Siamese Alignment Network for Weakly Supervised Video Moment Retrieval

Yunxiao Wang, Meng Liu, *Member, IEEE,* Yinwei Wei, *Member, IEEE,* Zhiyong Cheng, Yinglong Wang, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*—Video moment retrieval, i.e., localizing the specific video moments within a video given a description query, has attracted substantial attention over the past several years. Although great progress has been achieved thus far, most of existing methods are supervised, which require moment-level temporal annotation information. In contrast, weakly-supervised methods which only need video-level annotations remain largely unexplored. In this paper, we propose a novel end-to-end Siamese alignment network for weakly-supervised video moment retrieval. To be specific, we design a multi-scale Siamese module, which could progressively reduce the semantic gap between the visual and textual modality with the Siamese structure. In addition, we present a context-aware multiple instance learning module by considering the influence of adjacent contexts, enhancing the moment-query and video-query alignment simultaneously. By promoting the matching of both moment-level and video-level, our model can effectively improve the retrieval performance, even if only having weak video level annotations. Extensive experiments on two benchmark datasets, i.e., ActivityNet-Captions and Charades-STA, verify the superiority of our model compared with several state-of-the-art baselines.**

***Index Terms*—Weakly-supervised Video Moment Retrieval, Vision-Language Alignment, Siamese Alignment Network, Multiple Instance Learning**

I. Introduction

VIDEO moment retrieval, aiming to identify the specific start and end points within a video to precisely respond to the given query, has attracted increasing research interests due to its wide spectrum of applications in video surveillance [1]–[11]. Although existing approaches have achieved promising performance, most of them are supervised learning ones. In other words, they heavily depend on the temporal annotation information (i.e., the start and end points of the target moment). However, manually annotating temporal boundaries of target moments is time-consuming and very expensive. Moreover, different annotators may label different temporal boundaries for the same query due to the difference in understanding of the start and end of activities. As shown in Fig. 1, giving the same description query "The bikers begin the race", two human annotators give totally different labeling results. This may introduce bias to the training data and further influence the performance of the learnt model. In light of this, recent research attentions have been shifted to developing weakly-supervised models, which merely require video-level annotations as displayed in Fig. 1.

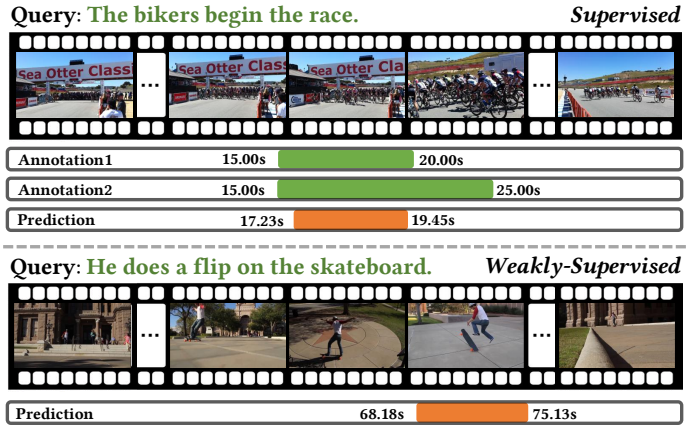

Fig. 1. Illustrations of both the supervised and weakly-supervised video moment retrieval tasks.

Compared with the supervised setting, weakly-supervised video moment retrieval is much more challenging due to the following reasons: 1) Video-query alignment. As illustrated in the bottom of Fig. 1, in the weakly-supervised setting, we merely know which video matches the given query instead of the specific moment, during training. However, in some real-world scenarios (e.g., autonomous driving and surveillance), the untrimmed videos usually contain complex scenes and involve a large number of objects and actions, whereby only some parts of the complex scene convey the desired cues or match the query. Therefore, how to exploit beneficial information from the video and well match the video with the query are two under-studied yet critical problems. 2) Moment-query alignment. Different moments in a video have varying durations and diverse spatial-temporal characteristics [12], thereby how to uncover the underlying moment and estimate the moment-query relevance are highly challenging.

With the remarkable success of deep neural networks in various computer vision tasks, several approaches have dedicated to solving weakly-supervised video moment retrieval by using deep learning models [13]–[16]. However, those methods often suffer from the following limitations: 1) Most existing methods [13], [14], [16] try to embed both the moment and query into a joint embedding space, and then

Yunxiao Wang and Liqiang Nie are with the School of Computer Science and Technology, Shandong University, China (e-mail: yunxiao.wang@mail.sdu.edu.cn; nieliqiang@gmail.com).

Meng Liu is with the School of Computer Science and Technology, Shandong Jianzhu University, China (e-mail: mengliu.sdu@gmail.com). (Corresponding authors: Meng Liu and Liqiang Nie.)

Yinwei Wei is with the School of Computing, National University of Singapore, Singapore (e-mail: weiyinwei@hotmail.com).

Zhiyong Cheng is with Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), China (e-mail: jason.zy.cheng@gmail.com).

Yinglong Wang is with the Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), China (e-mail: wangyl@sdas.org).

they execute the similarity estimation. For instance, [14] adopts the multi-modal processing module proposed in [1] to predict the alignment score. Although some progress has been achieved by those methods, the moment-query alignment is still far from being solved, because of the large semantic discrepancy between language and vision. Moreover, fine-grained interactions, such as frame-word interactions, needs further exploration. And 2) inspired by the astonishing success of the Multiple Instance Learning (MIL) [17] in abnormal activity detection, existing methods [13], [14] commonly adopt the MIL to enforce the video-query alignment due to lack of moment-level supervision. Concretely, both [14] and [13] obtain the video-query matching score by simply summarizing the corresponding scores of moments. However, there are plentiful query-unrelated moments in the video, directly aggregating all moments may induce noisy information and adversely affects the learning.

To address the above limitations, we present a Siamese Alignment Network (SAN) for weakly supervised video moment retrieval. As shown in Fig. 2, it mainly consists of two parts: a multi-scale Siamese module and a context-aware MIL module. The former aims to generate multi-scale moment candidates in one single pass and well match the moment candidate with the given query. The latter is designed to strengthen the alignment of the video-query pair. In particular, to enhance the fine-grained semantic alignment of the moment-query pair, the multi-scale Siamese module persistently projects visual and textual representations into a joint embedding space and constrains the model to capture consistent semantics from two modalities. Moreover, the context-aware MIL module promotes the video-level similarity estimation by considering contextual moments, therefore reinforcing the effect of weak supervision information. Extensive experimental results on two benchmark datasets, i.e., Charades-STA [1] and ActivityNet-Captions [18], validate the effectiveness and superiority of our proposed method as compared to several state-of-the-art baselines. In addition, we release our code to facilitate researchers in this community[1].

Our main contributions can be summarized as follows:

- We propose a novel video moment retrieval framework, i.e., SAN, which only requires video-level annotations for training. More importantly, it can enhance the alignment of both video-query and moment-query simultaneously.
- We design a multi-scale Siamese module, which is capable of promoting the semantic alignment of visual and textual information in different granularity.
- We devise a context-aware MIL module, which improves the estimation of the video-query similarity by considering contextual moment candidates as supplementary.

II. Related Work

In this section, we briefly review some studies related to video moment retrieval, Siamese network, and multiple instance learning.

[1]https://sancode.wixsite.com/san-model.

A. Video Moment Retrieval

Given an untrimmed video and a natural language query, the goal of video moment retrieval is to identify the specific start and end points within a video to precisely respond to the given query, as shown in Fig. 1. Over the past several decades, many techniques have been developed for video moment retrieval. They can be broadly divided into two main categories: supervised and weakly-supervised methods.

According to the key characteristics of supervised approaches, they could be further grouped into three categories: two-stage, one-stage, and reinforcement learning methods. Therein, the two-stage ones usually generate moment candidates in the off-line fashion, such as utilizing multi-scale sliding windows [1], [2] or designing proposal generation networks [4], [5], [19]. Differently, one-stage ones retrieve the golden moment (i.e., the moment that best matches the given query) in one single pass, such as [20] directly predicts the start and end points of the target moment via a boundary model. As the one- and two-stage studies inevitably suffer from inefficiency and unintelligent issues, reinforcement learning approaches are proposed. For instance, [21] formulates the video moment retrieval as a problem of sequential decision making and presents the first end-to-end reinforcement learning framework.

