that although the model without context representation can still achieve promising results, our model with context involved gains 2.35% and 2.53% improvement w.r.t tIoU@0.7 and tIoU@0.5 respectively, which demonstrates that contextual concepts helps for obtaining more content-aware results.

• **Effectiveness of Intra-video Ranking Loss.** To verify the effectiveness of the $\mathcal{L}_{intra}$, we construct a comparison variant that merely uses $\mathcal{L}_{inter}$ to optimize the evaluator, named as "Ours w/o $\mathcal{L}_{intra}$". Table 2 reveals that the grounding result suffers from an obvious drop without $\mathcal{L}_{intra}$. For example, tIoU@0.3 declines from 51.64% to 46.82%. Our approach with intra-video ranking loss manages to achieve more precise alignment scores and more accurate grounding results. To further demonstrate the effectiveness of the alignment score S obtained by our model, we additionally calculate the correlation coefficient (CC) between S and ground-truth IoU. It shows that CC can reach 0.79, which reveals the obtained S is reliable enough to correctly reflect the matching degree and infer the target-oriented rewards.

• **Analysis of Stop Signal.** We did not include an ending signal in the action space [11] as there is no absolutely reliable and stable internal segment-query matching that can help to effectively terminate the iteration. We further introduce an alignment threshold as a stopping signal (abbreviated as"Ours w/ stop"), which led to inferior results. In order to validate the significance of the length penalty strategy, we design a baseline that directly takes the score S^c_t to determine the termination time, denoted as "Ours w/o penalty". The results indicate that this baseline suffers from performance degradation. It may be due to the fact that "Ours w/o penalty" tends to provide an excessive score when the length of the video segment is too long or too short.

Figure 3 depicts the performance curves with varying δ or τ respectively in the procedure of cross-validation. We can see that a factor δ with too large or too small value will lead to obvious performance decline, which reveals that a video with suitable length is more likely to produce impressive results. A similar changing trend can be observed with varying τ. It demonstrates that an appropriate gaussian penalty encourages the model to perform better. we empirically observed that δ=0.35 and τ=0.5 contribute to obtaining the most promising performance in different levels of tIoU.

4.5 Efficiency

To further investigate the efficiency of this boundary adaptive refinement process, we compare BAR with TGA [20] in terms of average running time and number of candidate segments. As summarized in Table 3, BAR reduces the localization time and candidate boundaries by a sizeable margin. Please notice that the boundary in BAR is equivalent to the temporal proposal number to some extent, but the "boundary" here is more flexible and adaptable. BAR merely needs to refine an initial temporal boundary progressively, which manages to avoid redundant computations and employ a

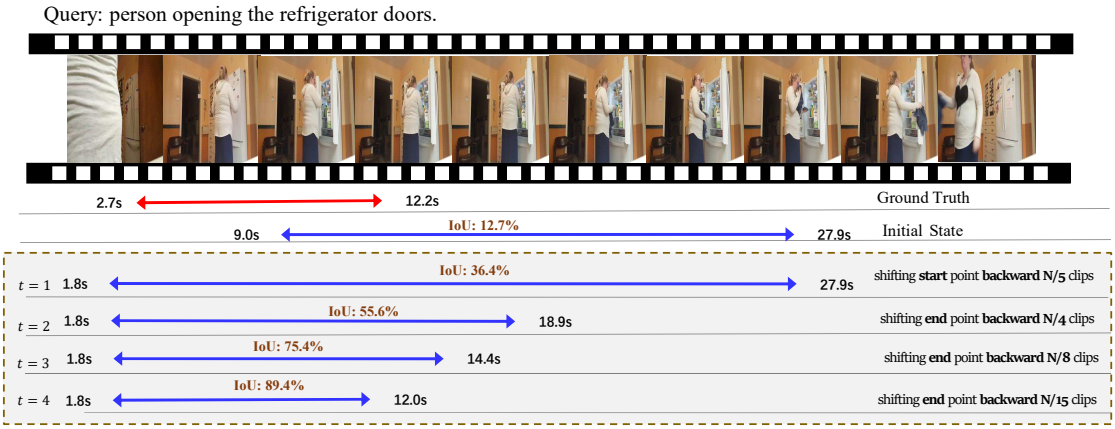


Figure 4: An illustration of how the proposed BAR framework accomplishes the task on Charades-STA.

