Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization

Da Cao
Hunan University
caoda0721@gmail.com

Yawen Zeng
Hunan University
yawenzeng11@gmail.com

Xiaochi Wei
Baidu Inc.
weixiaochi@baidu.com

Liqiang Nie
Shandong University
nieliqiang@gmail.com

Richang Hong
Hefei University of Technology
hongrc.hfut@gmail.com

Zheng Qin*
Hunan University
zqin@hnu.edu.cn

ABSTRACT

Retrieving video moments from an untrimmed video given a natural language as the query is a challenging task in both academia and industry. Although much effort has been made to address this issue, traditional video moment ranking methods are unable to generate reasonable video moment candidates and video moment localization approaches are not applicable to large-scale retrieval scenario. How to combine ranking and localization into a unified framework to overcome their drawbacks and reinforce each other is rarely considered. Toward this end, we contribute a novel solution to thoroughly investigate the video moment retrieval issue under the adversarial learning paradigm. The key of our solution is to formulate the video moment retrieval task as an adversarial learning problem with two tightly connected components. Specifically, a reinforcement learning is employed as a generator to produce a set of possible video moments. Meanwhile, a pairwise ranking model is utilized as a discriminator to rank the generated video moments and the ground truth. Finally, the generator and the discriminator are mutually reinforced in the adversarial learning framework, which is able to jointly optimize the performance of both video moment ranking and video moment localization. Extensive experiments on two well-known datasets have well verified the effectiveness and rationality of our proposed solution.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**; • **Theory of computation → Adversarial learning**; **Reinforcement learning**.

KEYWORDS

Video Moment Retrieval, Cross-Modal Retrieval, Adversarial Learning, Reinforcement Learning, Bayesian Personalized Ranking

ACM Reference Format:
Da Cao, Yawen Zeng, Xiaochi Wei, Liqiang Nie, Richang Hong, and Zheng Qin. 2020. Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization. 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/10.1145/3394171.3413841

1 INTRODUCTION

Owing to the quickening pace of modern life and ever-increasing information, it is urgently desired to quickly locate the most relevant information that matches people's real demands. When it comes to the video domain, people are more desired to browse a video moment that matches their interests instead of a whole video due to the massive cost of time as compared to other content (e.g., texts and images). To meet this demand, the task of video moment retrieval with language query has emerged [1, 9], which aims to locate the start and end points of a desired video moment that best matches a given language query.

In fact, great effort has been made to solve the video moment retrieval issue. The majority of existing works focus on generating video moment candidates with the help of multi-scale sliding window segmentation. They perform the retrieval process by exploiting cross-modal semantic expressions between video and query language, such as frame and word alignment [39], object and sentence expression [17], and temporal dependencies and reasoning between events [41]. Meanwhile, the technique of reinforcement learning (RL) has been utilized to locate desired video moments [12], which formulates the video moment retrieval task as a boundary localization process. To further improve the effectiveness of RL, object relationship [35] and graph [40] have been added to jointly process sentences and videos. Unfortunately, pioneer methods consider the video moment retrieval as either a ranking issue or a localization problem, which are two sides of a coin. Ranking is good at sorting numerous video moments but is incapable of forming reasonable candidates, while localization utilizes a fine-grained control to locate the boundary of a specific video moment but cannot be applied to large-scale retrieval scenario.

In this work, we pay special attention to the combination of ranking and localization in a unified framework, focusing on addressing the following challenges: **1) Reasonable video moment candidates.** The video moment retrieval heavily relies on the video moment candidates. However, traditional video moment ranking methods employ the sliding window strategy to randomly generate numerous candidates, which inevitably result in the

*Corresponding author.

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

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413841

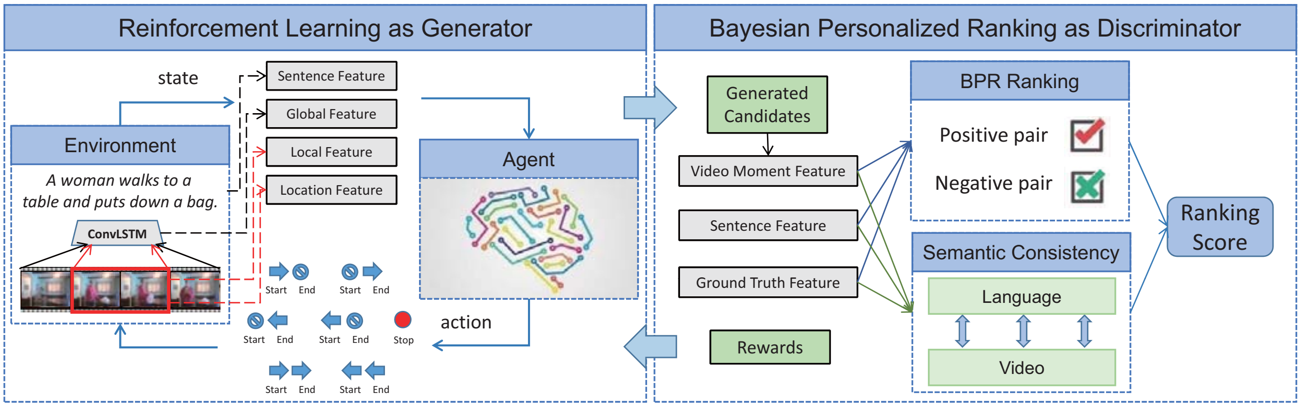


Figure 1: The graphical representation of our proposed AVMR framework. It is built upon two competing components: a RL-based generator produces a set of video moment candidates to confuse the discriminator, and a BPR-based discriminator distinguishes the generated video moment candidates and the ground truth.

less effective retrieval process. Hence, how to effectively produce reasonable video moment candidates is of great importance. **2) Robustness of RL.** Although RL-based approaches have shown great progress in the video moment localization task, the design of reward is less explored. Existing RL-based methods manually set the reward with the help of Interaction over Union (IoU), which are less reasonable and induce the slow and instable convergence. Thereby, how to design a flexible reward to improve the convergence efficiency of RL is a non-trivial task. **3) Mutual reinforcement of ranking and localization.** The tasks of both video moment ranking and video moment localization are crucial to the performance of video moment retrieval. Nevertheless, the correlation between ranking and localization is rarely considered in existing algorithms. Therefore, how to combine ranking and localization into a unified framework to reinforce each other is a tricky task.

