Multi-Modal Interaction Graph Convolutional Network for Temporal Language Localization in Videos

Zongmeng Zhang, Xianjing Han, Xuemeng Song, *Member, IEEE*,

Yan Yan, and Liqiang Nie, *Senior Member, IEEE*

Abstract—This paper focuses on tackling the problem of temporal language localization in videos, which aims to identify the start and end points of a moment described by a natural language sentence in an untrimmed video. However, it is non-trivial since it requires not only the comprehensive understanding of the video and sentence query, but also the accurate semantic correspondence capture between them. Existing efforts are mainly centered on exploring the sequential relation among video clips and query words to reason the video and sentence query, neglecting the other intra-modal relations (*e.g.*, semantic similarity among video clips and syntactic dependency among the query words). Towards this end, in this work, we propose a Multi-modal Interaction Graph Convolutional Network (MIGCN), which jointly explores the complex intra-modal relations and inter-modal interactions residing in the video and sentence query to facilitate the understanding and semantic correspondence capture of the video and sentence query. In addition, we devise an adaptive context-aware localization method, where the context information is taken into the candidate moments and the multi-scale fully connected layers are designed to rank and adjust the boundary of the generated coarse candidate moments with different lengths. Extensive experiments on Charades-STA and ActivityNet datasets demonstrate the promising performance and superior efficiency of our model.

Index Terms—Temporal language localization, graph convolutional network, video and language.

I. INTRODUCTION

IN RECENT years, the flourish of multimedia devices has promoted the unprecedented growth of videos in various domains, highlighting the necessity of automatic video processing. In particular, owing to its great potential in the

security domain, especially the video surveillance, temporal action localization in videos that aims to identify the start and end points of an action query has attracted more and more researchers [1]–[8]. Traditionally, the action queries are described by keywords from a pre-defined set, like running or jumping. Nevertheless, in real-world scenarios, long untrimmed videos usually involve a large number of objects and complex activities, which are hard to be pre-defined. In light of this, in this work, we focus on the task of temporal language localization in videos, where the activity query is described by the natural language.

In fact, some research efforts have been dedicated to the task of temporal language localization in videos [9]–[17]. Since the semantic correspondence capture between the video and sentence query plays a pivotal role in this context, earlier studies [9], [18]–[21] mainly adopt the sliding window strategy to generate dense candidate moments and then explore the inter-modal interactions between the candidate moments and sentence query with various attention mechanisms [22]–[24]. Though these methods have achieved promising performance, they mainly focus on the inter-modal interactions, ignoring the sequential dependencies among video clips or query words, which are also crucial to the understanding of the video or sentence query. Towards this end, some efforts [11], [13], [25], [26] have been made to capture the sequential relation among the video clips and query words with the recurrent neural networks (RNN) [27]. In fact, besides the sequential relation, there also exist other intra-modal relations in the video and sentence query, such as the semantic similarity among video clips and syntactic dependency among the query words, as shown in Figure 1. These relations are essential for the comprehensive understanding of the video and sentence query, yet they are overlooked by existing methods as it is challenging to capture these relations via a sequential manner. In addition, to facilitate the flexible moment localization, where the target moment length is unfixed, existing methods [9], [11], [18], [28] mainly utilize the pooling or sampling strategy to regularize the representations of candidate moments. Nevertheless, the pooling or sampling strategy tends to merely retain partial prominent information of the candidate moment, and hence they inevitably suffer from the enormous information loss regarding the candidate moments, resulting in the suboptimal performance.

To address the aforementioned issues, in this work, we devise a Multi-modal Interaction Graph Convolutional Network (MIGCN) for temporal language localization. As shown
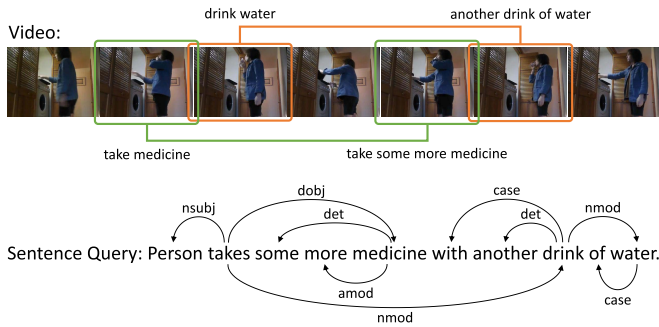
Fig. 1. An example of the intra-modal relations in the video and sentence query. The clips in boxes of the same color present high semantic similarity, *e.g.*, "drink water" and "another drink of water". The arrows between words show their syntactic dependencies, *e.g.*, "another" is the determiner of "drink". In a sense, these relations can facilitate the video reasoning and strengthen the action understanding of the sentence query, and hence promote the moment localization for the query "Person takes some more medicine with another drink of water".

in Figure 2, both the intra-modal relations and inter-modal interactions residing in the video and sentence query are comprehensively explored by the graph convolutional network (GCN) [29] which has been proven to be effective in propagating information among data with complex relations [30]–[37]. In particular, we first adopt BiGRU [38] to encode the video and sentence query, where we split the video into several clips without overlaps to reduce the computational complexity. To promote the representation learning of the video and sentence query by jointly modeling the intra-modal relations and inter-modal interactions, we then introduce the multi-modal interaction graph, comprising two types of nodes (*i.e.*, clip node and word node), and edges that compile intra-modal relations and inter-modal interactions among the clips and words. Specifically, we incorporate the temporally adjacent relation and semantic correlation among video clips, and the syntactic dependency among words as the intra-modal relations, and take semantic correspondence between the video and sentence query as the inter-modal interactions. Based on this graph, we utilize the graph convolution to fulfil both the intra- and inter-modal refinement over the node representation learning. To facilitate the target moment localization with flexible lengths, we employ sliding windows with different sizes to generate a set of coarse candidate moments with different lengths. As for the ranking and boundary adjustment of these coarse candidate moments, we devise an adaptive context-aware localization method, where the context information is considered to learn the ranking scores and boundary offsets of the coarse candidate moments with less information loss through the multi-scale fully connected layers.

The main contributions of this work can be summarized in threefold:

- We propose a multi-modal interaction graph convolutional neural network, where a graph comprising both video clip nodes and word nodes is constructed to jointly explore the complex intra-modal relations and inter-modal interactions residing in the video and sentence query. To the best of our knowledge, this is the first attempt

to construct a multi-modal graph to tackle the problem of temporal language localization in videos.
- We devise an adaptive context-aware localization method, which employs the multi-scale fully connected layers and considers the context information, to rank the variable-length candidate moments with less information loss and promote the accurate candidate moment boundary adjustment.
- Extensive experiments demonstrate the promising performance and efficiency of our MIGCN, compared with the state-of-the-art methods on two large datasets, Charades-STA [9] and ActivityNet [39]. As a byproduct, we have released the codes and involved parameters to benefit other researchers.[1]

II. RELATED WORK

A. Temporal Language Localization in Videos

The task of temporal language localization in videos is to determine the start and end points in an untrimmed video regarding the activity described in a sentence query, which was first introduced by Gao *et al.* [9] and Hendricks *et al.* [18]. The method in [9] concatenates the representations of the candidate moment and sentence query to estimate the moment-sentence alignment score and introduces temporal regression for the moment boundary adjustment, while [18] targets at measuring the similarity between the candidate moment and sentence query in a semantic space. Given that the semantic correlation capture of the given video and sentence query constitutes a pivotal part in this task, the researchers have resorted to various attention mechanisms to enhance the interaction modeling between the video and sentence query. For example, Liu *et al.* [19] developed a memory attention mechanism to emphasize the visual features that are highly correlated to the sentence query. It is worth noting that the intra-modal relations in video and sentence query are crucial to the representation learning, yet they are overlooked by these methods. In order to tackle this issue, Yuan *et al.* [25] encoded the sequential relations via Bi-directional LSTM to generate representations of the video and sentence query with the sequential contextual information. In fact, besides the temporal relations, there exist other intra-modal relations in the video and sentence query, *e.g.*, the semantic similarity among video clips and syntactic dependency among query words, which can strengthen the representation learning. Towards this end, we target at comprehensively exploring the intra-modal relations and inter-modal interactions residing in the video and sentence query to boost the model performance on the temporal language localization task.