As manually annotating temporal boundaries of target moments is time-consuming, recent research attentions have been shifted to developing weakly-supervised video moment retrieval models [13]–[16], [22]–[26], which merely require video-level annotations, as shown in Fig. 1. For example, Mithun et al. [13] proposed the first weakly-supervised video moment retrieval model TGA, which aims to learn a joint embedding space for video and query representations. In the same year, Gao et al. [14] presented a Weakly Supervised Natural Language Localization Network (WSLLN), which leverages a two-stream structure to measure moment-query consistency and conducts moment selection simultaneously. Although above methods have achieved promising performance, they are two-stage approaches, i.e., utilizing multi-scale sliding windows to generate moment candidates, therefore suffering from inferior effectiveness and efficiency. Inspired by this, Lin et al. [15] advanced a Semantic Completion Network (SCN), which scores all the moments sampled at different scales in a single pass. Wu et al. [16] developed a Boundary Adaptive Refinement (BAR) framework, which extends reinforcement learning to select candidates directly and guides the process of progressively refining the temporal boundary. Meanwhile, Wang et al. [22] introduced a Weakly Supervised Temporal Adjacent Network (WSTAN), which models temporal relations by a 2D feature map, where one dimension indicates the starting time of a moment and the other indicates the end time. These studies have significantly boosted the performance on existing datasets, nevertheless, weakly-supervised video moment retrieval is still far from being solved due to the large semantic discrepancy between textual and visual modality.

B. Action Localization

The goal of action localization is to locate and recognize a predefined action in an untrimmed video, which requires

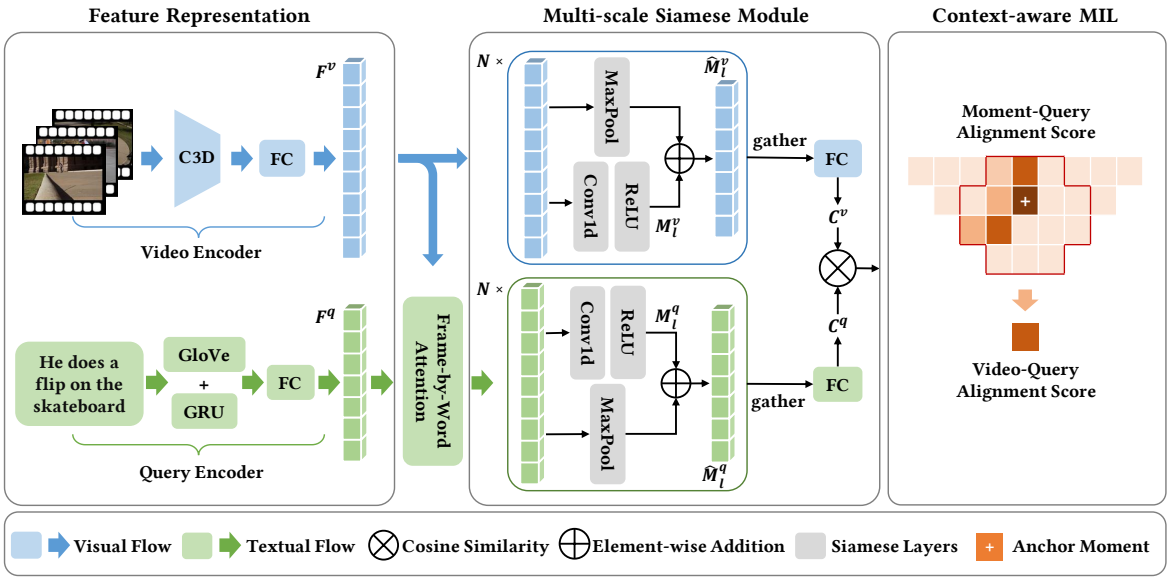

Fig. 2. Schematic illustration of the proposed SAN framework. The anchor moment is the candidate with highest moment-query alignment score, and moments within the red border represent the contextual moments of the anchor.

the recognition of the category and the location of temporal boundaries of the action. Similarly to video moment retrieval, the approaches pertaining to this task can be divided into two categories: supervised and weakly-supervised methods.

The models in supervised setting [27]–[31] are mainly structured in a two-stage manner, which first generate temporal action proposals through an elaborate proposal generation network, and then predict the action category for each proposal through a classification network. However, in the weakly-supervised setting, there are only video-level action category labels without exact temporal boundary annotations. To resolve this issue, most existing methods have made use of the principle of multiple instance learning. For instance, Wang et al. [32] proposed a two-branch structure model, including selection and classification branches. The selection branch is responsible for the assessment of the importance of each video clip using multiple instance learning, and the classification branch aims to predict which category each video clip belongs to. Although these methods are effective, Singh et al. [33] found that most weakly-supervised action localization networks tend to focus only on the most differentiated video moments, leading to incomplete localization. Therefore, how to localize more complete results becomes a new challenge for this task. To this end, many approaches have devised ingenious methods. For example, Zhong et al. [34] built a multi-layer classifier by deleting the recognizable segments step by step. The classifier trained at different steps can find different action segments and assemble into complete localization.

Even though action localization is similar to video moment retrieval task, it can merely localize action moments in a predefined category set. Differently, the video moment retrieval could retrieve any action moments depicted by the natural language query.

C. Siamese Network

Siamese network was first introduced by Bromley et al. [35] for signature verification. The term "Siamese" evolves from the concept of conjoined twins, analogously, a Siamese Neural Network is comprised of two or more identical sub-networks sharing parameters. It has been widely employed in different research communities, to measure the similarity between two instances. For instance, in image analysis domain, Chopra et al. [36] introduced a Siamese neural network to assess semantic similarity between two processed images for face verification. Varior et al. [37] applied the deep Siamese convolutional neural network (CNN) architecture to human re-identification, aiming to project similar paired images closer to each other and dissimilar paired images farther away from each other. In video analysis domain [38], Ryoo et al. [39] employed a multi-Siamese CNN to identify activities in low resolution videos. Tao et al. [40] designed a Siamese instance search tracker to match the target in the first frame with candidates in other frames. In language analysis domain, Mueller et al. [41] proposed a Siamese adaptation of the Long Short Term Memory (LSTM) network to assess semantic similarity between variable-length sentences.

Even though this concept has been studied before, it has not been studied in the field of weakly-supervised video moment retrieval. In this paper, we leverage a multi-scale Siamese module to enhance the visual and textual alignment.

D. Multiple Instance Learning

The multiple instance learning was originally proposed by Keeler et al [17] for handwritten digit recognition. It treats the element set as a "bag" and usually possesses weaker supervision, i.e., only have set-level labels. The ability to predict groups of elements with only weak supervision is a feature that makes MIL methods attractive for solving some real-world problems. The set-level predictions can be defined as a function defined for a particular example, or as a function defined for the association of all examples. Particularly, the standard MIL supposes a set is positive if it contains at least one positive element, otherwise the set is negative. Follow

this assumption, Wang et al. [42] leveraged a neural network to learn a set representation and directly carried out set classification without estimating instance-level probabilities or labels. Ilse et al. [43] proposed an attention mechanism to learn a pooling operation over instances. The weight learned from the attention mechanism of the instances can be used as an indicator of each instance's contribution to the final decision, resulting in interpretable predictions. Similarly, Liu et al. [44] adopted computational set representation to measure the distance between sets of images. Yan et al. [45] proposed to update the contributions of all instances of the collection by looking at all instances every predefined number of iterations.

Even though MIL has been studied in the field of weakly-supervised video moment retrieval, most existing approaches have not optimized MIL for the current task. Concretely, they simply summarized all the corresponding scores of moments as video-query matching score. This may induce noisy information due to the plentiful query-unrelated moments in the video. To tackle this defect, we utilized IoU scores as the prior knowledge to weaken the influence of query-unrelated moments and improve the retrieval performance.

III. Our Proposed Method

As Fig. 2 illustrates, our proposed SAN comprises of the following components: 1) the feature representation component encodes both visual and textual information and then projects the output embeddings into a joint embedding space; 2) the multi-scale Siamese module generates multi-scale moment candidates online and enhances fine-grained cross-modal semantic alignment; and 3) the context-aware MIL module leverages adjacent moment candidates to facilitate the video-query similarity estimation.

A. Problem Formulation

Let $\mathcal{V}$ and $\mathcal{Q}$ denote a video and a query, respectively. We represent a video as a sequence of frames $\mathcal{V} = \{v_i\}_{i=1}^{n_v}$, where v_i represents the i-th video frame and n_v denotes the number of video frames. Meanwhile, we represent the query as a sequence of words $\mathcal{Q} = \{q_i\}_{i=1}^{n_q}$, where q_i denotes the i-th word and n_q is the number of words. The query is affiliated with a temporal annotation $\tau = \{t_s, t_e\}$ (i.e., the timestamps of the target video moment), where t_s and t_e represent the start and end point of the target video moment, respectively. In this work, our task is to identify a golden moment $c = \{v_t\}_{t=\hat{t}_s}^{\hat{t}_e}$ from the video $\mathcal{V}$ corresponding to the given query $\mathcal{Q}$ under the weakly-supervised setting, whereby $(\hat{t}_s, \hat{t}_e) = (t_s, t_e)$. In other words, the temporal annotation information are unavailable during our training, but they are available during performance evaluation. As such, the weakly-supervised video moment retrieval problem can be formally defined as:

Input: A training set merely contains untrimmed videos and their corresponding queries.

