Video Moment Localization via Deep Cross-Modal Hashing

Yupeng Hu, *Member, IEEE*, Meng Liu, *Member, IEEE*, Xiaobin Su, *Member, IEEE*, Zan Gao, *Member, IEEE*, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*— Due to the continuous booming of surveillance and Web videos, video moment localization, as an important branch of video content analysis, has attracted wide attention from both industry and academia in recent years. It is, however, a non-trivial task due to the following challenges: temporal context modeling, intelligent moment candidate generation, as well as the necessary efficiency and scalability in practice. To address these impediments, we present a deep end-to-end cross-modal hashing network. To be specific, we first design a video encoder relying on a bidirectional temporal convolutional network to simultaneously generate moment candidates and learn their representations. Considering that the video encoder characterizes temporal contextual structures at multiple scales of time windows, we can thus obtain enhanced moment representations. As a counterpart, we design an independent query encoder towards user intention understanding. Thereafter, a cross-model hashing module is developed to project these two heterogeneous representations into a shared isomorphic Hamming space for compact hash code learning. After that, we can effectively estimate the relevance score of each "*moment-query*" pair via the Hamming distance. Besides effectiveness, our model is far more efficient and scalable since the hash codes of videos can be learned offline. Experimental results on real-world datasets have justified the superiority of our model over several state-of-the-art competitors.**

***Index Terms*— Video moment localization, multi-scale moment candidate generation, cross-modal hashing, temporal context modeling.**

I. Introduction

With the amount of videos growing exponentially, searching videos of interest from a large collection has been a hotspot in the field of information retrieval [1], [2]. Traditional studies mainly focus on searching for an entire video corresponding to a given query from a large-scale video collection [3], [4]. However, a video usually contains complex scenes and involves a large number of objects and actions, whereby only some moments may match the specific query while the rest may be redundant to the end-users [5], [6]. Taking the video demonstrated in Fig. 1 as an example, it depicts a scenario that a gangster robs and then flees. One may be only interested in the moment "*the man fired at the white vehicle before fleeing*", which starts at 30.5s and ends at 72.8s. Thereby, retrieving specific moments from a long untrimmed video via natural language queries, the so-called cross-modal moment localization, is highly desired in the real-world application scenarios, such as video surveillance [7], [8], robotic navigation [9], and autonomous driving [10].

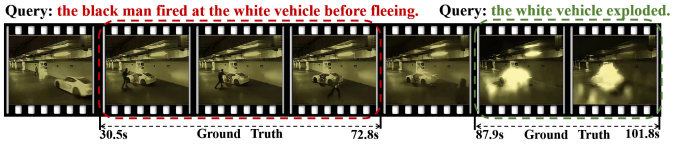


Fig. 1. An example of moment localization in a surveillance video.

Despite its importance and exciting prospect, localizing moments from a given video is non-trivial, since it faces the following research challenges. 1) **Temporal Context Modeling**. In practice, there are some temporal words in the given query, such as "first" and "before", and therefore modeling context information of video moments is essential for improving localization accuracy. However, most of the existing methods [8], [11], [12] adopt Bi-LSTM, Bi-GRU, or their variants to characterize contextual information of video moments. For instance, the final representations of video moments in [11] are computed by transforming the concatenation of the forward and backward LSTM outputs. It cannot precisely capture long-term and multi-scale semantic dependencies from relative long videos [13], and thus fails to model the temporal context. 2) **Moment Candidate Generation**. Existing methods [7], [14], [15] commonly adopt the sliding window strategy to densely divide the given video into segments with different lengths, and then treat them as the moment candidates. Such methods, however, are suboptimal not only for the high computational cost, but also the NP-hard search space. Moreover, regarding the offline moment generation, these networks cannot be trained in an end-to-end fash-

Manuscript received January 7, 2020; revised September 26, 2020; accepted March 29, 2021. Date of publication April 26, 2021; date of current version April 30, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant U1936203 and Grant 62006142, in part by the CCF-Baidu Open Fund under Grant CCF-BAIDU OF2020019, in part by the Shandong Provincial Natural Science Foundation under Grant ZR2019JQ23, in part by the Innovation Teams in Colleges and Universities in Jinan under Grant 2018GXRC014, in part by the New AI Project towards the Integration of Education and Industry in QLUT, in part by the Young Creative Team in Universities of Shandong Province under Grant 2020KJN012, and in part by the Project of Qingdao Postdoctoral Applied Research. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Aline Roumy. *(Corresponding authors: Meng Liu; Liqiang Nie.)*

Yupeng Hu, Xiaobin Su, and Liqiang Nie are with the School of Computer Science and Technology, Shandong University, Qingdao 266237, China (e-mail: nieliqiang@gmail.com).

Meng Liu is with the School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China (e-mail: mengliu.sdu@gmail.com).

Zan Gao is with the Shandong Computer Science Center, Shandong AI Institute, Jinan 250014, China.

Digital Object Identifier 10.1109/TIP.2021.3073867

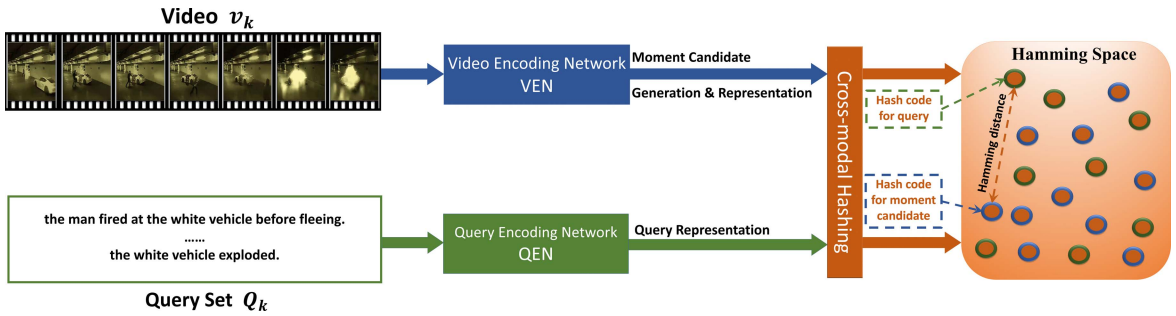


Fig. 2. Schematic illustration of our proposed CMHN approach, which seamlessly integrates the moment generation, representation and localization within a unified model.

ion, inducing poor availability when dealing with large-scale videos. And 3) **Efficiency and Scalability**. As shown in Fig. 1, to successfully take the criminal evidences of the gangster, policemen usually need to efficiently localize several moments from the surveillance video via different queries, such as "*the man fired at the white vehicle before fleeing*" and "*the white vehicle exploded*". But existing models [7], [8], [11], [12] typically utilize some query-aware attentive mechanism to estimate the correlation between "*moment-query*" pairs. We name such methods as iterative "*query-for-moment*" models. They have to repeat the similar process for any new query, even to the same video, resulting in inefficiency and low scalability.

To address the aforementioned challenges, we present an end-to-end Cross-Modal Hashing Network, dubbed CMHN, as shown in Fig. 2. Concretely, we first design a dual-path neural network, comprising two independent modules: the video encoding network (VEN) and the query encoding network (QEN). Thereinto, the VEN utilizes a bi-directional temporal convolutional network (Bi-TCN) to capture the multi-scale context information from the input video, and outputs the moment candidates and augmented moment representations. By contrast, the QEN adopts the Bidirectional Encoder Representation from Transformers (BERT) to deeply understand the semantics of the given queries and learn the corresponding feature representations of queries. Once the representations of moments and queries are learned, we develop a cross-modal hashing module to map them into a shared isomorphic Hamming space to generate their hash codes. Based upon hash codes, we estimate the similarities between the query and moment via Hamming distance. It is worth mentioning that with the well-trained model at hand, we can learn the hash codes of any upcoming videos offline and independently, which further improves the localization efficiency and scalability.

The main contributions of this work are three-fold:

- To the best of our knowledge, this is the first work on integrating cross-modal hashing into moment localization. Such a method enables query-video matching based upon hash codes and hence boosts the efficiency of moment localization remarkably.
- We present a novel Bi-TCN based VEN, which can automatically generate moment candidates and encode multi-scale context information towards moment representations. Most importantly, these two modules are independent. We are thus able to learn the hash codes of any new videos offline, which largely strengthens localization efficiency and scalability.
- We perform extensive experiments on two benchmark datasets, namely ActivityNet Captions [16] and TACoS [15], to justify the superiority of our model regarding accuracy, efficiency, and scalability compared to several state-of-the-art competitors. As a side contribution, we have released the data, codes, and parameter settings.[1]

II. Related Work

In this section, we briefly review the following two research directions highly related to ours: cross-modal hashing and moment localization in videos.

