IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 10, 2022, accepted June 3, 2022, date of publication June 14, 2022, date of current version June 22, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3183106

Cascaded MPN: Cascaded Moment Proposal Network for Video Corpus Moment Retrieval

SUNJAE YOON[1], (Member, IEEE), DAHYUN KIM[1], (Member, IEEE), JUNYEONG KIM[2], (Member, IEEE), AND CHANG D. YOO[1], (Senior Member, IEEE)

[1]School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea
[2]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea

Corresponding author: Chang D. Yoo (cd_yoo@kaist.ac.kr)

This work was partly supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-01381, Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments) and partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C201270611).

ABSTRACT Video corpus moment retrieval aims to localize temporal moments corresponding to textual query in a large video corpus. Previous moment retrieval systems are largely grouped into two categories: (1) anchor-based method which presets a set of video segment proposals (via sliding window) and predicts proposal that best matches with the query, and (2) anchor-free method which directly predicts frame-level start-end time of the moment related to the query (via regression). Both methods have their own inherent weaknesses: (1) anchor-based method is vulnerable to heuristic rules of generating video proposals, which causes restrictive moment prediction in variant length; and (2) anchor-free method, as is based on frame-level interplay, suffers from insufficient understanding of contextual semantics from long and sequential video. To overcome the aforementioned challenges, our proposed Cascaded Moment Proposal Network incorporates the following two main properties: (1) Hierarchical Semantic Reasoning which provides video understanding from anchor-free level to anchor-based level via building hierarchical video graph, and (2) Cascaded Moment Proposal Generation which precisely performs moment retrieval via devising cascaded multi-modal feature interaction among anchor-free and anchor-based video semantics. Extensive experiments show state-of-the-art performance on three moment retrieval benchmarks (TVR, ActivityNet, DiDeMo), while qualitative analysis shows improved interpretability. The code will be made publicly available.

INDEX TERMS Video corpus moment retrieval, cascaded moment proposal, multi-modal interaction, vision-language system.

I. INTRODUCTION

Comprehending visual context together with natural language has been a desiderata in the vision-language research societies. Numerous respectful works have made great strides in bridging computer vision and natural language processing including video/image captioning [30], [34], video moment retrieval [1], [8], video/image question answering [12], [25]. Especially, recent success of video streaming services (YouTube) has drowned interest in video search technologies at fine-grained level. Accordingly, Video Corpus Moment Retrieval (VCMR) [13] is a task to localize a moment in large video corpus, which includes two sub-tasks: (1) identifying relevant video in multiple videos and (2) searching for a specific moment in the identified video. To be concrete, the training of VCMR is given a single video-query pair and boundary label, so that system is trained to find the moment related to the query in the video. In the inference, system is given video corpus and query, where it is required to find moment in the video corpus-level. In this respect, VCMR perform a general format of single video moment retrieval.

In methodological aspect of moment retrieval, methods are typically grouped into two categories: (1) anchor-based method and (2) anchor-free method [16], [22], [36], [40]. The anchor-based method [16], [22] follows intuitive solution that first presets a set of video candidate proposals (via sliding

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

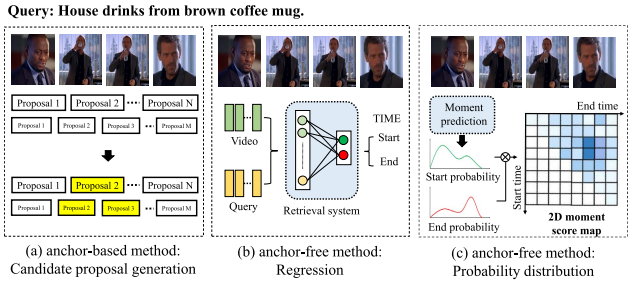


FIGURE 1. Examples of anchor-based and anchor-free moment retrieval system (best viewed in zoom).

window), and then performs classification for the proposals. Figure 1(a) depicts the concept of candidate proposal generation using sliding windows of different lengths, where the best proposal is selected that has the highest similarity to the given query. This anchor-based method is capable of understanding contextual semantics from long and sequential video frames, but it suffers from structural boundary limitation that the candidate proposals should be predefined in some heuristic manner. On the other hand, the anchor-free method [36], [40], has no burdens of generating predefined temporal boundaries, as it directly predicts the start-end time of moment pertinent to query. Figure 1(b) shows an example of anchor-free method [36], where it regresses the start and end time from the joint video-query embedding using multi-layer perceptron (MLP). The regression can be free to the preset boundary problem, but the MLP is easily overfitted and hard to understand heterogeneous vision-language semantics, so that the performance is still far from satisfactory.

As shown in figure 1(c), recent anchor-free method [40] devises two dimensional moment score map, where one dimension indicates start frame of a moment and the other indicates end frame of a moment. To build this score map, systems predict frame-level start time probability and end time probability, after then multiply the two probabilities to calculate joint start-end probability in the format of 2D moment score map. Therefore each element in the map contains frame-level moment score corresponding to its start and end frames. Although this 2D moment score map can leverage the aforementioned preset boundary problem, still they confront the inherent problem of anchor-free method, which is insufficient understanding of contextual semantics across the long and sequential frames. This is because existing anchor-free systems have utilized only frame-level video-query similarities, so that they were not aided from context-level understanding with different lengths in video. Especially, in a scene (drama, movie) based on multi-character interaction, this contextual semantic understanding can be more crucial, combined with auxiliary modalities such as subtitles and sounds.