Output: A retrieval model generating moment candidates and estimating the relevance score of each moment-query pair.

B. Feature Representation

1) Video Representation: Given the untrimmed video $\mathcal{V} = \{v_i\}_{i=1}^{n_v}$, we adopt the pre-trained C3D [46] network to extract frame features as [13] did. To be specific, we first extract frame features from the penultimate Fully-Connected (FC) layer, and then we sample the sequence of frame features to the fixed length by linear interpolation. Afterward, we utilize a FC layer to project these obtained frame features into the joint embedding space, obtaining the final frame representations, denoted as $\mathbf{F}^v = \{\mathbf{f}_i^v\}_{i=1}^{n_v}$, where $\mathbf{f}_i^v \in \mathbb{R}^d$ is the final representation of the i-th frame and d denotes the dimension of the joint embedding space.

2) Query Representation: For the given query $\mathcal{Q} = \{q_i\}_{i=1}^{n_q}$, we primarily leverage the pre-trained GloVe [47] to extract initial word embeddings. Furthermore, we feed the initial word embedding sequence into the Gated Recurrent Units (GRU) [48], to capture the contextual information of each word. Finally, we project these context-aware word features into the joint embedding space by using a FC layer with Batch Normalization and Dropout, obtaining the final word representations $\mathbf{F}^q = \{\mathbf{f}_i^q\}_{i=1}^{n_q}$, where $\mathbf{f}_i^q \in \mathbb{R}^d$.

C. Multi-scale Siamese Module

Most existing methods, such as [13] and [14], adopt the multi-scale sliding window strategy to generate moment candidates in an offline fashion, lacking of flexibility. Although the semantic alignment between the moment and the description query can be enhanced by existing work, the heterogeneity gap remains a challenging problem, which needs further studies to conquer. Inspired by these problems, we present a Multi-scale Siamese Module (MSM), comprising two branches: visual and textual branch, as shown in Fig. 2. We will detail each of branch in the following.

Visual Branch. Our task is to identify a golden moment corresponding to the description of the given query. Towards this end, we first need to obtain moment candidates. In our work, we introduce a hierarchical convolution module, which is comprised of multiple one-dimensional convolution layers, wherein each convolution layer could output a scale of video moment candidates as well as their corresponding representations. To guarantee the training process more consistent and enhance the representation of each candidate, we add a residual mapping on top of each convolution layer. The aforementioned process can be formulated as follows,

$$\begin{cases} \mathbf{M}_l^v = ReLU(Conv1d(\hat{\mathbf{M}}_{l-1}^v, \theta_l^k, \theta_l^s)), \\ \hat{\mathbf{M}}_l^v = \mathbf{M}_l^v + MaxPool(\hat{\mathbf{M}}_{l-1}^v, \omega_l^k, \omega_l^s), \end{cases} \tag{1}$$

where θ_l^k and θ_l^s respectively represent the kernel size and stride size of the l-th convolutional layer, ω_l^k and ω_l^s are the kernel size and stride size of the l-th MaxPool layer, $\hat{\mathbf{M}}_l^v$ is the final output of the l-th layer, as well $\hat{\mathbf{M}}_{l-1}^v$ ($l \geq 1$) is the input feature matrix of the l-th layer and $\hat{\mathbf{M}}_0^v = \mathbf{F}^v$. Note that both the kernel and stride size of the max pooling operation are the same as that of the convolution layer.

Actually, different convolution layer has different size of receptive field, $\hat{\mathbf{M}}_l^v$ hence stores moment candidates with

different temporal boundaries. In other words, we can obtain moment candidates with various temporal durations by simply adjusting θ_l^k and θ_l^s. Concretely, we could estimate the temporal boundary of each moment candidate as follows,

$$\begin{cases} w_l = \prod_{i=1}^{l} \theta_i^s, \\ r_l = r_{l-1} + (\theta_l^k - 1) * w_{l-1}, \\ t_{l,j}^s = w_l * j, \\ t_{l,j}^e = t_{l,j}^s + r_l, \end{cases} \tag{2}$$

where w_l is the accumulated stride size of previous l-th layer, r_l is the receptive field size of the l-th layer, as well $t_{l,j}^s$ and $t_{l,j}^e$ is the corresponding start and end points of the j-th moment in the l-th layer.

After executing the hierarchical convolution module, we acquire representations of moment candidates and their corresponding temporal boundaries. Thereafter, we gather moment candidates from all layers and then feed them into a FC layer with ReLU activation function, outputting representations of moment candidates $\mathcal{C}^v = \{\mathbf{m}_i^v\}_{i=1}^{n_m}$, and their corresponding temporal boundaries $\mathcal{T} = \{(t_i^s, t_i^e)\}_{i=1}^{n_m}$, where n_m is the number of moment candidates, $\mathbf{m}_i^v$ is the representation of the i-th moment candidate, as well as (t_i^s, t_i^e) represents the start and end point of the i-th moment candidate.

Textual Branch. To enhance the fine-grained semantic alignment between the moment and query, we introduce a Siamese structure. Particularly, we employ an identical hierarchical convolution module to process the query, which shares parameters with the visual branch. By doing so, we could capture aligned semantics between two branch, further reducing the semantic gap. To be specific, we first employ a frame-by-word attention mechanism to transfer the textual information into the visual space, obtaining the frame-specific textual representation. Thereby, the input of the textual branch has the same shape with that of the visual branch. Formally, we summarize the above process as follows,

$$\begin{cases} \hat{\mathbf{f}}_i^q = \sum_{j=1}^{n_q} a_{i,j} \mathbf{f}_j^q, \\ a_{i,j} = \frac{exp(z_{i,j})}{\sum_{k=1}^{n_q} exp(z_{i,k})}, \\ z_{i,j} = \mathbf{f}_i^{vT} \mathbf{f}_j^q, \end{cases} \tag{3}$$

where $a_{i,j}$ represents the the normalized attention weight of the i-th frame with respect to the j-th word and $\hat{\mathbf{f}}_i^q$ denotes the textual representation related to the i-th visual frame. From this knowable, the sequence of obtained textual representation has the same length as the input sequence of the visual branch. More importantly, each textual representation is semantically associated with the corresponding frame representation.

Afterwards, we feed the sequence of textual representations into an identical hierarchical convolution network, obtaining the moment-specific textual representation as follows,

$$\begin{cases} \mathbf{M}_l^q = ReLU(Conv1d(\hat{\mathbf{M}}_{l-1}^q, \theta_l^k, \theta_l^s)) \\ \hat{\mathbf{M}}_l^q = \mathbf{M}_l^q + MaxPool(\hat{\mathbf{M}}_{l-1}^q, \omega_l^k, \omega_l^s), \end{cases} \tag{4}$$

where $\hat{\mathbf{M}}_{l-1}^q$ is the input of the l-th layer, $\hat{\mathbf{M}}_0^q = \hat{\mathbf{F}}^q$, as well $\hat{\mathbf{M}}_l^q$ is the output feature map of the l-th layer. Likewise, we gather moment-specific textual representations from all layers into a set, and then feed them into a FC layer with ReLU activation function, which shares parameters with the visual branch. In this paper, we denote the corresponding output as $\mathcal{C}^q = \{\mathbf{m}_i^q\}_{i=1}^{n_m}$, where $\mathbf{m}_i^q$ is the textual representation of the i-th moment candidate.

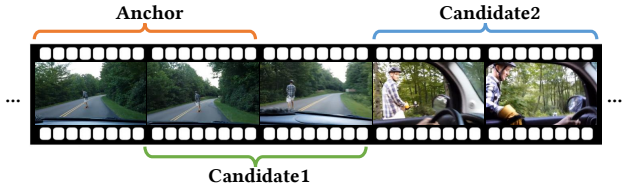

Fig. 3. Intuitively Illustrating the usefulness of contextual moments.

Taking full advantage of the Siamese structure, our model could enhance the semantic alignment of different granularity by persistently projecting visual and textual representations into a joint embedding space and constraining the model to capture consistent semantics from two modalities. Having obtained the representations of moment candidates and their corresponding query representation, i.e., $\mathcal{C}^v$ and $\mathcal{C}^q$, we could calculate the moment-query alignment score by adopting the cosine similarity as follows,

$$s_i = \frac{\mathbf{m}_i^{vT} \mathbf{m}_i^q}{\|\mathbf{m}_i^v\| \|\mathbf{m}_i^q\|}. \tag{5}$$

D. Context-aware MIL

To retrieve the target moment by merely considering video-level annotations, we propose a Context-aware MIL (C-MIL) module, which jointly enhances the alignment of the moment-query and video-query by considering contextual moment candidates.