B. Graph Convolutional Networks

In recent years, graph convolutional networks (GNN) have drawn increasing attention due to their successful applications in various tasks [30]–[37], [40]–[42]. At the beginning,

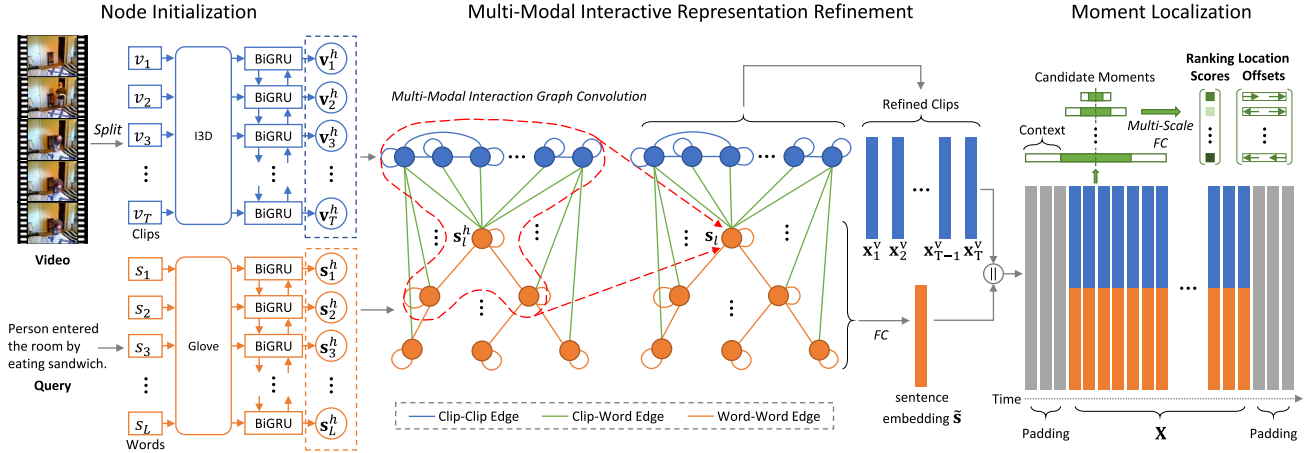[1] https://github.com/zmzhang2000/MIGCN/

Fig. 2. Pipeline of the proposed Multi-Modal Interaction Graph Convolutional Network (MIGCN). We first employ BiGRU to embed the clip and word features as the initial node representations of the multi-modal interaction graph. Then we utilize the graph convolution to refine the representation of the clip and word nodes with both the intra-modal relation and inter-modal interaction modeling. Based on the multi-modal fused clip representation matrix **X**, we use sliding windows to derive a set of coarse candidate moments with different lengths. Ultimately, we introduce the adaptive context-aware moment localization module with multi-scale fully connected layers to predict the ranking scores and location offsets of each candidate moment. The red dashed line represents the information aggregation for the node $\mathbf{s}_l^h$ by the graph convolution process.

Scarselli *et al.* [43] introduced the graph neural network for graph-focused and node-focused applications by extending recursive neural networks and random walk models. Inspired by this, Kipf *et al.* [29] presented the graph convolutional network (GCN) and defined the convolution on non-grid structures.

Since then, GCN has been employed in various tasks. For example, in natural language processing, Marcheggiani and Titov [44] employed GCN to model the syntactic dependency among words in sentence and learn the latent representations of words. In video understanding, Zhang *et al.* [45] put forward a temporal reasoning graph, which captures the temporal relation among video frames, to tackle the task of action recognition. Moreover, in the context of temporal language localization, Zeng *et al.* [3] introduced an action proposal graph to model the relations among different proposals, while Zhang *et al.* [46] presented an iterative graph adjustment network to exploit the graph-structured moment relations. Although these efforts have achieved promising performance, they only focus on enhancing the single-modal representation learning with GCN, while overlooking the potential of GCN in propagating information on multi-modal relations residing in the video and sentence query.

In fact, there have been several multi-modal based works employing GCN to promote the inter-modal interaction and achieving improved performance. For example, in the task of object grounding with textual description, Chen *et al.* [47] utilized the GCN to enhance the reasoning of the object and text, as well as boost the information passing among the object and text modalities. In addition, Bajaj *et al.* [48] proposed a grounding architecture with three connected graphs to tackle the language grounding, where a phrase graph and a visual graph are designed to boost the intra-modal representation, and then based on that a fusion graph is derived to enhance the inter-modal interaction. In this work, we focused on promoting

both the video and sentence representation learning via the multi-modal interaction graph to tackle the task of temporal language localization in videos. In addition, it is worth noting that different from the graphs with entirely learned edges in other vision-language tasks [47]–[49], our proposed multi-modal interaction graph explicitly leverages the temporally adjacent relation and semantic similarity among video clips, and the syntactic dependency in sentence by pre-defined adjacency matrices. Compared with other multi-modal graph methods, the pre-established adjacency matrices in our method are more interpretable and able to impose more powerful inductive biases to the model.

III. METHOD

A. Problem Formulation

In this work, we aim to tackle the problem of temporal language localization in videos. Suppose we have an untrimmed video $V = \{v_t\}_{t=1}^T$, which can be split into T parts, where v_t denotes the t-th video clip, and T is the number of video clips. Besides, we have a sentence query $S = \{s_l\}_{l=1}^L$, where s_l is the l-th word in the sentence and L represents the sentence length. For each sentence query, we have a ground truth start-end points of the target moment represented as (τ^s, τ^e). In a sense, our goal is to learn a mapping function $\mathcal{F}$ defined as follows:

$$\mathcal{F} : (V, S) \rightarrow (\tau^s, \tau^e). \tag{1}$$

B. Multi-Modal Interaction Graph Construction

Undoubtedly, it is essential for the temporal language localization method to comprehensively reason the given video and sentence. Existing efforts [9], [10], [18], [20], [25] mainly employ the sequential structures (*e.g.*, recurrent neural

networks) to reason the video and sentence due to their intrinsic sequential property. Nevertheless, apart from the sequential dependency, there are also other intra-model relations that can facilitate the reasoning of the video and sentence, *e.g.*, the semantic similarity among the video clips and syntactic dependency among the sentence words. To boost these intra-model relations learning, we resort to the graph neural network, where these relations can be explicitly modeled by the edges in the graph rather than implicitly excavated from training data. These edges in graph connect correlated clips and words, which impose more powerful inductive biases [50] (*i.e.*, more prior knowledge) on the neural network, and thus ease the relation learning. As the key of the temporal language localization task is to capture the semantic correspondence between the video clips and sentence query, in addition to the aforementioned intra-modal relations, we also model the inter-modal interaction among the video and sentence, and hence introduce a multi-modal interaction graph.

1) Node Initialization: The multi-modal interaction graph has two types of nodes: clip nodes and word nodes. To initialize the clip node, we employ the pre-trained model [51]–[53] to extract the visual feature $\mathbf{v}_t$ of the t-th clip. Then, to comprehensively explore the semantic information in the clip sequence, we employ Bi-directional Gated Recurrent Unit (BiGRU) to encode the whole video. In particular, the BiGRU network comprises a $\overrightarrow{GRU}^v$ moving forward from the start to the end of the video, a $\overleftarrow{GRU}^v$ moving in the opposite direction, and a fully connected layer f^v, which takes the concatenation of the t-th hidden states of the two GRUs as the input. Ultimately, the output is used as the initial clip node representation $\mathbf{v}_t^h$, which can be formulated as follows:

$$\begin{cases} \overrightarrow{\mathbf{h}}_t^v = \overrightarrow{GRU}^v(\mathbf{v}_t, \overrightarrow{\mathbf{h}}_{t-1}^v), \\ \overleftarrow{\mathbf{h}}_t^v = \overleftarrow{GRU}^v(\mathbf{v}_t, \overleftarrow{\mathbf{h}}_{t+1}^v), \\ \mathbf{v}_t^h = f^v(\overrightarrow{\mathbf{h}}_t^v \| \overleftarrow{\mathbf{h}}_t^v), \end{cases} \tag{2}$$

where $\overrightarrow{\mathbf{h}}_t^v$ and $\overleftarrow{\mathbf{h}}_t^v$ denote the t-th clip hidden states of the forward and backward GRUs, respectively. $\|$ signifies the concatenation operation and the Leaky ReLU active function is adopted for f^v.

The word node can be initialized in a similar manner. We first project each word s_l into the embedding $\mathbf{s}_l$ by Glove [54] and then utilize BiGRU to encode each word with the context information of the sentence, which can be defined as follows,

$$\begin{cases} \overrightarrow{\mathbf{h}}_l^s = \overrightarrow{GRU}^s(\mathbf{s}_l, \overrightarrow{\mathbf{h}}_{l-1}^s), \\ \overleftarrow{\mathbf{h}}_l^s = \overleftarrow{GRU}^s(\mathbf{s}_l, \overleftarrow{\mathbf{h}}_{l+1}^s), \\ \mathbf{s}_l^h = f^s(\overrightarrow{\mathbf{h}}_l^s \| \overleftarrow{\mathbf{h}}_l^s), \end{cases} \tag{3}$$

where $\overrightarrow{\mathbf{h}}_l^s$ and $\overleftarrow{\mathbf{h}}_l^s$ are the l-th word hidden states of the forward and backward GRUs, respectively. f^s is the fully connected layer with Leaky ReLU active function. $\mathbf{s}_l^h$ is the final initialization for the l-th word node.