To overcome aforementioned challenges of existing methods, our proposed Cascaded Moment Proposal Network (Cascaded MPN) incorporates following two main properties: (1) Hierarchical Semantic Reasoning (HSR) which provides video contextual understanding from anchor-free level to anchor-based level via building hierarchical video graph, and (2) Cascaded Moment Proposal Generation (Cascaded MPG) which precisely performs moment retrieval via devising cascaded multi-modal feature interaction among anchor-free and anchor-based video semantics. In overall pipeline, the HSR provides anchor-based level and anchor-free level semantics via building *bipartite* graph among video and subtitles, and this multi-level (anchor-based, anchor-free) semantics generates multi-level moment score maps based on a similarity with given query. Finally, the Cascaded MPG associates the contextual meanings from each level of moment score map in a recursive manner and predicts moment pertinent to the query. Cascaded MPN shows effectiveness on three challenging benchmarks (i.e., TVR, DiDeMo, and ActivityNet) and the code will be made publicly available.

II. RELATED WORK

A. VIDEO MOMENT RETRIEVAL

Video moment retrieval (VMR) is a task of localizing a moment pertinent to given natural language sentence. Gauged from the remarkable advancements of natural language processing [5], [23], VMR system has developed from localizing a temporal activity to a task that understands a natural language query and retrieves the relevant moment [6], [33], [35], [38]. Furthermore, improvements of video representation learning [3], [26] contribute to boosting performance of retrieval system from video retrieval to moment retrieval. The first attempts of retrieval [15], [24] were in temporal activity localization, which aims to predict start-end time of moment corresponding pre-defined actions. Henceforth, a large number of advancements of natural language processing bridges temporal activity localization to a language-based moment retrieval. Gao *et al.* [8] first proposes video moment retrieval, which localizes moments with a sentence describing actions. Hendricks *et al.* [1] proposed VMR with a simplified format for clip-level video understanding. In the meantime, Mithun *et al.* [21] proposed different type of retrieval system that finds a video related to text in multiple videos, which is called *video retrieval*. Recent efforts advance forward to a general form of video moment retrieval. Lei and Li *et al.* [13], [14] propose systems that perform moment retrieval in a video corpus level, which incorporates video retrieval and single video moment retrieval. This video corpus moment retrieval contains an insight that moment retrieval systems should be operated on a general situation given multiple videos. Inspired by another step of generality, we strive for enhanced interpretability in VCMR.

B. VIDEO PROPOSAL GENERATION

Video moment retrieval system estimates moments by predicting start-end time of the moments related to given query. Current literature for predicting moments can be largely grouped into two categories depending on the way

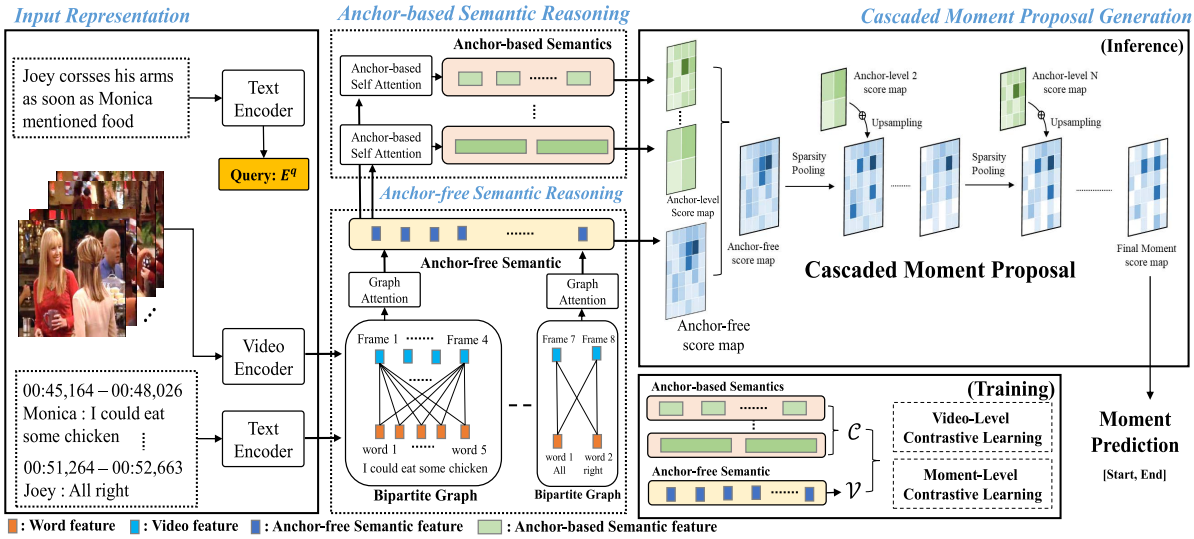


FIGURE 2. Illustration of cascaded MPN which is composed of (1) Input Representation, (2) Anchor-based Semantic Reasoning, (3) Anchor-free Semantic Reasoning and (4) Cascaded Moment Proposal Generation (best viewed in zoom).

of predicting the moments as like anchor-based methods and anchor-free methods. Anchor-based methods facilitate context-level video representation learning, which is suitable for learning sequential semantics in the video. Previous works in anchor-based methods generate several proposals with different sizes by sliding window and retrieve the most pertinent one. Lin *et al.* [16] applied an algorithm considering exploration and exploitation based on reinforcement learning to select top-K proposals. Ma *et al.* [19] proposes surrogate proposal selection which reduces the redundant proposals via selecting one surrogate from defined proposal group. In the case of anchor-free methods, they are versatile to predicting expected moments without temporal boundary constraints. Yuan *et al.* [36] proposed Attention Based Location Regression which regresses start-end points of moment related to query via a series of multi-modal co-attentions. Zhang *et al.* [40] designed two dimensional map with start time and end time as axes, which covers diverse video moments with different lengths. Xiao *et al.* [32] proposed two-stage candidate proposal generating method, which prepares the two-dimensional map as [40] and searches candidate moment proposals from the map. Wang *et al.* [31] also proposed two-stage coarse-to-fine grained multi-modal interaction between video and query. Although many novel thoughts have been proposed as above, there are still room for improvement in terms of converging both anchor-free and anchor-based manners, where we made an effort to perform moment retrieval in a fine-grained level in terms of integrating the beauty of these two methods.