A. Cross-Modal Hashing

Considering the advantages of low storage cost and fast retrieval speed, cross-modal hashing is widely adopted in the retrieval task. Existing methods can be roughly divided into two categories: unsupervised and supervised ones. The former [17]–[20] focuses on learning hash functions by exploiting the intra- and inter-modality relations. For example, Song *et al.* [17] proposed a novel inter-media hashing model to linearly project the heterogeneous data into a shared common Hamming space by co-regularizing the inter- and intra- media consistency. Zhou *et al.* [18] presented a latent semantic sparse hashing model, which utilizes sparse coding and matrix factorization to capture the high-level latent semantic information from images and texts. Moreover, Irie *et al.* [19] proposed an alternating co-quantization scheme that alternately seeks the binary quantizers for each modality by jointly solving the subspace learning and binary quantization. Cui *et al.* [20] utilized the semantics of related social tags to improve the discrimination of feature representation and effectiveness of hash code generation.

Different from the unsupervised methods, the supervised ones [21]–[25] work towards leveraging the semantic labels of training data to guide the hash code learning. For example, to capture the underlying semantic information, Yu *et al.* [21] introduced a bi-stage discriminative coupled dictionary hashing model to jointly learn the coupled dictionaries and hash functions for both modalities. Arguing that the semantic affinities can be used to guide hashing, Lin *et al.* [22]

[1] Our open source code: `https://github.com/Huyp777/CMHN`

formulated a semantic-preserving hashing paradigm where the probability distribution generated from semantic affinities is approximated via minimizing the Kullback-Leibler divergence. The above supervised methods mainly rely on hand-crafted features, which inevitably leads to the separate feature extraction and hash codes learning procedures. To overcome this drawback, Jiang and Li [23] established an end-to-end cross-modal hashing framework with deep neural networks, whereby hash code learning is performed on each modality from scratch. Lu *et al.* [24] proposed a hierarchical recurrent network to exploit both spatial details and semantic information for effective hash codes generation. Sun *et al.* [25] utilized the cross-modal hashing methods based upon hierarchical label comprehension for clothing recommendation in the fashion domain.

Although much progress has been made in cross-modal image retrieval, hashing for video retrieval is still limited to mono-modal retrieval task, such as near-duplicate video retrieval [26], [27]. Cross-modal hashing for video retrieval is still a largely untapped problem, not to mention the moment localization.

B. Moment Localization in Videos

To localize the target moment in a certain video related to the given query, Gao *et al.* [15], [28] designed two temporal unit regression networks, which can jointly predict action proposals and refine the temporal boundaries by temporal coordinate regression. Almost at the same time, Hendricks *et al.* [14] employed a Moment Context Network (MCN) to integrate local and global video features for query-based moment localization. Considering the fact that previous models ignore the spatial-temporal information within the multi-modal data, Liu *et al.* [7], [8] designed two different attention-based networks for moment localization. The former aims to capture the most important context information to enhance the moment representations, while the latter focuses on extracting useful keywords from the given query. Subsequently, Yuan *et al.* [11] designed an Attention Based Location Regression (ABLR). It first adopts a co-attention memory model to capture the spatial-temporal interactions between video segments and query, and then generates the temporal coordinate of the target moment via the attention based regression network. To better model the spatial-temporal information in both modalities, Xu *et al.* [29] introduced a multi-level model, which incorporates multi-modal data via early fusion, and then utilizes video captioning as an auxiliary task to further guide the temporal coordinates prediction of the moment candidates. As prior studies only focus on one aspect, such as contextual feature representation and spatial-temporal information modeling of this emerging task, Zhu *et al.* [12] proposed the Cross-Modal Interaction Network (CMIN). This model utilizes the graph convolution network and multi-head self-attention for fine-grained representation learning on each modality, which further captures the corresponding "frame-by-word" interactions among these modalities. Moreover, it predicts the alignment scores of the moment candidates and adjusts the start and end boundaries of the high-score moments to accomplish the moment localization task.

In summary, the aforementioned studies have dedicated great efforts to the video-query interaction and jointly learn their representations. Although promising retrieval accuracy has been achieved, they fail to deliver significant improvement in retrieval efficiency and scalability. More precisely, the aforementioned models can only identify one target moment based on each query at a time; whereas to localize all the target moments in a certain video regarding the relevant queries, they need to repeat the same operation iteratively until all the queries have been processed completely. Thereby, such iterative "*query-for-moment*" processing strategy seriously deteriorates the efficiency and scalability.

III. Our Proposed Method

As shown in Fig. 2, our proposed CMHN comprises two components: 1) a dual-path neural network, including two independent modules: VEN and QEN. The former one is designed to generate moment candidate set $\mathcal{C}$ and learn its representation $\widehat{\mathcal{H}}_{\mathcal{C}}$. The latter is utilized to extract representation $\widehat{\mathcal{H}}_{\mathcal{Q}}$ for the query set $\mathcal{Q}$. And 2) a cross-modal hashing module is built to learn the hash codes of both modalities. In what follows, we will detail them one by one.

A. Problem Formulation

Given a training video set containing N untrimmed videos $\mathcal{S}_v = \{v_1, \ldots, v_k, \ldots, v_N\}$, where v_k denotes the k-th video in the training set. Assuming that there are M_k queries with respect to the video v_k, we denote it as $\mathcal{Q}_k = \{s_{k,1}, \ldots, s_{k,j}, \ldots, s_{k,M_k}\}$. Meanwhile, we define $\mathcal{A}_k = \{[t_s^{k,1}, t_e^{k,1}], \ldots, [t_s^{k,M_k}, t_e^{k,M_k}]\}$ are the exact video moments in v_k corresponding to the queries in $\mathcal{Q}_k$ labeled by humans, where $[t_s^{k,j}, t_e^{k,j}] \in \mathcal{A}_k$ is the start and end time of the j-th target moment in v_k.

Based on the training data, we aim at learning a cross-modal hashing network. With the well-trained model at hand, given a new untrimmed video v and its query set $\mathcal{Q}$, we could first generate moment candidate set $\mathcal{C}$ and then the hash codes for each moment and query. Afterwards, we can localize the target moments from $\mathcal{C}$ based on Hamming distance such that their timestamps are equal to their annotation $\mathcal{A}$.

B. The VEN Module

To generate the moment candidate set $\mathcal{C}$ containing moments in various lengths from the given video v and learn their representations, we propose a novel video encoding network VEN, as illustrated in Fig. 3. VEN consists of two modules: temporal context modeling, as well as moment generation and representation.

1) Temporal Context Modeling: Given the untrimmed video v_k, we first utilize the 3D convolutional network (C3D) [30] to extract local feature sequence $\mathbf{X}_k = [\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,i}, \ldots, \mathbf{x}_{k,R_x}]$, where $\mathbf{x}_{k,i}$ represents the i-th local feature of video v_k and R_x denotes the length of local feature sequence. And then we build a model to enrich these local features with the consideration of

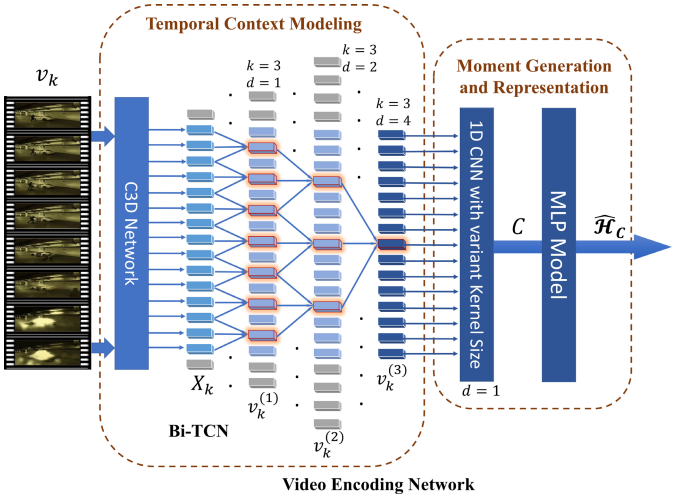


Fig. 3. The structure of our proposed VEN module. It first adopts the C3D model to obtain local features. And then it incorporates the Bi-TCN model to learn the local feature by capturing the corresponding pre-context and post-context information. Subsequently, a series of 1D regular convolution is adopted to generate moment candidates. Thereafter, a MLP model is applied to obtain feature representations of all the moment candidates.

context information. Most of the existing mainstream methods adopt Bi-LSTM or Bi-GRU model [31], [32] to extract the critical contextual information from the feature sequence $\mathbf{X}_k$. As each hidden state only memorizes part information from its input and the pre-hidden state, the center hidden state may merely retain very few information from the farther ones. Therefore, they cannot perfectly capture long-term semantic dependencies from a relatively long video [33].