1) Enhanced moment-query alignment: Previous work estimates alignment score of each moment-query pair by just measuring the similarity of their representations. Unlike them, in this work, we aggregate alignment scores of contextual moments to strengthen the alignment score of the current moment. This is because the moments with higher temporal Intersection-over-Union (IoU) values are complementary to some extent. As demonstrated in Fig. 3, compared with the Candidate2, the Candidate1 has higher IoU value with the anchor moment. Meanwhile, the visual information of Candidate1 is more similar with the anchor, and it is complementary to the comprehension of the anchor. In other words, the alignment scores of contextual moments can compensate for the current moment. Inspired by this, we first calculate the IoU score between a pair of moment candidate as follows,

$$\begin{cases} I_{i,j} = max[min(t_i^e, t_j^e) - max(t_i^s, t_j^s), 0], \\ U_{i,j} = (t_i^e - t_i^s) + (t_j^e - t_j^s) - I_{i,j}, \\ \delta_{i,j} = \frac{I_{i,j}}{U_{i,j}}, \end{cases} \tag{6}$$

where $I_{i,j}$ and $U_{i,j}$ are the intersection and union between i-th and j-th moment candidate, t_i^s and t_i^e are the start and end point of the i-th moment candidate, as well t_j^s and t_j^e are the start and end point of the j-th moment candidate. Afterwards,

we strengthen the alignment estimation of each moment-query pair by considering its contextual moments as follows,

$$\begin{cases} \alpha_{i,j} = \frac{exp(\delta_{i,j})}{\sum_{k=1}^{n_c} exp(\delta_{i,k})}, \\ \hat{s}_i = s_i + \lambda \sum_{j=1}^{n_c} \alpha_{i,j} s_j, \end{cases} \tag{7}$$

where s_i is the original alignment score of i-th moment, $\hat{s}_i$ is the strengthened score, λ is a hyper-parameter to control the balance of two scores, and n_c is the number of moment candidates except the i-th moment, i.e., $n_c = n_m - 1$.

2) Video-query alignment: To estimate the matching score between each video and the sophisticated query, we treat each video as a "bag" and each moment candidate as an instance of the "bag". A direct way to obtain "bag" level alignment score is to select the highest score among instances from the bag. Although feasible, solely considering the highest-scored moment candidate is inadequate. Because the model has poor learning ability at the beginning under the weakly-supervised setting, inducing imprecise predictions. Thereby, simply selecting highest-scored instance may introduce negative influence to the model. Although previous work [13], [14] tries to alleviate this issue by summarizing scores of all instance in the bag, they may introduce noise information due to they treat all instances equally.

To solve the above-mentioned problems, we present a novel video-query alignment score estimation strategy. Particularly, we first select the highest-scored moment as the anchor, namely choosing the most relevant candidate from the "bag". Afterwards, we utilize the IoU scores with respect to the anchor, as the guidance information, to assess the confidence of each moment candidate. More concretely, we first calculate IoU scores between the anchor moment and other moment candidates as follows,

$$\begin{cases} I_i = max[min(t^e, t_i^e) - max(t^s, t_i^s), 0], \\ U_i = (t^e - t^s) + (t_i^e - t_i^s) - I_i, \\ \delta_i = \frac{I_i}{U_i}, \end{cases} \tag{8}$$

where I_i and U_i are the intersection and union between i-th moment candidate and anchor moment, t^s and t^e are respectively the start and end point of the anchor moment, as well t_i^s and t_i^e are the start and end point of the i-th moment candidate.

Afterwards, we estimate the alignment score of the video-query pair by weighted aggregating candidates with IoU scores as follows,

$$\begin{cases} \alpha_i = \frac{exp(\delta_i)}{\sum_{k=1}^{n_c} exp(\delta_k)}, \\ S = \hat{s} + \mu \sum_{i=1}^{n_c} \alpha_i \hat{s}_i, \end{cases} \tag{9}$$

where $n_c = n_m - 1$ is the number of moment candidates exclude the anchor, $\hat{s}$ is the strengthened score of the anchor moment, S is the alignment score of the given video-query pair, and μ is a balance hyper-parameter.

E. Objective Function

To achieve semantic alignment of a given positive video-query pair (V, Q), we utilize the margin-based ranking loss for optimization, which is defined as,

$$\begin{aligned} L = \sum_{(V,Q)} \{ & \sum_{V^-} max[0, \triangle - S(V,Q) + S(V^-,Q)] \\ & + \sum_{Q^-} max[0, \triangle - S(Q,V) + S(Q^-,V)] \}, \end{aligned} \tag{10}$$

where $\triangle$ is the margin, as well as (Q^-, V) and (V^-, Q) are negative pairs. Specifically, (Q^-, V) is constructed by selecting any query from a different video in the mini-batch, while (V^-, Q) is built by adopting any video from a different query in the mini-batch.

IV. Experiment

In this section, we perform experiments on two benchmark datasets to verify the effectiveness of our proposed model. Specifically, we first introduce the datasets and the evaluation protocols followed by the implementation details. And then we report the results on the Charades-STA [1] and ActivityNet-Captions [18] datasets, respectively. Afterwards, we justify the usefulness of the Siamese network as well as the C-MIL module. Finally, we display several qualitative results.

A. Datasets

Charades-STA [1]. This dataset was first proposed by Sigurdsson et al. [49], including action interval annotations and video-level language descriptions. Later, Gao et al. [1] extended the annotation information for video moment retrieval. Specifically, they decomposed the video-level description into short sentences, and then assigned these sentences to video clips according to the activity category information. The average length of the query is 8.6 words, and the average duration of the video is 29.8 seconds. In total, the Charades-STA dataset contains 16,128 video-query pairs. 12,408 of them are set as the training data and the rest ones are testing data.

ActivityNet-Captions [18]. It is currently the largest dataset for video moment retrieval, consisting of 19,209 videos and 17,031 language queries. Each of video-query pair has corresponding time annotations. In this paper, we followed the settings in previous work, namely using val1 as the validation set and val2 as the testing set. The average length of the language query is 13.16 words, and the average duration of the video is 117.74 seconds.

B. Experimental Settings

1) Evaluation Protocols: To thoroughly measure our model and the baselines, we adopt "R@n, IoU=m" utilized in [13] as the evaluation metric. To be more specific, given a query, it is the percentage of top-n results having IoU larger than m, i.e., $R(n,m) = \frac{1}{k}\sum_{i=1}^{k} r(n,m,q_i)$, where $r(n,m,q_i)$is the recall rate of a query q_i, and k is the number of queries.

2) Implementation Details: We optimized our proposed model on 1 TITAN XP GPU using PyTorch library. The Adam optimizer is employed with 15 epochs. The learning rate is set as 5e-5 with cosine annealing adjustment [50], where the end learning rate is set as 5e-7. We set the batch size for both Charades-STA and ActivityNet-Captions to 32. Moreover, for

TABLE I
PERFORMANCE COMPARISON BETWEEN OUR PROPOSED MODEL AND THE STATE-OF-THE-ART BASELINES ON CHARADES-STA. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Supervision	Method	R@1			R@5		
		IoU = 0.7	IoU = 0.5	IoU = 0.3	IoU = 0.7	IoU = 0.5	IoU = 0.3
Supervised	CTRL [1]	8.89	23.63	-	29.52	58.92	-
	ROLE [2]	7.82	21.74	37.69	30.06	70.37	92.79
	QSPN [4]	15.80	35.60	54.70	45.40	79.40	**95.60**
	2D-TAN [5]	23.25	39.81	-	52.15	79.33	-
	DPIN [6]	26.96	47.98	-	55.00	85.53	-
	PFGA [8]	33.74	52.02	**67.53**	-	-	-
	FIAN [10]	**37.72**	**58.55**	-	**63.52**	**87.80**	-
Weakly-supervised	TGA [13]	5.73	15.13	27.74	25.22	56.83	83.23
	SCN [15]	9.97	23.58	42.96	38.87	71.80	**95.56**
	BAR [16]	12.23	27.04	44.97	-	-	-
	WSTAN [23]	12.28	29.35	43.39	41.53	**76.13**	93.04
	SAN	**13.12**	**31.02**	**51.02**	**41.75**	72.56	89.95