III. METHOD

Cascaded MPN takes video and textual query as inputs and produces a moment score map that includes scores in terms of how similar each moment is to the query. The figure 2 presents overall pipeline of Cascaded MPN, where we define anchor-based and anchor-free semantic features from Hierarchical Semantic Reasoning and explain how they are associated to predict best-matched moment in Cascaded MPG. In training, the system is given a single video-query pair and trained to find the moment in the paired video. In inference, it is required to find moment in a video corpus.

A. HIERARCHICAL SEMANTIC REASONING

1) INPUT REPRESENTATION

VCMR systems are given video, subtitles and single sentence query as $V = \{v_i\}_{i=1}^{N_v}$, $S = \{s_i\}_{i=1}^{N_s}$ and Q, where N_v and N_s is the number of frames and subtitles in a single video. To reason hierarchical semantics in anchor-based and anchor-free level, we first define (1) anchor-free semantic features and construct (2) anchor-based semantic features founded on the anchor-free semantics.

2) ANCHOR-FREE SEMANTIC REASONING

As shown in Anchor-free Semantic Reasoning in figure 2, video frames that share the same subtitle have common contextual meaning from that subtitle. To give this common meaning on video frames, we build a bipartite graph between the shared frames and subtitle. To this, we reorganize the video frames to be aligned with single subtitle s_i as $V^{s_i} = \{v_j\}_{j=1}^{M_v^{s_i}}$, where $M_v^{s_i}$ is the number of frames including a subtitle s_i and also define words in that subtitle s_i as $W^{s_i} = \{w_j\}_{j=1}^{M_w^{s_i}}$, where $M_w^{s_i}$ is the number of words in the subtitle s_i. The query Q is defined as d-dimensional sentence features $\mathbf{E}^q \in \mathbb{R}^d$. The final video and subtitle features are embedded into d-dimensional space as follows:

$$\mathbf{E}_v^{s_i} = \mathrm{LN}(\phi_v(V^{s_i}) + \mathrm{PE}(V^{s_i})) \in \mathbb{R}^{M_v^{s_i} \times d}, \quad (1)$$

$$\mathbf{E}_w^{s_i} = \mathrm{LN}(\phi_w(W^{s_i}) + \mathrm{PE}(W^{s_i})) \in \mathbb{R}^{M_w^{s_i} \times d}, \quad (2)$$

where ϕ_v and ϕ_w is d-dimensional embedder. LN is layer normalization [2] and PE is the positional encoding [27]. The frame embedding $\mathbf{E}_v^{s_i}$ and word embedding $\mathbf{E}_w^{s_i}$ contain common semantic of subtitle s_i and in order to hold this semantic in each video feature, we formally construct video-subtitle graph $\mathcal{G}_{s_i} = (\mathcal{H}^{s_i}, \mathcal{E}^{s_i})$ by regarding $\mathbf{E}_w^{s_i}$ and $\mathbf{E}_v^{s_i}$ as nodes group $\mathcal{H}^{s_i}$ in Equation 3. For the edges $\mathcal{E}^{s_i}$ of video-subtitle graph, we design *bipartite* graph between the words and frames.

$$\mathcal{H}^{s_i} = [\mathbf{E}_v^{s_i} \,||\, \mathbf{E}_w^{s_i}] \in \mathbb{R}^{M^{s_i} \times d}, \tag{3}$$

where $M^{s_i} = M_v^{s_i} + M_w^{s_i}$ is number of nodes in node group $\mathcal{H}^{s_i}$ and $[\cdot||\cdot]$ denotes concatenation along with frame and word axis. To help understanding, in the section of Anchor-free Semantic Reasoning in figure 2, we depict diagram of bipartite graph showing connectivity between nodes in the graph. To associate these frames embedding with the words embedding, we conduct multi-head graph attention [29]. In each head, we use attention coefficient α_{mn}^k to give association between any linked two nodes m and n within node group $\mathcal{H}^{s_i}$, and k in α_{mn}^k means k-th head like below:

$$\alpha_{mn}^k = \frac{\exp(\text{LeakyReLU}(w_k^\top [W^k \mathcal{H}_m^{s_i} || W^k \mathcal{H}_n^{s_i}])}{\sum_{l \in \mathcal{N}_m} \exp(\text{LeakyReLU}(w_k^\top [W^k \mathcal{H}_m^{s_i} || W^k \mathcal{H}_l^{s_i}]))}, \tag{4}$$

$w_k \in \mathbb{R}^{2d}$ is weight vector and $W^k \in \mathbb{R}^{d \times d}$ are shared embedding. $\mathcal{H}_m^{s_i} \in \mathbb{R}^d$ is m-th node feature in $\mathcal{H}^{s_i}$ and $\mathcal{H}_n^{s_i}, \mathcal{H}_l^{s_i}$ follow the same meaning. $\mathcal{N}_m$ is the set of all nodes linked to node m in the bipartite graph. All nodes are updated with this attention coefficients α_{mn}^k and we define *video-subtitle* features $\mathcal{Z}^{s_i}$ by averaging of this updated nodes features. Here, we use *video-subtitle*, because frames and words in one subtitle get the shared semantic by attention. We define final anchor-free level semantic features by adding this $\mathcal{Z}^{s_i}$ to original $\mathcal{H}^{s_i}$:

$$\mathcal{Z}_m^{s_i} = \frac{1}{K} \sum_k^K \sum_{n \in \mathcal{N}_m} \alpha_{mn}^k W^k \mathcal{H}_n^{s_i}, \tag{5}$$

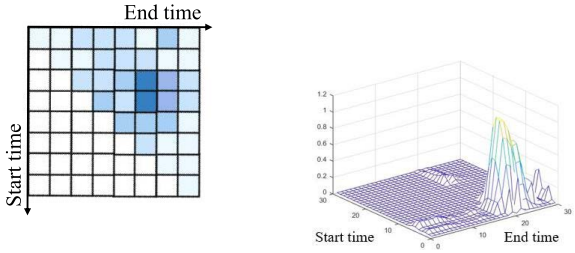
$$\mathcal{V}^{s_i} = (\mathcal{Z}^{s_i} + \mathcal{H}^{s_i})[: M_v^{s_i}] \in \mathbb{R}^{M_v^{s_i} \times d}, \tag{6}$$

where K is the number of attention heads. Here, we only used video features in $(\mathcal{Z}^{s_i} + \mathcal{H}^{s_i})$ as $\mathcal{V}^{s_i}$, supposing that subtitle semantics are involved in frames by *video-subtitle* features $\mathcal{Z}^{s_i}$, where $[: i]$ is slicing operation along node-axis.