2) Clip-Clip Edge: Intuitively, the sequence order of clips in a video reflects the temporally adjacent relation among these clips. Meanwhile, in the context of temporal language localization, the given sentence query may correspond to

multiple clips over the whole video, where these clips tend to be visually similar and semantically correlated. Consequently, to explore both the temporal and semantic correlations and enhance the representation learning of clips, we devise two types of clip-clip edges. On the one hand, for the temporally adjacent clip pairs, we define a set of temporal correlated edges as follows:

$$\mathcal{E}^t = \{(v_i, v_{i+1}) | i \in \{1, 2, \ldots, T-1\}\}, \tag{4}$$

where (v_i, v_{i+1}) represents the edge between the i-th and the next clip nodes. We set the weight of each edge $(v_i, v_{i+1}) \in \mathcal{E}^t$ as 1. On the other hand, to model the semantic correlation, we link two clip nodes if they share the similar visual content. In particular, we define a set of semantic correlated edges as follows:

$$\mathcal{E}^s = \{(v_i, v_j) | d_c(\mathbf{v}_i, \mathbf{v}_j) > \theta \wedge i \neq j\}, \tag{5}$$

where $i, j \in \{1, 2, \ldots, T\}$. $d_c(\mathbf{v}_i, \mathbf{v}_j)$ is the cosine similarity between the i-th and j-th clip features extracted by the pre-trained model, which can be computed by:

$$d_c(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i^T \mathbf{v}_j}{\|\mathbf{v}_i\|_2 \cdot \|\mathbf{v}_j\|_2}. \tag{6}$$

θ is the pre-defined threshold. We set the weight of the edge $(v_i, v_j) \in \mathcal{E}^s$ as $d_c(\mathbf{v}_i, \mathbf{v}_j)$. Finally, we can obtain the clip-clip edge set $\mathcal{E}^{vv} = \mathcal{E}^t \cup \mathcal{E}^s \cup \mathcal{E}^l$, where $\mathcal{E}^l$ is the set of self-loop edges utilized to maintain the information of the node itself. We set the weight of each self-loop edge as 1.

3) Word-Word Edge: Considering the success of the syntactic dependency graph in the query semantic understanding [44], we extract the syntactic dependency among words by Stanford CoreNLP [55] and represent each dependency relation as an edge. Accordingly, we derive a set of word-word edges denoted as follows:

$$\mathcal{E}^{ss} = \{(s_i, s_j) | \langle i, j \rangle \in \Omega \vee i = j\}, \tag{7}$$

where Ω denotes the syntactic dependency relation set extracted from the sentence query, while $\langle i, j \rangle$ indicates that there are the syntactic dependency between the i-th and the j-th words in the sentence, $i, j \in \{1, 2, \ldots, L\}$. The weight of each edge $(s_i, s_j) \in \mathcal{E}^{ss}$ is set as 1. Notably, self-loop edges are also included in $\mathcal{E}^{ss}$ to preserve the information of the words themselves.

4) Clip-Word Edge: To promote the information propagation among different modalities, apart from the intra-modal edges, we also bridge each clip and each word to constitute the inter-modal edges. In particular, we connect each clip with each word and get the clip-word edge set as follows:

$$\mathcal{E}^{vs} = \{(v_i, s_j) | i \in \{1, 2, \ldots, T\}, \quad j \in \{1, 2, \ldots, L\}\}. \tag{8}$$

Given that our goal is to localize the moment described by the sentence query, the inter-modal interactions are particularly essential. Different from the static weight setting for clip-clip edges and word-word edges, the weight for each clip-word edge (v_i, s_j) is dynamically updated according to the node similarity defined as $d_s(v_i, s_j) = (\mathbf{v}_i^h)^T \mathbf{s}_j^h$.

C. Multi-Modal Interactive Representation Refinement

Over the constructed multi-modal graph, we adopt the graph convolution process which is capable of propagating information among nodes with complex relations to enhance the clip and word representation learning. Basically, the general graph convolution can be implemented as follows:

$$\mathbf{H} = \sigma(\mathbf{AXW}), \tag{9}$$

where $\mathbf{H}$ represents the hidden representations of the nodes, and $\mathbf{A}$ is the adjacency matrix. $\mathbf{X}$ denotes the input node features, and $\mathbf{W}$ is the to-be-learned weight matrix. σ represents the non-linear operation. In a sense, each node representation can be refined according to the representations of itself and its adjacent nodes with a graph convolution operation. In our context, the graph convolution refinement over the clip and word representations consists of two aspects: intra- and inter-modal refinements.

1) Intra-Modal Refinement: According to Eqn.(9), the intra-modal node representation refinement can be implemented by:

$$\begin{cases} \tilde{\mathbf{V}} = ReLU(\mathbf{A}_{vv}\mathbf{V}\mathbf{W}_{vv}), \\ \tilde{\mathbf{S}} = ReLU(\mathbf{A}_{ss}\mathbf{S}\mathbf{W}_{ss}), \end{cases} \tag{10}$$

where $\mathbf{V} = [\mathbf{v}_1^h; \mathbf{v}_2^h; \ldots; \mathbf{v}_T^h]$ and $\mathbf{S} = [\mathbf{s}_1^h; \mathbf{s}_2^h; \ldots; \mathbf{s}_L^h]$. $\tilde{\mathbf{V}} \in \mathbb{R}^{T \times d}, \tilde{\mathbf{S}} \in \mathbb{R}^{L \times d}$ are the refined clip and word node representations, respectively. $\mathbf{W}_{vv} \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_{ss} \in \mathbb{R}^{d \times d}$ are the learnable parameters. $ReLU$ denotes the leaky ReLU function. $\mathbf{A}_{vv} \in \mathbb{R}^{T \times T}$ represents the clip node adjacency matrix, which is constructed according to the clip-clip edge set $\mathcal{E}^{vv}$, while $\mathbf{A}_{ss} \in \mathbb{R}^{L \times L}$ is the word node adjacency matrix constructed with the word-word edge set $\mathcal{E}^{ss}$. Note that if $(v_i, v_j) \notin \mathcal{E}^{vv}$, we set the value of $\mathbf{A}_{vv}(i, j)$ as 0. Similarly, if $(s_i, s_j) \notin \mathcal{E}^{ss}$, we set the value of $\mathbf{A}_{ss}(i, j)$ as 0.

2) Inter-Modal Refinement: To fulfil the inter-modal refinement, one simple method is to update the node representations of each modality according to the inter-modal adjacent relation as follows:

$$\begin{cases} \mathbf{X}^v = ReLU(\mathbf{A}_{sv}\tilde{\mathbf{S}}\mathbf{W}_{sv}), \\ \mathbf{X}^s = ReLU(\mathbf{A}_{vs}\tilde{\mathbf{V}}\mathbf{W}_{vs}), \end{cases} \tag{11}$$

where $\mathbf{X}^v$ and $\mathbf{X}^s$ are the refined clip and word node representations, respectively. $\mathbf{A}_{sv}$ and $\mathbf{A}_{vs}$ are respectively the word-clip and clip-word adjacency matrices, both constructed based on the clip-word edge set $\mathcal{E}^{vs}$. $\mathbf{W}_{sv}$ and $\mathbf{W}_{vs}$ are the parameters to be learned. Apparently, this simple graph convolution operation only refers the node representations from another modality (*e.g.*, the sentence words) to refine one modality (*e.g.*, the video clips), totally ignoring the modality inherent information, which may hinder the thorough capture of the semantic interactions between the two modalities and hence hurt the performance. Therefore, to realize the comprehensive inter-modal refinement, we resort to the gate mechanism (*i.e.*, gated graph convolution [47]), and the inter-modal refinement over the clip node representation is then formulated as follows:

$$\begin{cases} \mathbf{H}_s = \mathbf{A}_{sv}\tilde{\mathbf{S}}\mathbf{W}_{sv}, \\ \mathbf{Z}_v = sigmoid([\tilde{\mathbf{V}}\|\mathbf{H}_s]\mathbf{W}_{gate,v}), \\ \mathbf{X}^v = ReLU(\mathbf{Z}_v \circ \tilde{\mathbf{V}} + (1 - \mathbf{Z}_v) \circ \mathbf{H}_s), \end{cases} \tag{12}$$

where $\circ$ denotes the Hadamard multiplication of two matrices. $\mathbf{H}_s \in \mathbb{R}^{T \times d}$ is the word information obtained from the word node via the graph convolution, and $\mathbf{Z}_v \in \mathbb{R}^{T \times d}$ denotes the retain ratio matrix of clip representations. $\mathbf{W}_{sv} \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_{gate,v} \in \mathbb{R}^{2d \times d}$ are parameters to be learned. $\mathbf{A}_{sv} \in \mathbb{R}^{T \times L}$ is the word-clip adjacency matrix constructed according to the clip-word edge set $\mathcal{E}^{vs}$.[2]

Similarly, we summarize the inter-modal refinement over the word node representation as follows:

$$\begin{cases} \mathbf{H}_v = \mathbf{A}_{vs}\tilde{\mathbf{V}}\mathbf{W}_{vs}, \\ \mathbf{Z}_s = sigmoid([\tilde{\mathbf{S}}\|\mathbf{H}_v]\mathbf{W}_{gate,s}), \\ \mathbf{X}^s = ReLU(\mathbf{Z}_s \circ \tilde{\mathbf{S}} + (1 - \mathbf{Z}_s) \circ \mathbf{H}_v), \end{cases} \tag{13}$$

where $\mathbf{H}_v \in \mathbb{R}^{L \times d}$ is the clip information obtained from the clip node via the graph convolution, and $\mathbf{Z}_s \in \mathbb{R}^{L \times d}$ denotes the retain ratio matrix of word representations. $\mathbf{W}_{vs} \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_{gate,s} \in \mathbb{R}^{2d \times d}$ are parameters to be learned. $\mathbf{A}_{vs} \in \mathbb{R}^{L \times T}$ is the clip to word adjacency matrix.