TABLE II
PERFORMANCE COMPARISON BETWEEN OUR PROPOSED MODEL AND THE STATE-OF-THE-ART BASELINES ON ACTIVITYNET-CAPTIONS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Supervision	Method	R@1			R@5		
		IoU = 0.7	IoU = 0.5	IoU = 0.3	IoU = 0.7	IoU = 0.5	IoU = 0.3
Supervised	QSPN [4]	13.60	27.70	45.30	38.30	59.20	75.70
	PFGA [8]	19.26	33.04	51.28	-	-	-
	2D-TAN [5]	27.38	44.05	58.75	**62.26**	76.65	85.65
	DPIN [6]	28.31	47.27	62.40	60.03	77.45	87.52
	CSMGAN [11]	29.15	**49.11**	**68.52**	59.63	77.43	**87.68**
	FIAN [10]	**29.81**	47.90	64.10	59.66	**77.64**	87.59
Weakly-supervised	WSLLN [14]	-	22.70	42.80	-	-	-
	SCN [15]	-	29.22	47.23	-	55.69	71.45
	BAR [16]	-	**30.73**	49.03	-	-	-
	WSTAN [23]	-	30.01	**52.45**	-	63.42	79.38
	SAN	**13.85**	30.54	48.44	**29.96**	**64.52**	**82.41**

each dataset, we employed linear interpolation to ensure the sequence length of videos are same. Specifically, the length of Charades-STA is 31, while that of the ActivityNet-Captions is 127. Besides, for Charades-STA, we set the layer number of multi-scale Siamese module to 3, all kernel sizes to 3, and stride sizes to 2. For ActivityNet-Captions, the corresponding values are 4, 7, and 2, respectively. For both datasets, the dimension of joint embedding space is 512. The output dimension of all FC layers and convolution layers of multi-scale Siamese module is 512. The margin of the triplet loss is set as 0.2 for Charades-STA and 0.1 for ActivityNet-Captions. The hyper-parameter λ and μ is set to 0.7 and 0.3 for Charades-STA. For ActivityNet-Captions, the corresponding values are 0.7 and 1.0, respectively. Note that we only strengthened moment-query alignment scores during training.

During inference, we first fed video-query pairs into SAN model and obtained moment-query alignment scores. Afterwards, a soft Non-Maximum Suppression (NMS) is applied to filter out redundant moment candidates. Finally, we ranked all remaining candidates to obtain the target moment.

3) Baselines: Following [22], in this paper, we compared SAN with the following state-of-the-art baselines to justify the effectiveness of our proposal:

- **TGA** [13]: This method first generates moment candidates by sliding windows. And then it introduces a text-guided attention module to obtain the attention weight, which is used to calculate the query-wise video feature. Afterwards, both obtained video and query representations are projected into a joint embedding space by two FC layers.
- **WSLLN** [14]: It is a two branch network including the alignment branch and detection branch, which trains with pseudo segment-level labels.
- **SCN** [15]: It scores all the moments sampled at different scales in a single pass. Specifically, it introduces the semantic completion module to measure the semantic similarity between moments and the query. Meanwhile, it designs a reconstruction loss to train the semantic completion module, enforcing the model to extract key visual information.
- **BAR** [16]: This work resorts to reinforcement learning to guide the process of progressively refining the temporal boundary. Besides, it employs a MIL-based alignment evaluator to measure the alignment score of each moment-query pair, providing tailor-designed rewards.
- **WSTAN** [22]: It learns cross-modal semantic alignment by exploiting temporal adjacent network, with a whole description paragraph as input.
- **Supervised Methods:** For a more complete comparison, we consider some representative supervised learning methods, including CRTL [1], ROLE [2], QSPN [4], 2D-TAN [5], DPIN [6], PFGA [8], FIAN [10], and CSMGAN [11].

Note that we directly quoted the results of these baselines from their original papers due to the source codes and involved parameters are not released by the authors except TGA.

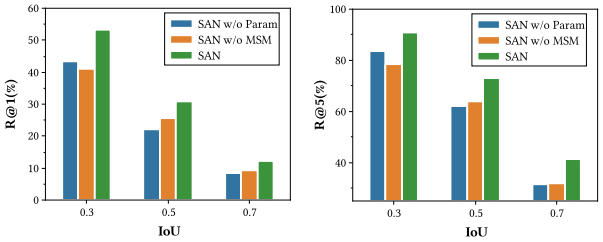

Fig. 4. Performance comparison among MSM-related variants on the Charades-STA dataset.

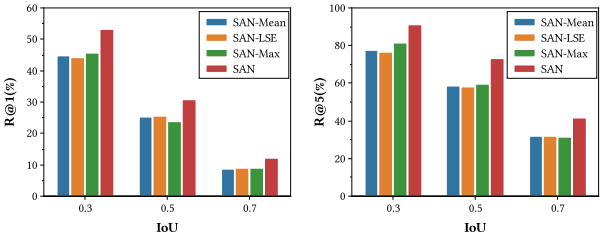

Fig. 5. Performance comparison among our model variants related to the C-MIL module on Charades-STA dataset.

C. Performance Comparison

The comparison results on Charades-STA and ActivityNet-Captions are summarized in TABLE I and TABLE II, respectively. By jointly analyzing these tables, we gained the following observations:

- TGA achieves poor performance on Charades-STA since it overlooks the fine-grained interactions modeling between visual and textual modalities. Moreover, WSLLN scored worst on ACticityNet Captions. This mainly because it summarizes the scores of all moment-query pairs to obtain the video-query alignment score, introducing noise information.
- SCN shows consistent improvements over TGA and WSLLN, verifying the importance of considering the context information. BAR outperforms TGA and WSLLN on two datasets. The observed results make sense since it resorts to a tailor-designed reinforcement learning paradigm to adaptively optimize the temporal boundary towards shrinking the cross-modal semantic gap. WSTAN models the temporal relations by a 2D feature map, where one dimension indicates the starting time of a moment and the other indicates the end time, and obtains best performance in all baseline models on two datasets.
- Our SAN achieves superior performance, and the results are competitive to these baselines on two datasets. Specially, compared with TGA, SCN, BAR and WSTAN, our SAN obtains relative "R@1 IoU=0.7" gains with 48.42%, 31.60%, 7.28% and 6.84%on Charades-STA, respectively. On ActivityNet-Captions, we also have an improvement of 15.86% as compared to SCN in terms of "R@5 IoU=0.5".
- Compared with weakly-supervised approaches, supervised ones yield significant gains in retrieval performance on both two datasets. Because they have more abundant supervise information. Nevertheless, we can see that SAN still outperforms some supervised methods. For instance, on Charades-STA, it has an improvement of 47.58% and 67.78% compared to CTRL and ROLE in terms of "R@1 IoU=0.7", respectively.

D. Study of SAN

In this section, we carried out several experiments on Charades-STA to further analyze the effectiveness of our model. Specifically, we first explored how the multi-scale Siamese module affects the moment retrieval results. We then displayed how the context-aware MIL influences the retrieval performance.

TABLE III
EFFECTS OF THE RESIDUAL MAPPING ON CHARADES-STA DATASET.

Method	R@1		
	IoU = 0.7	IoU = 0.5	IoU = 0.3
w residual	13.12	31.02	51.02
w/o residual	6.80	19.70	41.91

1) The Multi-scale Siamese Module: To justify the effectiveness of MSM discussed in Section 3.3, we experimented with two variants: 1) **SAN w/o Param**, without sharing parameters for both the visual and textual branches; and 2) **SAN w/o MSM**, removing the textual branch from the multi-scale Siamese module[2].

We tested these model variants on the Charades-STA dataset. And the component-wise comparison results are shown in Fig. 4. From these results, we have the following findings:

- Compared with our model, the performance of SAN w/o Param degrades dramatically. Particularly, It drops absolutely by 18.66%, 28.50%, and 31.23% on R@1, respectively. This demonstrates the vital importance of sharing parameters for both the visual and textual branch, namely the Siamese structure. Because the Siamese Network can enforce the alignment between the visual and textual information, therefore improving the similar estimation between the moment and the given query.
- Our model achieves better results than SAN w/o MSM, revealing that the Siamese structure can well capture the textual information related to the corresponding video moment and boost the model performance.
- In general, our proposed model largely exceeds all variants on all evaluation metrics, verifying the effectiveness and necessity of our multi-scale Siamese module for weakly-supervised video-moment retrieval.

As mentioned above, we added max pooling layers as residual mapping on top of each convolution layer to make training consistent. To verify the effect of the residual mapping, we experimented a variant of MSM without residual. As TABLE III shown, getting rid of the residual mapping will

[2]Concretely, having obtained multi-scale moment candidates from the visual branch, we first utilized the frame-by-word attention introduced in Section 3.3 to learn the visual-specific query representation for each candidate. Afterwards, we directly utilized cosine similarity to estimate the moment-query alignment score.

TABLE IV
ABLATION STUDY OF THE C-MIL IN TGA ON CHARADES-STA DATASET.

Methods	R@1		
	IoU = 0.7	IoU = 0.5	IoU = 0.3
TGA	5.73	15.13	27.74
TGA-CMIL	6.21	16.10	29.54

cause a 48% drop on "R@1 IoU=0.7". This shows that residual mapping plays an important role in keeping training consistent.