3) ANCHOR-BASED SEMANTIC REASONING

In the anchor-based semantic reasoning, we first collect all the anchor-free level features $\mathcal{V} = \{\mathcal{V}^{s_i}\}_{i=1}^{N_s} \in \mathbb{R}^{N_v \times d}$ and uniformly divide $\mathcal{V}$ into N segments. From this segments, we build N anchor-based semantics $\mathcal{C}^N \in \mathbb{R}^{N \times d}$. In one segment, we perform multi-head self-attention to $\mathcal{V}$ using Transformer [28] and treat average of the segment along frame-axis like below:

$$\mathcal{V}[i \times N : (i+1)N] = \text{Head}(\mathcal{V}[i \times N : (i+1)N]), \tag{7}$$



Moment Score Map Moment Score Map (3D view)

FIGURE 3. Illustration of moment score map.

$$\mathcal{C}_i^N = \frac{1}{N} \sum_{j=i \times N}^{(i+1)N} \mathcal{V}_j, \tag{8}$$

where Head denotes multi head self-attention of the Transformer. The following Cascaded MPG operate with this anchor-free semantic features $\mathcal{V} \in \mathbb{R}^{N_v \times d}$ and anchor-based semantic features $\mathcal{C}^N \in \mathbb{R}^{N \times d}$.

B. CASCADED MOMENT PROPOSAL GENERATION

Cascaded Moment Proposal Generation (Cascaded MPG) is introduced to perform moment prediction considering different-level (anchor-free, anchor-base) multi-modal interaction, where it takes inputs as these two semantics, and produces a two-dimensional map containing query-moment similarity score in figure 3(a), where one dimension represents start time of moment and the other dimension represents end time. Casaceded MPG assumes a score map for frame-wise moment retrieval as the same score maps in previous works [13], [14], but also performs contextual reasoning associated with anchor-based semantics. To this, Cascaded MPG includes two main processes: (1) Conditional Moment Score Map generation and (2) Sparsity Pooling, which contribute to multi-modal feature interaction between anchor-based and anchor-free semantics.

1) CONDITIONAL MOMENT SCORE MAP

Conditional Moment Score Map (CMSM) produces two-dimensional moment score map via containing query-moment similarity score in figure 3(a). To build CMSM, we define conditional moment score map generator f_{cond} by multiplying start time probability of moment $P(t_{st}|\mathbf{v}, \mathbf{q}) \in \mathbb{R}^{L \times 1}$ and conditional end time probability of moment $P(t_{ed}|I_{st}; \mathbf{v}, \mathbf{q}) \in \mathbb{R}^{L \times L}$. Given d-dimensional video feature $\mathbf{v} = [v_1, \ldots, v_L] \in \mathbb{R}^{L \times d}$ with the number of frames L and sentence feature $\mathbf{q} \in \mathbb{R}^d$, the start time probability $P(t_{st}|\mathbf{v}, \mathbf{q})$ calculates the probability along the frame axis using video-query similarities as:

$$P(t_{st}|\mathbf{v}, \mathbf{q}) = \text{Softm ax}(\text{Conv}_{st}^{1D}(\mathbf{vq})) \in \mathbb{R}^{L \times 1}, \tag{9}$$

where the Conv_{st}^{1D} and Conv_{ed}^{1D} below denote 1D convolution layer embedding into start-end probabilities and t_{st}, t_{ed} are frame level start-end time. For the conditional end time probability, we first define conditional probability $P(t_{ed}|i_{st}; \mathbf{v}^\star, \mathbf{q}^\star)$,

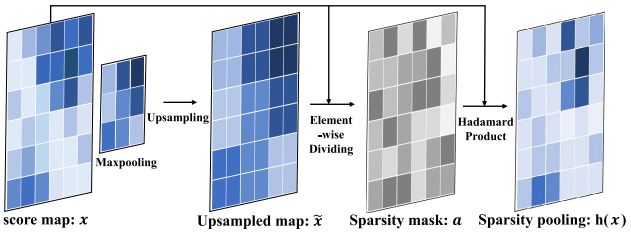


FIGURE 4. Illustration of sparsity pooling process.

that one of the start-frame indices $i_{st} \in \{0, 1, \ldots, L-1\}$ is given as conditional prior. The $\mathbf{v}^\star, \mathbf{q}^\star$ includes start time information from $[\mathbf{v}||(i_{st}/L)] \in \mathbb{R}^{L\times(d+1)}$ and $[\mathbf{q}||(i_{st}/L)] \in \mathbb{R}^{d+1}$, so that they are utilized for generating conditional end time probability as follows:

$$\mathbf{v}^\star = [\mathbf{v}||(i_{st}/L)]_{\text{axis=1}} W_{\mathbf{v}} \in \mathbb{R}^{L\times d} \tag{10}$$

$$\mathbf{q}^\star = [\mathbf{q}||(i_{st}/L)]_{\text{axis=0}} W_{\mathbf{q}} \in \mathbb{R}^{d} \tag{11}$$

$$P(t_{ed}|i_{st}; \mathbf{v}^\star, \mathbf{q}^\star) = \text{Softmax}(\text{Conv}^{1D}_{ed}(\mathbf{v}^\star\mathbf{q}^\star))^\top \in \mathbb{R}^{1\times L}, \tag{12}$$