To address aforementioned issues, we propose a novel Adversarial Video Moment Retrieval (AVMR) solution to investigate the video moment retrieval task comprehensively. The adversarial learning framework of AVMR is illustrated in Fig. 1. Firstly, a RL is utilized as a generator to produce a set of video moments. Through this way, limited and reasonable video moment candidates are generated. Furthermore, a bayesian personalized ranking (BPR) [32] is employed as a discriminator to rank the generated video moment and the ground truth in a pairwise comparison manner. It is able to provide the RL with adaptive reward and further improve the retrieval performance. Ultimately, an adversarial optimization method is proposed, which could jointly optimize the tasks of both video moment ranking and video moment localization. By conducting experiments on two real-world datasets, we validate that our proposed framework is superior to other state-of-the-art competitors on both overall performance comparison and micro-scope analyses.

The main contributions of this work are summarized as follows:

- The promising yet challenging problem of video moment retrieval is explored. To the best of our knowledge, this is the first work that attempts to solve this problem by jointly considering video moment ranking and video moment localization.
- We propose a novel AVMR solution, which unifies RL and BPR into an adversarial learning framework. As such, the tasks of video moment ranking and video moment localization are mutually reinforced.
- Extensive experiments are conducted on two well-known datasets, which demonstrate the rationality and effectiveness of our method. Meanwhile, we have released the datasets and implementation to facilitate the research community[1].

2 RELATED WORK

In this section, we briefly review some literatures that are tightly related to our work, namely, video moment retrieval and adversarial learning for retrieval.

2.1 Video Moment Retrieval

As an intersection of multimedia and information retrieval, video moment retrieval [1, 9] has drawn great attention in the research community. Technically, existing methods can be divided into two categories — candidate-based and RL-based approaches. Candidate-based retrieval methods attempt to bridge the semantic gap between language and video by obtaining pre-segmented video moments and further perform the cross-modal retrieval. Benefiting from the merit of neural attention network [2–4], the methods of attentive cross-modal retrieval network [22], cross-modal temporal moment localization [23], and spatial and language-temporal attention model [17] are introduced. By jointly learning the embedding space of visual and textual content, the dependency between vision and language are well learnt to boost the language-video moment retrieval performance [39, 41, 42]. Besides, the techniques of graph structure [40], cross-gated attended recurrent networks [5], and weakly supervised method [24] are further employed to enhance the performance.

Inspired by the recent advance of RL [8, 37], RL-based methods formulate the video moment retrieval task as a problem of sequential decision by learning an agent, which fits naturally into the RL paradigm. Specifically, He et al. [12] modeled the

[1] https://github.com/yawenzeng/AVMR

video moment retrieval task as controlling an agent to read the description, to watch the video as well as the current localization, and then to move the temporal grounding boundaries iteratively to find the best matching clip. Thereafter, Wang et al. [35] proposed a recurrent neural network-based RL model for language-driven temporal activity localization, which dynamically observes a sequence of video frames conditioned on the given language query and finally outputs temporal boundaries.

However, candidate-based methods regard the video moment retrieval as a ranking issue which fails to generate reasonable candidates, while RL-based approaches regard the video moment retrieval as a localization problem which fails to manipulate large-scale retrieval scenario. In this work, we aim to perform the video moment retrieval under the adversarial learning paradigm, which is able to make the best of both categories and avoid their drawbacks.

2.2 Adversarial Learning for Retrieval

This work is inspired by the recent advance of generative adversarial network (GAN) [10]. It employs two neural networks, pitting one against the other (thus the “adversarial”), to generate newly synthesized instances of data. By utilizing the merit of adversarial learning, GAN has been widely applied to multiple retrieval applications, such as expert retrieval [20] and image retrieval [11, 31]. Besides, as an extension of GAN, IRGAN [34] is proposed to manipulate multiple information retrieval tasks, such as web search, item recommendation, and question answering.

In addition to directly adopting GAN, some efforts have been made to address various retrieval tasks by designing some sophisticated adversarial algorithms [7, 36]. Reasonably alternating the training process of two completing components can boost the retrieval performance. Specifically, Wang et al. [33] presented an adversarial cross-modal retrieval method, which seeks an effective common subspace based on adversarial learning. Thereafter, He et al. [14] contributed a new learning method for optimizing recommender models, which enhances the pairwise ranking method by performing adversarial training. In the domain of point-of-interest (POI) recommendation, Zhou et al. [43] unified RL and matrix factorization methods into an adversarial learning framework. Furthermore, Wang et al. [38] proposed an adversarial learning framework for collaborative ranking to learn with a dynamic scenario.

Inspired by these pioneering efforts, the intension of our AVMR framework is to take full advantage of the merits of both candidate-based and RL-based video moment retrieval methods under the adversarial learning paradigm.

3 OUR PROPOSED FRAMEWORK

The framework of AVMR is presented in Fig. 1. In general, it consists of two complementary components: 1) RL model acts as the generator to generate (or more precisely, locate) reasonable video moment candidates; and 2) BPR combined with semantic consistency acts as the discriminator to rank the generated video moment and the ground truth. Specifically, we first present the problem formulation in Section 3.1. Thereafter, we introduce the two key components of our proposed framework — RL as generator in Section 3.2, and BPR as discriminator in Section 3.3. Ultimately, the adversarial learning is introduced in Section 3.4.

3.1 Problem Formulation

Let v and q denote a long untrimmed video and a query sentence, respectively. The query sentence q is affiliated with a temporal annotation $\tau = (\tau_s, \tau_e)$, where τ_s and τ_e are the start and end points of the target video moment. Given a video v and its query sentence q, the goal of adversarial video moment retrieval is to identify a desired video moment with the boundary of $l = (l_s, l_e)$ to be close to the ground truth τ.

3.2 RL as Generator

Aiming at localizing the boundary of desired video moment, RL-based algorithms have been proposed [12, 35], which formulate the video moment localization issue as a sequential decision making problem. In particular, at each step, the RL generates an action policy according to the state of the current environment, which defines the probability distribution of all actions in the action space. Once the action selection has been executed, the state is updated. The network update is sampled by the agent according to the policy. And the localized boundary is repeatedly adjusted until the RL is converged or the maximum number of steps has been reached.

