Coarse-to-Fine Semantic Alignment for Cross-Modal Moment Localization

Yupeng Hu, *Member, IEEE*, Liqiang Nie, *Senior Member, IEEE*, Meng Liu, *Member, IEEE*, Kun Wang, Yinglong Wang, and Xian-Sheng Hua, *Fellow, IEEE*

***Abstract*— Video moment localization, as an important branch of video content analysis, has attracted extensive attention in recent years. However, it is still in its infancy due to the following challenges: cross-modal semantic alignment and localization efficiency. To address these impediments, we present a cross-modal semantic alignment network. To be specific, we first design a video encoder to generate moment candidates, learn their representations, as well as model their semantic relevance. Meanwhile, we design a query encoder for diverse query intention understanding. Thereafter, we introduce a multi-granularity interaction module to deeply explore the semantic correlation between multi-modalities. Thereby, we can effectively complete target moment localization via sufficient cross-modal semantic understanding. Moreover, we introduce a semantic pruning strategy to reduce cross-modal retrieval overhead, improving localization efficiency. Experimental results on two benchmark datasets have justified the superiority of our model over several state-of-the-art competitors.**

***Index Terms*— Cross-modal moment localization, coarse-to-fine semantic alignment, hierarchical semantic pruning.**

I. Introduction

CROSS-MODAL video retrieval, aiming to search for a whole video from a large-scale video collection via a given query, has attracted increasing research interest [1], [2]. In short, it mainly focuses on determining whether a specific video contains the semantic of the given query. However, an untrimmed video usually contains complex scenes and events. This induces only part of the video content matches the semantic of the given query, while the rest is irrelevant and may be not desired by users [3], [4]. As illustrated in Fig. 1, the video depicts a scenario that a bartender is inviting a customer to taste various drinks. One may only pay close attention to the moment "*the man is tasting this third drink*". Consequently, localizing specific moments from a long and untrimmed video via natural language queries, i.e., **Cross-modal Moment Localization**, is essential [5].

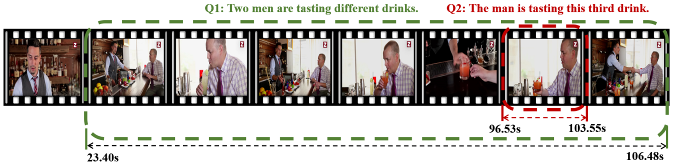

Q1: Two men are tasting different drinks.
Q2: The man is tasting this third drink.
96.53s
103.55s
23.40s
106.48s

Fig. 1. Examples of localizing moments via two queries within an untrimmed video.

Despite the intense interest in the task of cross-modal moment localization, it remains a highly challenging problem. The primary reasons are as follows: **1) Effective Cross-modal Semantic Alignment**. Given an untrimmed video, people may pose different queries to locate their desired moments. These queries can be roughly grouped into two types: summary queries and detailed queries. As shown in Fig. 1, there are two queries (i.e., Q1 and Q2) referring to moments marked with green and red bounding boxes, respectively. The former (summary query) "*two men are tasting different drinks.*" is utilized to localize the moment containing a series of successive actions and complex interactions; while the latter (detailed query) "*the man is tasting this third drink.*" is designed for the moment corresponding to the specific clue word "third". Obviously, moments, localized by two types of queries, involve different visual semantic information. In light of this, how to effectively align the diverse textual (query) semantic information with the visual semantic exploited from the given video is a crucial problem. Although much progress has been made in bridging visual and textual semantic information [6], [7], yet they suffer from several critical shortcomings. On the one hand, some methods [3], [6] conduct coarse-grained (clip-by-sentence) semantic matching between two modalities, ignoring the semantic correlation between key video frames and temporal words. Thereby, they fail to localize the moment depicted by the detailed query (e.g., Q2). On the other hand, certain methods [7], [8] [9]

Manuscript received December 1, 2020; revised May 10, 2021; accepted May 31, 2021. Date of publication June 24, 2021; date of current version June 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 toward 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 Prof. Guo-Jun Qi. *(Corresponding authors: Liqiang Nie; Meng Liu.)*

Yupeng Hu, Liqiang Nie, and Kun Wang 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).

Yinglong Wang is with the Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China.

Xian-Sheng Hua is with the Artificial Intelligence Center, Alibaba Group, Hangzhou 311121, China, and also with the City Brain Lab of DAMO Academy, Alibaba Group, Hangzhou 311121, China.

Digital Object Identifier 10.1109/TIP.2021.3090521

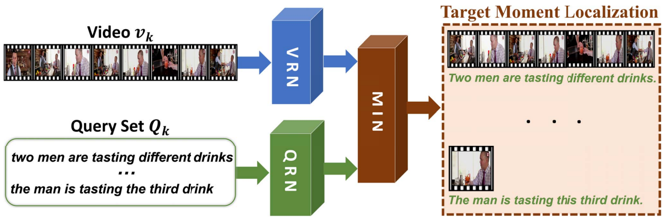

Video v_k
VRN
Query Set Q_k
two men are tasting different drinks
...
the man is tasting the third drink
QRN
MIN
Target Moment Localization
Two men are tasting different drinks.
. . .
The man is tasting this third drink.

Fig. 2. Schematic illustration of our proposed CLEAR model.

exploit fine-grained (frame-by-word) semantic interaction to implement moment localization for detailed queries. However, for summary queries (e.g., Q1), such operation may bury meaningful global visual-textual interaction into trivial details, and hence induces a sharp degradation on localization performance. **2) Efficient Cross-modal Moment Localization**. As shown in Fig. 1, to localize these two target moments, most existing methods [4], [8] [7], [9] [10] first consider the current video and each given query as a "video-query" pair, and then typically utilize the query-aware attentive mechanism to localize the desired moment. In other words, they have to repeat the prior steps until all "video-query" pairs have been processed, named as iterative "*query-for-moment*" models [5], thereby resulting in inefficiency. Consequently, how to build an efficient cross-modal moment localization model is still a largely unsolved problem.

To address the aforementioned challenges, we present an end-to-end **C**oarse-to-fine cross-moda**L** s**E**mantic **A**lignment netwo**R**k, dubbed as **CLEAR**, as illustrated in Fig. 2. Concretely, we first design a dual-path neural network, comprising two independent modules: the video representation network (VRN) and the query representation network (QRN). Thereinto, VRN is designed to learn moment representations and model their semantic relevance, while QRN is utilized to embed queries. We then develop a multi-granularity interaction network (MIN) for cross-modal semantic alignment. We finally utilize the coarse-grained semantic pruning to filter out irrelevant moments, and the fine-grained semantic alignment for moment localization. Compared with previous models, our coarse-to-fine approach effectively reduces retrieval overhead and hence boosts the efficiency of moment localization remarkably.

The main contributions of our work are three-fold:

- We propose a cross-modal semantic alignment network, which can realize efficient moment localization in a coarse to fine manner. To the best of our knowledge, this is the first work that integrates both semantic pruning and semantic alignment into the task of cross-modal moment localization.
- We present a novel video encoder to synchronously model the hierarchical semantic and temporal-spatial position relation among video moments. By this means, the enhanced moment features can be well adapted to diverse queries for effective cross-modal semantic alignment, therefore improving localization accuracy.
- We perform extensive experiments on two benchmark datasets to justify the superiority of our model on both accuracy and efficiency. As a side contribution, we have released the codes and parameter settings.[1]

The rest of this paper is organized as follows. Section II reviews the related work. Section III and Section IV detail our proposed CLEAR model and its justification, respectively. We conclude the work and discuss the future direction in Section V.