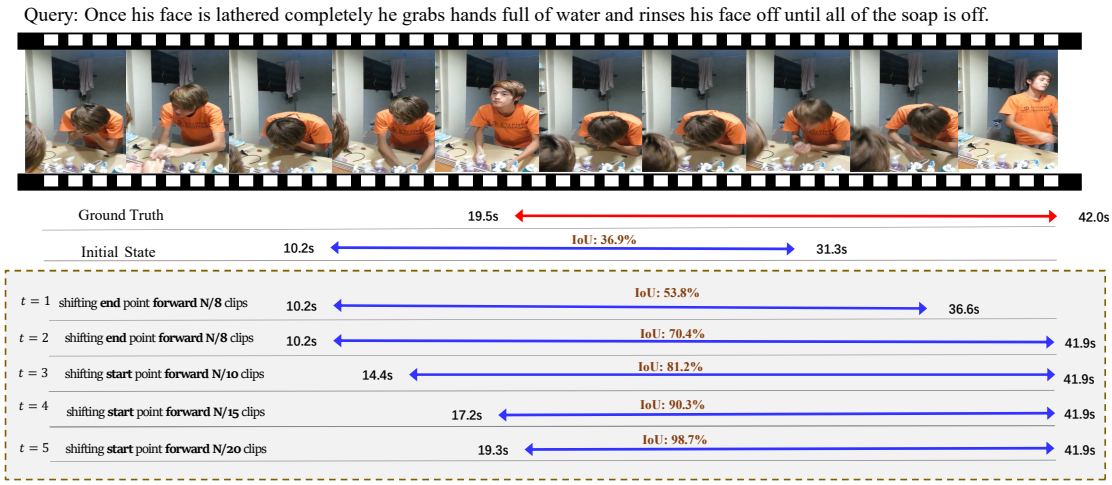


Figure 5: An illustration of how the proposed BAR framework accomplishes the task on ActivityNet.

time-efficient and space-efficient manner. Based on the above discussion, we can conclude that BAR is better than the previous competitive methods in both accuracy and efficiency.

4.6 Qualitative Visualizations

We illustrate two qualitative results in Figure 4, 5 to show the whole process of how BAR obtains the described event location. We observe that our algorithm mainly performs optimization from coarse to fine. The agent will choose a larger movement adjustment at the initial stage of the iteration to quickly narrow the semantic difference between language and vision, and as the iteration progresses, the adjustment range of the movement will change rapidly to achieve local fine-tuning, this is also more consistent with humans performing cross-modal target retrieval.

5 CONCLUSIONS

We propose a Boundary Adaptive Refinement framework that resorts to reinforcement learning to address the task of weakly-supervised temporal grounding of natural language in videos. This refinement scheme completely abandons traditional sliding window-based solution patterns and contributes to obtaining more efficient, boundary-flexible and content-aware grounding results. Extensive experiments show that our approach establishes new state-of-the-art performance on the widely used Charades-STA and ActivityNet datasets. Furthermore, our method even achieves a better result than some competitive fully-supervised methods.

Table 3: The average running time and number of candidate proposals to localize a moment in a video on Charades-STA.

Methods	Time(s)	Candidate Proposal Number
TGA [20]	0.104	65.11
BAR (Ours)	**0.068**	**1**

REFERENCES

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

[2] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 162–171.

[3] Shaoxiang Chen and Yu-Gang Jiang. 2019. Semantic Proposal for Activity Localization in Videos via Sentence Query. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[4] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. 2018. Recurrent attentional reinforcement learning for multi-label image recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[6] 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.

[7] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, and Jiebo Luo. 2018. Video re-localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 51–66.

[8] 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*. 5267–5275.

[9] Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. 2019. WSLLN: Weakly Supervised Natural Language Localization Networks. (2019).

[10] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. MAC: Mining Activity Concepts for Language-based Temporal Localization. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 245–253.

[11] 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 AAAI Conference on Artificial Intelligence*.

[12] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 706–715.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[14] Guanbin Li, Yuan Xie, and Liang Lin. 2018. Weakly supervised salient object detection using image labels. In *Thirty-second AAAI conference on artificial intelligence*.

[15] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network. (2020).

[16] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 15–24.

[17] 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 26th ACM international conference on Multimedia*. 843–851.

[18] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. 2019. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the 27th ACM International Conference on Multimedia*. 539–547.

[19] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

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

[21] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. 1928–1937.

[22] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. 2018. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision*. 563–579.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*. 1532–1543.

[24] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732* (2015).

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[26] Yukai Shi, Guanbin Li, Qingxing Cao, Keze Wang, and Liang Lin. 2019. Face hallucination by attentive sequence optimization with reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[27] 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 *European Conference on Computer Vision*. Springer, 510–526.

[28] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[29] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[30] 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.

[31] Jiajie Wang, Jiangchao Yao, Ya Zhang, and Rui Zhang. 2018. Collaborative learning for weakly supervised object detection. *arXiv preprint arXiv:1802.03531* (2018).

[32] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[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] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*. Springer, 5–32.

[35] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. 2020. Tree-Structured Policy based Progressive Reinforcement Learning for Temporally Language Grounding in Video. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[36] 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 AAAI Conference on Artificial Intelligence*, Vol. 2. 7.

[37] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4145–4154.

[38] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE International Conference on Computer Vision*. 4644–4653.

[39] Sibei Yang, Guanbin Li, and Yizhou Yu. 2020. Graph-Structured Referring Expression Reasoning in The Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9952–9961.

[40] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. 2016. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2678–2687.

[41] Tong Yu, Yilin Shen, Ruiyi Zhang, Xiangyu Zeng, and Hongxia Jin. 2019. Vision-language recommendation via attribute augmented multimodal reinforcement learning. In *Proceedings of the 27th ACM International Conference on Multimedia*. 39–47.

[42] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[43] 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.

[44] Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019. Exploiting Temporal Relationships in Video Moment Localization with Natural Language. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1230–1238.