In this paper, we employ Deep Deterministic Policy Gradient (DDPG) [21], instead of previously utilized Actor-Critic (AC) [25], to perform the video moment retrieval task. Compared to AC, DDPG is able to accelerate the convergence and increase the diversity of generated video moments, which is mainly because of the following superiorities: deep learning for Q-value function approximation, the utilization of experience replay, and the implementation of dual target networks.

State, Action, and Reward. At each time step t, the agent takes an action according to its current state. To make the agent sensitive to the learning environment, the state $\mathbf{s}^t$ is defined as the combination of the query sentence feature $\mathbf{f}_q$, the global video feature $\mathbf{f}_g$, the current temporal grounding boundaries $\mathbf{l}^t$ and its corresponding local feature $\mathbf{f}_l^t$. The state $\mathbf{s}^t$ is formally formulated as:

$$\mathbf{s}^t = [\mathbf{f}_q, \mathbf{f}_g, \mathbf{l}^t, \mathbf{f}_l^t]. \tag{1}$$

According to this state $\mathbf{s}^t$, the agent takes an action a^t to update the grounding boundary $\mathbf{l}^t$ and its local feature $\mathbf{f}_l^t$. The action space $\mathcal{A}$ consists of seven actions, namely, moving start/end point forward, moving start/end point backward, shifting both start and end points backward/forward, and a STOP action. The initial position is set as $\mathbf{l}^0 = [0.25 * h, 0.75 * h]$, where h is the length of image frames in the video. The step size is set as $h/2\epsilon$, which ensures that ϵ steps can traverse the whole video. Once an action is executed, the localizated boundary moves accordingly and the state changes.

After the execution of an action a^t, the updated boundary $\mathbf{l}^t$ is utilized to generate a video moment candidate, and the action is evaluated to obtain a reward which measures whether the generated video moment candidate matches the query. In our method, the reward r^t of each step t is determined by the discriminator as revealed Eqn.(8), and is then utilized to update the parameters in DDPG.

Actor, Critic, and Target Networks. DDPG is composed of actor, critic, and their target networks. Specifically, critic is used for value function approximation, actor is employed for parameterization policy, and target networks are utilized to do off-policy updates. At each time step, $(\mathbf{s}, a, r, \mathbf{s}')$ is stored in the experience replay memory M, where superscript $'$ denotes new time step. Then a random minibatch of transition i is sampled from M. For the optimal policy π, the action-value $Q(\mathbf{s}^i, a^i)$ is composed of the reward and the value function. The critic updates temporal-difference error with target network $Q^*(\mathbf{s}^{i+1}, a^{i+1})$:

$$\mathbb{E}_{(\mathbf{s}^i, a^i, r^i, \mathbf{s}^{i+1})\sim M}[(Q(\mathbf{s}^i, a^i) - r^i - \gamma maxQ^*(\mathbf{s}^{i+1}, a^{i+1}))^2], \quad (2)$$

where γ is the discount factor, $(\mathbf{s}^i, a^i, r^i, \mathbf{s}^{i+1})$ is randomly sampled from M. The actor is updated to make the critic's Q-value higher under the deterministic policy μ:

$$\nabla J = \mathbb{E}_{\mathbf{s}^i\sim M}[\nabla\mu(a^i|\mathbf{s}^i)\nabla_a(\mathbf{s}^i, a^i)|_{a=\mu(\mathbf{s}^i)}]. \quad (3)$$

The target network stores parameters. Besides, a soft update strategy [21] is employed, in which the target network is updated in each step of both source networks.

3.3 BPR as Discriminator

The majority of existing video moment retrieval methods employ IoU as the loss function or reward to optimize the retrieval performance. However, the difference between a generated video moment and the ground truth is relatively small, the same goes for the values of IoU, resulting in insufficient discrimination. Especially for RL-based methods, the performance is relatively sensitive to the design of reward [15]. Utilizing indistinguishable values of IoU as reward is unstable and difficult to convergence. This will be discussed in Section 4.2. Besides, during the exploration process, even if RL has found the optimal boundary, it cannot stop because it does not know whether there are better boundaries at subsequent steps. This will result in the missing of optimal boundary, which would be discussed in Section 4.5.

To address above mentioned issue, we introduce a discriminator to provide flexible reward. Specifically, we resort to a pairwise comparison method, namely BPR, to maximize the difference between a generated video moment and the ground truth. BPR models a triplet of a query sentence and two video moments, where one of the video moments is a generated video moment and the other one is the ground truth. Before the realization of BPR, the representations of video and language are projected into a common space under the constrain of semantic consistency, which is able to regularize the modalities' embeddings and boost the retrieval performance.

Semantic Consistency. Given the target video moment feature $\mathbf{f}_\tau$, the generated video moment feature $\mathbf{f}_l$, and the query sentence feature $\mathbf{f}_q$, the correlation among them is essentially depend on the semantics they shared in the high-level semantic space. We first project video and language modalities of varying lengths into a common space,

$$\begin{cases} \widetilde{\mathbf{f}}_\tau = o_v(\mathbf{f}_\tau) \\ \widetilde{\mathbf{f}}_l = o_v(\mathbf{f}_l) \\ \widetilde{\mathbf{f}}_q = o_l(\mathbf{f}_q) \end{cases}, \quad (4)$$

where $o_v(\cdot)$ and $o_l(\cdot)$ are projection functions with the approximation of multi-layer perceptron. $\widetilde{\mathbf{f}}_\tau$, $\widetilde{\mathbf{f}}_l$, and $\widetilde{\mathbf{f}}_q$ are projected features with the same dimension. In the common space, the modalities are forced to be close under the constraint of semantic consistency. It can be formulated as follows,

$$\mathcal{L}_{sc} = \|\widetilde{\mathbf{f}}_\tau - \widetilde{\mathbf{f}}_q\|^2. \quad (5)$$