II. Related Work

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

A. Temporal Action Localization

Temporal action localization aims to predict the start and end points of a specific action from an untrimmed video. In 2016, Singh *et al.* [11] presented a multi-stream Bi-LSTM to effectively obtain the long-term temporal correlations for fine-grained action detection. Meanwhile, Shou *et al.* [12] proposed an end-to-end segment-based 3D convolutional neural network (S-CNN) to capture spatio-temporal information, thereby improving action localization performance. Considering the sliding window based proposal generation would lead to an increase in processing time overhead, Gao *et al.* [13] presented a novel temporal unit regression network (TURN) for fast temporal action proposal generation. Relevant experimental results verify that integrating TURN, as the proposal generation sub-model, into S-CNN can further improve localization performance. However, the above models still fail to fully explore the temporal context information and multi-stream features, resulting in limited performance improvement. In light of this, Chao *et al.* [14] introduced a novel temporal action localization network (TAL-Net) based on Faster R-CNN to deeply exploit the temporal context information and multi-stream feature fusion, improving the overall performance. Since this multi-tower dilated temporal convolution based TAL-Net may not fully improve localization efficiency, Zhao *et al.* [15] proposed a structural temporal pyramid pooling based structured segment network for more efficient temporal action localization.

Although the aforementioned approaches have achieved promising performance, they are still limited to the pre-defined simple action list, such as "drinking and smoking", "open the door and sit down", failing to localize complex activities expressed in arbitrary natural language.

B. Cross-Modal Moment Localization

Different from temporal action localization, cross-modal moment localization works towards localizing the target moment in a certain video related to the given query. In 2017, Gao *et al.* [6] designed a temporal coordinate regression network, which can jointly predict action proposals and refine the temporal boundaries. In the same year, Hendricks *et al.* [3] employed a moment context network (MCN) to integrate local and global video features for query guided moment

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

localization. As previous models ignore the spatial-temporal information within the visual and textual modalities, Liu *et al.* [4], [8] designed two different attention-based networks for moment localization. The former aims to capture the most important visual context information to enhance the moment representations, while the latter focuses on extracting useful keywords from the given query. Subsequently, Yuan *et al.* [7] 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 the query, and then generates the temporal coordinate of the target moment via the attention based regression network. To better exploit the spatial-temporal information in both modalities, Xu *et al.* [16] introduced a multi-level model and utilized video captioning as an auxiliary task to further guide the temporal coordinate prediction of the target moment. As prior studies only focus on one aspect, such as contextual feature representation or spatial-temporal information modeling, Zhang *et al.* [9], Lin *et al.* [10] proposed the cross-modal interaction network (CMIN) to utilize the graph convolution network and multi-head self attention for fine-grained representation learning on each modality. Moreover, it adopts query reconstruction strategy to further strengthen the cross-modal representations. Considering that CMIN cannot sufficiently perceive corresponding contextual information on the fused features, Zhang *et al.* [17] introduced a 2D temporal adjacent network (2D-TAN) to model the temporal relationships of multi-modalities in the 2D feature map for more deeply cross-modal semantic understanding.

Although much progress has been made in developing multi-modal representation and cross-modal fusion, yet they suffer from two critical shortcomings: 1) they ignored the necessary hierarchical similarity evaluation [18] towards video-query interaction, and hence fails to deliver significant improvement on localization accuracy; 2) they did not consider any information pruning strategy to accelerate localization. Namely, to localize all 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, resulting in low efficiency.

III. Our Proposed Method

As shown in Fig. 2, our proposed CLEAR model mainly comprises two components: 1) a dual-path neural network including VRN and QRN. The former is designed to generate moment candidates, learn their representations and model their semantic relevance. The latter is utilized to extract representations for given queries. And 2) a multi-granularity interaction module is built to implement the cross-modal semantic alignment between two modalities. In what follows, we will detail them sequentially.

A. Problem Formulation

Let $\mathcal{V} = \{v_1, \ldots, v_k, \ldots, v_N\}$ be a training set of N untrimmed videos, where v_k denotes the k-th video. Each video v_k is annotated by M_k queries (represented as $\mathcal{Q}_k = \{q_{k,1}, \ldots, q_{k,j}, \ldots, q_{k,M_k}\}$), and each query $q_{k,j}$ corresponds to a specific moment that starts at t_s^j and ends at t_e^j. Based on the training data, we aim at learning a cross-modal semantic alignment network. During the inference, given a new untrimmed video v_x and its corresponding query set $\mathcal{Q}_x$, we could first generate its moment candidate set $\mathcal{C}_x$ and then obtain representations ($\widehat{\mathcal{C}}_x$ and $\widehat{\mathcal{Q}}_x$) for both modalities. Afterwards, we can utilize multi-granularity cross-modal interaction to effectively localize the target moments.

B. Video Representation Module

For the given video v_k, to generate the moment candidate set $\mathcal{C}_k$ and learn the representations of these moments, we propose a novel video representation network VRN. It consists of three parts: enhanced moment representing, position relation determining, and hierarchical semantic modeling.

1) Enhanced Moment Representing: Given an untrimmed video v_k, we first segment it into L video units with a fixed temporal step. And then we utilize a pre-trained 3D convolutional network (C3D) [19] to extract their features, obtaining the corresponding feature sequence $\mathbf{U}_k = [u_{k,1}, \ldots, u_{k,l}, \ldots, u_{k,L}]$, where $u_{k,l}$ is the representation of the l-th video unit. Afterward, we intend to build a model to enhance these local features via context information.

Concretely, we utilize Bi-TCN model [20], instead of Bi-LSTM or Bi-GRU [21], [22], to capture long-term semantic dependencies for each unit. As illustrated in Fig. 3, each $u_{k,i}$ can integrate the contextual information from two directions (pre/post-context), therefore obtaining more comprehensive representations. Generally, inputting $\mathbf{U}_k$ into the Bi-TCN with R layers, the output can be formulated as,

$$\begin{cases} \widetilde{\mathbf{U}}_k^{(1)} = \theta_1(\mathbf{U}_k, \delta^1, d^1), \\ \quad\vdots \\ \widetilde{\mathbf{U}}_k^{(r)} = \theta_r(\widetilde{\mathbf{U}}_k^{(r-1)}, \delta^r, d^r), \\ \quad\vdots \\ \widetilde{\mathbf{U}}_k^{(R)} = \theta_R(\widetilde{\mathbf{U}}_k^{(R-1)}, \delta^R, d^R), \end{cases} \tag{1}$$

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

Having obtained $\widetilde{\mathbf{U}}_k^{(R)}$, we could generate moment candidates for localization. As illustrated in Fig. 3, we first adopt multi-scale temporal pooling [3] to generate representations for all possible enumerated moments. For example, the first l unit features are max pooled to form the representation for the moment starting at the 1-th unit and ending at the l-th unit, i.e., $\widetilde{c}^k_{1,l} = maxpool(\widetilde{u}_{k,1}, \widetilde{u}_{k,2}, \ldots, \widetilde{u}_{k,l})$. Similarly, the whole ($L$) unit features can be max pooled to generate the representation for the moment starting at the 1-th unit and ending at the L-th unit, denoted by $\widetilde{c}^k_{1,L} = maxpool(\widetilde{u}_{k,1}, \widetilde{u}_{k,2}, \ldots, \widetilde{u}_{k,L})$. In this way, the basic moment representation set $\mathcal{C}_k$, containing representations of $\Delta = \frac{L(1+L)}{2}$ moment candidates, is generated. Subsequently, we adopt the multi-layer perception (MLP) network to output the enhanced moment representation set $\widehat{\mathcal{C}}_k \in \mathbb{R}^{\Delta \times d}$ for all moment candidates, where d is the feature dimension of each moment candidate.