Recently, a new context modeling strategy TCN [13] has been introduced, yet it only focuses on the correlation between current local feature and its pre-context information, thoroughly ignoring the post-context information. Motivated by this, we design a novel Bi-TCN network, which can capture long-term contextual dependencies of each $\mathbf{x}_{k,i}$ from both pre-context and post-context information, and hence effectively enhance the contextual representation. For instance, Fig. 3 shows an example of the Bi-TCN with three layers.[2] After the three-layer 1D dilated convolution processing, each element of $\mathbf{v}_k^{(3)}$ integrates the contextual information from two directions to form a more comprehensive feature representation. Generally, inputting $\mathbf{X}_k$ into the Bi-TCN with $E-1$ layers, the output can be formulated as,

$$\begin{cases} \mathbf{v}_k^{(1)} = \theta_1(\mathbf{X}_k, \delta^1, d^1), \\ \quad \vdots \\ \mathbf{v}_k^{(e)} = \theta_e(\mathbf{v}_k^{(e-1)}, \delta^e, d^e), \\ \quad \vdots \\ \mathbf{v}_k^{(E-1)} = \theta_{E-1}(\mathbf{v}_k^{E-2}, \delta^{E-1}, d^{E-1}), \end{cases} \tag{1}$$

where θ_e refers to the 1D dilated convolution of the e-th layer, d^e and δ^e respectively denote the dilation factor and filter kernel size of θ_e.

[2]The dilation factors d of two hidden layers and the output layer are respectively 1, 2 and 4. Besides, the kernel size of all filters is 3. Notably, to ensure the output feature sequence of each layer equals to the input length, the zero padding (i.e., gray rectangle) with length (2,4,8) is added to different layers.

2) Moment Generation and Representation: After obtaining long-term dependencies of each local feature, a series of 1D regular (i.e., $d = 1$) convolution operations with different kernel sizes δ^* are applied to $\mathbf{v}_k^{(E-1)}$ for moment candidates generation,[3] which can be formulated as follows,

$$\begin{aligned} \mathcal{C}_k &= \{c_{k,1}, \ldots, c_{k,p}, \ldots, c_{k,N_k}\} \\ &= \theta_E\left(\mathbf{v}_k^{(E-1)}, \delta^*, d = 1\right), \end{aligned} \tag{2}$$

where θ_E denotes the 1D regular convolution, $c_{k,p}$ refers to the p-th moment candidate, and N_k stands for the number of candidates.

Afterwards, we utilize the multi-layer perception (MLP) network[4] to obtain the corresponding feature representations for all the moment candidates in $\mathcal{C}_k$,

$$\begin{cases} \mathcal{H}_{\mathcal{C}_k}^1 = \sigma_c^1(\mathbf{W}_c^1 \mathcal{C}_k + \mathbf{b}_c^1), \\ \quad \vdots \\ \mathcal{H}_{\mathcal{C}_k}^t = \sigma_c^t(\mathbf{W}_c^t \mathcal{H}_{\mathcal{C}_k}^{t-1} + \mathbf{b}_c^t), \\ \quad \vdots \\ \mathcal{H}_{\mathcal{C}_k}^T = \sigma_c^T(\mathbf{W}_c^T \mathcal{H}_{\mathcal{C}_k}^{T-1} + \mathbf{b}_c^T), \end{cases} \tag{3}$$

where $\mathbf{W}_c^t$, $\mathbf{b}_c^t$, and $\mathcal{H}_{\mathcal{C}_k}^t$ respectively denote the weight matrix, bias vector, and output sequence of the t-th hidden layers. Meanwhile, σ_c^t is the Randomized Leaky Rectified Linear Units (RReLU) function [34], and $\widehat{\mathcal{H}}_{\mathcal{C}_k} = \mathcal{H}_{\mathcal{C}_k}^T = [\mathbf{h}_{c_{k,1}}, \ldots, \mathbf{h}_{c_{k,p}}, \ldots, \mathbf{h}_{c_{k,N_k}}] \in \mathbb{R}^{N_k \times L}$ refers to the representations of moment candidates, where L denotes the dimension of each candidate.

C. The QEN Module

To learn query representations, most existing methods utilize Bi-LSTM with word embeddings as input to obtain the final sentence-level representation. However, Bi-LSTM cannot obtain the long-term semantic dependencies for the relatively long query, resulting in poor feature representation. Fortunately, the cutting-edge language representation model, i.e., Bidirectional Encoder Representations from Transformers (BERT) [33], has gained promising performance in the field of natural language processing. It can deeply understand the semantics of the given sentence through the multi-layer bi-directional representation. In light of this, we employ the off-the-shelf language representation model BERT to perform query encoding. More importantly, the pre-trained BERT can be utilized directly in our QEN and fine-tuned accordingly to better represent queries. As shown in Fig. 4, all queries of $\mathbf{Q}_k$ are processed by BERT to output corresponding representations $\widetilde{\mathcal{Q}}_k = \{\widetilde{\mathbf{s}}_{k,1}, \ldots, \widetilde{\mathbf{s}}_{k,j}, \ldots, \widetilde{\mathbf{s}}_{k,M_k}\}$, where $\widetilde{\mathbf{s}}_{k,j}$ denotes the feature representation of the j-th query related to v_k.

[3]In our work, δ^* ranges from 1 to Δ, where Δ is a pre-defined maximum kernel size for candidates generation.

[4]In our experiments, the layers of MLP are set to two, i.e., $T = 2$.

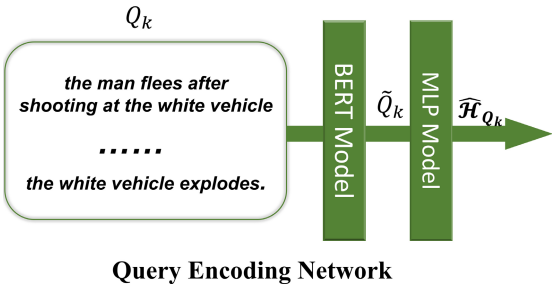


Fig. 4. The pipeline of our QEN module. It first utilizes BERT to perform feature encoding for each query in $\mathcal{Q}$. Afterwards, it adopts the MLP model to obtain their final representations $\widehat{\mathcal{H}}_{\mathcal{Q}}$.

Subsequently, QEN utilizes a three-layer MLP to obtain the corresponding embeddings for all queries,

$$\begin{cases} \mathcal{H}^1_{\mathcal{Q}_k} = \sigma^1_s(\mathbf{W}^1_s \widetilde{\mathcal{Q}}_k + \mathbf{b}^1_s), \\ \quad \vdots \\ \mathcal{H}^t_{\mathcal{Q}_k} = \sigma^t_s(\mathbf{W}^t_s \mathcal{H}^{t-1}_{\mathcal{Q}_k} + \mathbf{b}^t_s), \\ \quad \vdots \\ \mathcal{H}^T_{\mathcal{Q}_k} = \sigma^T_s(\mathbf{W}^T_s \mathcal{H}^{T-1}_{\mathcal{Q}_k} + \mathbf{b}^T_s), \end{cases} \tag{4}$$

where $\mathbf{W}^t_s$, $\mathbf{b}^t_s$, and $\mathcal{H}^t_{\mathcal{Q}_k}$ respectively denote the weight matrix, bias vector, and output vector of the t-th hidden layers. Symbol σ^t_s is the RReLU function [34], and $\widehat{\mathcal{H}}_{\mathcal{Q}_k} = \mathcal{H}^T_{\mathcal{Q}_k} = [\mathbf{h}_{s_{k,1}}, \ldots, \mathbf{h}_{s_{k,j}}, \ldots, \mathbf{h}_{s_{k,M_k}}] \in \mathbb{R}^{M_k \times L}$ refers to the representations of queries, where L denotes the dimension of each query.