To learn the interplay of video and language, we combine the information embedded in both modalities by utilizing element-wise multiplication (×), element-wise addition (+), and a fully connected (FC) layer,

$$\begin{cases} \widetilde{\mathbf{f}}_{\tau q} = [\widetilde{\mathbf{f}}_\tau \times \widetilde{\mathbf{f}}_q, \widetilde{\mathbf{f}}_\tau + \widetilde{\mathbf{f}}_q, FC([\widetilde{\mathbf{f}}_\tau, \widetilde{\mathbf{f}}_q])] \\ \widetilde{\mathbf{f}}_{lq} = [\widetilde{\mathbf{f}}_l \times \widetilde{\mathbf{f}}_q, \widetilde{\mathbf{f}}_l + \widetilde{\mathbf{f}}_q, FC([\widetilde{\mathbf{f}}_l, \widetilde{\mathbf{f}}_q])] \end{cases}. \quad (6)$$

Customized BPR. Given the target video moment's mapping feature $\widetilde{\mathbf{f}}_\tau$, the generated video moment's mapping feature $\widetilde{\mathbf{f}}_l$, and the query sentence's mapping feature $\widetilde{\mathbf{f}}_q$, the triplet $(\widetilde{\mathbf{f}}_\tau, \widetilde{\mathbf{f}}_l, \widetilde{\mathbf{f}}_q)$ is defined as the training sample. $(\widetilde{\mathbf{f}}_\tau, \widetilde{\mathbf{f}}_q)$ is treated as a positive pair, while $(\widetilde{\mathbf{f}}_l, \widetilde{\mathbf{f}}_q)$ is considered as a negative pair. The purpose of BPR is to make the positive pair has a highly matching score than its negative counterpart. The optimization objective for BPR is based on the maximum posterior estimator. In particular, by applying the above learnt features, our customized BPR model is formulated as,

$$\mathcal{L}_{bpr} = ln\sigma(o_s(\widetilde{\mathbf{f}}_{\tau q}) - o_s(\widetilde{\mathbf{f}}_{lq}) - \Delta), \quad (7)$$

where $\sigma(\cdot)$ is the sigmoid function, o_s is the score function with the approximation of multi-layer perceptron, and Δ is the margin. Through this way, the score of positive pair $(\widetilde{\mathbf{f}}_\tau, \widetilde{\mathbf{f}}_q)$ is able to be larger than its negative counterpart $(\widetilde{\mathbf{f}}_l, \widetilde{\mathbf{f}}_q)$, which effectively distinguishes the ground truth from various generated video moment candidates.

3.4 Adversarial Learning

The key to adversarial learning is how to jointly promote the performance of two competing components, namely the generator and the discriminator. Let G_θ and D_ϕ denote the generator and the discriminator, which are parameterized by θ and ϕ, respectively. The reward of generator is formulated as,

$$r = -\mathcal{L}_{bpr} - \lambda_s\mathcal{L}_{sc} + \lambda_j\mathcal{L}_{joint}, \quad (8)$$

where λ_s and λ_j are balance factors. The intuition of adding the component of $\mathcal{L}_{joint}$ is to jointly consider the influence of both ranking and localization,

$$\mathcal{L}_{joint} = o_s(\widetilde{\mathbf{f}}_{lq})\text{IoU}(\tau, l), \quad (9)$$

where $o_s(\widetilde{\mathbf{f}}_{lq})$ is utilized to measure the ranking performance and $\text{IoU}(\tau, l)$ is used to evaluate the localization performance. Moreover, $\text{IoU}(\tau, l)$ is formulated as,

$$\text{IoU}(\tau, l) = \frac{min(\tau_e, l_e) - max(\tau_s, l_s)}{max(\tau_e, l_e) - min(\tau_s, l_s)}. \quad (10)$$

Under the adversarial learning, the generator and the discriminator are jointly optimized.Specifically, the generator G_θ maximizes the total reward of all generated samples to improve its ability of imitating the ground truth,

$$\theta = \underset{\theta}{argmax} \sum_{k=1}^{K} r^k, \tag{11}$$

where K is the number of generated video moments. As for each iteration, the generator's parameters are updated with the help of Eqn.(3). The discriminator D_ϕ optimizes the BPR ranking loss and the semantic consistency loss,

$$\phi = \underset{\phi}{argmin} \sum_{k=1}^{K} -\mathcal{L}_{bpr} - \lambda_s \mathcal{L}_{sc}. \tag{12}$$

Through this way, generator G_θ and discriminator D_ϕ are alternatively optimized until convergence, leading to the mutual reinforcement of video moment ranking and video moment localization.

4 EXPERIMENTS

In this section, we conducted extensive experiments on two real-world datasets to answer the following four research questions:

RQ1 How does our proposed AVMR framework perform as compared to other state-of-the-art competitors?

RQ2 How do different components in adversarial learning contribute to the performance of AVMR?

RQ3 How do different predefined settings (e.g., the discount factor γ, the margin Δ, and the balance factors λ_s and λ_j) affect our framework?

RQ4 Can we visualize the retrieval performance of our method and other baselines?

4.1 Experimental Settings.

4.1.1 Datasets. We experimented with two publicly accessible datasets, one is related to daily activities at home and the other one is related to cooking activities in lab kitchen.

1. Charades-STA. Charades-STA [9] is utilized for temporal activity localization via query language, which is constructed on the top of original Charades dataset [30] and contains $6,672$ videos. Since Charades dataset only contains video-level paragraph description, Charades-STA further annotates video moment with query language. We downloaded original videos[2] and their corresponding caption annotations[3]. To narrow down the searching space, we further segmented each video by employing multi-scale sliding windows with the size of $[64, 128, 256, 512]$ frames with 80% overlap on adjacent video moments. In summary, we ultimately obtained $12,541$ video moment-query sentence pairs.

2. TACoS. The TACoS dataset is constructed by [27] on the top of MPII-Compositive dataset [28] and contains 127 cooking videos. Each video is affiliated with two kinds of annotations. One is fine-grained activity labels with temporal location (i.e., start and end time). The other kind of annotation is natural language descriptions with temporal locations. We downloaded original videos[4] and further employed multi-scale sliding windows to segment each video with the size of $[64, 128, 256, 512]$ frames with 80% overlap on adjacent video moments. Based upon these criteria, $7,463$ video moment-query sentence pairs are obtained.