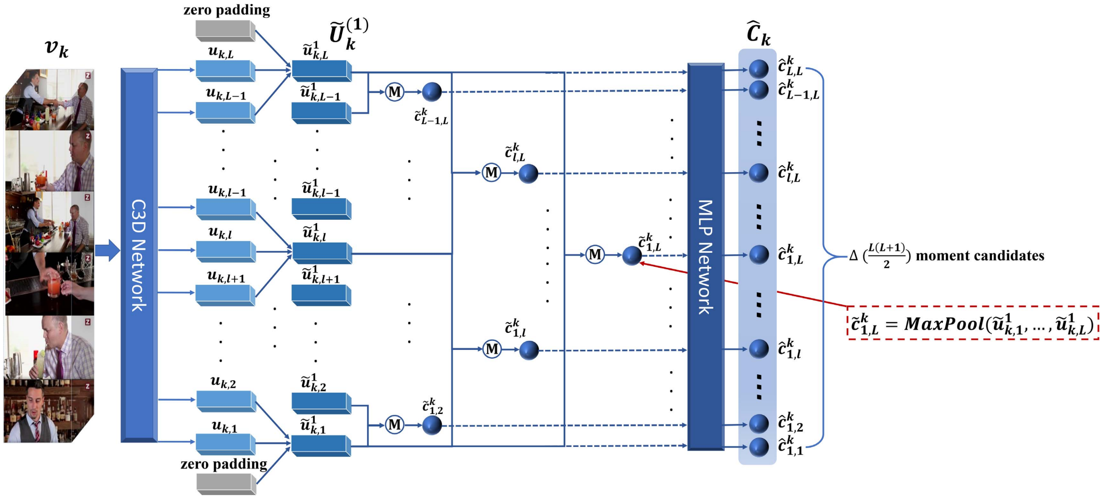


Fig. 3. The pipeline of the enhanced moment representing module. It first adopts the C3D model to obtain unit features. And then it incorporates the Bi-TCN model to enhance the unit feature by capturing its pre-context and post-context information. Concretely, the kernel size is 3 and the necessary zero paddings are also added to ensure the length of the output sequence is consistent with that of input units. Subsequently, a series of multi-scale max pooling operations are adopted to generate the basic moment representation set $\widetilde{\mathcal{C}}_k$. Finally, a MLP model is applied to obtain the enhanced moment representation set $\widehat{\mathcal{C}}_k$.

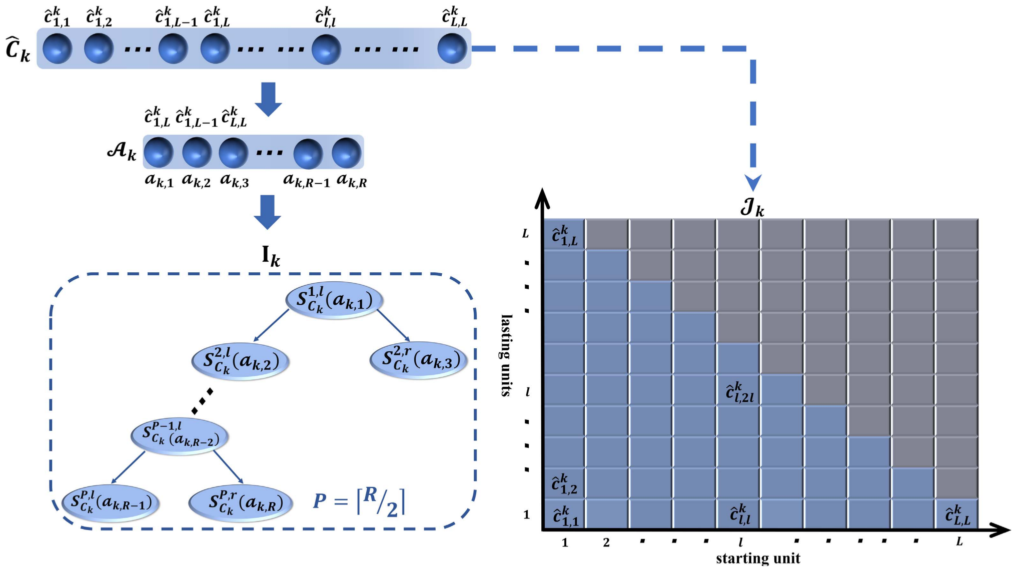
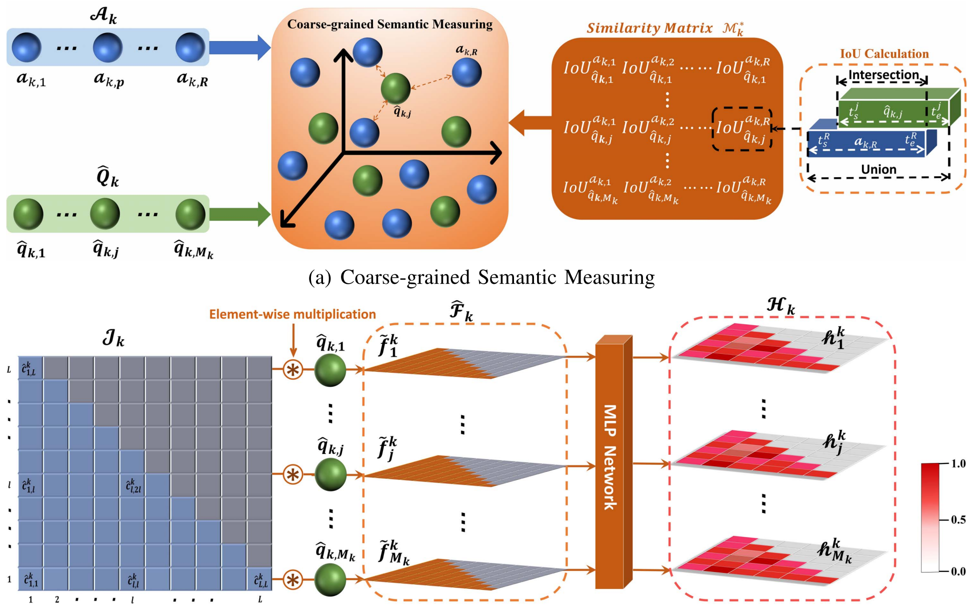


Fig. 4. Illustration of hierarchical semantic building. It first selects R representative moment candidates from $\widehat{\mathcal{C}}_k$ as anchor moments $\mathcal{A}_k$. And then it builds the video semantic tree $\mathcal{I}_k$ to indicate the semantic structure of $\widehat{\mathcal{C}}_k$.

2) Position Relation Determining: To well exploit relations [17] among moment candidates, we project $\widehat{\mathcal{C}}_k$ into a temporal-spatial position map, dubbed as $\mathcal{J}_k$. As illustrated in Fig. 4, the horizontal and vertical axes of $\mathcal{J}_k$ separately represent the starting and lasting units of each moment. The l-th column of $\mathcal{J}_k$ place enhanced representations of moment candidates starting at the l-th video unit. Note that, $\mathcal{J}_k \in \mathbb{R}^{L\times L\times d}$ is indeed a lower triangular matrix, where the first two dimensions indicate the maximum indexes of the starting and lasting points. In addition, with the help of $\mathcal{J}_k$, the real start and end time of each moment can be easily obtained. Without loss of generality, supposing the total duration of video v_k is T_k, the duration of each video unit would be $\frac{T_k}{L}$, and the real temporal range of the moment $\hat{c}^k_{l,l}$ is $\left[(l-1)\times\frac{T_k}{L}, l\times\frac{T_k}{L}\right]$.

3) Hierarchical Semantic Modeling: Having obtained the enhanced moment representations and determined their temporal-spatial relations, we explore the semantic relationship between different moments. Particular, in this paper, we assume that if two moments have an obvious temporal entailment relation, they indeed have an inclusive semantic relationship. As illustrated in Fig. 1, two moments (in green and red boxes) have an obvious overlap in temporal span, while the semantic of the former one ("Two men are tasting different drinks") indeed covers the semantic of the latter one ("The man is tasting this third drink"). Because tasting the third drink is an instance of tasting different drinks.