D. Cross-Modal Hashing

Having obtained $\widehat{\mathcal{H}}_{\mathcal{C}_k}$, we adopt the element-wise sign function $sgn(\cdot)$ to generate the final hash codes for $\mathcal{C}_k$ as,

$$\mathcal{B}_{\mathcal{C}_k} = sgn(\widehat{\mathcal{H}}_{\mathcal{C}_k}) = [\mathbf{b}_{c_{k,1}}, \ldots, \mathbf{b}_{c_{k,p}}, \ldots, \mathbf{b}_{c_{k,N_k}}], \tag{5}$$

where $\mathbf{b}_{c_{k,p}} \in \{-1, +1\}^L$ denotes the final hash code of the p-th moment candidate with length L. Analogously, the element-wise sign function $sgn(\cdot)$ is also utilized to transform the $\widehat{\mathcal{H}}_{\mathcal{Q}_k}$ into the final hash codes of query set $\widetilde{\mathcal{Q}}$, as

$$\mathcal{B}_{\mathcal{Q}_k} = sgn(\widehat{\mathcal{H}}_{\mathcal{Q}_k}) = [\mathbf{b}_{s_{k,1}}, \ldots, \mathbf{b}_{s_{k,j}}, \ldots, \mathbf{b}_{s_{k,M_k}}], \tag{6}$$

where $\mathbf{b}_{s_{k,j}} \in \{-1, +1\}^L$ denotes the final hash code of the j-th query with length L.

To ensure each "*moment-query*" pair in the Hamming space maintains the intrinsic similarity in the original real-valued feature space, a loss function Γ_1 for semantic similarity preserving is proposed as follows,

$$\Gamma_1 = \sum_k \left(\left\| \widehat{\mathcal{H}}^T_{\mathcal{C}_k} \widehat{\mathcal{H}}_{\mathcal{Q}_k} - L\mathcal{M} \right\|^2_F \right), \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, L refers to the length of hash codes, and $\mathcal{M}$ is the cross-modal similarity matrix to ensure that the similarity in the Hamming space coincides with that in the original space.

Similar to the mainstream cross-modal similarity matrix [23], [25], the straightforward approach is to adopt 0 or 1 to represent the similarity of each "*moment-query*" pair by the following two steps: 1) We first adopt the general evaluation indicator Intersection over Union (IoU) [35] to evaluate the semantic similarities of the above $N_k \times M_k$ "*moment-query*" pairs. Specifically, the IoU between two moments is calculated as

$$IoU^{c_{k,p}}_{s_{k,j}} = \frac{\min\left(t^j_e, \tau^p_e\right) - \max\left(t^j_s, \tau^p_s\right)}{\max\left(t^j_e, \tau^p_e\right) - \min\left(t^j_s, \tau^p_s\right)}, \tag{8}$$

where τ^p_s and τ^p_e respectively denote the start and end time points of the moment $c_{k,p}$, and t^j_s and t^j_e are the start and end time points of the target moment depicted by the query $s_{k,j}$. And 2) each $IoU^{c_{k,p}}_{s_{k,j}}$ is then converted into $\mathcal{N}_{p,j} \in \{0, 1\}$ based on a pre-defined threshold λ. In particular, if $IoU^{c_{k,p}}_{s_{k,j}} \geq \lambda$, $\mathcal{N}_{p,j}$ is set to 1; otherwise, $\mathcal{N}_{p,j}$ is set to 0. Therefore, the $\mathcal{N}_{p,j}$ are arranged orderly to construct the similarity matrix $\mathcal{T} \in \{0, 1\}^{N_k \times M_k}$.

Although feasible, solely adopting this relatively rough metric approach (1-or-0) cannot accurately reflect the complex intrinsic semantic similarities of different "*moment-query*" pairs. As demonstrated in Fig. 5, the $IoU^{c_{k,1}}_{s_{k,1}}$ and $IoU^{c_{k,2}}_{s_{k,1}}$ are larger than the threshold $\lambda = 0.5$, thereby the corresponding $\mathcal{N}_{1,1}$ and $\mathcal{N}_{2,1}$ are all set as 1, i.e., $c_{k,1}$ and $c_{k,2}$ have identical semantic similarities with $s_{k,1}$. However, there are obvious differences between $c_{k,1}$ and $c_{k,2}$. Consequently, the above straightforward $\mathcal{T}$ has limited ability, resulting in information loss and sub-optimal retrieval performance.

To address this issue, we propose a novel approach to constructing the cross-modal similarity matrix $\mathcal{M}$, as shown in Fig. 5. More specifically, each element of $\mathcal{M}$ is directly set as the IoU value to reflect the corresponding similarity of each *"moment-query"* pair. For example, the (p, j)-th entry is represented as $\mathcal{M}_{p,j} = IoU^{c_{k,p}}_{s_{k,j}}$ $(0 \leq \mathcal{M}_{p,j} \leq 1)$. Accordingly, with the help of this soft-value $\mathcal{M}$, our proposed CMHN could well retain the intrinsic similarity of each "*moment-query*" pair in the Hamming space.

Apart from the semantic preserving regularization, we further respectively regularize the binarization difference of $(\mathbf{h}_{c_{k,p}}, \mathbf{b}_{c_{k,p}})$ and $(\mathbf{h}_{s_{k,j}}, \mathbf{b}_{s_{k,j}})$ to obtain the optimal continuous surrogates of the binary hash codes. It is formulated as,

$$\Gamma_2 = \sum_k \left(\left\| \widehat{\mathcal{H}}_{\mathcal{C}_k} - \mathcal{B}_{\mathcal{C}_k} \right\|^2_F + \left\| \widehat{\mathcal{H}}_{\mathcal{Q}_k} - \mathcal{B}_{\mathcal{Q}_k} \right\|^2_F \right). \tag{9}$$

Besides, to balance the learnt hash codes and maximize the information conveyed by each bit of the codes [25], $\widehat{\mathcal{H}}_{\mathcal{C}_k}$ and $\widehat{\mathcal{H}}_{\mathcal{Q}_k}$ need to be further regularized as follows,

$$\Gamma_3 = \sum_k \left(\left\| \widehat{\mathcal{H}}_{\mathcal{C}_k} \mathcal{U} \right\|^2_F + \left\| \widehat{\mathcal{H}}_{\mathcal{Q}_k} \mathcal{Z} \right\|^2_F \right), \tag{10}$$

where $\mathcal{U} \in \mathbb{R}^{L \times N_k}$ and $\mathcal{Z} \in \mathbb{R}^{L \times M_k}$ respectively denote a matrix whose elements are ones.

The final objective function of our CMHN model is the combination of the above three (Γ_1, Γ_2, and Γ_3),

$$\Psi = \Gamma_1 + \alpha\Gamma_2 + \beta\Gamma_3, \tag{11}$$

where α and β are the non-negative trade-off parameters.

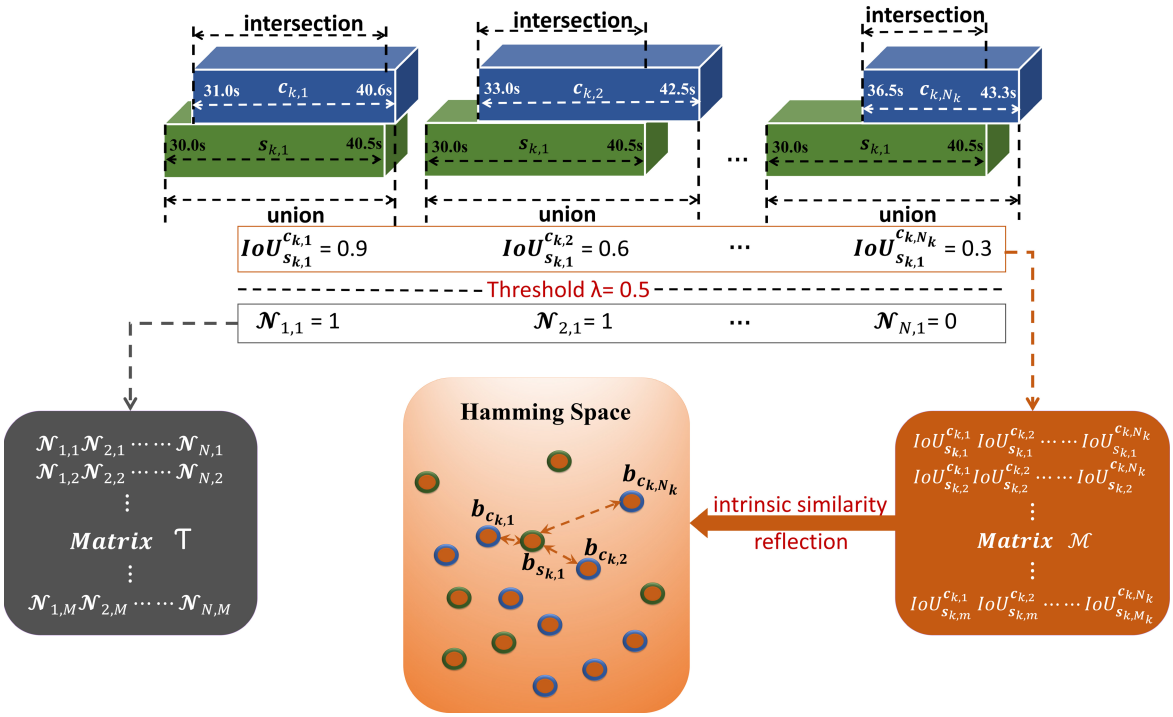


Fig. 5. Two kinds of similarity matrices, i.e., $\mathcal{M}$ and $\mathcal{T}$. Compared with relatively rough similarity metric (1-or-0) of $\mathcal{T}$, $\mathcal{M}$ is built based on IoU values to ensure that the intrinsic similarity of each "*moment-query*" pair can be maintained in Hamming space.