where $W_{\mathbf{v}}, W_{\mathbf{q}} \in \mathbb{R}^{(d+1)\times d}$ are weight matrix and the operation $[\cdot||\cdot]_{\text{axis=n}}$ is concatenation along the axis n. We stack all the $P(t_{ed}|i_{st}; \mathbf{v}, \mathbf{q})$ along the column axis like Equation 13 and build conditional end time probability $P(t_{ed}|I_{st}; \mathbf{v}, \mathbf{q}) \in \mathbb{R}^{L\times L}$. Finally, $f_{cond}(\mathbf{v}, \mathbf{q}) \in \mathbb{R}^{L\times L}$ builds CMSM by multiplying start time and conditional end time probability, where $\odot$ is column wise and $\cdot$ is element wise multiplication:

$$P(t_{ed}|I_{st}; \mathbf{v}, \mathbf{q}) = \{P(t_{ed}|0; \mathbf{v}^\star, \mathbf{q}^\star); \cdots; P(t_{ed}|(L-1); \mathbf{v}^\star, \mathbf{q}^\star)\} \tag{13}$$

$$f_{cond}(\mathbf{v}, \mathbf{q}) = \mathrm{U}_m \cdot (P(t_{st}|\mathbf{v}, \mathbf{q}) \odot P(t_{ed}|I_{st}; \mathbf{v}, \mathbf{q})) \in \mathbb{R}^{L\times L}, \tag{14}$$

where, we give upper triangular mask $\mathrm{U}_m \in \mathbb{R}^{L\times L}$ composed of 1 to remove the score in moments, where end-time comes before start-time. Therefore, the anchor-free score map $f_{cond}(\boldsymbol{\mathcal{V}}, \mathbf{E}^q)$ and anchor-level score map $f_{cond}(\boldsymbol{\mathcal{C}}^N, \mathbf{E}^q)$ in figure 2 are defined by regarding video feature $\mathbf{v}$ as anchor-free features $\boldsymbol{\mathcal{V}}$ and anchor-based features $\boldsymbol{\mathcal{C}}^N$.

2) SPARSITY POOLING

Sparsity pooling is introduced to mitigate redundantly overlapping moments of frame-level moment score map. The Figure 3(b) shows moment score map in a 3-dimensional view, the high overlapping candidate moments in the map keep similar scores in local region of video, which loses the chance of retrieval in various areas and degrades retrieval performance. To resolve this, our proposed sparsity pooling $\mathbf{h}(x)$ makes the distribution of score map to be sporadic, which allows the retrieval systems to explore diverse moments in the positions and lengths. In detail, the sparsity pooling $\mathbf{h}(x)$ takes input of moment score map $x \in \mathbb{R}^{L\times L}$ and outputs of the same score map but that holds sparsity in the distribution. To this, we first build sparsity mask $\mathbf{a} \in \mathbb{R}^{L\times L}$ in Figure 4, which includes the following processes: (1) calculating 2D max pooling outputs x_N from original score map x with kernel size of $N \times N$ and stride of N, (2) generating 2D upsampled map $\tilde{x}$ by nearest neighbor upsampling up to the original size of x and (3) finally, preparing sparsity mask from element-wise dividing x by $\tilde{x}$. The aforementioned processes can be described as follows:

$$x_N = \text{MaxPool2D}(x) \in \mathbb{R}^{\frac{L}{N}\times\frac{L}{N}} \tag{15}$$

$$\tilde{x} = \text{Upsample}(x_N) \in \mathbb{R}^{L\times L} \tag{16}$$

$$\mathbf{a} = x./\tilde{x} \tag{17}$$

$$\mathbf{h}(x) = \mathbf{a} \cdot x \in \mathbb{R}^{L\times L}, \tag{18}$$

where $./$ and $\cdot$ are element wise dividing and multiplication. In Equation 15 and 16, the $\tilde{x}$ contains a local maximum score of original score map x. In this process, sparsity pooling $\mathbf{h}(x)$ maintains the maximum score in the $N\times N$ window and builds sparse distribution within the windows.

3) CASCADED MOMENT PROPOSAL GENERATION

This section introduces cascaded moment proposal generation (Cascaded MPG) algorithm in detail. Based on the anchor-free semantic $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{N_v\times R}$ and anchor-based semantic $\boldsymbol{\mathcal{C}}^N \in \mathbb{R}^{N\times d}$, Cascaded MPG produces 2D moment score map for moment prediction, where the algorithm relies on the conditional moment score map generator f_{cond} and sparsity pooling $\mathbf{h}(x)$ in a recursive manner. Figure 5 summarizes the pipeline of Cascaded MPG. In the first stage, $f_{cond}(\boldsymbol{\mathcal{V}}, \mathbf{E}^q)$ builds anchor-free moment score map $\mathbb{M}$. The sparsity pooling $\mathbf{h}(x)$ bridges to the next stage by performing sparsity masking on the score map. At the last stage, $f_{cond}(\boldsymbol{\mathcal{C}}^N, \mathbf{E}^q)$ builds a anchor-based moment score map $\mathbb{N}^N$, after then the map is up-sampled to the original anchor-free score map and added to the output of the sparsity pooling. The whole pipeline of Cascaded MPG is described in Algorithm 1.