To model the hierarchical semantic based on $\widehat{\mathcal{C}}_k$, we first select R anchor moments $\mathcal{A}_k = \{a_{k,r}\}_{r=1}^{R}$, where each $a_{k,r}$ is


$\mathcal{A}_k$
$a_{k,1}$ $a_{k,p}$ $a_{k,R}$
$\widehat{\mathcal{Q}}_k$
$\widehat{q}_{k,1}$ $\widehat{q}_{k,j}$ $\widehat{q}_{k,M_k}$
Coarse-grained Semantic Measuring
$a_{k,R}$
$\widehat{q}_{k,j}$
Similarity Matrix $\mathcal{M}_k^*$
$IoU^{a_{k,1}}_{\widehat{q}_{k,1}}$ $IoU^{a_{k,2}}_{\widehat{q}_{k,1}}$ …… $IoU^{a_{k,R}}_{\widehat{q}_{k,1}}$
$IoU^{a_{k,1}}_{\widehat{q}_{k,j}}$ $IoU^{a_{k,2}}_{\widehat{q}_{k,j}}$ …… $IoU^{a_{k,R}}_{\widehat{q}_{k,j}}$
$IoU^{a_{k,1}}_{\widehat{q}_{k,M_k}}$ $IoU^{a_{k,2}}_{\widehat{q}_{k,M_k}}$ …… $IoU^{a_{k,R}}_{\widehat{q}_{k,M_k}}$
IoU Calculation
Intersection
t_s^j $\widehat{q}_{k,j}$ t_e^j
t_s^R $a_{k,R}$ t_e^R
Union
Element-wise multiplication
$\mathcal{J}_k$
$\widehat{c}^k_{1L}$ $\widehat{c}^k_{11}$ $\widehat{c}^k_{l2l}$ $\widehat{c}^k_{11}$ $\widehat{c}^k_{ll}$ $\widehat{c}^k_{LL}$
L l 1
1 2 l L
$\widehat{q}_{k,1}$ $\widehat{q}_{k,j}$ $\widehat{q}_{k,M_k}$
$\widetilde{\mathcal{F}}_k$
$\widetilde{f}^k_1$ $\widetilde{f}^k_j$ $\widetilde{f}^k_{M_k}$
MLP Network
$\mathcal{H}_k$
h^k_1 h^k_j $h^k_{M_k}$
1.0
0.5
0.0

(a) Coarse-grained Semantic Measuring

(b) Fine-grained Semantic Fusing

Fig. 5. Illustration of MIN. (a) The coarse-grained semantic measuring between anchor moments $\mathcal{A}_k$ and $\widehat{\mathcal{Q}}_k$ for semantic similarity metric learning; (b) The fine-grained semantic fusing between $\mathcal{J}_k$ and $\widehat{\mathcal{Q}}_k$ to form cross-modal representation $\widetilde{\mathcal{F}}_k$ and calculate the confidence map $\mathcal{H}_k$.

chosen as follows,

$$a_{k,r} = \begin{cases} \widehat{c}^k_{1,L}, & r = 1, \\ \widehat{c}^k_{1,L-\frac{r}{2}}, & 1 < r \leq R \ \& \ r \ mod \ 2 = 0, \\ \widehat{c}^k_{L-(\lfloor \frac{r}{2} \rfloor -1),L-(\lfloor \frac{r}{2} \rfloor -1)}, & 1 < r \leq R \ \& \ r \ mod \ 2 \neq 0. \end{cases} \tag{2}$$

Based on $\mathcal{A}_k$, we then build up a visual semantic tree, dubbed as $\mathcal{I}_k$. It divides the overall semantic of v_k into P (i.e., $P = \lceil R/2 \rceil$) layers. As shown in Fig. 4, the first anchor moment $a_{k,1}$ ($\widehat{c}^k_{1,L}$) is treated as the root of $\mathcal{I}_k$, since it can cover the semantic of all moment candidates. In the following, we rename it as $\mathcal{S}^{1,l}_{\mathcal{C}_k}$. Subsequently, two adjacent anchor moments $a_{k,2}$ ($\widehat{c}^k_{1,L-1}$) and $a_{k,3}$ ($\widehat{c}^k_{L,L}$) are considered as the left and right semantic nodes of the 2-th layer, denoted as $\mathcal{S}^{2,l}_{\mathcal{C}_k}$ and $\mathcal{S}^{2,r}_{\mathcal{C}_k}$, respectively. Therein, $\mathcal{S}^{2,l}_{\mathcal{C}_k}$ can cover the semantic of moments $\widehat{c}^k_{i,j}$ ($1 \leq i \leq j \leq L-1$). Analogously, $a_{k,R-1}$ and $a_{k,R}$ are set as two semantic nodes of the P-th layer. Moreover, for two adjacent semantic layers, $\mathcal{S}^{p,l}_{\mathcal{C}_k}$ and $\mathcal{S}^{p,r}_{\mathcal{C}_k}$ are both subsets of $\mathcal{S}^{p-1,l}_{\mathcal{C}_k}$, i.e., $\mathcal{S}^{p,l}_{\mathcal{C}_k} \subset \mathcal{S}^{p-1,l}_{\mathcal{C}_k}$, $\mathcal{S}^{p,r}_{\mathcal{C}_k} \subset \mathcal{S}^{p-1,l}_{\mathcal{C}_k}$ ($1 \leq p \leq P$).

C. Query Representation Module

To learn query representations, we directly utilize the pre-trained Bidirectional Encoder Representations from Transformers (BERT) [23], [24], instead of the word embedding based Bi-LSTM and its variants [25], [26], to obtain the corresponding semantic features for all queries, expressed as $\widetilde{\mathcal{Q}}_k = \{\widetilde{q}_{k,1}, \ldots, \widetilde{q}_{k,j}, \ldots, \widetilde{q}_{k,M_k}\}$, where $\widetilde{q}_{k,j}$ denotes the feature representation of the j-th query related to v_k. Subsequently, we utilize a one-layer fully connected network to obtain the final query representation $\widehat{\mathcal{Q}}_k$, formulated as follows,

$$\begin{aligned} \widehat{\mathcal{Q}}_k &= \sigma_q(\mathbf{W}_q \widetilde{\mathcal{Q}}_k + \mathbf{b}_q), \\ &= \{\widehat{q}^k_1, \ldots, \widehat{q}^k_j, \ldots, \widehat{q}^k_{M_k}\}, \end{aligned} \tag{3}$$

where Symbol σ_q, $\mathbf{W}_q$ and $\mathbf{b}_q$ respectively denote the ReLU function [27], weight matrix and bias vector. $\widehat{q}^k_j$ refers to the final representation of the j-th query related to v_k.

D. Multi-Granularity Interaction Module

Having obtained representations $\widehat{\mathcal{C}}_k$ and $\widehat{\mathcal{Q}}_k$, we develop the MIN to explore cross-modal semantic correlation.

1) Coarse-Grained Semantic Alignment: We project the anchor moments $\mathcal{A}_k \subset \widehat{\mathcal{C}}_k$ and the corresponding $\widehat{\mathcal{Q}}_k$ into a shared isomorphic representation space for semantic similarity metric learning. As shown in Fig. 5 (a), we utilize Intersection over Union (IoU) [28] to obtain the cross-modal semantic similarity of each "*moment-query*" pair. Specifically, the IoU between each "*moment-query*" pair is calculated as follows,

$$IoU^{a_{k,p}}_{\widehat{q}_{k,j}} = \frac{\min\left(t_e^j, t_e^p\right) - \max\left(t_s^j, t_s^p\right)}{\max\left(t_e^j, t_e^p\right) - \min\left(t_s^j, t_s^p\right)}, \tag{4}$$

where t_s^p and t_e^p respectively denote the start and end points of the p-th anchor moment $a_{k,p}$, and t_s^j and t_e^j are the ground truth start and end points of the target moment depicted by the j-th query $\widehat{q}_{k,j}$.

Furthermore, based on the IoU scores of "*moment-query*" pairs, we construct the similarity matrix $\mathcal{M}_k^*$ for preserving the intrinsic semantic similarities of them. Subsequently, we propose a loss function Γ_1 for semantic similarity metric learning,

$$\Gamma_1 = \sum_k \left(\left\| \mathcal{A}_k^T \widehat{\mathcal{Q}}_k - d\mathcal{M}_k^* \right\|_F^2 \right), \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, d refers to the dimension of multi-modality (i.e., anchor moment and query) features, and $\mathcal{M}_k^*$ is the cross-modal similarity matrix.

2) Fine-Grained Semantic Alignment: We adopt element-wise multiplication $\circledast$ to fuse multi-modal features. As shown in Fig. 5 (b), the moment based temporal-spatial position map $\mathcal{J}_k$ is successively fused with each $\widehat{q}_j^k \in \widehat{\mathcal{Q}}_k$ to form the corresponding cross-modal representation map $\widehat{\mathcal{F}}_k = \{\widetilde{f}_j^k\}_{j=1}^{M_k}$. Concretely, for each $\widetilde{f}_j^k \in \widehat{\mathcal{F}}_k$, it is calculated as follows,

$$\widetilde{f}_j^k = \left[\widehat{q}_j^k \circledast \widehat{c}_{m,n}^k\right]_{1\le m\le n}^{L,L}, \tag{6}$$

where $\circledast$ denotes the element-wise multiplication.