The experimental datasets are further divided into 3 disjoint sets, with 70%, 10%, and 20% randomly selected video moment-query sentence pairs for training, validation and testing, respectively. The validation set is leveraged to tune hyperparameters and the final performance comparison is conducted on the testing set. As both datasets only contain positive instances, namely the triplets (v, q, τ), we utilized the RL to generate negative samples, namely the triplets (v, q, l). In the training phase, as revealed in Algorithm **??**, the generator G_θ and the discriminator D_ϕ respectively generate $K = 20$ samples in each iteration. In the validation and testing stages, the positive instance with its newly generated 20 negative samples are utilized to evaluate the retrieval performance.

4.1.2 Evaluation Protocol. To evaluate the performance of our proposed method and other competitors, we adopted "$R@n$, IoU $= m$" proposed by [16] as the evaluation metric. Specifically, for each query sentence, we first calculated the temporal IoU between the generated video moments and the ground truth. Thereafter, we measured the percentage of top-k results having IoU larger than m. In the rest of this article, we use $R(n, m)$ to denote "$R@n$, IoU $= m$". This metric itself is on the query level and the overall performance is the average of all query sentences as follows,

$$R(n, m) = \frac{1}{N_q} \sum_{i=1}^{N_q} r(n, m, q_i), \tag{13}$$

where $r(n, m, q_i)$ is the recall for a given language query q_i, N_q denotes the total number of language queries, and $R(n, m)$ is the averaged overall performance.

4.1.3 Baselines. To justify the effectiveness of our method, we compared it to the following methods.

- **CTRL [9].** This is a cross-modal temporal regression localizer that jointly models query description and video moments, and outputs alignment scores and action boundary regression results for the moment candidates.
- **ROLE [23].** This is a cross-modal temporal moment localization approach, which is able to adaptively encode complex and significant language query information for localizing desired video moments.
- **RWM [12].** This is a RL-based method, which formulates the problem of video moment localization as a sequential decision making process. Meanwhile, it further combines the supervised learning in a multi-task learning framework.
- **SM-RL [35].** This RL-based model selectly observes a sequence of frames and associates the given sentence with video content in a matching-based manner. In addition, semantic concepts of videos are further extracted to enhance the performance.
- **APR [14].** This method improves the quality of generated samples by adding adversarial noises and uses BPR for ranking. We realized this adversarial learning scheme based on the CTRL model to demonstrate its positive effect on video moment retrieval.
- **IRGAN [34].** This method combines two types of models via adversarial training, a generative model that generates instances

[2] https://allenai.org/plato/charades

[3] https://github.com/jiyanggao/TALL

[4] http://www.coli.uni-saarland.de/projects/smile/tacos

Table 1: Performance comparison of various state-of-the-art baselines on Charades-STA and TACoS datasets (Section 4.2).

	Charades-STA						TACoS					
Method	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5
CTRL	80.30%	64.67%	37.02%	88.39%	73.48%	53.21%	76.24%	50.46%	34.86%	81.35%	70.42%	49.37%
ROLE	85.48%	67.14%	41.69%	90.00%	76.56%	57.39%	79.49%	56.71%	37.80%	87.54%	74.33%	54.96%
RWM	87.80%	70.66%	42.15%	-	-	-	83.46%	58.03%	38.45%	-	-	-
SM-RL	90.83%	72.81%	47.67%	-	-	-	87.09%	64.11%	42.37%	-	-	-
APR	82.65%	67.94%	40.10%	89.17%	75.77%	57.43%	78.91%	51.60%	35.73%	82.11%	71.54%	50.23%
IRGAN	88.17%	71.23%	45.53%	92.25%	81.21 %	62.74%	86.37%	63.52%	39.22%	90.03%	75.91%	57.16%
GVMR	88.56%	71.10%	44.81%	-	-	-	85.72%	61.17%	40.01%	-	-	-
DVMR	84.72%	65.79%	40.77%	88.68	75.46	55.34	80.67%	54.48%	36.29%	85.31	72.42	54.80
AVMR	**93.20%**	**77.72%**	**54.59%**	**98.87%**	**88.92%**	**72.78%**	**89.77%**	**72.16%**	**49.13%**	**94.26%**	**83.37%**	**64.40%**

Table 2: Performance comparison of various component combinations on Charades-STA and TACoS datasets (Section 4.3).

	Charades-STA						TACoS					
Method	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5
G+S	77.24%	58.58%	32.50%	79.45%	71.67%	50.58%	70.55%	52.68%	30.87%	72.60%	63.91%	41.02%
G+J	89.65%	70.88%	45.34%	-	-	-	86.28%	63.32%	40.19%	-	-	-
G+B	90.67%	71.27%	46.46%	93.72%	83.37%	63.61%	87.14%	65.47%	43.81%	87.98%	76.26%	56.64%
G+S+J	90.88%	71.32%	48.71%	94.37%	84.02%	63.77%	87.23%	68.45%	44.29%	88.07%	77.26%	57.88%
G+B+S	91.91%	74.14%	50.59%	96.69%	86.54%	68.40%	88.96%	70.58%	47.31%	91.82%	80.32%	61.91%
G+B+J	92.30%	75.52%	51.42%	98.17%	87.34%	69.83%	89.03%	71.36%	47.46%	92.54%	81.04%	62.48%
ALL	**93.20%**	**77.72%**	**54.59%**	**98.87%**	**88.92%**	**72.78%**	**89.77%**	**72.16%**	**49.13%**	**94.26%**	**83.37%**	**64.40%**

for a given query and a discriminative model that determines whether the instance is from real data or generated.

- **GVMR.** GVMR is short for "Generator-based Video Moment Retrieval". It is a variant of our AVMR method by retaining the generator and employing IoU as the reward to feedback the RL-based generator. This is to demonstrate the effect of the discriminator in providing reasonable reward to RL.
- **DVMR.** DVMR indicates "Discriminator-based Video Moment Retrieval", which removes the generator and employs BPR-based discriminator to rank the pre-segmented video moment candidates. It is utilized to demonstrate the effect of the generator in providing reasonable video moment candidates.