2) The Context-aware MIL: We experimented with variants of our model to verify the effectiveness of the C-MIL:

- **SAN-Mean**: We utilized the average of all moment-query scores to represent the video-query alignment score.
- **SAN-LSE**: We leveraged LogSumExp (LSE) pooling [51] to weighted average the moment-query scores, obtaining the video-query alignment score.
- **SAN-Max**: We directly selected the highest score of moment-query pairs as the video-query matching score.

We evaluated these model variants on the Charades-STA dataset. And the component-wise comparison results are shown in the Fig. 5. From Fig. 5, we observed that:

- SAN outperforms SAN-Mean by a large margin in terms of all metrics. It reveals that simply operating average aggregation for moment-query pairs is insufficient to represent the video-query alignment score. As average aggregation assumes that the moments are linearly independent and equally contributing to the final relevance estimation. It hence fails to identify the adaptive importance of each moment and hardly weaken the irrelevant even noisy information. The improvement achieved by SAN verifies the effectiveness of the C-MIL.
- Although SAN-LSE executes the weighted summarize operation to obtain the video-query similarity score, it achieves similar results to the SAN-Mean. This may be because that it uses the probability distribution of scores as weights, leading to some visually dissimilar and query-unrelated moments are given the equal weight. The performance drop of SAN-LSE indicate that it is crucial to consider the context information as the guidance to enhance the video-query alignment.
- The performance drop of SAN-Max can be observed, revealing that solely considering the highest-scored moment candidate is inadequate. Because the model has poor learning ability at the beginning under the weakly-supervised setting, inducing the predictions are imprecise. Thereby, simply selecting highest-scored instance may lead to a negative impact on model training.
- SAN shows consistent improvement over all variants on Charades-STA, verifying the crucial influence of considering context information. Particularly, in this paper, we not only consider context information to enhance the moment-query alignment but also promote the video-query alignment. All these comparison results fully demonstrate the availability and necessity of our proposed C-MIL.

To further verify the effect of C-MIL, we evaluated a variant of TGA on Charades-STA dataset, named TGA-CMIL. To be specific, TGA introduces a text-guided attention module to weight all candidate features, obtaining the global video feature. As the replacement, TGA-CMIL only weights the largest-scored moment and the adjacent moments which have non-zero IoU scores with it. As TABLE IV shown, TGA-CMIL obtains relative "R@1 IoU=0.7" gains with 8.4% on Charades-STA. It proves the effectiveness of considering contextual information. The slightly poor performance of TGA may due to it summarizes information from all candidates to estimate video-query similarity. In contrast, the C-MIL module utilizes IoU scores as prior knowledge to weaken the influence of query-unrelated moments.

TABLE V
PERFORMANCE OF OUR MODEL WITH DIFFERENT LAYER NUMBERS ON CHARADES-STA DATASET.

Method	R@1		
	IoU = 0.7	IoU = 0.5	IoU = 0.3
3	13.12	31.02	51.02
2	12.07	30.97	43.79
1	1.34	8.52	31.59

E. Parameter Analysis

In this section, we explored the influence of four pivotal parameters.

1) Kernel Size: We first explored the influence of kernel sizes in MSM on Charades-STA. To be specific, we first empirically initialized the kernel sizes of all layers to 3. Afterwards, we verified four kinds of kernel sizes for each layer by setting the kernel sizes of all its previous layers to the best verified values and others to the default values. In this paper, we select the "R@1 IoU=0.7" as the main evaluation indicator. The experimental results are shown in Fig. 6. From Fig. 6, we can see that setting all kernel sizes to 3, the model obtains the best performance in terms of "R@1 IoU=0.7". Moreover, the performance of SAN changes within small ranges nearby this setting, demonstrating that SAN is non-sensitive to the parameters around optimal settings.

2) Layer Number: In addition, we further explored the effect of the layer number of MSM. Due to the stride sizes are set to 2, the maximum number of layers is 3 on Charades-STA. The experimental results of different layer numbers are shown in TABLE V. We can see that the performance drops sharply as the layer number decreases. This may be because reducing the number of layers decreases the number of candidates as well as the duration scales of moment candidates, therefore influencing the retrieval accuracy.

3) Balance Parameters: We also explored the effect of balance parameters. We first set λ randomly based on experience and then explored μ. After that, we set μ to the best value and then explored λ. Fig. 7 shows the performance of our model regarding the two parameters, which is accomplished by varying one and fixing another. We found that the performance increases first and then decreases with the increase of parameters. It is observed that the setting of $\mu = 0.3$ and $\lambda = 0.7$ works well on Charades-STA.

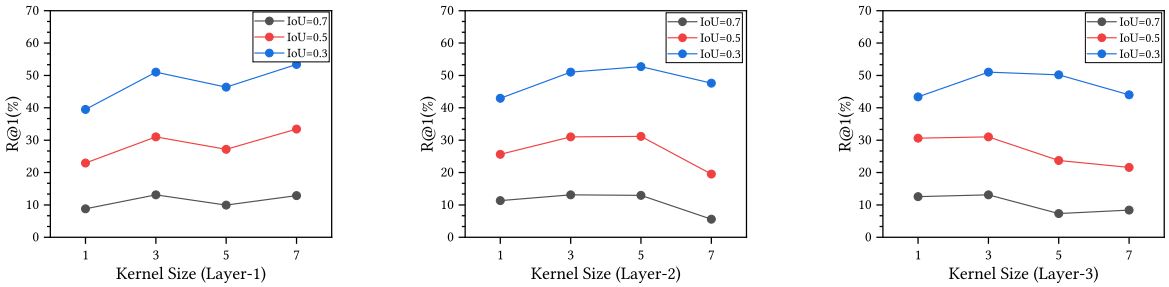

Fig. 6. Justification the influence of the kernel size in the multi-scale Siamese module. The first figure displays the R@1 results over various kernel sizes of Layer-1 by setting kernel sizes of Layer-2 and Layer-3 to 3, respectively. The second one shows the R@1 results over various kernel sizes of Layer-2 by setting kernel size of Layer-1 to the best setting in the first figure and that of Layer-3 to 3. The last one illustrates the R@1 results over various kernel sizes of Layer-3 by setting kernel sizes of Layer-1 and Layer-2 to the best setting in the first and second figures, respectively.

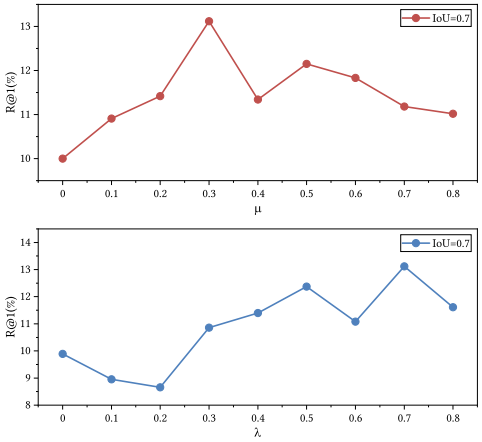

Fig. 7. Parameter tuning and sensitivity analysis of balance hyper-parameters μ and λ on Charades-STA. This is implemented by varying one parameter and fixing another.

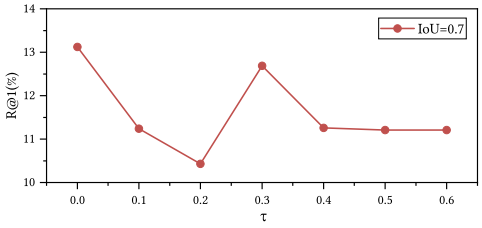

Fig. 8. Parameter tuning and sensitivity analysis of IoU threshold τ on Charades-STA.

4) IoU Threshold: In C-MIL module, all candidates are considered as contextual moments to strengthen the video-query alignment. To further explore the influence of the range of contextual moments, we introduced an extra IoU threshold τ. In particular, we dropped the candidates whose IoU score is less than τ when calculating the aggregating weights. As shown in Fig. 8, the performance shows a decreasing trend with the increase of τ, and the setting of $\tau = 0$ (i.e., considering all candidates) achieves the best performance on Charades-STA.

F. Visualization Results

1) Frame-by-word Attention Weights: To validate the effectiveness of the frame-by-word attention, we visualized an example from the Charades-STA in Fig. 9. As mentioned above, the sequence length of a video is 31 when it feeds into the frame-by-word attention module. For visualization purposes, we utilized max pooling operation to down sampling the attention weights of video and obtained six weights for each word. Each row in Fig. 9 represents the weights of a video moment, and we selected the central frame as the overlay to be displayed. From Fig. 9, we can see that the moment which has the largest IoU score with the ground truth obtains higher attention weights on "walks", "over" and "refrigerator", which can help to retrieve the query-related moment. Moreover, the forth moment which the person walks back to the camera also has higher attention weight on the word "walks". This reflects the frame-by-word attention has keen ability to capture the movements of persons. This may profit from the pre-training on the action detection task.