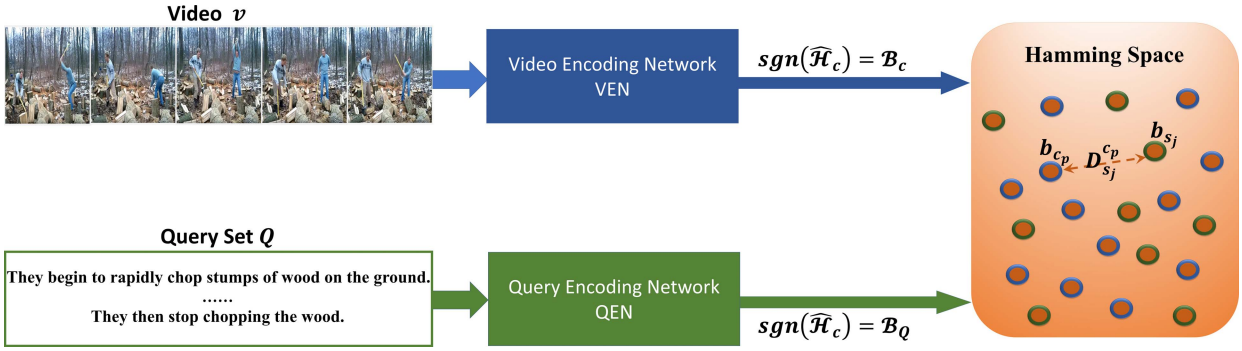


Fig. 6. The inference process of our moment localization model CMHN. After completing the model training, given the video v and its related queries $\mathcal{Q}$, CMHN not only generates the corresponding hash codes ($\mathcal{B}_\mathcal{C}$, $\mathcal{B}_\mathcal{Q}$) for different modalities, but also implements the moment retrieval based on Hamming distance.

E. Inference

As shown in Fig. 6, for an untrimmed video v and its related query set $\mathcal{Q}$, CMHN first separately generates the corresponding feature representations of moment candidates and queries, i.e., $\widehat{\mathcal{H}}_\mathcal{C}$ and $\widehat{\mathcal{H}}_\mathcal{Q}$. And then it projects all these representations into the same Hamming space to obtain the corresponding hash codes $\mathcal{B}_\mathcal{C}$ and $\mathcal{B}_\mathcal{Q}$. After that, the similarity can be measured via Hamming distance,

$$D_{s_j}^{c_p} = \frac{1}{2}(L - \mathbf{b}_{c_p}^T \mathbf{b}_{s_j}), \tag{12}$$

where $\mathbf{b}_{c_p}$ and $\mathbf{b}_{s_j}$ refer to the corresponding hash codes of the p-th moment candidate and the j-th query, respectively. The smaller the $D_{s_j}^{c_p}$ is, the more similar c_p and s_j will be, and vice versa. Finally, after obtaining the similarity values for all "*moment-query*" pairs according to Eq. (12), the target moment of each query can be efficiently localized by ranking the corresponding similarity values.

IV. Experiments

To thoroughly justify the effectiveness of our proposed model, we carried out extensive experiments to answer the following three research questions (RQs):

- **RQ1**: Is our proposed CMHN able to outperform several state-of-the-art competitors for moment localization?
- **RQ2**: Is each component of our model helpful for boosting the performance?
- **RQ3**: Is CMHN much more efficient and scalable than the state-of-the-art competitors?

A. Experimental Settings

1) Datasets: In this paper, we adopted two benchmark datasets, namely ActivityNet Captions [16] and TACoS [15], to evaluate our proposed model. The former contains 20,000 untrimmed videos, and each video has 3 natural language queries on average along with temporal annotations.

TABLE I

DATA STATISTICS OF ACTIVITYNET CAPTIONS AND TACOS, INCLUDING THE NUMBER OF VIDEOS (#VIDEO) AND QUERIES (#QUERY), ADV (SECONDS), ADM (SECONDS), AND THE AVERAGE LENGTH OF QUERY (ALQ)

ActivityNet Captions dataset					
	#Video	#Query	ADV	ADM	ALQ
training	10,009	37,421	117.30	35.45	13.48
validation	4,917	17,505	118.23	37.73	13.58
testing	4,885	17,031	118.21	40.25	12.02
all	19,811	71,957	117.74	37.14	13.16
TACoS dataset					
	#Video	#Query	ADV	ADM	ALQ
training	75	10,146	224.16	5.70	8.69
validation	27	4,589	387.46	6.23	9.12
testing	25	4,083	367.70	6.96	9.00
all	127	18,818	296.21	6.10	8.86

Specifically, each query is related to one target moment within the video. To facilitate the comparison, we followed the validation split introduced in [12], whereby val_1 is used for validating and val_2 for testing. As to the other dataset with 127 videos, each video involves an average of 148 natural language queries and temporal annotations. Compared with the ActivityNet Captions, the average duration of video (ADV) in TACoS is generally longer, while that of moment (ADM) is relatively shorter. The details of the two datasets are summarized in Table I.

2) Evaluation Metrics: To thoroughly measure our model and the baselines, we selected "R@n, IoU = m" designed by [35] as the evaluation metric of localization accuracy. In the following, we utilized $R(n, m)$ to denote "*R@n, IoU=m*". Moreover, we employed the total run time (TRT) and the average run time (ART) as the efficiency evaluation metrics. To be more specific, TRT denotes the overall run time to complete moment localization for all the queries; ART refers to the average run time to complete the moment localization based on each query,

$$ART = \frac{TRT}{N_q}, \tag{13}$$

where N_q is the total number of queries.

3) Implementation Details: For each video in these datasets, we considered 16 continuous frames as a unit with 8 frames overlapping between adjacent units. Subsequently, all the units are input into the pre-trained C3D [30] to produce a 4,096-d feature for each unit. In particular, for each unit of ActivityNet Captions, the feature is further reduced from 4,096-d to 500-d by the PCA strategy [16]. Accordingly, these 500-d and 4,096-d features are adopted respectively as the local features of ActivityNet Captions and TACoS. Moreover, all the parameters of VEN and QEN are initialized randomly. The adam optimizer [36] is adopted to minimize the multi-task loss. The grid search strategy is used to determine the hyper-parameters α and β in Eq. (11). Specially, in this paper, we set $\alpha = \beta = 1$ for the subsequent experiments. During training, for ActivityNet Captions, each batch packs an average of 1,648 "*moment-query*" pairs and the learning rate is set to 0.001; for TACoS, each batch contains an average of 4,865 "*moment-query*" pairs and the learning rate is set to 0.0001. And we empirically set the maximum number of epochs as 500 to ensure the convergence. Moreover, all the experiments are conducted over a computer equipped with Ubuntu 16.04.6 LTS, Intel Xeon CPU E5-2620, 128 GB Memory and NVidia TITAN Xp GPU.