Based on $\widehat{\mathcal{F}}_k$, the confidence map set ($\mathcal{H}_k$) can be obtained by a MLP network. In Fig. 5 (b), each $\hbar_j^k \in \mathcal{H}_k$ indicates the matching scores between the j-th query $\widehat{q}_j^k$ and $\varDelta$ moment candidates in $\mathcal{J}_k$. Specially, the darker the color is, the more similar the semantics will be, and vice versa. To generate confidence scores more accurately, we also adopt corresponding IoU values as the supervision information and propose a binary cross entropy loss function for semantic fusing learning between $\mathcal{J}_k$ and $\widehat{\mathcal{Q}}_k$, formulated as follows,

$$\Phi_k = \sum_{j=1}^{M_k} \frac{1}{\varDelta} \sum_{i=1}^{\varDelta} \left[o_i^j \log p_i^j + \left(1 - o_i^j\right) \log\left(1 - p_i^j\right)\right], \tag{7}$$

where o_i^j denotes the IoU value between the i-th moment candidate and the j-th query, p_i^j is the corresponding similarity score of them in confidence map $\hbar_j^k$. Therefore, the overall loss function for N videos and their related queries is formulated as follows,

$$\Gamma_2 = \sum_{k=1}^{N} \Phi_k. \tag{8}$$

The final objective function of our CLEAR model is the combination of the above two (Γ_1 and Γ_2),

$$\Psi = \Gamma_1 + \alpha\Gamma_2, \tag{9}$$

where α is the non-negative trade-off parameter. Obviously, multi-granularity (coarse/fine-grained) semantic alignment is integrated into the same cross-modal learning framework.

E. Inference

After the model training has been completed, our proposed CLEAR could harvest the learned cross-modal semantic knowledge for moment localization. For the given untrimmed video v_x and its related query set $\mathcal{Q}_x$, we first encode them by VRN and QRN, respectively. Afterwards, we utilize MIN to obtain the semantic similarity between $\widehat{\mathcal{Q}}_x$ and $\mathcal{I}_x$, as well as the confidence score map $\mathcal{H}_x$. Finally, we can localize the target moments corresponding to the given queries efficiently through coarse-to-fine semantic aligning. The overall procedure is briefly summarized in Algorithm 1.

Algorithm 1 Coarse-to-Fine Moment Localization

```
   Input: $\widehat{\mathcal{Q}}_x$, $\mathcal{A}_x$, $\mathcal{I}_x$ and $\mathcal{H}_k$.
   Output: Target moments corresponding to $\mathcal{Q}_x$.
1  Initialize: Set intra-layer similarity threshold $\mu$, $\rho$
     ($\mu < \rho$); Set $R$ to the number of anchor moments of
     $\mathcal{A}_x$;
2  for each $\widehat{q}_j^x$ in $\widehat{\mathcal{Q}}_x$ do
3      i = 1; cur = $\mathcal{S}_{\mathcal{C}_x}^{1,l}$; //Set the current semantic node to
         the root of $\mathcal{I}_x$ ;
4      while $i \le R - 2$ do
5          $x$ = ED($\widehat{q}_j^x, a_{x,i}$); // Calculate the Euclidean
             distance between query $\widehat{q}_j^x$ and anchor
             moment $a_{x,i}$;
6          $y$ = ED($\widehat{q}_j^x, a_{x,i+1}$);
7          $z$ = ED($\widehat{q}_j^x, a_{x,i+2}$);
8          if $Max(y, z) < x$ then
9              $\eta = \frac{z-y}{y}$; // Judge the intra-layer similarity;
10             if $\eta > \rho$ then
11                 cur = $\mathcal{S}_{\mathcal{C}_x}^{i+1,r}$ and break;
12             else if $\eta \le \mu$ then
13                 cur = $\mathcal{S}_{\mathcal{C}_x}^{i+1,l}$;
14                 i = i+1;
15             else
16                 cur = $\mathcal{S}_{\mathcal{C}_x}^{i,l} - \mathcal{S}_{\mathcal{C}_x}^{i+1,l} - \mathcal{S}_{\mathcal{C}_x}^{i+1,r}$ and break;
17         else
18             cur = $\mathcal{S}_{\mathcal{C}_x}^{i,l}$ and break;
19     Obtain the matching scores from $\hbar_j^x$ related to the
         refined semantic node: $cur$;
20     Output the target moment of $\widehat{q}_j^x$ by ranking
         matching scores;
```

Taking $\widehat{q}_2^x$ from $\mathcal{Q}_x$ in Fig. 6 as an example, after completing the corresponding similarity evaluation between $\widehat{q}_2^x$ (in yellow) and three anchor moments (i.e., $a_{x,1}$, $a_{x,2}$, and $a_{x,3}$ in blue), we have two observations: 1) compared to $a_{x,1}$, $\widehat{q}_2^x$ has more similar semantics to $a_{x,2}$ and $a_{x,3}$. We then continue searching for more appropriate semantic node from the second layer of $\mathcal{I}_x$; 2) the semantics of $\widehat{q}_2^x$ and $a_{x,2}$ are closer, we hence obtain the refined semantic node $\mathcal{S}_{\mathcal{C}_x}^{2,l}$ from $\widehat{\mathcal{I}}_x$, corresponding to the 13-th line in Algorithm 1. Furthermore, guided by $\mathcal{S}_{\mathcal{C}_x}^{2,l}$, the target moment is localized from the yellow area of $\hbar_2^x$. Analogously, we can effectively localize the target moments related to other three queries from the relatively small areas (highlighted in corresponding color) of their confidence maps.