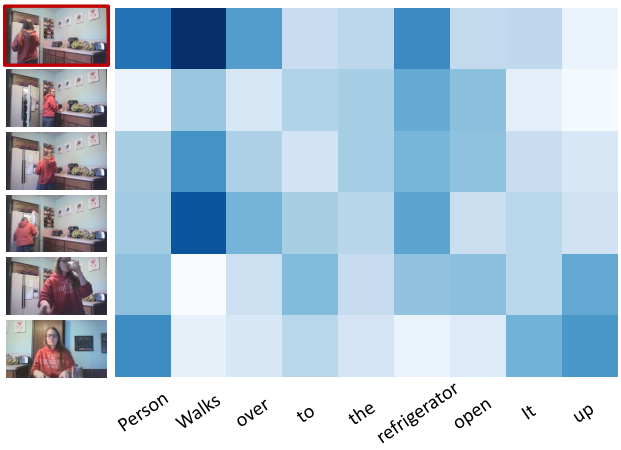

Fig. 9. Visualization of frame-by-word attention weights. Each row represents the weights of one down-sampled video moment over all words, and each column represents the weights of all moments over one specific word. The darker color indicates the higher weight. The moment has the largest IoU score with the ground truth is marked in red box.

2) Moment Alignment Scores: One key advantage of SAN over other methods is that its C-MIL module is able to enhance the moment-query alignment. Towards this end, we selected one video-query pair from the Charades-STA dataset, and visualized moment-query alignment values estimated by Eqn. (5) and Eqn. (7) in Fig. 10. We could find that: 1) compared to moment alignment scores shown in Fig. 10 (a), the moment alignment scores displayed in Fig. 10 (b) are more continuous in temporal and the position of the high-

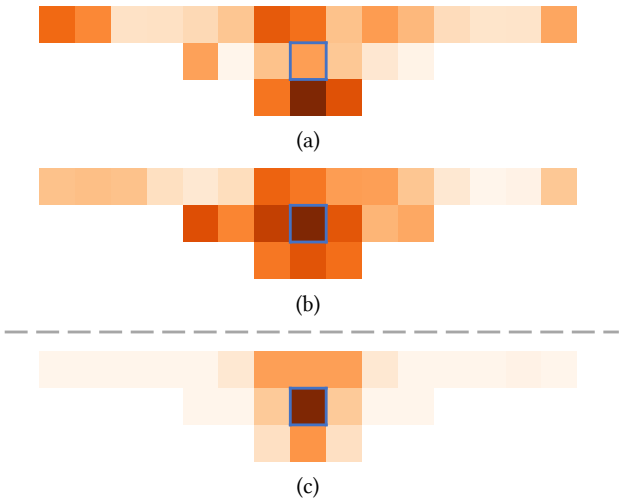

Fig. 10. (a) shows the initial moment alignment scores predicted by Eqn. (5). (b) shows the context-enhanced moment alignment scores predicted by Eqn. (7). And (c) displays the IoU-based weights of moment candidates with respect to the ground truth moment marked in blue box.

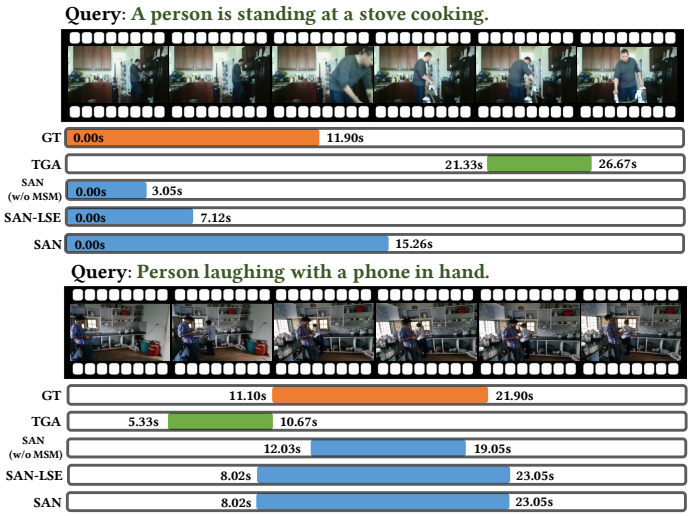

Fig. 11. Moment retrieval results on Charades-STA. All of the above figures are the R@1 results.

scored moments are more concentrated; 2) for the anchor moment (i.e., the ground truth moment) in Fig. 10 (c), it merely leverages its adjacent moments, which have similar and complementary visual information with the anchor moment, to enhance the corresponding moment-query alignment score; and 3) the context-enhanced alignment scores of all moment candidates are diverse. These findings are consistent with our expectation, and further demonstrate that aggregating complementary information from contextual moments would not diminish the distinction of different moment candidates. Hence, this verifies the effectiveness of our C-MIL module.

G. Qualitative Results

To qualitatively validate the effectiveness of our SAN method, we displayed two examples on video moment retrieval in Fig. 11. To analyze the effectiveness of each component, we also displayed the retrieval results of two variants: SAN-LSE and w/o MSM. Meanwhile, we displayed the retrieval results of first weakly-supervised approach TGA [13] for comparison.

From the results shown in Fig. 11, we observe that:

- Both SAN w/o MSM and SAN-LSE achieve better performance as compared to TGA [13]. This reflects that not only MSM but also context-aware-MIL could reduce the semantic gap between different modalities, therefore further improving the accuracy of retrieval.
- Compared with SAN w/o MSM, SAN-LSE achieves superior performance, indicating that the MSM module plays a more prominent role in enhancing the semantic alignment between the visual and textual modality.
- SAN achieves the best performance, substantially surpassing other baselines. This verifies the importance of jointly considering the context-aware MIL and the MSM module to enhance the semantic alignment.

V. Conclusion and Future Work

In this paper, we propose a Siamese Alignment Network for weakly supervised video moment retrieval, which aims to search the most relevant moment according to the given natural language query in a video. To well match the moment candidates with the given query, we design a multi-scale Siamese module to reduce the semantic gap and enhance the fine-grained semantic alignment. Moreover, we advance a context-aware multiple instance learning module to strengthen the video-query alignment, which facilitates the video-query similarity estimation by jointly considering adjacent moments. Extensive experimental results on two benchmark datasets validate the effectiveness of our proposed method.

In future, we plan to enhance visual representations by considering multiple pre-training tasks. Therefore, the generated visual representations would be more sensitive to temporal information.

Acknowledgment

This work is supported by the National Natural Science Foundation of China, No.:62006142, No.:61872270, No.:U1936203; the Shandong Provincial Key Research and Development Program, No.:2019JZZY010118; the Shandong Provincial Natural Science Foundation for Distinguished Young Scholars, No.:ZR2019JQ23, No.:ZR2021JQ26; the Major Basic Research Project of Natural Science Foundation of Shandong Province, No.:ZR2021ZD15; the Young creative team in universities of Shandong Province, No.:2020KJN012; Science and Technology Innovation Program for Distinguished Young Scholars of Shandong Province Higher Education Institutions, No.:2021KJ036.

References

[1] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: temporal activity localization via language query," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2017, pp. 5277–5285.

[2] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T. Chua, "Cross-modal moment localization in videos," in *Proceedings of the International Conference on Multimedia*. ACM, 2018, pp. 843–851.

[3] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T. Chua, "Attentive moment retrieval in videos," in *Proceedings of the International SIGIR Conference on Research Development in Information Retrieval*. ACM, 2018, pp. 15–24.

[4] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proceedings of the American Association for Artificial Intelligence*. AAAI, 2019, pp. 9062–9069.

[5] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proceedings of the American Association for Artificial Intelligence*. AAAI, 2020, pp. 12 870–12 877.

[6] H. Wang, Z. Zha, X. Chen, Z. Xiong, and J. Luo, "Dual path interaction network for video moment localization," in *Proceedings of the International Conference on Multimedia*. ACM, 2020, pp. 4116–4124.

[7] D. Cao, Y. Zeng, M. Liu, X. He, M. Wang, and Z. Qin, "STRONG: spatio-temporal reinforcement learning for cross-modal video moment localization," in *Proceedings of the International Conference on Multimedia*. ACM, 2020, pp. 4162–4170.

[8] C. R. Opazo, E. Marrese-Taylor, F. S. Saleh, H. Li, and S. Gould, "Proposal-free temporal moment localization of a natural-language query in video using guided attention," in *Proceedings of the Winter Conference on Applications of Computer Vision*. IEEE, 2020, pp. 2453–2462.

[9] D. Cao, Y. Zeng, X. Wei, L. Nie, R. Hong, and Z. Qin, "Adversarial video moment retrieval by jointly modeling ranking and localization," in *Proceedings of the International Conference on Multimedia*. ACM, 2020, pp. 898–906.