TABLE II

PERFORMANCE OF CMHN WITH DIFFERENT HASH CODE LENGTH L ON ACTIVITYNET CAPTIONS AND TACOS. SPECIFICALLY, WE SET "*R@n, IoU = m*" WITH $n \in \{1, 5\}$ AND $m \in \{0.3, 0.5, 0.7\}$

CMHN	ActivityNet Captions						
	R@1 IoU=0.7	R@1 IoU=0.5	R@1 IoU=0.3	R@5 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.3	TRT(s)
$L = 32$	17.66%	38.42%	59.57%	46.89%	70.54%	84.02%	127.70
$L = 64$	19.72%	40.59%	61.06%	48.86%	71.92%	84.43%	128.29
$L = 128$	22.12%	42.45%	61.73%	51.27%	73.08%	85.45%	129.29
$L = 256$	24.02%	43.47%	62.49%	53.16%	73.42%	85.37%	130.17
$L = 512$	24.12%	43.87%	63.34%	53.26%	73.52%	85.58%	131.03
CMHN	TACoS						
	R@1 IoU=0.7	R@1 IoU=0.5	R@1 IoU=0.3	R@5 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.3	TRT(s)
$L = 32$	8.55%	19.94%	27.23%	20.94%	32.87%	39.80%	3.61
$L = 64$	12.98%	23.17%	27.72%	25.32%	34.63%	42.79%	3.62
$L = 128$	18.44%	25.58%	30.04%	28.24%	35.23%	44.05%	3.65
$L = 256$	19.98%	26.08%	30.91%	28.53%	36.69%	45.87%	3.73
$L = 512$	20.72%	26.23%	30.98%	29.37%	36.72%	45.93%	3.84

TABLE III

LOCALIZATION ACCURACY COMPARISON BETWEEN OUR PROPOSED MODEL AND SEVERAL STATE-OF-THE-ART BASELINES ON ACTIVITYNET CAPTIONS DATASET. (P-VALUE*: P-VALUE OVER $R(1, 0.5)$)

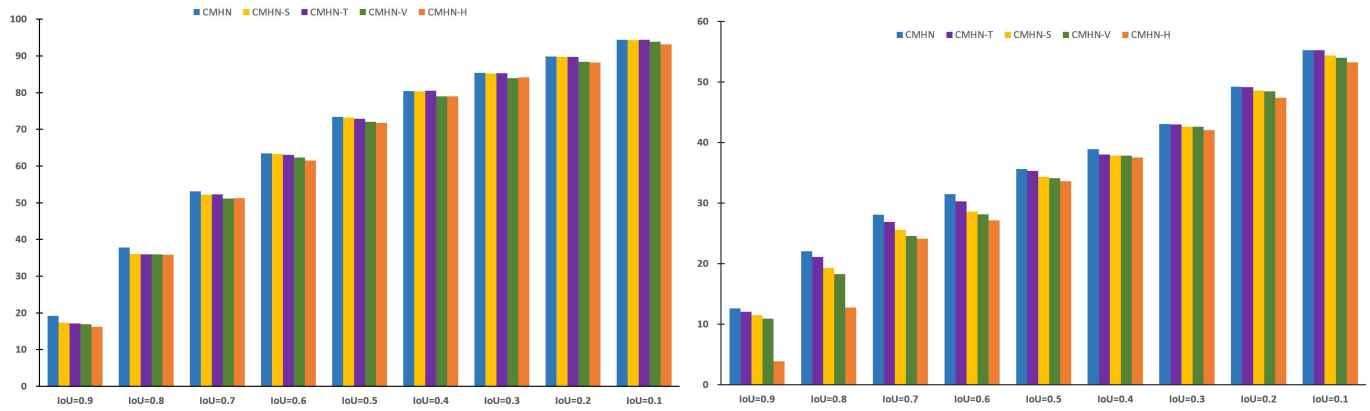
Method	R@1 IoU=0.7	R@1 IoU=0.5	R@1 IoU=0.3	R@5 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.3	p-value*
MCN	6.43%	21.36%	39.35%	29.70%	53.23%	68.12%	4.16E-17
CTRL	10.34%	29.01%	47.43%	37.54%	59.17%	75.32%	1.83E-15
ACRN	11.25%	31.67%	49.70%	38.57%	60.34%	76.50%	1.11E-14
QSPN	13.43%	33.26%	52.13%	40.78%	62.39%	77.72%	8.38E-14
ABLR	15.71%	36.79%	55.67%	-	-	-	1.68E-12
CMIN	23.88%	43.40%	**63.61%**	50.73%	67.95%	80.54%	4.68E-03
CMHN	**24.02%**	**43.47%**	62.49%	**53.16%**	**73.42%**	**85.37%**	**-**

4) Hash Code Length Setting: The length of hash code, i.e., L, is a crucial hyper-parameter of our proposed CMHN. Therefore, we should find the optimal setting before conducting formal performance comparisons. We studied the impact of different L on localization accuracy and efficiency. Experimental results on ActivityNet Captions and TACoS are summarized in Table II. Obviously, with the hash code length L increasing, the localization accuracy of CMHN is improved continuously, but the corresponding running time is also increasing concurrently. This fact reflects that longer hash codes could retain more information to improve the accuracy, yet it may deteriorate the localization efficiency. Accordingly, to further balance the accuracy and efficiency of CMHN, the hash code length of CMHN is set as 256 on ActivityNet Captions and 128 on TACoS, respectively.

B. On Model Comparison (RQ1)

In order to justify the effectiveness of our proposed CMHN, we compared it with six state-of-the-art baselines: MCN [14], CTRL [15], ACRN [7], ABLR [11], QSPN [29], and CMIN [12].

Table III and Table IV report the overall localization accuracy comparisons of our CMHN and baselines on ActivityNet Captions and TACoS datasets, respectively. For the above two datasets, we set the unified evaluation criteria "*R@n, IoU=m*" with $n \in \{1, 5\}$ and $m \in \{0.3, 0.5, 0.7\}$.



(a) R@5 w.r.t IoU on ActivityNet Captions

(b) R@5 w.r.t IoU on TACoS

Fig. 7. Localization accuracy comparison between our CMHN and its variants on ActivityNet Captions and TACoS datasets. (a) The experimental results based on R@5 w.r.t IoU $\in \{0.1, \ldots, 0.9\}$ on ActivityNet Captions dataset; (b) The experimental results based on R@5 w.r.t IoU $\in \{0.1, \ldots, 0.9\}$ on TACoS dataset.

TABLE IV

LOCALIZATION ACCURACY COMPARISON BETWEEN OUR PROPOSED MODEL AND SEVERAL STATE-OF-THE-ART BASELINES ON TACOS DATASET. (P-VALUE*: P-VALUE OVER $R(1, 0.5)$)

Method	R@1 IoU=0.7	R@1 IoU=0.5	R@1 IoU=0.3	R@5 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.3	p-value*
MCN	1.00%	1.25%	1.64%	1.01%	1.25%	2.03%	6.11E-14
CTRL	6.98%	13.08%	18.17%	15.60%	26.13%	35.97%	2.40E-11
ACRN	8.28%	14.11%	20.06%	15.43%	26.30%	38.11%	5.18E-11
QSPN	7.56%	15.23%	20.15%	14.35%	25.30%	36.72%	1.30E-10
ABLR	6.13%	9.30%	18.90%	-%	-%	-%	2.25E-12
CMIN	7.64%	18.05%	24.64%	11.81%	27.02%	38.46%	2.20E-09
CMHN	**18.44%**	**25.58%**	**30.04%**	**28.24%**	**35.23%**	**44.05%**	-

From Table III and Table IV, we have the following observations:

- Both MCN and CTRL deliver inferior performance, because they incorporate contextual information by either roughly fusing the entire video features or extending the moment boundaries within a limited scale. More concretely, the former fuses too much contextual information, which may bring in noise information and hurt the discriminative context representation. The latter merely considers limited pre- and post-moment extension as context, it hence fails to model the longer-term dependencies.
- ACRN, QSPN, and ABLR obtain higher accuracy than MCN and CTRL. This reflects that the attention mechanism indeed highlight crucial modality information to enhance the corresponding feature representation. Note that ABLR directly regresses the temporal locations based on visual-textual co-attention weights/features, it hence merely generates one prediction for a query.
- CMIN has relatively higher accuracy than other baselines do, which verifies that the long-range semantic dependency modeling is critical for moment localization.
- CMHN achieves the highest localization accuracy, substantially surpassing all state-of-the-art baselines, especially on TACoS. These results demonstrate the effectiveness of our proposed model. Furthermore, we also conducted a significance test between CMHN and each of the baselines regarding R(1,0.5) based on the 20-round results. All the p-values are smaller than 0.05, indicating that the advantage of our CMHN is statistically significant.

Interestingly, the accuracy of all the methods on TACoS is much lower than that on ActivityNet Captions. The main reasons can be summarized as follows: 1) the longer ADV and the shorter ADM on TACoS bring greater challenges on effective moment representation. And 2) the more queries and less ALQ may also lead to difficulty in differentiating similar moments with different queries. In general, moment localization on TACoS is more challenging than that on ActivityNet Captions.

C. On Component Analysis (RQ2)

We conducted some ablation studies on video encoding, query encoding, and cross-modal hashing. Concretely, we omitted one component once to generate the corresponding ablation models as follows.