Algorithm 1 Cascaded MPG

Input: conditional moment score map generator f_{cond}, sparsity pooling $\mathbf{h}$, anchor-based semantics $\boldsymbol{\mathcal{C}}$, anchor-free semantics $\boldsymbol{\mathcal{V}}$, query $\mathbf{E}^q$.
Output: 2D moment score map $\mathbb{M}$ Initialize anchor-free score map $\mathbb{M}_0 = f_{cond}(\boldsymbol{\mathcal{V}}, \mathbf{E}^q)$
for $i \leftarrow 1$ **to** T **do**
  Perform sparsity pooing: $\mathbb{M}_{i-1} \leftarrow \mathbf{h}(\mathbb{M}_{i-1})$
  Update the number of anchors: $N = 2^i$
  Get anchor-level score map: $\mathbb{N}^N = f_{cond}(\boldsymbol{\mathcal{C}}^N, \mathbf{E}^q)$
  Update score map: $\mathbb{M}_i = \sigma(\mathbb{M}_{i-1} + \text{Upsample}(\mathbb{N}^N))$ (σ is sigmoid.)
end
Update output $\mathbb{M} = \mathbb{M}_T$

4) TRAINING

The anchor-free semantics $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{N_v\times d}$ and anchor-based semantics $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{N\times d}$ are trained under two types of loss

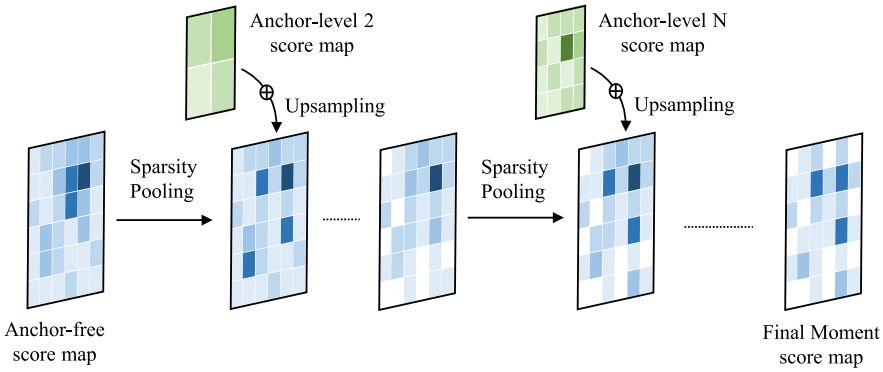


FIGURE 5. Cascaded moment proposal generation.

as follows: (1) video-level loss; and (2) moment-level loss. In video-level loss, we use hinge loss in terms of cosine similarity with query feature $\mathbf{E}^q \in \mathbb{R}^d$ like below:

$$\mathcal{L}_{\text{v}}^{\mathcal{V}} = \max[0, \Delta_c - p(s(\mathcal{V}^+, \mathbf{E}^q)) + p(s(\mathcal{V}^-, \mathbf{E}^q))], \quad (19)$$

$$\mathcal{L}_{\text{v}}^{\mathcal{C}} = \max[0, \Delta_c - p(s(\mathcal{C}^+, \mathbf{E}^q)) + p(s(\mathcal{C}^-, \mathbf{E}^q))], \quad (20)$$

where $+$ is positive from video-query pair and $-$ is negative from other videos in a batch. $p(\cdot)$ is 1D max-pooling and $s(\cdot, \cdot)$ is cosine similarity. $\Delta_c = 0.1$ is a margin and we select surrogate cosine similarity by max-pooling among anchor-based and anchor-free semantics. In moment-level loss, we use cross-entrophy loss CE in terms of ground-truth start-end time (g_{st}, g_{ed}) and predicted start-end time probabilities as:

$$\mathcal{L}_{\text{m}} = \text{CE}(g_{st}, P(t_{st}|\mathcal{V}, \mathbf{E}^q)) + \text{CE}(g_{ed}, P(t_{ed}|t_{st} = g_{st}; \mathcal{V}, \mathbf{E}^q)), \quad (21)$$

$$\mathcal{L}_{\text{v}} = \mathcal{L}_{\text{v}}^{\mathcal{V}} + \mathcal{L}_{\text{v}}^{\mathcal{C}} \quad (22)$$

$$\mathcal{L} = \alpha\mathcal{L}_{\text{v}} + \beta\mathcal{L}_{\text{m}}, \quad (23)$$

Total loss $\mathcal{L}$ is defined with $\mathcal{L}_{\text{v}}$ and $\mathcal{L}_{\text{m}}$ using hyperparameters α and β.

IV. EXPERIMENTS

A. DATASETS

We validate our prospoed Cascaded MPN on three recent benchmarks (TVR, DiDeMo, ActivityNet Captions) as follows: (1) TV show Retrieval (TVR) dataset [13] is constructed under 6 TV shows across 3 genres: medical dramas, sitcoms and crime dramas. TVR contains 109K queries from 21.8K videos with subtitles and each video is about multi character interactions with 60-90 seconds in length. For the fair comparison [13], [14], [37], We also split TVR into 80% train, 10% val, 5% test-private, 5% test-public. The test-public is prepared for official leaderboard. (2) The Distinct Describable Moments (DiDeMo) dataset [1] covers over 10,000 unedited, personal videos in diverse scenarios with pairs of trimmed video segments and referring expressions. DiDeMo is split into about 80% train, 10% val, and 10% test, where each video is pruned up to maximum of 30 seconds. To relieve the complexity, all videos are uniformly divided into 5-second segments, so that labels are of 21 possible moments from a single video. (3) ActivityNet Captions [11] contains 20k videos with 100k temporal descriptions. The average length of the video is about 117 seconds, and the queries are about 14.8 words. 10,009 videos are available for training and 4,917 for validation (val_1). We evaluate models on the val_1 split.

B. EXPERIMENTAL DETAILS

1) EVALUATION METRIC

For the evaluation of VCMR, prediction is correct if: (1) a predicted video matches the ground-truth video; and (2) the predicted moment has high overlap with the ground-truth moment. Average recall at K (R@K) over all queries is used as the evaluation metric, where temporal Intersection over Union (tIoU) is used to measure the overlap between the predicted moment and the ground-truth. We first predict top-100 videos from video corpus by measuring $p(s(\mathcal{V}, \mathbf{E}^q))$ in Equation 19 as and Cascade MPG localizes the best matched moment among the videos.