Ultimately, we can obtain the refined clip and word representations $\mathbf{X}^v = [\mathbf{x}_1^v, \mathbf{x}_2^v, \ldots, \mathbf{x}_T^v] \in \mathbb{R}^{T \times d}$ and $\mathbf{X}^s = [\mathbf{x}_1^s, \mathbf{x}_2^s, \ldots, \mathbf{x}_L^s] \in \mathbb{R}^{L \times d}$, which encode both the intra-modal relations and inter-modal semantic interactions.

D. Moment Localization

Based on the refined clip and word representations, we resort to sliding windows with different sizes to first generate a set of coarse candidate moments with different lengths to facilitate the target moment localization with flexible lengths. To accurately localize the target moment, we propose an adaptive context-aware localization method to rank the candidate moments and adjust their boundaries, where the context information of the candidate moments is taken into account and the multi-scale fully connected layers are devised to adaptively tackle candidate moments with different lengths.

1) Candidate Moment Representation: We first generate variable-length candidate moments by a set of sliding windows with sizes of $\{\omega_m\}_{m=1}^M$. In particular, we slide the window with each size ω_m on the T video clips with a stride of δ to generate candidate moments. Be aware that we discard the candidate moment whose boundary (either the start point or the end point) exceeds the video clip range. We denote the set of candidate moments generated by ω_m as $\{(t_c^{s,m}, t_c^{e,m})\}_{c=1}^{C_m}$, where $t_c^{s,m}$ and $t_c^{e,m}$ are respectively the start and end points of the c-th candidate moment generated by the sliding window with size ω_m. In addition, considering that the information near the candidate moment is essential for adjusting the moment boundary, different from [11], [56], we take the context information into account to get more accurate boundary offsets of candidate moments. In particular, we expand both the start and end points of the c-th candidate moment, which is formulated as:

$$\begin{cases} \tilde{t}_c^{s,m} = t_c^{s,m} - \dfrac{\omega_m}{2}, \\ \tilde{t}_c^{e,m} = t_c^{e,m} + \dfrac{\omega_m}{2}. \end{cases} \tag{14}$$

[2]To keep $\mathbf{H}_s$ and $\tilde{\mathbf{S}}$ in the same scale and balance the impact of $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{S}}$ on $\mathbf{X}^v$, we perform the row-wise normalization on $\mathbf{A}_{sv}$ to facilitate the information propagation.

In a sense, each candidate moment length is doubled by this context extension. In order to measure the matching degree between each candidate moment $(t_c^{s,m}, t_c^{e,m})$ and the sentence query, we first employ a non-linear transformation to get the sentence embedding $\tilde{\mathbf{s}} \in \mathbb{R}^d$ based on the refined word representations as follows:

$$\tilde{\mathbf{s}} = ReLU(\mathbf{W}_s \mathbf{X}^s), \tag{15}$$

where $\mathbf{W}_s \in \mathbb{R}^{1 \times L}$ is the transformation parameter. Thereafter, we concatenate the sentence embedding to each clip representation of the candidate moment, i.e., $\mathbf{x}_t^v$, where $v_t \in (\tilde{t}_c^{s,m}, \tilde{t}_c^{e,m})$ and then derive the moment representation as $\mathbf{X}_c^m \in \mathbb{R}^{2\omega_m \times 2d}$. Note that we pad zero to $\mathbf{X}_c^m$ if the extended start point or end point exceeds the video clip range, i.e., $[1, T]$.

2) Adaptive Context-Aware Localization: Based on the candidate moment set, the task of temporal language localization can be converted to the candidate moment ranking problem. Similar to the previous work [9], apart from the candidate moment ranking, we also predict the offsets of the start and end point of the candidate moment with a sibling output layer to refine the location of candidate moment. Towards this end, existing methods [9], [11], [18], [28] mainly employ the pooling or sampling strategy to unify the representation dimensions of candidate moments with different lengths. Nevertheless, the pooling or sampling strategy mainly focuses on retaining the prominent information of the candidate moment, which inevitably suffers from the information loss to some extent. Towards this end, to retain the information as much as possible, we devise the multi-scale fully connected layers to adaptively process the candidate moments with different lengths. In particular, we calculate the ranking score r_c^m and offset $(d_c^{s,m}, d_c^{e,m})$ of the candidate moment $(t_c^{s,m}, t_c^{e,m})$ as follows:

$$\begin{cases} r_c^m = sum(\mathbf{X}_c^m \circ \mathbf{W}_r^m) + b_r^m, \\ d_c^{s,m} = sum(\mathbf{X}_c^m \circ \mathbf{W}_s^m) + b_s^m, \\ d_c^{e,m} = sum(\mathbf{X}_c^m \circ \mathbf{W}_e^m) + b_e^m, \end{cases} \tag{16}$$

where $sum()$ is the sum operator over all elements in a matrix. $\mathbf{W}_r^m, \mathbf{W}_s^m$ and $\mathbf{W}_e^m$ are learnable weight matrices of the multi-scale fully connected layers, which have the same dimensions with the candidate moment representation $\mathbf{X}_c^m$. b_r^m, b_s^m and b_e^m are the corresponding biases. It is worth noting that as the dimension of $\mathbf{X}_c^m$ is decided by the sliding window size, all the candidate moments generated by a specific window would share the identical parameters in the multi-scale fully connected layers. Ultimately, the rectified boundary $(\hat{\tau}_c^{s,m}, \hat{\tau}_c^{e,m})$ is computed by:

$$\begin{cases} \hat{\tau}_c^{s,m} = t_c^{s,m} + d_c^{s,m}, \\ \hat{\tau}_c^{e,m} = t_c^{e,m} + d_c^{e,m}. \end{cases} \tag{17}$$

E. Learning

As for the optimization, we utilize the alignment loss [56], ranking loss [16] and regression loss [9], where the former loss is used for encouraging the model to evaluate the alignment degree between each candidate moment and the ground truth moment, while the latter two losses are used for adjusting the ranking scores and boundaries of candidate moments.

1) Alignment Loss: Similar with [56], we employ the alignment loss to assign high ranking scores to the candidate moments in line with the ground truth moment and lower ranking scores to the other moments. In this work, we employ Intersection over Union (IoU) between each candidate moment $(t_c^{s,m}, t_c^{e,m})$ and the ground truth moment (τ^s, τ^e), denoted as γ_c^m, to represent their alignment degree.

Accordingly, we utilize binary cross entropy and formulate the alignment loss as follows:

$$\begin{cases} \hat{r}_c^m = sigmoid(r_c^m), \\ \mathcal{L}_{aln} = -\dfrac{1}{C} \displaystyle\sum_{m=1}^{M} \sum_{c=1}^{C_m} \gamma_c^m \log(\hat{r}_c^m) + (1 - \gamma_c^m) \log(1 - \hat{r}_c^m), \end{cases} \tag{18}$$

where $\hat{r}_c^m$ is the normalized ranking score, which is regarded as the predicted IoU of the c-th candidate moment generated by the sliding window with size ω_m. C_m is the number of the candidate moments generated by the sliding window with size ω_m, and C is the total number of the candidate moments. In particular, we set γ_c^m to 0, if γ_c^m is less than a pre-defined threshold λ, to promote the identification of the high-score candidate moment [56].

2) Ranking Loss: As there can be plenty of candidate moments having similar locations and lengths with the target moment, it is hard to select the optimal candidate moment among them. Therefore, to promote the model to distinguish the optimal candidate moment, similar with [16], we adopt the multi-class cross entropy loss as the ranking loss, which can be formulated as follows:

$$\begin{cases} c^*, m^* = \arg\max_{c,m} \gamma_c^m, \\ \mathcal{L}_{rank} = -\log\left(\dfrac{\exp(r_{c^*}^{m^*})}{\sum_{m=1}^{M} \sum_{c}^{C_m} \exp(r_c^m)}\right), \end{cases} \tag{19}$$

where $r_{c^*}^{m^*}$ is the ranking score of the optimal candidate moment, which has the highest IoU with the ground truth moment. The ranking loss promotes the model to distinguish the optimal candidate moment from other candidates by raising its ranking scores while decreasing that of others.