- **CMHN-V**: We utilized traditional Bi-LSTM to replace our proposed Bi-TCN in VEN for video encoding.
- **CMHN-T**: We used traditional TCN to replace our proposed Bi-TCN in VEN for video encoding.
- **CMHN-S**: We adopted Glove word2vec [37] and Bi-LSTM to replace BERT in QEN for query encoding.
- **CMHN-H**: We eliminated our proposed cross-modal similarity matrix (CSM) and set the threshold $\lambda = 0.5$ to obtain the hard similarity matrix $\mathcal{T}$ for evaluating the similarity of each "*moment-query*" pair.

We conducted component-wise evaluation on ActivityNet Captions and TACoS datasets, respectively. The results are summarized in Fig. 7. By jointly analyzing the experimental results, we have the following findings:

- CMHN-H performs the worst on the two datasets. Especially in the case of relatively large IoU (IoU $\in \{0.6, 0.7, 0.8, 0.9\}$), the disadvantage of CMHN-H in localization accuracy becomes more apparent. This phenomenon reveals that this hard similarity representation (1-or-0) cannot accurately reflect the complex intrinsic semantic similarities of "*moment-query*" pairs and hence fails to identify the appropriate target moment related

TABLE V

RUNNING TIME COMPARISON BETWEEN OUR PROPOSED CMHN AND THE STATE-OF-THE-ART BASELINES ON ACTIVITYNET CAPTIONS AND TACOS DATASETS

Method	Dataset			
	ActivityNet Captions		TACoS	
	ART (s)	TRT (s)	ART (s)	TRT (s)
MCN	6.97	118,706.07	15.68	64,021.44
CTRL	3.41	58,135.38	5.52	26,515.75
ACRN	4.42	75,235.08	6.94	33,457.43
ABLR	0.06	1,021.86	0.23	958.50
CMIN	0.72E-2	123.18	1.03E-2	41.93
CMHN	0.76E-2	130.17	**0.90E-3**	**3.65**

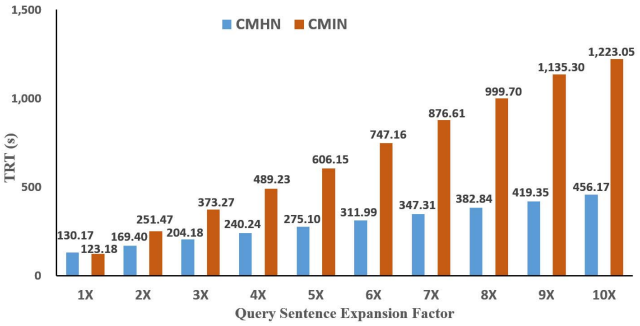
to the corresponding query. Accordingly, the considerable improvement achieved by CMHN verifies the effectiveness of our proposed soft similarity matrix for cross-modal hashing.

- Compared with CMHN, the localization accuracy of CMHN-S, CMHN-V, and CMHN-T is relatively low. This situation implies that: 1) word2vec based query encoding cannot understand the corresponding queries more deeply than BERT; 2) Bi-LSTM is incapable of temporal context modeling for relatively long videos; and 3) TCN cannot simultaneously capture the pre-context and post-context information for enhanced feature modeling, therefore failing to achieve more effective video encoding than Bi-TCN does.
- Our proposed CMHN model outperforms all variants on both datasets, which adequately demonstrates that the Bi-TCN, BERT, and CSM are all helpful for moment localization.

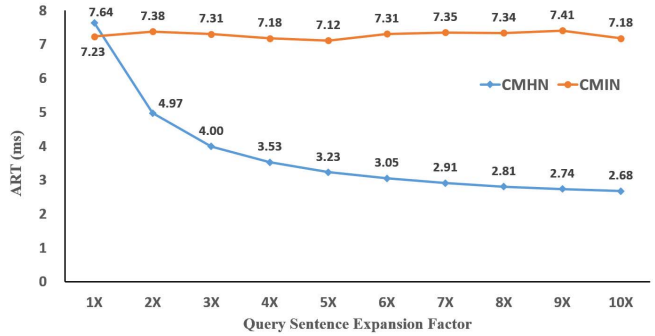
D. On Efficiency and Scalability Analysis (RQ3)

1) Localization Efficiency Comparison: Table V reports overall efficiency comparison between our CMHN and baselines on ActivityNet Captions and TACoS datasets. Through analyzing Table V, we could find that:

- ART and TRT of the ABLR model are smaller than the two stage "scan and localize" architecture, such as the MCN, CTRL and ACRN. This is because the latter models generate densely overlapped moment candidates by the sliding window approach for cross-modal localization. However, to avoid redundant computations, the ABLR evenly divides the original video into the corresponding moment candidates and performs temporal coordinates regression.
- The ART and TRT of CMIN model are much smaller than that of other baselines. The main reason is that CMIN further downsamples the 500-d feature sequences to 200-d, after using PCA to reduce the original video feature dimension from 4,096-d to 500-d. Compared with other baselines, these operations not only accelerate the video encoding, but also improve the overall efficiency.
- The efficiency of CMHN is better than that of all the baselines. Especially on the TACoS dataset, the efficiency of CMHN is at least one order of magnitude higher than that of CMIN. These results verify the high efficiency of our proposed model.



(a) TRT on ActivityNet Captions

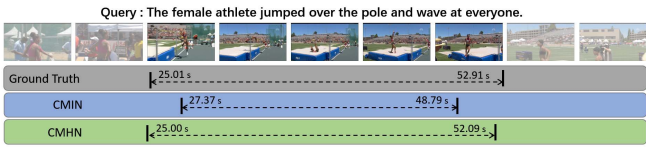


(b) ART on ActivityNet Captions

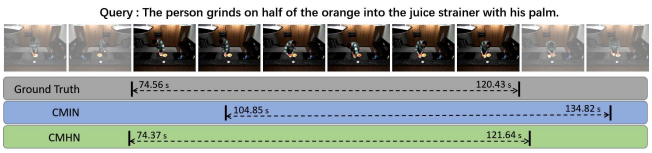
Fig. 8. Scalability comparison between our proposed CMHN and the CMIN on ActivityNet Captions with different numbers of queries. We use "2X" to denote that the queries in the testing set are copied once, i.e., the current number of queries in the testing set is $34,062 = 2 * 17,031$. (a) The TRT results regarding {1X, 2X, ..., 10X} queries; and (b) the ART results regarding {1X, 2X, ..., 10X} queries.

2) Localization Scalability Comparison: Through the previous efficiency analyses, we could find that there is a large gap between the workload of cross-modal moment localization on the two datasets. Specifically, the average number of moments to be localized for each video in ActivityNet Captions is 3, i.e., nearly three queries related to one video. But the average number of moments to be localized for each video in TACoS dataset is 148. Since the localization efficiency (ART) of CMIN is slightly better than our proposed CMHN on ActivityNet Captions (7.2 ms vs 7.6 ms), we further analyzed the scalability of them on ActivityNet Captions. In particular, both two models are completely trained offline using the identical training set, we then compared their running time when the number of online queries is multiplied.

The comparison results are shown in Fig. 8. Regarding the TRT results in Fig. 8(a), it is obvious that the corresponding TRT results of both models continuously rise along with the increase of queries. However, the growth rate of our proposed CMHN is much lower than that of CMIN. This is because the attentive aggregation module of the CMIN model relies heavily on video information. In other words, given different queries, CMIN must re-execute the video encoder and the attentive aggregation module to extract the corresponding enhanced representations. Therefore, the continuous increase of queries would lead to a sharp rise of the TRT. On the contrary, the VEN and QEN module of our model are completely



(a) Moment localization result on ActivityNet Captions.



(b) Moment localization result on TACoS.

Fig. 9. Moment localization results on ActivityNet Captions and TACoS datasets. All the above figures are the R@1 results.

independent, we hence able to learn the hash codes of any new videos and queries offline, which promotes the efficiency of localization. In light of this, our proposed CMHN is less susceptible than that of CMIN with the increasing number of queries.

In Fig. 8(b), along with the increasing number of queries, the ART results of CMIN are slightly fluctuating around the average of 7.28. This illustrates that the average running time for localizing one moment in a certain video is basically the same. In contrast, the ART results of our model CMHN are continuously decreasing, verifying that the growth rate of TRT is much smaller than the increase amplitude of queries, as defined in Eq. (13). According to the above analyses, we can find that with the increase of queries, the TRT growth rate of CMHN is much smaller than that of CMIN, i.e., our proposed CMHN has superior scalability compared with the state-of-the-art baseline.