2) TRAINING DETAILS

We used same video features with [14] using SlowFast [7] pre-trained on Kinetics [10] and ResNet-101 [9] pre-trained on ImageNet [4]. The text features are contextualized token features from pre-trained on RoBERTa [17]. Our model is trained on NVIDIA Quadro RTX 8000 (48GB of memory) GPU. The dimensoin of hidden layer is $d = 768$ and the number of attention heads in hierarchical semantic reasoning is $K = 8$. We use AdamW optimizer [18] with a learning rate of $3e-5$ weight decay of 0.01 to train the model. The training hyperparameters in Equation 23 are $\alpha = 8$ and $\beta = 0.01$.

C. BENCHMARKING RESULTS

Table 1 and Table 2 summarize the experimental results on TVR, DiDeMo and ActivityNet comparing best performed methods, including CAL, XML, HERO, HAMMER, ReLoCLNet. Table 1 presents experimental results from

TABLE 1. Performance comparisons for VCMR on TVR (test-public), ActivityNet and DiDeMo. ⋆: reconstruction-based results, †: without pre-training.

Method	TVR†			ActivityNet Caption†			DiDeMo		
	tIoU=0.7			tIoU=0.7			tIoU=0.7		
	R@1	R@10	R@100	R@1	R@10	R@100	R@1	R@10	R@100
XML [13]	2.76	9.08	15.97	-	-	-	1.59	6.17	25.44
HERO [14]	2.98	10.65	18.25	1.06⋆	6.54⋆	15.34⋆	2.14	11.43	36.09
HAMMER [37]	5.13	11.38	16.71	1.74	8.75	19.08	-	-	-
ReLoCLNet [39]	4.15	14.06	32.42	1.82	6.91	18.33	-	-	-
Cascaded MPN	**5.27**	**16.12**	**35.11**	**2.02**	**9.56**	**23.42**	**3.02**	**13.57**	**41.31**

TABLE 2. Performance comparisons for VCMR with large data pre-training (HowTo100M) on TVR (test-public) and two sub-tasks: Single video moment retrieval (SVMR⋆) and video retrieval (VR⋆) on TVR (validation). ⋆: without pre-training.

Method	VCMR			SVMR⋆		VR⋆	
	tIoU=0.7			tIoU=0.7		-	
	R@1	R@10	R@100	R@1	R@10	R@1	R@10
CAL [6]	-	-	-	4.68	20.17	-	-
XML [13]	3.25	11.38	29.51	13.41	31.11	16.54	50.41
HERO [14]	6.21	14.06	36.66	3.76	9.59	19.44	52.43
ReLoCLNet [39]	-	-	-	15.04	45.24	22.13	57.25
Cascaded MPN	**7.14**	**23.03**	**45.18**	**16.23**	**46.86**	**28.54**	**61.73**

TABLE 3. Ablation study on model variants of cascaded MPN on TVR (validation). (SP: sparsity poolng, AFr: anchor-free semantic reasoning, ABr: anchor-based semantic reasoning, CMSM: Conditional moment score map).

SP	AFr	ABr	CMSM	tIoU=0.7 R@1
✓	✓			4.34
✓		✓		4.96
	✓	✓		5.02
✓	✓	✓		5.21
✓	✓	✓	✓	**5.32**

TABLE 4. Ablation study on cascaded MPN layer.

# of Cascaded layers	tIoU=0.7 R@1
Cascaded layer $n = 1$	4.03
Cascaded layer $n = 2$	4.89
Cascaded layer $n = 3$	**5.32**
Cascaded layer $n = 4$	5.30
Cascaded layer $n = 5$	5.13

scratch on TVR and ActivityNet. The Cascaded MPN consistently outperforms the runner models without pre-training. As the subtitles are unavailable in the ActivityNet and DiDeMo, we utilize video features from video encoder instead of anchor-free semantic features. To the further experiment of DiDeMo, we complement the subtitles with the auxiliary features using Audio Speech Recognition (ASR) from [14], which makes anchor-free semantics available and gives performance gain up to 3.31 in the measure of tIoU=0.7,R@1 on DiDeMo. As reported in [14], the previous results from DiDeMo and TVR are also conducted under pre-training with large-scale dataset HowTo100M [20]. For the fair comparison, we also presents the results from pre-training of HowTo100M on DiDeMo and TVR from Table 1 and 2. Besides, for the two sub-tasks of VCMR : SVMR and VR, Table 2 presents the results on TVR, where the Cascaded MPN also validate the effectiveness.

D. ABLATION STUDY

We perform ablation studies with several variants of Cascaded MPN. Table 3 summarizes ablative results of sparsity pooling (SP), anchor-free semantic reasoning (AFr), anchor-based semantic reasoning (ABr) and conditional moment score map (CMSM). The absence of anchor-based semantic features gives large performance drop, which implies the contextual understanding is crucial in the video with multi character interactions. For the ablation of AFr, we substitute video-subtitle semantics with original video features embedded into d-dimensional space. For the absence of CMSM, we utilize $P(t_{ed}|\mathbf{v},\mathbf{q})$ in stead of $P(t_{ed}|I_{st};\mathbf{v},\mathbf{q})$ and define score map generator as $f_{cond}(\mathbf{v},\mathbf{q}) = \mathrm{U}_m(P(t_{st}|\mathbf{v},\mathbf{q})P(t_{ed}|\mathbf{v},\mathbf{q})^\top)$. CMSM is also worth that it saves about half of training time by early saturation. The Table 4 presents experimental results according to the variants of cascaded layer length. The cascaded layer $n = 3$ shows highest performance with kernel size $N = 2, 4, 8$ in sparsity pooling and more long layers give a slight deterioration in performance. This is because in the early stage of cascaded proposal generation, it is effective to remove many redundant candidates, but after the layers longer than 5, as there are not many redundant candidates, it may damage the proposal scores in way of dropping performance.