3) Regression Loss: Considering that the candidate moments generated by sliding windows with fixed lengths may not exactly align with the ground truth moment, we propose to predict the candidate moment offset to adjust the boundary and employ the offset regression presented in [9] to optimize the location of candidate moment. The regression loss is formulated as:

$$\mathcal{L}_{reg} = SL_1(\tau^s - \hat{\tau}_{c^*}^{s,m^*}) + SL_1(\tau^e - \hat{\tau}_{c^*}^{e,m^*}), \tag{20}$$

where $SL_1()$ is the Smooth L_1 function [57]. $(\hat{\tau}_{c^*}^{s,m^*}, \hat{\tau}_{c^*}^{e,m^*})$ is the rectified boundary of the candidate moment that has the highest IoU with the ground truth moment.

Hence we obtain the final objective function of the proposed model as follows:

$$\mathcal{L} = \mathcal{L}_{aln} + \alpha \mathcal{L}_{rank} + \beta \mathcal{L}_{reg}, \tag{21}$$

where α and β are hyper-parameters to balance these items.

IV. EXPERIMENTS

A. Datasets

To evaluate the proposed method, we conducted experiments on two benchmark datasets.

1) Charades-STA [9]: The Charades-STA dataset is built based on the Charades [58] dataset, comprising 6,672 videos of indoor activities. Totally, there are 12,408 moment-sentence pairs in the training set and 3,720 pairs in the testing set. The videos in Charades-STA are 30 seconds on average and the annotated moments are 8 seconds on average.

2) ActivityNet [39]: We evaluate our model on another benchmark ActivityNet Captions built upon the videos from ActivityNet [59], to demonstrate the robustness of the proposed model. The dataset contains 20,000 videos, 37,421 moment-sentence pairs for training and 34,536 pairs for testing. On average, the videos in ActivityNet are 180 seconds and each annotated moment lasts for 36 seconds.

B. Experimental Settings

1) Evaluation Metric: Similar with [9], we adopted the metric of "R@1, IoU@n" to measure our model. Specifically, this metric represents the percentage of the top one candidate moments for all sentence queries with IoU larger than n for all the given sentence queries, where the IoU is calculated with the boundaries of the predicted moment and the ground truth moment.

2) Implementation Details: For the video representation, we split each video in Charades-STA into 75 clips and that in ActivityNet into 300 clips. For Charades-STA, based on the previous studies, we utilized three mainstream frameworks: I3D network [51], C3D network [53] and Two-Stream network [52] to extract the visual features of 1,024-D, 4,096-D, and 8,192-D, respectively. For ActivityNet, we directly utilized the publicly available 500-D C3D features,[3] which are derived by PCA over the original 4,096-D visual features extracted by C3D network.

For the sentence representation, we first tokenized the sentence query and extracted the syntactic dependency graph using Stanford CoreNLP [55] toolkit. Then we employed the Glove [54] model pre-trained on Wikipedia to obtain the 300-D embedding feature for each word token. The max length of the sentence query in Charades-STA and ActivityNet are respectively set to 10 and 50. We truncated the sentences exceeding the maximum length and padded the shorter ones with zeros.

In the training phase, the batch size is set to 128 and Adam optimizer is used for optimization. The learning rate is set to 0.001 and 0.0003 for Charades-STA and ActivityNet, respectively. Furthermore, we added a weight decay item with factor 0.00001 and a dropout probability 0.5 to improve the performance. The node dimension d is set as 256 and 512 for Charades-STA and ActivityNet, respectively, while the dimension of GRU hidden state is half of the node dimension. The threshold θ and λ are set to 0.7 and 0.3. The trade-off α and β are set to 0.1 and 0.001 for Charades-STA, 1 and 0.001 for ActivityNet, respectively. We set 6 window sizes of [6, 12, 18, 24, 30, 36] for Charades-STA and 7 window sizes of [6, 12, 24, 48, 96, 192, 288] for ActivityNet. Window stride δ is set to 3. More details are in our released code.

C. Comparison Among Methods

We compared the proposed MIGCN with other state-of-the-art methods on Charades-STA and ActivityNet. Among these methods, MCN [18], CTRL [9], ACRN [19], TGN [11], SAP [14], QSPN [13], CBP [60], MAN [46], 2D-TAN [62] and DRN [63] are based on the proposal generation and ranking. To improve the computation efficiency, ABLR [25] and ExCL [26] directly predict the location without the dense proposal generation. In addition, RWM [12], TripNet [21], and TSP-PRL [61] regard the task as a sequential decision making process and employ the reinforcement learning paradigm to iteratively adjust the boundary of the moment. According to the original papers, ExCL [26], MAN [46], and DRN [63] adopt the I3D features, TSP-PRL [61] selects both the Two-Stream and C3D features, while DRN [63] and other methods use the C3D features.

Table I shows the performance of different methods, from which we could observe that: 1) The proposed MIGCN exhibits superiority over other methods in most scenarios, demonstrating the effectiveness and robustness of our model. 2) On Charades-STA dataset, MIGCN is inferior to DRN with C3D feature but outperforms DRN with I3D feature. One possible reason is that the I3D network has higher temporal resolution and deeper architecture [51] than C3D and hence is able to learn the representation of activity with more concrete semantics. Compared with the feature pyramid construction in DRN, which may collapse the semantics in the I3D feature, the graph convolution in MIGCN is more likely to completely capture the feature semantics and thus result the better performance. 3) Both MIGCN and TSP-PRL perform better on the Two-Stream features than C3D features. One possible reason is that compared with the C3D features, the Two-Stream with optical flow features are more powerful in capturing temporal information and recognizing the actions in the clips and can benefit the temporal language localization task. Form the observation 2) and 3), we found that MIGCN exhibits different degrees of improvement over other methods with different visual features, but it is difficult to provide intuitive explanations. Therefore, we will dive into the effects of various visual and linguistic representations, *e.g.*, VL-BERT [64] and BERT [65], on vision-language tasks in our future work. And 4) MIGCN, MAN and 2D-TAN that have graph or graph-like components achieve better performance than most purely sequence-based networks, *e.g.*, CTRL, TGN and CBP. This could be an evidence that the graph architecture may have advantage in this task and benefit the performance in general.

D. Ablation Study

We conducted the ablation study to demonstrate the effects of the components in MIGCN.

---

[3]http://activity-net.org/challenges/2016/download.html#c3d/

TABLE I

PERFORMANCE COMPARISON ON CHARADES-STA AND ACTIVITYNET DATASETS IN TERMS OF R@1, IoU@n (%). THE RESULTS OF OTHER METHODS ARE REPORTED ACCORDING TO THEIR PAPERS OR EXISTING REIMPLEMENTS

Feature	Method	Charades-STA [9]		ActivityNet [39]	
		R@1, IoU@0.5	R@1, IoU@0.7	R@1, IoU@0.3	R@1, IoU@0.5
C3D	MCN [18]	17.46	8.01	21.37	9.58
	CTRL [9]	23.63	8.89	28.70	14.00
	ACRN [19]	-	-	31.29	16.17
	TGN [11]	-	-	43.81	27.93
	ABLR [25]	24.36	9.01	55.67	36.79
	SAP [14]	27.42	13.36	-	-
	QSPN [13]	35.60	15.80	45.30	27.70
	RWM [12]	36.70	13.74	53.00	36.90
	CBP [60]	36.80	18.87	54.30	35.76
	TripNet [21]	38.29	16.07	48.42	32.19
	TSP-PRL [61]	37.39	17.69	56.08	38.76
	2D-TAN [62]	-	-	59.45	44.51
	DRN [63]	**45.40**	**26.40**	-	43.97
	MIGCN (Ours)	42.26	22.04	**60.03**	**44.94**
Two-Stream	TSP-PRL [61]	45.30	24.73	-	-
	MIGCN (Ours)	**51.80**	**29.33**	-	-
I3D	ExCL [26]	44.10	23.30	-	-
	MAN [46]	46.63	22.72	-	-
	DRN [63]	53.09	31.75	-	-
	MIGCN (Ours)	**57.10**	**34.54**	-	-

1) Ablation on Multi-Modal Interaction Graph: To exploit the effect of the multi-modal interaction graph convolution, we designed the following derivations of MIGCN:

- *MIGCN-w/o-GCN:* We removed the total process of multi-modal interactive representation refinement and generated the candidate moment representation merely with the output of BiGRU, *i.e.*, $\mathbf{v}_t^h$ and $\mathbf{s}_l^h$.
- *MIGCN-G3G:* To verify the effectiveness of the multi-modal interaction graph, we replaced our graph refinement procedure with that in G^{3}RAPHGROUND [48]. MIGCN-G3G first executes intra-modal refinement to refine the video clip and sentence representations. Then MIGCN-G3G concatenates the sentence representations to each clip representation and conducts the graph convolution using clip-clip edges to fuse the two modalities.
- *MIGCN-w/o-Inter:* We disabled the inter-modal refinement and only used the intra-modal refinement in the graph convolution refinement process to validate the importance of the inter-modal interactions.
- *MIGCN-w/o-Temp:* We removed the temporal correlated edges in the intra-modal refinement to investigate the impact of the temporally adjacent relation.
- *MIGCN-w/o-Seman:* We removed the semantic correlated edges in the intra-modal refinement to validate the influence of the semantic similarity relation.
- *MIGCN-w/o-Syntac:* We removed the syntactic dependency edges in the intra-modal refinement to demonstrate the effectiveness of the syntactic dependency.
- *MIGCN-w/o-Gate:* To explore the effect of the gated graph convolution in inter-modal refinement, we replaced it with the naive graph convolution in Eqn.(11).