[10] X. Qu, P. Tang, Z. Zou, Y. Cheng, J. Dong, P. Zhou, and Z. Xu, "Fine-grained iterative attention network for temporal language localization in videos," in *Proceedings of the International Conference on Multimedia*. ACM, 2020, pp. 4280–4288.

[11] D. Liu, X. Qu, X. Liu, J. Dong, P. Zhou, and Z. Xu, "Jointly cross- and self-modal graph attention network for query-based moment localization," in *Proceedings of the International Conference on Multimedia*. ACM, 2020, pp. 4070–4078.

[12] B. Peng, J. Lei, H. Fu, Y. Jia, Z. Zhang, and Y. Li, "Deep video action clustering via spatio-temporal feature learning," *Neurocomputing*, vol. 456, pp. 519–527, 2021.

[13] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury, "Weakly supervised video moment retrieval from text queries," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 11 592–11 601.

[14] M. Gao, L. Davis, R. Socher, and C. Xiong, "WSLLN: weakly supervised natural language localization networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2019, pp. 1481–1487.

[15] Z. Lin, Z. Zhao, Z. Zhang, Q. Wang, and H. Liu, "Weakly-supervised video moment retrieval via semantic completion network," in *Proceedings of the American Association for Artificial Intelligence*. AAAI, 2020, pp. 11 539–11 546.

[16] J. Wu, G. Li, X. Han, and L. Lin, "Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos," in *Proceedings of the International Conference on Multimedia*. ACM, 2020, pp. 1283–1291.

[17] J. D. Keeler, D. E. Rumelhart, and W. K. Leow, "Integrated segmentation and recognition of hand-printed numerals," in *Proceedings of the Conference on Advances in Neural Information Processing Systems*. Morgan Kaufmann, 1990, pp. 557–563.

[18] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2017, pp. 706–715.

[19] J. Chen, X. Chen, L. Ma, Z. Jie, and T. Chua, "Temporally grounding natural sentence in video," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2018, pp. 162–171.

[20] J. Chen, L. Ma, X. Chen, Z. Jie, and J. Luo, "Localizing natural language in videos," in *Proceedings of the American Association for Artificial Intelligence*. AAAI, 2019, pp. 8175–8182.

[21] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos," in *Proceedings of the American Association for Artificial Intelligence*. AAAI, 2019, pp. 8393–8400.

[22] Y. Wang, J. Deng, W. Zhou, and H. Li, "Weakly supervised temporal adjacent network for language grounding," *IEEE Transactions on Multimedia*, 2021.

[23] Y. Song, J. Wang, L. Ma, Z. Yu, and J. Yu, "Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos," *CoRR*, 2020.

[24] Z. Chen, L. Ma, W. Luo, P. Tang, and K. K. Wong, "Look closer to ground better: Weakly-supervised temporal grounding of sentence in video," *CoRR*, 2020.

[25] Z. Zhang, Z. Lin, Z. Zhao, J. Zhu, and X. He, "Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos," in *Proceedings of the International Conference on Multimedia*. ACM, 2020, pp. 4098–4106.

[26] Z. Zhang, Z. Zhao, Z. Lin, J. Zhu, and X. He, "Counterfactual contrastive learning for weakly-supervised vision-language grounding," in *Proceedings of the Conference on Neural Information Processing Systems*. MIT, 2020.

[27] Z. Shou, D. Wang, and S. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 1049–1058.

[28] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2017, pp. 2933–2942.

[29] Y. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 1130–1139.

[30] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: boundary sensitive network for temporal action proposal generation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2018, pp. 3–21.

[31] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: boundary-matching network for temporal action proposal generation," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2019, pp. 3888–3897.

[32] L. Wang, Y. Xiong, D. Lin, and L. V. Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 6402–6411.

[33] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2017, pp. 3544–3553.

[34] J. Zhong, N. Li, W. Kong, T. Zhang, T. H. Li, and G. Li, "Step-by-step erasion, one-by-one collection: A weakly supervised temporal action detector," in *Proceedings of the International Conference on Multimedia*. ACM, 2018, pp. 35–44.

[35] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *Proceedings of the Advances in Neural Information Processing Systems*. Morgan Kaufmann, 1993, pp. 737–744.

[36] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2005, pp. 539–546.

[37] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 791–808.

[38] B. Peng, J. Lei, H. Fu, C. Zhang, T. Chua, and X. Li, "Unsupervised video action clustering via motion-scene interaction constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 131–144, 2020.

[39] M. S. Ryoo, K. Kim, and H. J. Yang, "Extreme low resolution activity recognition with multi-siamese embedding learning," in *Proceedings of the American Association for Artificial Intelligence*. AAAI, 2018, pp. 7315–7322.

[40] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proceedings of the Conference on Computer Vision and Pattern Recognition,*. IEEE, 2016, pp. 1420–1429.

[41] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proceedings of the American Association for Artificial Intelligence*. AAAI, 2016, pp. 2786–2792.

[42] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[43] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2018, pp. 2132–2141.

[44] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 4694–4703.

[45] Y. Yan, X. Wang, X. Guo, J. Fang, W. Liu, and J. Huang, "Deep multi-instance learning with dynamic pooling," in *Proceedings of the Asian Conference on Machine Learning*. PMLR, 2018, pp. 662–677.
[46] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2015, pp. 4489–4497.
[47] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2014, pp. 1532–1543.
[48] K. Cho, B. V. Merrienboer, Çaglar Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2014, pp. 1724–1734.
[49] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 510–526.
[50] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2017.
[51] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision*. Springer, 2018, pp. 212–228.

Yunxiao Wang received the B.E. degree from China University of Petroleum (East China) in 2019. He is currently a master with the School of Computing Science and Technology, Shandong University. His research interests include multimedia computing and information retrieval.

Meng Liu is currently a Professor with the School of Computer Science and Technology, Shandong Jianzhu University. She received the Ph.D. degree in computer science and technology from Shandong University, China, in 2019. Her research interests are multimedia computing and information retrieval. Various parts of her work have been published in top forums and journals, such as SIGIR, MM, and IEEE TIP. She has served as reviewers and subreviewers for various conferences and journals, such as ACM MM 2018/2019/2020, CVPR 2021, AAAI 2021, IEEE TIP, IEEE TKDE, and INS.

Yinwei Wei received his MS degree from Tianjin University and Ph.D. degree from Shandong University, respectively. Currently, he is a research fellow with National University of Singapore. His research interests include multimedia computing and recommendation. Several works have been published in top forums, such as MM and TIP. Dr. Wei has served as the PC member for several conferences, such as MM, AAAI, and IJCAI, and the reviewer for TMM, TKDE, and TIP.

Zhiyong Cheng is currently a Professor with Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences). He received the Ph.D degree in computer science from Singapore Management University in 2016, and then worked as a Research Fellow in National University of Singapore. His research interests mainly focus on large-scale multimedia content analysis and retrieval. His work has been published in a set of top forums, including ACM SIGIR, MM, WWW, TOIS, IJCAI, TKDE, and TCYB. He has served as the PC member for several top conferences such as SIGIR, MM, IJCAI, AAAI. and the regular reviewer for journals including TKDE, TIP, TMM.

Yinglong Wang is currently a Researcher, a Ph.D. Supervisor, and the Party Secretary of the Qilu University of Technology (Shandong Academy of Sciences). In recent years, he has taken charge of more than 20 national, provincial, and ministerial projects. Moreover, he organized the compilation of three volumes of national standards. He has published over 60 top academic articles and owns more than 20 authorized invention patents. His main research interests include medical artificial intelligence and high-performance computing. He is granted as a Young and Middle-aged Expert with outstanding contributions to Shandong Province, a High-End Think Tank Expert of Shandong Province, and enjoys special government allowances from the State Council. He serves as the President for the Shandong Internet of Things Association, a member for the Shandong Informatization Expert Group and the Shandong Informatization Expert Advisory Committee, the Director for the China-Australia International Health Technology Joint Laboratory, the Vice Chairman for the Shandong Science and Technology Association, and the Deputy Chairman for the Shandong Information Standardization Technical Committee. The scientific research projects led by him won two First Prizes, four Second Prizes, and two Third Prizes of the Shandong Science and Technology Progress Award.

Liqiang Nie received the B.Eng. degree from Xi'an Jiaotong University in 2009 and the Ph.D. degree from the National University of Singapore (NUS) in 2013. After the Ph.D., he continued his research in NUS as a Research Fellow for more than three years. He is currently a Professor with the School of Computer Science and Technology, Shandong University. Meanwhile, he is the Adjunct Dean with the Shandong AI Institute. His research interests lie primarily in multimedia computing and information retrieval. He has published around 100 papers in the top conferences or journals, with 6,000 plus Google Scholar citations as of Dec. 2019. He is an AE of information science and an Area Chair of ACM MM 2018/2019.