IV. Experiments

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

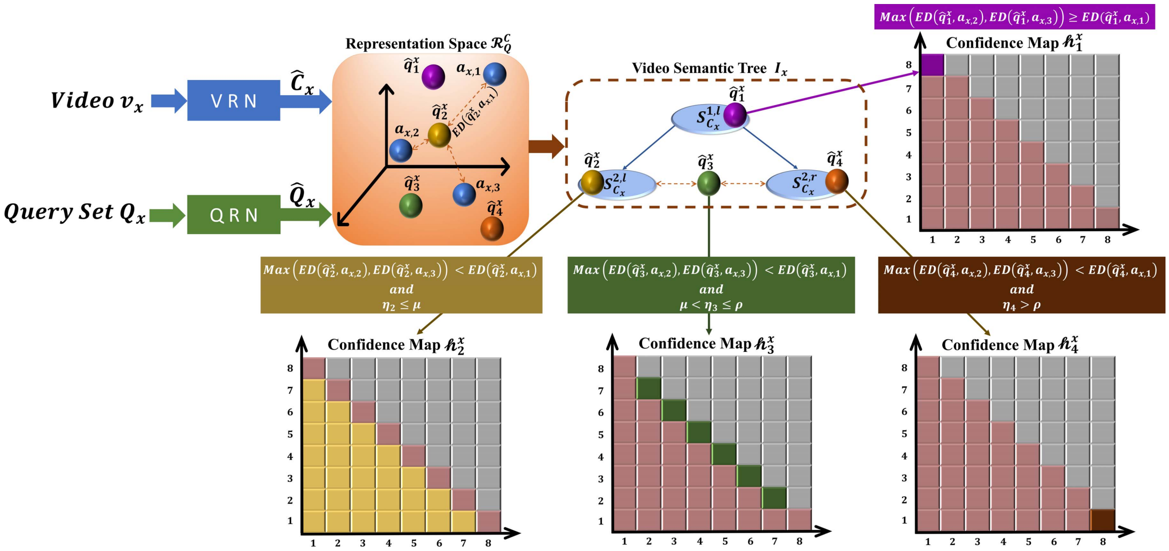

Video v_x
VRN
$\widehat{C}_x$
Query Set Q_x
QRN
$\widehat{Q}_x$
Representation Space $\mathcal{R}_Q^C$
$\widehat{q}_1^x$
$a_{x,1}$
$\widehat{q}_2^x$
$a_{x,2}$
$ED(\widehat{q}_2^x, a_{x,1})$
$\widehat{q}_3^x$
$a_{x,3}$
$\widehat{q}_4^x$
Video Semantic Tree I_x
$S_{C_x}^{1,l}$
$S_{C_x}^{2,l}$
$S_{C_x}^{2,r}$
$Max\left(ED(\widehat{q}_1^x, a_{x,2}), ED(\widehat{q}_1^x, a_{x,3})\right) \geq ED(\widehat{q}_1^x, a_{x,1})$
Confidence Map $\hbar_1^x$
$Max\left(ED(\widehat{q}_2^x, a_{x,2}), ED(\widehat{q}_2^x, a_{x,3})\right) < ED(\widehat{q}_2^x, a_{x,1})$ and $\eta_2 \leq \mu$
$Max\left(ED(\widehat{q}_3^x, a_{x,2}), ED(\widehat{q}_3^x, a_{x,3})\right) < ED(\widehat{q}_3^x, a_{x,1})$ and $\mu < \eta_3 \leq \rho$
$Max\left(ED(\widehat{q}_4^x, a_{x,2}), ED(\widehat{q}_4^x, a_{x,3})\right) < ED(\widehat{q}_4^x, a_{x,1})$ and $\eta_4 > \rho$
Confidence Map $\hbar_2^x$
Confidence Map $\hbar_3^x$
Confidence Map $\hbar_4^x$
1 2 3 4 5 6 7 8

Fig. 6. A simple illustration of our inference process. VRN first divides the given v_x into 8 equal video units, and then obtains the enhanced moment representation set $\widehat{\mathcal{C}}_x$ and the two-layer semantic tree $\mathcal{I}_x$. Meanwhile, QRN encodes queries into $\widehat{\mathcal{Q}}_x$. Thereafter, the cross-modal semantic similarity and corresponding confidence score maps $\mathcal{H}_x$ can be obtained by MIN. According to Algorithm 1, we can determine the appropriate range in $\mathcal{I}_x$ and obtain the refined $\hbar_j^x$ for moment localization.

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

A. Experimental Settings

1) Datasets: In this paper, we adopted two benchmark datasets, namely ActivityNet Captions [29] and TACoS [6] to evaluate our proposed model. ActivityNet Captions contains 20,000 untrimmed videos and more than 70,000 natural language queries along with temporal annotations. As to TACoS, it is derived from MPII Cooking Composite Activities dataset [30] and contains 127 videos related to more than 18,000 natural language queries and temporal annotations. Compared to ActivityNet Captions, the video contents of TACoS are limited to single cooking scenes, but each video involves much more target moments to be localized, thus may make moment localization more difficult.

2) Evaluation Metrics: To thoroughly measure our model and the baselines, we selected "R@n, IoU $= m$" designed by [28] as the evaluation metric. In the following, we utilized $R(n, m)$ to denote "*R@n, IoU* $= m$" and set the unified evaluation criteria $R(n, m)$ with $n \in \{1, 5\}$ and $m \in \{0.3, 0.5, 0.7\}$. Moreover, we employed the total localization time ($\mathcal{R}_T$) and the average localization time ($\mathcal{R}_A$) as the efficiency evaluation metrics.

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 units are input into the pre-trained C3D [19] to produce features for video units. These 500-d and 4,096-d features are adopted as the local features for ActivityNet Captions and TACoS, respectively. Moreover, all parameters of CLEAR are initialized randomly. The adam optimizer [31], [32] is adopted to minimize the multi-task loss. Specially, in this paper, we set 2 layers (i.e., $P = 2$) semantic tree for ActivityNet Captions and 4 layers (i.e., $P = 4$) semantic tree for TACoS. Besides, we set hyper-parameter α to 1 for the subsequent experiments. During the training, for ActivityNet Captions, the batch size and learning rate are separately set to 32 and 0.0005; for TACoS, the batch size and learning rate are set to 16 and 0.001, respectively. In addition, we empirically set the maximum number of epochs as 200 with the necessary early stopping strategy to ensure convergence. All experiments are conducted over a workstation equipped with Ubuntu 16.04.6 LTS, Intel Xeon E7 CPU, 1 TB Memory and 4 × NVIDIA RTX 2080Ti GPUs.

B. On Model Comparison (RQ1)

In order to justify the effectiveness of our proposed CLEAR, we compared it with seven state-of-the-art baselines[2]: MCN [3], ACRN [4], CTRL [6], ABLR [7], CMIN [10], QSPN [16], and 2D-TAN [17].

Table I and Table II report the overall localization accuracy results of our CLEAR and baselines on ActivityNet Captions and TACoS datasets, respectively. According to these results, 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

[2] ABLR directly regresses the temporal coordinates based on maximum co-attention weights/features, it hence merely generates R@1 results. Moreover, we adopted the latest extended version of CMIN [10]. In addition, there are two variants of 2D-TAN model: 2D-TAN Pool and 2D-TAN Conv, therefore we took the average $R(n, m)$ of them as the overall performance of 2D-TAN for fair comparison.

TABLE I

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%	1.28E-32
CTRL	10.34%	29.01%	47.43%	37.54%	59.17%	75.32%	1.90E-29
ACRN	11.25%	31.67%	49.70%	38.57%	60.34%	76.50%	5.58E-28
QSPN	13.43%	33.26%	52.13%	40.78%	62.39%	77.72%	5.85E-27
ABLR	15.71%	36.79%	55.67%	-	-	-	4.14E-24
CMIN	24.48%	44.62%	**64.41%**	52.96%	69.66%	82.39%	2.40E-05
2D-TAN	26.96%	44.28%	59.10%	62.11%	76.89%	85.59%	3.77E-06
CLEAR	**28.05%**	**45.33%**	59.96%	**62.13%**	**77.26%**	**85.83%**	-

TABLE II

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%	7.12E-29
CTRL	6.98%	13.08%	18.17%	15.60%	26.13%	35.97%	6.78E-25
ACRN	8.28%	14.11%	20.06%	15.43%	26.30%	38.11%	1.94E-24
QSPN	7.56%	15.23%	20.15%	14.35%	25.30%	36.72%	6.55E-24
ABLR	6.13%	9.30%	18.90%	-	-	-	2.19E-26
CMIN	10.21%	19.57%	27.33%	18.16%	28.53%	43.35%	1.79E-21
2D-TAN	12.47%	25.23%	36.23%	25.21%	44.64%	57.40%	1.72E-16
CLEAR	**15.54%**	**30.27%**	**42.18%**	**26.75%**	**51.76%**	**63.61%**	-

considers limited pre- and post-moment extension as context, it hence fails to model the longer-term dependencies.

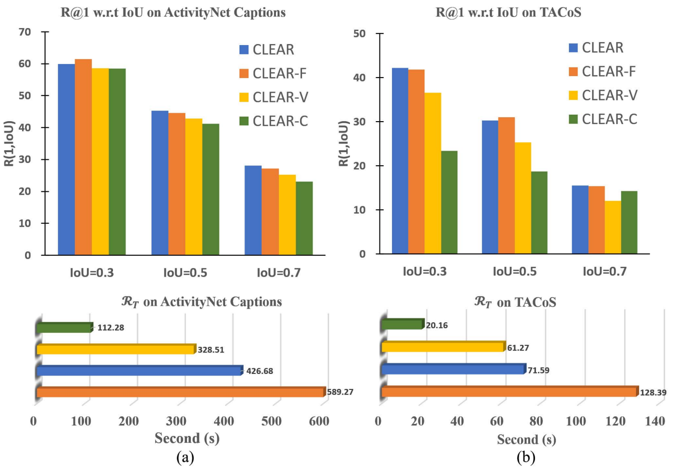
- ACRN, QSPN, and ABLR obtain higher localization accuracy than MCN and CTRL. This reflects that the attention mechanism indeed highlight crucial modality information to enhance the corresponding representation and overall localization performance.
- CMIN and 2D-TAN have relatively higher localization accuracy than other baselines. The observed results verify that the multi-stage cross-modal interaction and long-range semantic dependency modeling are critical for improving localization performance.
- CLEAR achieves the highest localization accuracy, surpassing all state-of-the-art baselines, especially on TACoS. As mentioned above, although TACoS does bring greater challenges on moment localization, CLEAR achieves the best performance, verifying the effectiveness of our coarse-to-fine semantic alignment.