Table II presents the ablation study results, from which we could observe that: 1) MIGCN shows superiority over MIGCN-w/o-GCN and MIGCN-G3G, demonstrating the

TABLE II

PERFORMANCE COMPARISON AMONG THE PROPOSED MIGCN AND ITS DERIVATIONS IN TERMS OF R@1, IoU@n (%) BASED ON THE I3D FEATURES OF CHARADES-STA AND C3D FEATURES OF ACTIVITYNET

Method	Charades-STA [9]		ActivityNet [39]	
	R@1, IoU@0.5	R@1, IoU@0.7	R@1, IoU@0.3	R@1, IoU@0.5
MIGCN-w/o-GCN	42.26	24.27	54.21	37.46
MIGCN-G3G	44.30	25.78	57.34	41.17
MIGCN-w/o-Inter	43.15	24.81	59.03	41.96
MIGCN-w/o-Temp	56.61	33.52	59.91	44.76
MIGCN-w/o-Seman	53.68	32.31	57.76	42.38
MIGCN-w/o-Syntac	55.19	33.66	59.71	43.99
MIGCN-w/o-Gate	54.73	31.67	58.44	43.64
MIGCN-MP	36.21	16.32	43.50	27.06
MIGCN-Sample	50.81	29.27	44.57	28.87
MIGCN-w/o-Context	**57.82**	34.30	53.97	37.97
MIGCN	57.10	**34.54**	**60.03**	**44.94**

effectiveness of the multi-modal interaction graph in MIGCN. As for the two different multi-modal graph architectures, MIGCN performs the modality fusion after the intra-modal and inter-modal refinement, while MIGCN-G3G fuses the two modalities before the inter-modal refinement. The early modal fusion in MIGCN-G3G may affect the correlation capture among the video and sentence query in the temporal language localization task and thus results in the inferior performance. 2) MIGCN outperforms MIGCN-w/o-Inter, indicating the effectiveness of the inter-modal refinement in our method. The reason may be that, the inter-modal refinement is the key of the temporal language localization task to capture the semantic correspondence between the video and sentence query, and thus is pivotal to ensure the performance. 3) MIGCN surpasses MIGCN-w/o-Temp, MIGCN-w/o-Seman, and MIGCN-w/o-Syntac, suggesting the necessity of employing the intra-modal refinement in the task of temporal language localization in

TABLE III
EFFICIENCY COMPARISON AMONG DIFFERENT METHODS ON CHARADES-STA AND ACTIVITYNET DATASETS

Feature	Method	Charades-STA [9]				ActivityNet [39]			
		Time ↓	FPS ↑	FLOPs ↓	Params ↓	Time ↓	FPS ↑	FLOPs ↓	Params ↓
C3D	CTRL [9]	2541.06ms	0.28K	0.82G	21.60M	-	-	-	-
	MAC [28]	2141.36ms	0.33K	**0.61G**	23.10M	-	-	-	-
	DRN [63]	162.97ms	4.31K	4.05G	35.66M	-	-	-	-
	2D-TAN [62]	-	-	-	-	1228.91ms	3.02K	498.71G	91.59M
	TSP-PRL [61]	**49.53ms**	**14.17K**	139.81G	182.46M	-	-	-	-
	MIGCN (Ours)	51.45ms	13.64K	3.88G	**10.94M**	**136.50ms**	**27.15K**	**1.18G**	**8.74M**
I3D	**MIGCN (Ours)**	**28.03ms**	**25.04K**	**0.32G**	**2.26M**	-	-	-	-

videos. The possible explanation is that the temporally adjacent relation and semantic correlation among video clips can promote the excavating of the sequential information and semantic clues among video clips, respectively, and hence boost the reasoning of the video. Moreover, the syntactic dependency among words is able to parse the sentence query and thus strengthen the representations learning of objects and actions in the sentence query. In a sense, these intra-modal relations are able to benefit the comprehensive understanding of the video and sentence query, which undoubtedly promote the inter-modal semantic correlation capture between the two modalities. And 4) MIGCN outperforms MIGCN-w/o-Gate, which implies that the modality inherent information retained by the gate mechanism can contribute to the inter-modal refinement and performance improvement.

2) Ablation on Adaptive Context-Aware Localization: To investigate the effect of the adaptive context-aware localization method, we devised the following derivations:

- *MIGCN-MP:* To check the impact of multi-scale fully connected layers in information maintaining, we employed the max-pooling operation to process the candidate moments with different lengths and then used a general fully connected layer to score the candidate moments and predict their offsets.
- *MIGCN-Sample:* Similar with [56], we sampled the center vector of the candidate moment representation as its final representation and then scored the candidate moment and predicted its offsets same as MIGCN-MP.
- *MIGCN-w/o-Context:* To evaluate the effect of the context information, we scored the candidate moments and predicted the offsets without extending the candidate moments.

From Table II, we observe that 1) a large performance decrease in both MIGCN-MP and MIGCN-Sample compared with MIGCN, which suggests that the information loss caused by MIGCN-MP and MIGCN-Sample can seriously deteriorate the performance. 2) MIGCN outperforms MIGCN-w/o-Context in most scenarios, confirming the importance of the context information in the candidate moment scoring and offset prediction. The underlying philosophy is that the context of candidate moment may convey the background or surrounding clues of the query moment, which can promote the localization accuracy. And 3) the performance improvement by context information on ActivityNet is more significant than that on Charades-STA. This phenomenon may be caused by the longer video duration of ActivityNet,

where the video context are more informative for the localization.

E. Efficiency Analysis

To gain the comprehensive understanding of our proposed MIGCN, we also analyzed its efficiency. We calculated the computation time (Time) [25], the frames per second (FPS) [11], the floating point operations (FLOPs) [45], and the number of parameters (Params) [45] of different methods. In particular, Time is defined as the average time to localize one sentence query in the video. FPS is calculated as dividing the number of frames in the video by its computation time. FLOPs is defined as the average floating point operations to localize one sentence query in the video. Params is the total number of learnable parameters in the model. Among these indicators, the model efficiency is mainly reflected by Time and FPS. FLOPs is the reference of the required computation of the method and calculated without considering the concrete implementation of model. Params is the reference of the required space. For a fair comparison, the feature extraction and data loading procedure are excluded in the Time, FPS and FLOPs calculation. It is worth noting that Time and FLOPs are not proportionally correlated. The underlying reason is that different methods are implemented with different parallel degree as they contain various network structures, *e.g.*, linear layers and convolution layers.

All the experiments are conducted on a single NVIDIA GeForce RTX 2080 Ti GPU. For a fair comparison, we only compare our MIGCN with those methods that have code and hyper-parameters publicly released. Except for the comparison, to show the highest efficiency of our method, we also display the results of MIGCN on Charades-STA with I3D feature, which has the minimum feature dimension compared with C3D and Two-Stream features.