E. Qualitative Results

To qualitatively validate the effectiveness of our proposed CMHN model, we presented several examples of moment localization by different queries. In particular, Fig. 9 shows the results of CMHN and the best baseline CMIN on ActivityNet Captions and TACoS datasets, respectively. From the results shown in Fig. 9, it is obvious that our proposed CMHN model obtains more accurate results, i.e., the moment returned by our model has the larger IoU with the ground truth moment. The reasons for these comparison results may be two-folds 1) our proposed Bi-TCN based CMHN can well capture the long-term contextual semantic dependencies than the Bi-GRU based CMIN; and 2) CMHN utilizes IoU based cross-modal similarity matrix to reflect the semantic similarity of each "moment-query" pair for the subsequent cross-modal hash code learning, which could help CMHN return better moment localization results to some extent.

V. Conclusion and Future Work

Given a natural language query and a video, in this paper, we present an end-to-end deep cross-modal hashing network to localize the matched video moment, the so-called moment localization. Specifically, we first integrate a bidirectional temporal convolutional network into the video encoder, which could characterize temporal contextual structures at multiple scales of time windows, to simultaneously generate moment candidates and their enhanced representations. As a counterpart, we design an independent query encoder to well understand the user requirements. Thereafter, a cross-model hashing module is introduced for compact hash code learning. Based upon the hash codes, we can effectively estimate the relevance score of each "*moment-query*" pair via the Hamming distance. To justify the effectiveness, efficiency and scalability, we conducted extensive experiments on two public benchmark datasets compared with several state-of-the-art competitors. As a byproduct, we have released the data, codes, and parameter settings to facilitate research in this community.

In the future, we plan to deepen and widen our work from the following two aspects: 1) As shown in Fig. 7, when IoU is larger than 0.7, the corresponding accuracy of CMHN is relatively low. Motivated by this, we intend to integrate the necessary spatial information into our model to improve localization accuracy. And 2) we will incorporate query-guided moment proposal network into our model to adaptively generate the corresponding moment candidates related to the given query for reducing the searching space while boosting the localization efficiency.

References

[1] R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," in *Proc. 12th Annu. ACM Int. Conf. Multimedia (MULTIMEDIA)*, 2004, pp. 548–555.

[2] D. Vallet, F. Hopfgartner, J. M. Jose, and P. Castells, "Effects of usage-based feedback on video retrieval: A simulation-based study," *ACM Trans. Inf. Syst.*, vol. 29, pp. 11–32, 2011.

[3] K. Schoeffmann and F. Hopfgartner, "Interactive video search," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1321–1322.

[4] M. H. T. D. Boer, Y.-J. Lu, H. Zhang, K. Schutte, C.-W. Ngo, and W. Kraaij, "Semantic reasoning in zero example video event retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 4, pp. 1–17, Oct. 2017.

[5] M. Liu, L. Nie, X. Wang, Q. Tian, and B. Chen, "Online data organizer: Micro-video categorization by structure-guided multimodal dictionary learning," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1235–1247, Mar. 2019.

[6] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Trans. Image Process.*, vol. 29, pp. 1–14, 2020.

[7] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua, "Attentive moment retrieval in videos," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 15–24.

[8] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 843–851.

[9] X. Wang *et al.*, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6629–6638.

[10] H. Huang *et al.*, "Transferable representation learning in vision-and-language navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7404–7413.

[11] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *Proc. Amer. Assoc. Artif. Intell.*, 2019, pp. 9159–9166.

[12] Z. Zhang, Z. Lin, Z. Zhao, and Z. Xiao, "Cross-modal interaction networks for query-based moment retrieval in videos," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 655–664.

[13] S. Bai, J. Z. Kolter, and V. Koltun, "Convolutional sequence modeling revisited," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–14.

[14] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5803–5812.

[15] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: Temporal activity localization via language query," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5267–5275.

[16] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 706–715.

[17] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2013, pp. 785–796.

[18] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2014, pp. 415–424.

[19] G. Irie, H. Arai, and Y. Taniguchi, "Alternating co-quantization for cross-modal hashing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1886–1894.

[20] H. Cui, L. Zhu, J. Li, Y. Yang, and L. Nie, "Scalable deep hashing for large-scale social image retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 1271–1284, 2020.

[21] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2014, pp. 395–404.

[22] Z. Lin, G. Ding, J. Han, and J. Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4342–4355, Dec. 2017.

[23] Q. Jiang and W. Li, "Deep cross-modal hashing," *Comput. Res. Repository*, vol. abs/1602.02255, pp. 1–12, 2016.

[24] X. Lu, Y. Chen, and X. Li, "Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 106–120, Jan. 2018.

[25] C. Sun, X. Song, F. Feng, W. X. Zhao, H. Zhang, and L. Nie, "Supervised hierarchical cross-modal hashing," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 725–734.

[26] Y. Hao, T. Mu, J. Y. Goulermas, J. Jiang, R. Hong, and M. Wang, "Unsupervised t-distributed video hashing and its deep hashing extension," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5531–5544, Nov. 2017.

[27] S. Li, Z. Chen, J. Lu, X. Li, and J. Zhou, "Neighborhood preserving hashing for scalable video retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8212–8221.

[28] J. Gao, Z. Yang, C. Sun, K. Chen, and R. Nevatia, "TURN TAP: Temporal unit regression network for temporal action proposals," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3628–3636.

[29] 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 *Proc. Amer. Assoc. Artif. Intell.*, 2019, pp. 9062–9069.

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[31] M. Liu, L. Nie, M. Wang, and B. Chen, "Towards micro-video understanding by joint sequential-sparse modeling," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 970–978.

[32] Y. Li, M. Liu, J. Yin, C. Cui, X.-S. Xu, and L. Nie, "Routing micro-videos via a temporal graph-guided recommendation system," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1464–1472.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 4171–4186.

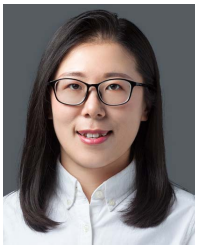
[34] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1–5.

[35] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4555–4564.

[36] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.

[37] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

Yupeng Hu (Member, IEEE) received the Ph.D. degree in software engineering from Shandong University in 2018. He is currently a Postdoctoral Fellow with the School of Computer Science and Technology, Shandong University. Various parts of his work have been published in famous journals and forums, such as *Science China Information Sciences*, *Neural Computing and Applications*, *Computer Networks*, and *Soft Computing*. His research interests include information retrieval, data mining, and explainable AI. He has served as a PC member for ACM MM, AAAI, and ChinaMM, and a Reviewer for IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS and IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.



Meng Liu (Member, IEEE) is currently a Professor with the School of Computer Science and Technology, Shandong Jianzhu University. Various parts of her work have been published in top forums and journals, such as SIGIR, MM, and IEEE TRANSACTIONS ON IMAGE PROCESSING. Her research interests are multimedia computing and information retrieval. She has served as a Reviewer and a Subreviewer for various conferences and journals, such as *MMM*, *MM*, *PCM*, *JVCI*, and *INS*.

Xiaobin Su (Member, IEEE) is currently pursuing the bachelor's degree with the School of Computing Science and Technology, Shandong University. His current research interests include machine learning, multimedia computing, and information retrieval.

Zan Gao (Member, IEEE) received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2011. From September 2009 to September 2010, he was a Visiting Scholar with the School of Computer Science, Carnegie Mellon University, USA, and worked with Prof. Alexander G. Hauptmann. From July 2016 to January 2017, he worked with the School of Computing, National University of Singapore, as a Visiting Scholar, with Prof. Chua Tat Seng. He is currently a Full Professor with the Shandong Computer Science Center, Shandong AI Institute, Qilu University of Technology (Shandong Academy of Sciences). His research interests include artificial intelligence, multimedia analysis and retrieval, and machine learning.

Liqiang Nie (Senior Member, IEEE) 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 Ph.D., he continued his research at 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 also the Adjunct Dean with the Shandong AI Institute. He has published around 100 papers in the top conferences or journals, with 9,600 plus Google Scholar citations. His research interests lie primarily in multimedia computing and information retrieval. He was granted several awards, like the SIGIR 2019 Best Paper Honorable Mention, the ACM SIGMM Rising Star 2020, the Green Orange Award 2020 of Alibaba DAMO Academy, the AI 2000 The Most Influential Scholar In Artificial Intelligence, and the MIT TR35 China. He is also an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA (IEEE TMM), IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (IEEE TKDE), *ACM Transactions on Multimedia Computing, Communications, and Applications (ACM TOMM)*, and *Information Science*. He serves as the PC Chair of ICIMCS 2017, PCM 2018, and ChinaMM 2020, and the Area Chair of ACM MM 2018–2021.