Furthermore, we conducted a significance test between CLEAR and each baseline regarding R(1,0.5) based on 20-round results. All the p-values are smaller than 0.05, indicating that the advantage of our CLEAR is statistically significant.

C. On Component Analysis (RQ2)

We conducted ablation studies on our proposed model. Concretely, we omitted specific components to generate the corresponding ablation models as follows:

- CLEAR-V: We removed Bi-TCN in VRN and directly adopted feature pooling for video encoding.
- CLEAR-C: We eliminated the fine-grained component in MIN and adopted coarse-grained semantic alignment, i.e., semantic similarity evaluation in the shared isomorphic representation space for moment localization.



(a)
(b)

Fig. 7. Localization accuracy and efficiency comparison between our CLEAR and its variants on ActivityNet Captions and TACoS datasets. (a) The reported results based on R@1 w.r.t IoU $\in \{0.3, 0.5, 0.7\}$ and $\mathcal{R}_T$ on ActivityNet Captions dataset; (b) The reported results based on R@1 w.r.t IoU $\in \{0.3, 0.5, 0.7\}$ and $\mathcal{R}_T$ on TACoS dataset.

- CLEAR-F: We removed our proposed coarse-grained component and utilized fine-grained semantic alignment to sequentially retrieve the highest similarity score in the entire confidence map for moment localization.

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

- CLEAR-C has the worst localization accuracy on two datasets, especially on the more challenging dataset TACoS. This phenomenon reveals that only relying on the coarse-grained semantic metrics cannot accurately capture the complex semantic similarity, therefore failing to complete effective moment localization. However, since CLEAR-C discards the relatively time-consuming fine-grained semantic matching, CLEAR-C is more efficient than the other three competitors.
- CLEAR-V achieves relatively low accuracy but relatively high efficiency. This indicates that on the one hand, ignoring Bi-TCN does improve processing efficiency; on the other hand, due to the lack of long-term dependencies acquisition for video encoding, the accuracy may be seriously affected. As shown in Fig. 7 (b), in the case of $IoU = 0.7$ on TACoS, the accuracy is severely degraded.
- CLEAR-F delivers relatively high accuracy but the lowest efficiency, which demonstrates that the lack of coarse-grained semantic pruning cannot effectively improve the overall localization efficiency.
- CLEAR has the relatively high localization accuracy and efficiency on both datasets. The experimental results not only demonstrate that CLEAR achieves the best performance balance between accuracy and efficiency, but also verify that the Bi-TCN, coarse-grained, and fine-grained semantic alignment are all helpful for moment localization.

TABLE III

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

Method	Dataset			
	ActivityNet Captions		TACoS	
	$\mathcal{R}_A$ (s)	$\mathcal{R}_T$ (s)	$\mathcal{R}_A$ (s)	$\mathcal{R}_T$ (s)
MCN	5.23	89,021.56	11.53	47,095.03
CTRL	2.57	43,715.49	4.90	19,989.97
ACRN	3.31	56,425.11	6.14	25,087.73
ABLR	0.04	765.98	0.18	717.68
CMIN	0.67E-2	114.51	0.96E-2	39.02
2D-TAN	0.74E-1	1,256.41	0.29	1,204.24
CLEAR	0.25E-1	426.68	1.75E-2	71.59

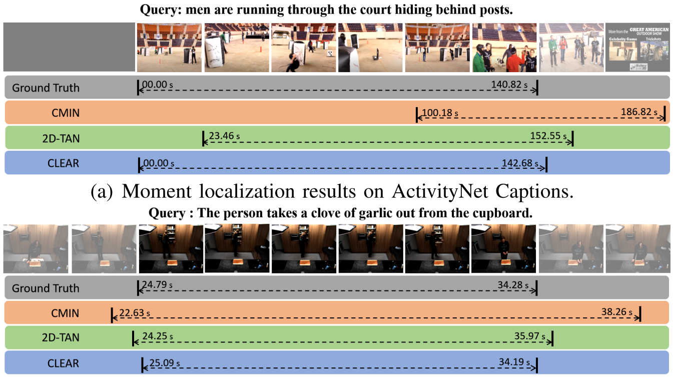
D. On Efficiency Analysis (RQ3)

Table III reports localization efficiency comparison results between our CLEAR and baselines on ActivityNet Captions and TACoS datasets. By analyzing Table III, we could find that:

- The efficiency of MCN, CTRL, and ACRN are relatively low. Because they all adopt densely sliding window based moment candidates generation and iteratively traversal retrieving for the target moment related to the corresponding query. They hence fail to complete moment localization efficiently.
- The $\mathcal{R}_A$ and $\mathcal{R}_T$ of ABLR and 2D-TAN are smaller than the above three models. This may be ascribed to 1) ABLR evenly divides the original video into the corresponding moment candidates and directly performs temporal coordinates regression; 2) 2D-TAN utilizes the sparse sampling strategy for efficient cross-domain feature correlation modeling.
- CMIN obtains relatively higher efficiency than other baselines. The main reason is that CMIN separately adopts PCA and downsample strategies to perform corresponding reduction operations on the feature dimension and length. This not only accelerates the cross-modal interaction, but also improves the overall efficiency.
- CLEAR has higher efficiency than all the baselines except CMIN. This may be because CLEAR does not adopt feature dimensionality reduction to accelerate representation encoding. Moreover, it indeed takes certain computational overhead to realize the multi-granularity semantic alignment, resulting in lower efficiency than CMIN. However, by jointly analyzing the experimental results of $\mathcal{R}_T$ in Fig. 7 and Table III, the efficiency of the ablation model CLEAR-C is higher than that of CMIN. Therefore, our model CLEAR comprehensively achieves high accuracy and efficiency.

E. Qualitative Results

To qualitatively validate the effectiveness of our proposed CLEAR model, we compared our model with two best baselines (CMIN, 2D-TAN) on two examples from ActivityNet Captions and TACoS datasets, respectively. As shown in Fig. 8, our proposed CLEAR achieves more accurate localization results than the other two competitors, i.e., the target moments returned by our model have the larger IoU with the corresponding ground truth moments. The reasons for these results may be two-folds: 1) our proposed CLEAR can well capture hierarchical semantic dependency and temporal-spatial position relationships as compared to CMIN and 2D-TAN; and 2) CLEAR utilizes multi-granularity semantic alignment for cross-modal correlation modeling, which can significantly improve the localization performance.



(a) Moment localization results on ActivityNet Captions.

(b) Moment localization results on TACoS.

Fig. 8. Moment localization results on ActivityNet captions and TACoS datasets. All the above figures are the R@1 results.

V. CONCLUSION AND FUTURE WORK

In this paper, we present a deep cross-modal semantic alignment network CLEAR for moment localization. Concretely, we first design a dual-path neural network for video moment semantic modeling and query intention understanding, respectively. Thereafter, we develop a multi-granularity interaction module to perform effective cross-modal semantic alignment for target moment localization. Finally, we conduct extensive experiments on two public benchmark datasets to justify the superiority of our CLEAR model over several state-of-the-art competitors. As a byproduct, we have released codes and parameter settings to facilitate research in this community.

In the future, we plan to deepen and widen our work from the following four aspects: 1) we intend to integrate the necessary knowledge distillation strategies [33], [34] and query-guided attentive graph convolution networks [35], [36] into our model for semantic consistency and complimentary modeling, thereby improving localization accuracy. 2) We will incorporate the multi-view metric learning [37], [38] into our model to quickly filter out the irrelevant moment candidates for reducing the searching overhead, whereby boosting the overall efficiency. 3) we plan to integrate the deep collaborative embedding [39] and semantic guided feature selection approaches [40], [41] into our model, while achieving weakly supervised moment localization [42]. And 4) we desire to adopt CLEAR as a useful multi-modal representation learning model into a wide range of multimedia application scenarios, such as surveillance video analysis [43], [44], video recommendation [45], [46] and multi-modal dialog system [47], [48].

REFERENCES

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

[2] 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.

[3] 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.*, Oct. 2017, pp. 5803–5812.

[4] 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.