From Table III, we observe that: 1) Among all the proposal-based methods (*i.e.*, CTRL, MAC, DRN, 2D-TAN and MIGCN), the Time of MIGCN, 2D-TAN and DRN are much shorter than that of CTRL and MAC. This may due to the fact that the former methods localize one moment with a single run of the model, while the latter ones require multiple times of model running to score all the candidate moments and then localize the target moment, which is time-consuming. 2) Compared with other single run models (*i.e.*, 2D-TAN and DRN), our MIGCN spends less Time and has lower FLOPs and Params. 3) Although MIGCN is based on the dense proposal generation and ranking, the Time and FPS of
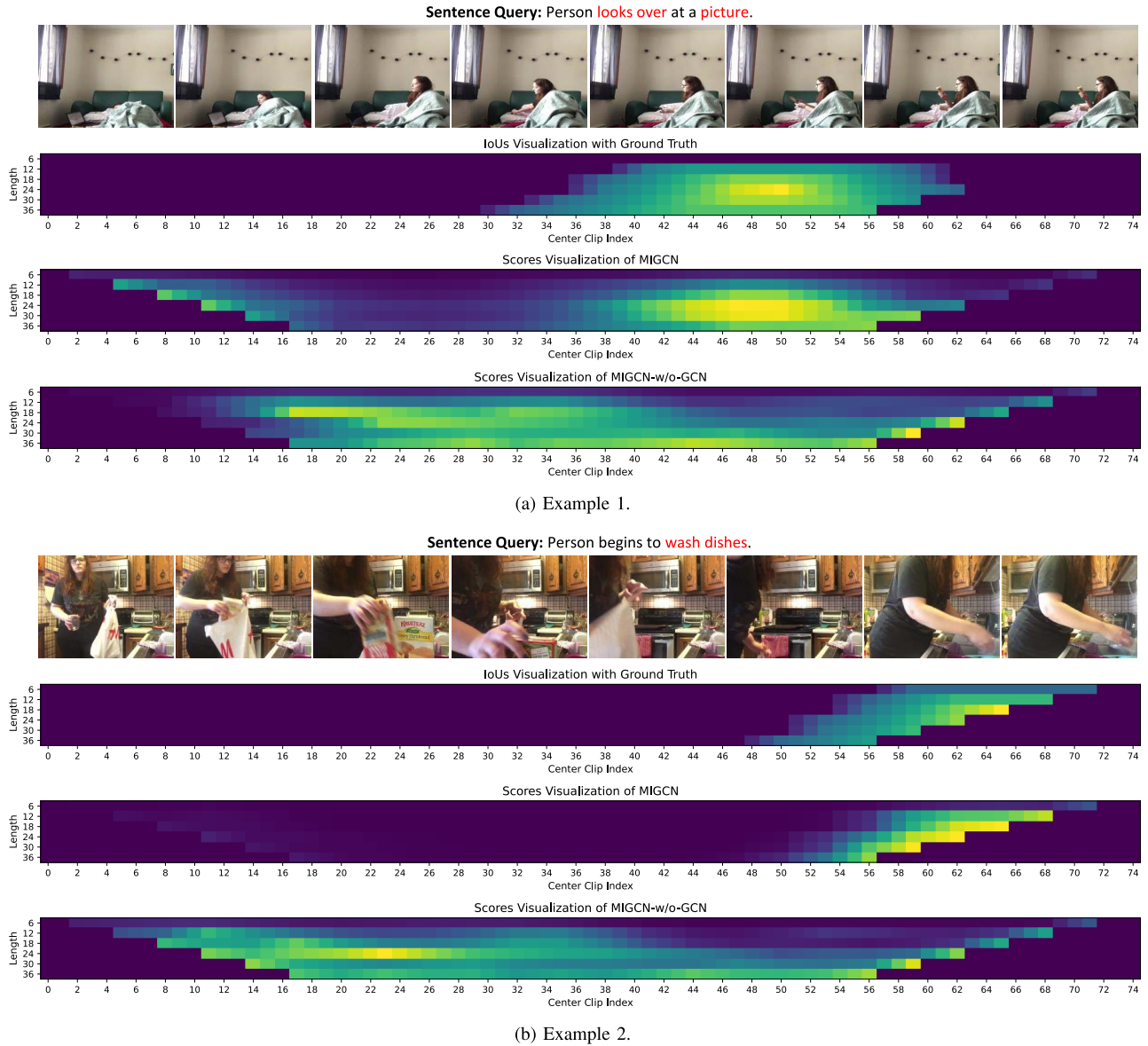
(a) Example 1.

(b) Example 2.

Fig. 3.   Visualization of the candidate moment scores of MIGCN and MIGCN-w/o-GCN with two examples in Charades-STA dataset, where the IoUs of candidate moments with the ground truth are also provided for comparison. The horizontal and vertical coordinates stand for the center clip index and length of candidate moment, respectively. Scores in each row are smoothed by linear interpolation. Lighter colors indicate higher candidate moment scores.

MIGCN on Charades-STA with C3D are comparable with the proposal-free reinforcement learning based TSP-PRL. Therefore, we can draw the conclusion that our MIGCN has not only the promising performance but also the superior efficiency.

F. Result Visualization

In order to gain more deep insights regarding the effects of intra-modal relations as well as inter-modal interactions, we visualized the scores of candidate moments predicted by MIGCN and MIGCN-w/o-GCN with examples in Figure 3. In Figure 3(a), given the sentence query "Person looks over at a picture", MIGCN-w/o-GCN highly scores all the candidates that contain the action of a person looking over at some objects, such as laptop, curtain and bottle. Beyond that, MIGCN gives the high score to the candidate moment only

when the person looks over at a picture, and thus obtains a more similar score distribution with the ground truth. This may be attributed to the fact that MIGCN not only understands the more concrete action "looks over at the picture" due to the intra-modal relation, *i.e.*, syntactic dependency, but also comprehensively considers the inter-modal interactions between the sentence query and video, and giving more rational candidate moment scores. As to the slightly high score predictions of MIGCN at the beginning of the video, one possible explanation is that these candidate moments contain the action that the person looks over at the laptop, which is rather indistinguishable with the picture. Similarly, in Figure 3(b), MIGCN gives high scores to the candidates where the person "wash dishes", which is more accurate than MIGCN-w/o-GCN, since the latter only sees the "dishes" in the video while neglecting the key action "wash".

Sentence Query: Person starts running in place.

(a) Example 1.

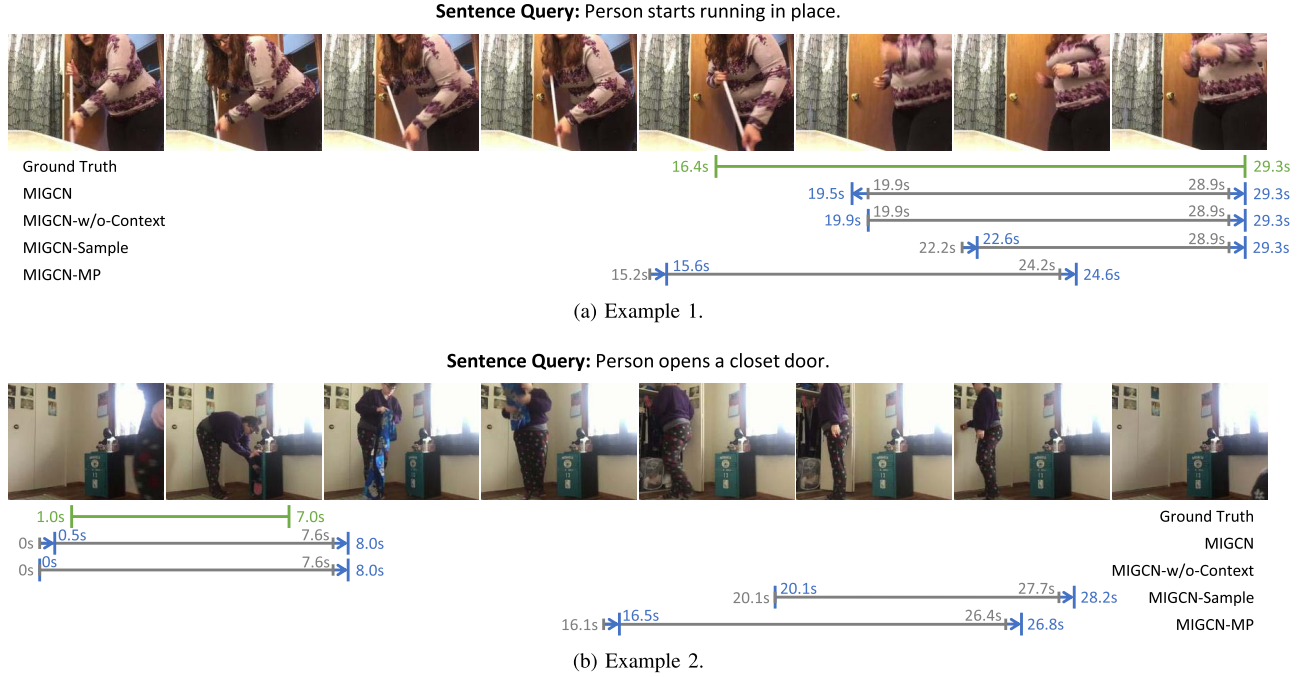**Sentence Query:** Person opens a closet door.

(b) Example 2.

Fig. 4. Prediction examples of MIGCN, MIGCN-w/o-Context, MIGCN-Sample and MIGCN-MP on Charades-STA dataset. The annotations in green color represent the ground truth moments. Those in gray color denote the retrieved coarse candidate moments without boundary adjustment, while those in blue represent the predicted locations after boundary adjustment.

Moreover, we provided some prediction results of MIGCN, MIGCN-w/o-Context, MIGCN-Sample and MIGCN-MP to intuitively show the effect of the adaptive context-aware localization method in Figure 4. As we could see from the two examples, with the multi-scale fully connected layers, MIGCN selects the more ideal coarse candidate moments than MIGCN-Sample and MIGCN-MP, indicating that compared with the sampling and pooling methods, the multi-scale fully connected layers of MIGCN are able to rank the variable-length candidate moments with less information loss and thus accurately locate the target moment. In addition, we observed that although MIGCN and MIGCN-w/o-Context select the same candidate moment, MIGCN obtains a higher IoU result with the more precise boundary adjustment. Similar observation can be found in the second example, which confirms the effect of the context information in the boundary adjustment.