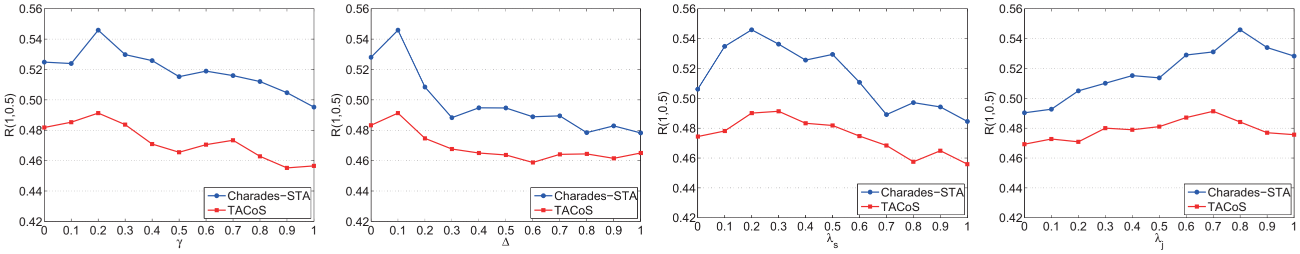
4.1.4 Implementation Details. We implemented our solution based on the PyTorch framework[5] on a server equipped with a NVIDIA 2080TI-11G GPU. In the feature representation, $\mathbf{f}_q$ is a $4,800$-dimensional vector extracted by skip-thought [19], while $\mathbf{f}_g$, $\mathbf{f}_\tau$, and $\mathbf{f}_l$ are $2,048$-dimensional vectors obtained by employing ConvLSTM [26] on consecutive image frames where the image features are extracted by ResNet [13]. In the common space, the dimension of projected features ($\widetilde{\mathbf{f}}_q$, $\widetilde{\mathbf{f}}_\tau$, and $\widetilde{\mathbf{f}}_l$) is set as $1,024$. To initialize the hidden layers in our method, we randomly set their parameters with a Gaussian distribution (a mean of 0 and a stand deviation of 0.1). We used the Adam optimizer [18] for all gradient-based methods, where the mini-batch size and the learning rate were searched in $[128, 256, 512, 1024]$ and $[0.001, 0.005, 0.01, 0.05, 0.1]$, respectively. For specific hyper-parameters in our framework, the step size ϵ, the discount factor γ, the margin Δ, the balance factor λ_s, and the balance factor λ_j are set as $(20, 0.2, 0.1, 0.2, 0.8)$ and $(20, 0.2, 0.1, 0.3, 0.7)$ on Charades-STA and TACoS datasets, respectively.

4.2 Overall Performance Comparison (RQ1)

To demonstrate the effectiveness of our proposed AVMR method, we compared it with several state-of-the-art approaches: 1) CTRL; 2) ROLE; 3) RWM; 4) SM-RL; 5) APR; 6) IRGAN; 7) GVMR; and 8) DVMR. CTRL, ROLE, and DVMR are classified as the candidate-based video moment retrieval algorithms, RMW, SM-RL, and GVMR belong to RL-based video moment localization approaches, and APR and IRGAN are adversarial learning-based retrieval methods. It worth to mention that RWM, SM-RL, and GVMR are designed for boundary localization, which only returns a boundary value. Therefore, their performance is only compared on $R(1, m)$.

Experimental results are shown in Table 1. We have the following observations: 1) Our AVMR approach achieves the best performance on both Charades-STA and TACoS datasets, significantly outperforming state-of-the-art baselines. It is mainly because AVMR model employs the adversarial learning scheme to jointly optimize the performance of both video moment ranking and video moment localization. 2) RL-based algorithms, RWM, SM-RL, and GVMR outperform candidate-based approaches, CTRL, ROLE, and DVMR, by a great margin. The dynamic localization of video moment boundary is crucial to the performance of video moment retrieval, which is ignored in traditional candidate-based methods. 3) The experimental results of APR are superior to that of CTRL on both datasets. APR improves the performance of CTRL by adding the adversarial learning strategy. It well demonstrates the

[5] http://www.pytorch.org



(a) Discount factor γ **(b) Margin Δ** **(c) Balance factor λ_s** **(d) Balance factor λ_j**

Figure 2: Performance of AVMR w.r.t. various hyper-parameters.

effectiveness of adversarial learning in solving the video moment retrieval problem. 4) SM-RL achieves the best performance among all baselines on $R(1, m)$, while IRGAN beats other baselines on $R(5, m)$. This is because SM-RL is designed for video moment localization and considers the semantic gap between video and language. Meanwhile, IRGAN is designed for solving information retrieval issues, which is also suitable for our video moment retrieval scenario. 5) The performance of AVMR is superior to that of GVMR. It well demonstrates the rationality of adding the discriminator to provide reasonable reward to the RL-based generator instead of the inflexible value of IoU. Meanwhile, AVMR outperforms DVMR to a large extent. It manifests the importance of the RL-based generator in providing reasonable video moment candidates instead of the pre-segmented clips.

4.3 Ablation Study (RQ2)

To investigate the effectiveness of our proposed adversarial learning scheme, especially the design of reward for RL-based generator and the loss function for discriminator, we performed the study on various component combinations. In particular, we employed G to denote the RL-based generator, B to denote the BPR component as revealed in Eqn.(7), J to denote the joint influence of ranking and localization as revealed in Eqn.(9), and S to denote the semantic consistency between video and language as revealed in Eqn.(5).

The performance of different component combinations is shown in Table 2[6]. We have the following observations: 1) The method of ALL (i.e., AVMR) achieves the best performance among various component combinations. This validates the effectiveness of our AVMR method, more specifically, the efficiency in aggregating multiple components. 2) Jointly observing the performance of G+S, G+J, G+B, G+S+J, G+B+S, and G+B+J, we can infer that B is more important than J, followed by S. This is mainly because B not only provides a reward to RL-based generator, but also optimizes the ranking performance of discriminator. Meanwhile, J provides a reasonable reward to RL-based generator, but the semantic consistency provided by S is less important for both generator and discriminator. That is why J is more important than S. 3) It is obviously observed that the more components are incorporated, the better performance we achieved. This verifies that there are consistent rather than conflicting relationships among multiple components.