[5] Y. Hu, M. Liu, X. Su, Z. Gao, and L. Nie, "Video moment localization via deep cross-modal hashing," *IEEE Trans. Image Process.*, vol. 30, no. 347, pp. 4667–4677, Apr. 2021.

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

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

[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] 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.

[10] Z. Lin, Z. Zhao, Z. Zhang, Z. Zhang, and D. Cai, "Moment retrieval via cross-modal interaction networks with query reconstruction," *IEEE Trans. Image Process.*, vol. 29, no. 283, pp. 3750–3762, Jan. 2020.

[11] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1961–1970.

[12] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1049–1058.

[13] 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.*, Oct. 2017, pp. 3628–3636.

[14] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1130–1139.

[15] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," *Int. J. Comput. Vis.*, vol. 128, no. 1, pp. 74–95, Jan. 2020.

[16] 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. AAAI Conf. Artif. Intell.*, 2019, pp. 9062–9069.

[17] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2D temporal adjacent networks for moment localization with natural language," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1–9.

[18] S. Chang, G.-J. Qi, C. C. Aggarwal, J. Zhou, M. Wang, and T. S. Huang, "Factorized similarity learning in networks," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 60–69.

[19] 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.

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

[21] 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.

[22] 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. ACM Int. Conf. Multimedia*, 2019, pp. 1464–1472.

[23] 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.

[24] L. Qu, M. Liu, D. Cao, L. Nie, and Q. Tian, "Context-aware multi-view summarization network for image-text matching," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1047–1055.

[25] X. Chen, X. Song, S. Cui, T. Gan, Z. Cheng, and L. Nie, "User identity linkage across social media via attentive time-aware user modeling," *IEEE Trans. Multimedia*, early access, Nov. 2, 2020, doi: 10.1109/TMM.2020.3034540.

[26] X. Chen, X. Song, R. Ren, L. Zhu, Z. Cheng, and L. Nie, "Fine-grained privacy detection with graph-regularized hierarchical attentive representation learning," *ACM Trans. Inf. Syst.*, vol. 38, no. 4, pp. 1–26, 2020.

[27] 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.

[28] 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.*, Jun. 2016, pp. 4555–4564.

[29] 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.

[30] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 144–157.

[31] 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.

[32] Y. Hu, P. Zhan, Y. Xu, J. Zhao, Y. Li, and X. Li, "Temporal representation learning for time series classification," *Neural Comput. Appl.*, vol. 33, no. 8, pp. 3169–3182, Apr. 2021.

[33] X. Song, F. Feng, X. Han, X. Yang, W. Liu, and L. Nie, "Neural compatibility modeling with attentive knowledge distillation," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 5–14.

[34] X. Han, X. Song, Y. Yao, X.-S. Xu, and L. Nie, "Neural compatibility modeling with probabilistic knowledge distillation," *IEEE Trans. Image Process.*, vol. 29, no. 70, pp. 871–882, Aug. 2020.

[35] F. Liu, Z. Cheng, C. Liu, and L. Nie, "An attribute-aware attentive GCN model for attribute missing in recommendation," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 26, 2020, doi: 10.1109/TKDE.2020.3040772.

[36] F. Feng, X. He, J. Tang, and T.-S. Chua, "Graph adversarial training: Dynamically regularizing based on graph structure," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2493–2504, Jun. 2021.

[37] Y. Guo, Z. Cheng, J. Jing, Y. Lin, L. Nie, and M. Wang, "Enhancing factorization machines with generalized metric learning," *IEEE Trans. Knowl. Data Eng.*, early access, Oct. 29, 2020, doi: 10.1109/TKDE.2020.3034613.

[38] F. Liu, Z. Cheng, C. Sun, Y. Wang, L. Nie, and M. Kankanhalli, "User diverse preference modeling by multimodal attentive metric learning," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1526–1534.

[39] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2070–2083, Sep. 2019.

[40] Z. Li, J. Tang, L. Zhang, and J. Yang, "Weakly-supervised semantic guided hashing for social image retrieval," *Int. J. Comput. Vis.*, vol. 128, nos. 8–9, pp. 2265–2278, Sep. 2020.

[41] Z. Li and J. Tang, "Semi-supervised local feature selection for data classification," in *Science China Information Sciences*. China: Science in China Press, 2021.

[42] W. Yang, T. Zhang, Y. Zhang, and F. Wu, "Local correspondence network for weakly supervised temporal sentence grounding," *IEEE Trans. Image Process.*, vol. 30, no. 242, pp. 3252–3262, Feb. 2021.

[43] M. Liu, L. Qu, L. Nie, M. Liu, L. Duan, and B. Chen, "Iterative local-global collaboration learning towards one-shot video person re-identification," *IEEE Trans. Image Process.*, vol. 29, no. 700, pp. 9360–9372, Oct. 2020.

[44] 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. 18, pp. 1235–1247, Oct. 2019.

[45] G.-J. Qi, "Hierarchically gated deep networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2267–2275.

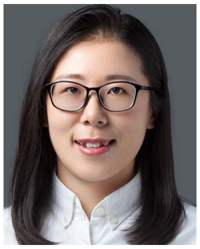
[46] G.-J. Qi, W. Liu, C. Aggarwal, and T. Huang, "Joint intermodal and intramodal label transfers for extremely rare or unseen classes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1360–1373, Jul. 2017.

[47] L. Nie, Y. Li, F. Feng, X. Song, M. Wang, and Y. Wang, "Large-scale question tagging via joint question-topic embedding learning," *ACM Trans. Inf. Syst.*, vol. 38, no. 2, pp. 1–23, Mar. 2020.

[48] Y. Guo, Z. Cheng, L. Nie, Y. Liu, Y. Wang, and M. Kankanhalli, "Quantifying and alleviating the language prior problem in visual question answering," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 75–84.

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 IEEE TRANSACTIONS ON IMAGE PROCESSING, *Science China Information Sciences*, and *Neural Computing and Applications*. His research interests include information retrieval, data mining, and explainable AI. He has served as a PC Member for ACM MM and AAAI, and a Reviewer for IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS (TII) and IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE).

Liqiang Nie (Senior Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong University and the Ph.D. degree from the National University of Singapore (NUS). 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. He is also the Adjunct Dean of the Shandong AI Institute. He has published around 100 papers in the top conferences or around 100 articles in the top journals, with 10,000 Google Scholar citations. His research interests lie primarily in multimedia computing and information retrieval. He was granted several awards, such as the SIGIR 2019 (best paper honorable mention), the ACM SIGMM Rising Star 2020, the DAMO Academy Young Fellow in 2020, and the AI 2000 the most Influential Scholar in artificial intelligence and MIT TR35 China. He serves as the PC Chair for ICIMCS 2017, PCM 2018, ChinaMM 2020, and the Area Chair for ACM MM 2018–2021. He is an AE of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), *ACM TOMM*, and *Information Science*.



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 (TIP). Her research interests include 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*.

Kun Wang 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.

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.

Xian-Sheng Hua (Fellow, IEEE) received the B.S. degree and the Ph.D. degree in applied mathematics from Peking University, Beijing, China, in 1996 and 2001, respectively. In 2001, he joined Microsoft Research Asia, as a Researcher. He was a Principal Research and Development Lead in multimedia search for the Microsoft search engine, Bing, USA, from 2011 to 2013. He was a Senior Researcher with Microsoft Research Redmond from 2013 to 2015. He is currently the VP/a Distinguished Engineer of Alibaba Group, leading the Artificial Intelligence Center and the City Brain Lab of DAMO Academy. He is also an ACM Distinguished Scientist. He has authored or coauthored more than 200 research articles and has 60 granted patents. His research interests include big multimedia data search, advertising, understanding, and mining, and pattern recognition and machine learning. He was one of the recipients of the 2008 MIT Technology Review TR35 Young Innovator Award for his outstanding contributions on video search. He was a recipient of the Best Paper Awards at ACM Multimedia 2007 and the Best Paper Award of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (CSVT) in 2014. He served as the Program Co-Chair for IEEE ICME 2013, ACM Multimedia 2012, and IEEE ICME 2012. He also served as the General Co-Chair for ACM Multimedia 2020. He served or is serving as an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA and *ACM Transactions on Intelligent Systems and Technology*.