V. CONCLUSION

In this work, we propose a multi-modal interaction graph convolutional network to tackle the task of temporal language localization in videos, which promotes the comprehensive understanding of the video and sentence query and facilitates their semantic correspondence capture with both the intra-modal relation and inter-modal interaction modeling. Moreover, we devise an adaptive context-aware localization method to calculate the ranking scores and boundary offsets of the coarse candidate moments, which considers the context and retains the information of candidate moments as much as possible with the multi-scale fully connected layers. Extensive experiments on Charades-STA and ActivityNet datasets verify the superior effectiveness and efficiency

of our method compared with the state-of-the-art methods. Furthermore, the ablation study confirms the virtues of intra-modal relations, inter-modal interactions and adaptive context-aware localization method on this task. As for the future work, on the one hand, we plan to investigate the effect of visual and linguistic representations on vision-language task. On the other hand, we would enhance the generalization of the model to adapt to various datasets (*i.e.*, TACoS [66]).

REFERENCES

[1] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based LSTM networks for 3D action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3459–3471, Jul. 2018.

[2] R. Zeng, C. Gan, P. Chen, W. Huang, Q. Wu, and M. Tan, "Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5797–5808, Dec. 2019.

[3] R. Zeng *et al.*, "Graph convolutional networks for temporal action localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7093–7102.

[4] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in LSTMs for activity detection and early detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1942–1950.

[5] 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.

[6] J. Yuan, Z. Liu, and Y. Wu, "Discriminative subvolume search for efficient action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2442–2449.

[7] L. Yang, H. Peng, D. Zhang, J. Fu, and J. Han, "Revisiting anchor mechanisms for temporal action localization," *IEEE Trans. Image Process.*, vol. 29, pp. 8535–8548, 2020.

[8] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal localization and categorization of human actions in unsegmented image sequences," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 1126–1140, Apr. 2011.

[9] 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.

[10] 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.

[11] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 162–171.

[12] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2019, pp. 8393–8400.

[13] 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.* Palo Alto, CA, USA: AAAI Press, 2019, pp. 9062–9069.

[14] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2019, pp. 8199–8206.

[15] W. Wang, Y. Huang, and L. Wang, "Language-driven temporal activity localization: A semantic matching reinforcement learning model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 334–343.

[16] J. Chen, L. Ma, X. Chen, Z. Jie, and J. Luo, "Localizing natural language in videos," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2019, pp. 8175–8182.

[17] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury, "Weakly supervised video moment retrieval from text queries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11592–11601.

[18] 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. 1380–1390.

[19] 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.

[20] B. Jiang, X. Huang, C. Yang, and J. Yuan, "Cross-modal video moment retrieval with spatial and language-temporal attention," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 217–225.

[21] M. Hahn, A. Kadav, J. M. Rehg, and H. P. Graf, "Tripping through time: Efficient localization of activities in videos," 2019, *arXiv:1904.09936*. [Online]. Available: http://arxiv.org/abs/1904.09936

[22] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[23] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2016, pp. 289–297.

[24] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov, "Gated-attention architectures for task-oriented language grounding," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2018, pp. 2819–2826.

[25] 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.* Palo Alto, CA, USA: AAAI Press, 2019, pp. 9159–9166.

[26] S. Ghosh, A. Agarwal, Z. Parekh, and A. Hauptmann, "ExCL: Extractive clip localization using natural language descriptions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 1984–1990.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] R. Ge, J. Gao, K. Chen, and R. Nevatia, "MAC: Mining activity concepts for language-based temporal localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 245–253.

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

[30] A. Fout, J. Byrd, B. Shariat, and A. Ben-Hur, "Protein interface prediction using graph convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6530–6539.

[31] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2018, pp. 7444–7452.

[32] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2019, pp. 7370–7377.

[33] X. Qian, Y. Zhuang, Y. Li, S. Xiao, S. Pu, and J. Xiao, "Video relation detection with spatio-temporal graph," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 84–93.

[34] Y. Bi, A. Chadha, A. Abbas, E. Bourtsoulatze, and Y. Andreopoulos, "Graph-based spatio-temporal feature learning for neuromorphic vision sensing," *IEEE Trans. Image Process.*, vol. 29, pp. 9084–9098, 2020.

[35] B. Wang, W. Liu, G. Han, and S. He, "Learning long-term structural dependencies for video salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 9017–9031, 2020.

[36] G. Chen, X. Song, H. Zeng, and S. Jiang, "Scene recognition with prototype-agnostic scene layout," *IEEE Trans. Image Process.*, vol. 29, pp. 5877–5888, 2020.

[37] C. Yan, Q. Zheng, X. Chang, M. Luo, C.-H. Yeh, and A. G. Hauptman, "Semantics-preserving graph propagation for zero-shot object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 8163–8176, 2020.

[38] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555

[39] 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.

[40] X. He, Q. Liu, and Y. Yang, "MV-GNN: Multi-view graph neural network for compression artifacts reduction," *IEEE Trans. Image Process.*, vol. 29, pp. 6829–6840, 2020.

[41] C. Li, Z. Cui, W. Zheng, C. Xu, R. Ji, and J. Yang, "Action-attending graphic neural network," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3657–3670, Jul. 2018.

[42] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-TAD: Sub-graph localization for temporal action detection," 2019, *arXiv:1911.11462*. [Online]. Available: http://arxiv.org/abs/1911.11462

[43] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[44] D. Marcheggiani and I. Titov, "Encoding sentences with graph convolutional networks for semantic role labeling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1506–1515.

[45] J. Zhang, F. Shen, X. Xu, and H. T. Shen, "Temporal reasoning graph for activity recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 5491–5506, 2020.

[46] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "MAN: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1247–1257.

[47] L. Chen, M. Zhai, J. He, and G. Mori, "Object grounding via iterative context reasoning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1407–1415.

[48] M. Bajaj, L. Wang, and L. Sigal, "G3raphGround: Graph-based language grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4281–4290.

[49] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10313–10322.

[50] P. W. Battaglia *et al.*, "Relational inductive biases, deep learning, and graph networks," 2018, *arXiv:1806.01261*. [Online]. Available: http://arxiv.org/abs/1806.01261

[51] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[52] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 20–36.

[53] 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.

[54] 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.

[55] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60.

[56] 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.

[57] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[58] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2016, pp. 510–526.

[59] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[60] J. Wang, L. Ma, and W. Jiang, "Temporally grounding language queries in videos by contextual boundary-aware prediction," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2020, pp. 12168–12175.

[61] J. Wu, G. Li, S. Liu, and L. Lin, "Tree-structured policy based progressive reinforcement learning for temporally language grounding in video," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2020, pp. 12386–12393.

[62] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2D temporal adjacent networks formoment localization with natural language," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2020, pp. 12870–12877.

[63] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, and C. Gan, "Dense regression network for video grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10287–10296.

[64] W. Su *et al.*, "VL-BERT: Pre-training of generic visual-linguistic representations," 2019, *arXiv:1908.08530*. [Online]. Available: http://arxiv.org/abs/1908.08530

[65] 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*, 2019, pp. 4171–4186.

[66] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 25–36, Dec. 2013.

Xuemeng Song (Member, IEEE) received the B.E. degree from the University of Science and Technology of China in 2012 and the Ph.D. degree from the School of Computing, National University of Singapore, in 2016. She is currently an Associate Professor with Shandong University, Jinan, China. She has published several papers in the top venues, such as ACM SIGIR, *MM*, and *TOIS*. Her research interests include the information retrieval and social network analysis. In addition, she has served as a reviewer for many top conferences and journals.

Yan Yan received the Ph.D. degree in computer science from the University of Trento. He was an Assistant Professor at Texas State University and a Research Fellow at the University of Michigan and the University of Trento. He is currently Gladwin Development Chair Assistant Professor with the Department of Computer Science, Illinois Institute of Technology. His research interests include computer vision, machine learning, and multimedia.

Zongmeng Zhang received the B.E. degree from Shandong University in 2021. He is currently pursuing the master's degree with the University of Science and Technology of China. His research interests include computer vision and multimedia computing.

Xianjing Han received the B.E. degree from Northeastern University, China, in 2017. She is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Shandong University, under the supervision of Prof. L. Nie and Prof. X. Song. Her research interests include multimedia computing and computer vision.

Liqiang Nie (Senior Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong University in July 2009 and the Ph.D. degree from the National University of Singapore (NUS) in 2013. After the Ph.D. degree, he continued his research at NUS as a Research Follow for more than three years. He is currently a Professor with the School of Computer Science and Technology, Shandong University. Meanwhile, he is the Adjunct Dean with Shandong AI institute. He has coauthored more than 140 papers, received more than 7,800 Google Scholar citations as of June 2019. Meanwhile, he is supported by the Program of "Thousand Youth Talents Plan 2016," "Qilu Scholar 2016," and "The Shandong Province Science Fund for Distinguished Young Scholars 2018." In 2017, he co-founded "Qilu Intelligent Media Forum." His research interests lie primarily in multimedia computing and information retrieval. He is the Area Chair of ACM MM 2018, the Special Session Chair of PCM 2018, the PC Chair of ICIMCS 2017, and an AE of *Information Science*.