4.4 Sensitivity of Hyper-parameters (RQ3)

In order to demonstrate the robustness and effectiveness of our proposed AVMR framework, we investigated the sensibility of several factors, namely, the discount factor γ, the margin Δ, and the balance factors λ_s and λ_j with respect to $R(1, 0.5)$ on both Charades-STA and TACoS datasets.

Impact of Discount Factor. The discount factor is the measurement of how far ahead the RL should look. To prioritise rewards in the distant future, the value should be closer to 1. A discount factor closer to 0 indicates that only rewards in the immediate future are being considered, implying a shallow lookahead. The parameter tuning results of γ are revealed in Fig. 2a. The performance on both datasets are relatively stable and reach their maximum values when $\gamma = 0.2$. It indicates that the immediate influence should be carefully considered.

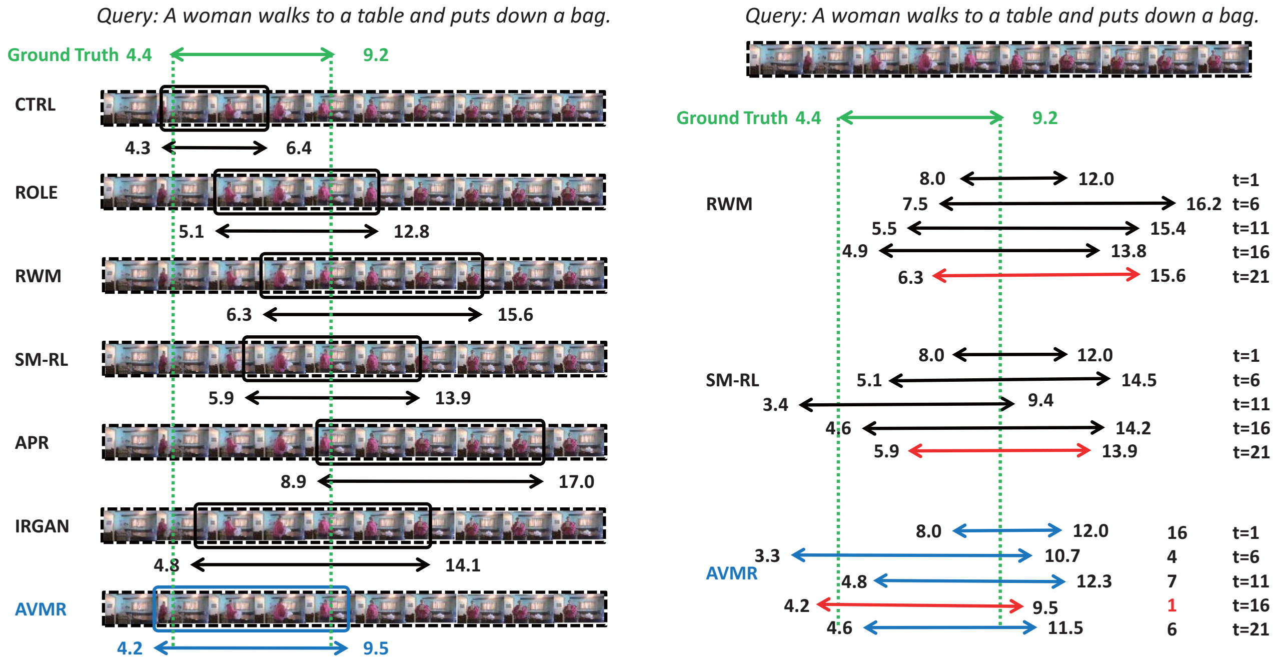
Impact of Margin. The strategy of utilizing margin to perform the retrieval process has been proven rational and effective in [6, 29]. It measures the gap between the positive pair and the negative pair. To illustrate the impact of margin for AVMR, we show the performance of AVMR w.r.t. different margin settings on both datasets in Fig. 2b. The values of $R(1, 0.5)$ increase first and then decrease along with the increasing of Δ, reaching their maximum values when $\Delta = 0.1$. Our finding is consistent with traditional retrieval algorithms [6, 29], which are inclined to use a margin value between 0.1 and 0.5.

Impact of balance factors. Balance factors λ_s and λ_j determine the importance of the components of S and J, where S manifests the semantic consistency between video and language and J indicates the joint optimization of ranking and localization. Higher values indicate greater importance. As can been seen from Fig. 2c and Fig. 2d, the optimal values of $R(1, 0.5)$ have been reached when $\lambda_s = 0.2$ and $\lambda_j = 0.8$ on the Charades-STA dataset, and $\lambda_s = 0.3$ and $\lambda_j = 0.7$ on the TACoS dataset. The parameter tuning results demonstrate that our model cares more about J than S, which is consistent with the conclusion in Section 4.3.

4.5 Retrieval Visualization (RQ4)

To gain deep insight into the effectiveness of our proposed framework and other baselines, we exploited some micro-level case studies. Specifically, we randomly selected a new video-language pair accompanied with the ground truth of video moment, and cast it into all methods to observe both retrieval performance and retrieval process. As revealed in Fig. 3, the randomly selected video depicts a sequence of scenes — a woman carries a bag into the

[6] It worth to mention that performance of G+J is only compared on $R(1, m)$ since it only provides reward to the RL-based generator and the loss function for the discriminator is omitted, which degenerates to a RL-based localization method.



(a) Retrieval performance of various methods.

(b) Retrieval process of RL-based methods.

Figure 3: The visualization of video moment retrieval w.r.t both performance and process. In figure (a), only finally retrieved video moments are revealed. In figure (b), the red lines indicate the retrieved video moments (Section 4.5).

room, puts her bag on the table, walks to the camera, and picks up a notebook. Given the language query "A woman walks to a table and puts down a bag", the scenes of "carrying a bag into the room" and "putting her bag on the table" are closely matched.

Fig. 3a shows the retrieval performance of various methods. We have the following observations: 1) Our AVMR method achieves the best result as compared to other baselines, which manifests the rationality of our proposed framework. 2) The performance of CTRL is relatively acceptable, but it is restricted by the fixed-length window size. 3) ROLE focuses on capturing the occurrence of actions, and that is why it is able to localize the action of "putting down". 4) RWM and SM-RL employ RL for video moment localization, which improves the search range and query results. 5) Although the performance of APR is unsatisfied, the result of IRGAN still proves that the adversarial learning is effective.

Fig. 3b manifests the retrieval process of RL-based methods and our proposed solution. We have the following observations: 1) RWM and SM-RL achieve their best performance at the 16th time step and the 11th time step, respectively. But they still keep exploring until the terminal condition has been reached. In fact, when the localized video moments are close to the ground truth, RL-based methods still not converge. That is why they keep exploring until the maximum step has been reached. 2) Under the AVMR framework, RL-based generator explores the boundary of desired video moment, while BRP-based discriminator scores the generated video moment and works as a reward to feed back the generator. Ultimately, the 16th exploration with the optimal ranking performance is returned. It well demonstrates the rapid convergence ability of AVMR under the adversarial learning paradigm.

5 CONCLUSIONS

In this work, we propose to address the video moment retrieval issue under the adversarial learning paradigm. Under the AVMR framework, there are three key components that contribute to the improvement of the retrieval task: 1) RL-based generator; 2) BPR-based discriminator; and 3) adversarial learning. Specifically, a RL-based generator is proposed to generate various video moment candidates. Thereafter, a BPR-based discriminator is employed to rank the generated video moments with the ground truth in a pairwise comparison manner. Ultimately, the generator and the discriminator are mutually reinforced under the adversarial learning paradigm. The experimental results show that AVMR achieves state-of-the-art performance for the video moment retrieval task; further micro-level analyses demonstrate how different components in adversarial learning contribute to the performance of AVMR, how AVMR is sensitive to predefined hyper-parameters, and how AVMR beats other competitors via visualization.

6 ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Project of China (No. 2018YFB0704000), the National Natural Science Foundation of China (No. 61802121, No. 61772191, No. 61932009, and No. 61722204), the Natural Science Foundation of Hunan Province (No. 2019JJ50057), the Science and Technology Key Projects of Hunan Province (No. 2015TP1004 and No. 2019GK2082), the Science and Technology Project of Changsha City (No. kq1902051), and the Fundamental Research Funds for the Central Universities.

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*. IEEE, 5803–5812.

[2] Da Cao, Xiangnan He, Lianhai Miao, Yahui An, Chao Yang, and Richang Hong. 2018. Attentive group recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 645–654.

[3] Da Cao, Xiangnan He, Lianhai Miao, Guangyi Xiao, Hao Chen, and Jiao Xu. 2019. Social-enhanced attentive group recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[4] Da Cao, Zhiwang Yu, Hanling Zhang, Jiansheng Fang, Liqiang Nie, and Qi Tian. 2019. Video-based cross-modal recipe retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1685–1693.

[5] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 8175–8182.

[6] Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. 2018. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1020–1028.

[7] Juntong Cheng, Yi-Ping Phoebe Chen, Minjun Li, and Yu-Gang Jiang. 2019. TC-GAN: Triangle cycle-consistent GANs for face frontalization with facial features preserved. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 220–228.

[8] Jingtao Ding, Yuhan Quan, Xiangnan He, Yong Li, and Depeng Jin. 2019. Reinforced negative sampling for recommendation with exposure data. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI, 2230–2236.

[9] 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, 5267–5275.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. MIT Press, 2672–2680.

[11] Longteng Guo, Jing Liu, Yuhang Wang, Zhonghua Luo, Wei Wen, and Hanqing Lu. 2017. Sketch-based image retrieval using generative adversarial networks. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1267–1268.

[12] 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*. AAAI, 8393–8400.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[14] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 355–364.

[15] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 3207–3214.

[16] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4555–4564.

[17] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. 2019. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the ACM SIGMM International Conference on Multimedia Retrieval*. ACM, 217–225.

[18] P.Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*. ACM.

[19] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. MIT Press, 3294–3302.

[20] Shangsong Liang. 2019. Unsupervised semantic generative adversarial networks for expert retrieval. In *Proceedings of the International Conference on World Wide Web*. ACM, 1039–1050.

[21] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[22] 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.

[23] 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.

[24] 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*. IEEE, 11592–11601.

[25] 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 *Proceedings of the International Conference on Machine Learning*. ACM, 1928–1937.

[26] Joe Yue-Hei Ng, Matthew J.Hausknecht, Sudheendra Vi jayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. (2015), 4694–4702.

[27] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics* 1 (2013), 25–36.

[28] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script data for attribute-based recognition of composite activities. In *Proceedings of the European Conference on Computer Vision*. Springer, 144–157.

[29] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3020–3028.

[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*. Springer, 510–526.

[31] Jingkuan Song, Tao He, Lianli Gao, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2018. Binary generative adversarial networks for image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 394–401.

[32] Rendle Steffen, Freudenthaler Christoph, Gantner Zeno, and Schmidt-Thieme Lars. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 452–461.

[33] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 154–162.

[34] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A minimax game for unifying generative and discriminative information retrieval models. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 515–524.

[35] 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*. IEEE, 334–343.

[36] Wenxuan Wang, Qiang Sun, Yanwei Fu, Tao Chen, Chenjie Cao, Ziqi Zheng, Guoqiang Xu, Han Qiu, Yu-Gang Jiang, and Xiangyang Xue. 2019. Comp-GAN: Compositional generative adversarial network in synthesizing and recognizing facial expression. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 211–219.

[37] Xiang Wang, Yaokun Xu, Xiangnan He, Yixin Cao, Meng Wang, and Tat-Seng Chua. 2020. Reinforced Negative Sampling over Knowledge Graph for Recommendation. In *Proceedings of the International Conference on World Wide Web*. ACM.

[38] Zitai Wang, Qianqian Xu, Ke Ma, Yangbangyan Jiang, Xiaochun Cao, and Qingming Huang. 2019. Adversarial preference learning with pairwise comparisons. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 656–664.

[39] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 9062–9069.

[40] 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*. IEEE, 1247–1257.

[41] Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1230–1238.

[42] 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*. ACM, 655–664.

[43] Fan Zhou, Ruiyang Yin, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Jin Wu. 2019. Adversarial point-of-interest recommendation. In *Proceedings of the International Conference on World Wide Web*. ACM, 3462–3468.