E. QUALITATIVE RESULTS

Figure 6 represents moment prediction, conditional stat-end probability, and Figure 7 represents moment score map after sparsity pooling. In the Figure 6, the red curve is the start-probability distribution and the blue curve is the end-probability distribution. From these two distributions, final

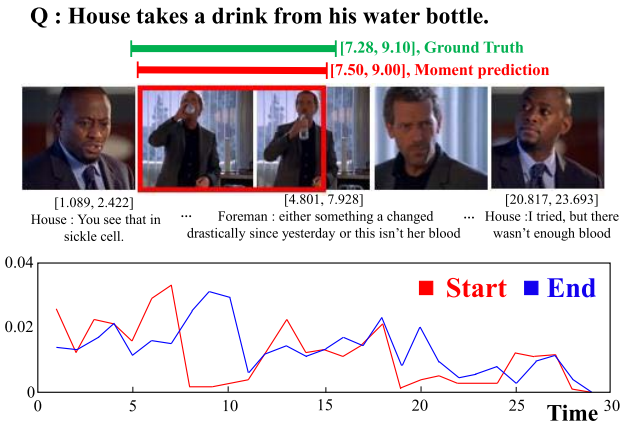


FIGURE 6. **Moment prediction (up), Conditional start-end probability (down).**

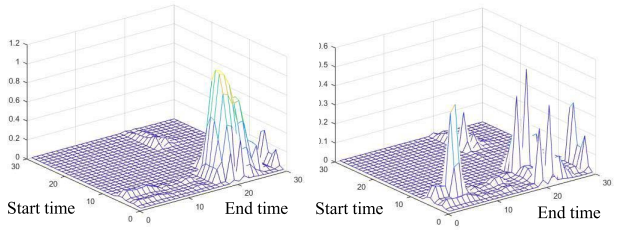


FIGURE 7. **Original score map (left), Sparsity pooling (right).**

moment prediction is performed. Since end-probabilities have the start-probabilities as conditional prior, they have high values right behind the start time. In Figure 7, we can see that the intensive score distribution in a specific moments is alleviated into a sporadic distribution through sparsity masking, which gives the chances of retrieval in various areas and boosts performance in recall. From these, the conditional moment score probability generation and sparsity pooling have a positive effect on the retrieval.

F. LIMITATIONS

We think that Cascaded MPN used many attention weights to represent the two different types of video representations: (1) anchor-free semantic and (2) anchor-based semantic, which took a lot of time to fully learn each features. In this regard, further study would be possible to generate this two hierarchical representations in a more efficient way (weight sharing, model pruning) or another types of representation. We believe that many valuable researches will be built under motivation of overcoming these limitations.

V. CONCLUSION

We propose Cascaded MPN for video corpus moment retrieval to overcome two main challenges: (1) anchor-based method is vulnerable to heuristic rules of generating video proposals, which incurs restrictive moment prediction in length; and (2) anchor-free method systemically suffers from insufficient understanding of long and sequential video semantics. Therefore, our proposed cascaded MPN incorporates following two properties: (1) Hierarchical Semantic Reasoning which gives video understanding from anchor-free level to anchor-based level by building hierarchical video graph, and (2) Cascaded Moment Proposal Generation which precisely performs moment retrieval by devising cascaded multi-modal interaction among anchor-free and anchor-based level video semantics. Experimental results on three benchmarks show effectiveness of our Cascaded MPN.

REFERENCES

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

[2] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[3] 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. 6299–6308.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[6] V. Escorcia, M. Soldan, J. Sivic, B. Ghanem, and B. Russell, "Temporal localization of moments in video collections with natural language," 2019, *arXiv:1907.12763*.

[7] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6202–6211.

[8] 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.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[11] 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.

[12] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "TVQA: Localized, compositional video question answering," 2018, *arXiv:1809.01696*.

[13] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVR: A large-scale dataset for video-subtitle moment retrieval," 2020, *arXiv:2001.09099*.

[14] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "HERO: Hierarchical encoder for video+language omni-representation pre-training," 2020, *arXiv:2005.00200*.

[15] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[16] Z. Lin, Z. Zhao, Z. Zhang, Q. Wang, and H. Liu, "Weakly-supervised video moment retrieval via semantic completion network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11539–11546.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[19] M. Ma, S. Yoon, J. Kim, Y. Lee, S. Kang, and C. D. Yoo, "VLANet: Video-language alignment network for weakly-supervised video moment retrieval," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 156–171.

[20] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2630–2640.
[21] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2018, pp. 19–27.
[22] 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.
[23] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
[24] 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.
[25] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "MovieQA: Understanding stories in movies through question-answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4631–4640.
[26] 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.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–11.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.
[29] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
[30] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence—Video to text," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4534–4542.
[31] H. Wang, Z.-J. Zha, L. Li, D. Liu, and J. Luo, "Structured multi-level interaction network for video moment localization via language query," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7026–7035.
[32] S. Xiao, L. Chen, S. Zhang, W. Ji, J. Shao, L. Ye, and J. Xiao, "Boundary proposal network for two-stage natural language video localization," 2021, *arXiv:2103.08109*.
[33] H. Xu, K. He, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9062–9069.
[34] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4584–4593.
[35] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," 2019, *arXiv:1910.14303*.
[36] 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. (AAAI)*, vol. 33, 2019, pp. 9159–9166.
[37] B. Zhang, H. Hu, J. Lee, M. Zhao, S. Chammas, V. Jain, E. Ie, and F. Sha, "A hierarchical multi-modal encoder for moment localization in video corpus," 2020, *arXiv:2011.09046*.
[38] 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.
[39] H. Zhang, A. Sun, W. Jing, G. Nan, L. Zhen, J. T. Zhou, and R. S. M. Goh, "Video corpus moment retrieval with contrastive learning," 2021, *arXiv:2105.06247*.
[40] 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.*, vol. 34, 2020, pp. 12870–12877.

SUNJAE YOON (Member, IEEE) received the B.S. degree from the Daegu Gyeongbuk Institute of Science and Technology, in 2019, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, in 2021. He is currently pursuing the Ph.D. degree with the Korea Advanced Institute of Science and Technology. His research interests include video search technologies, visual-language reasoning, and causal reasoning.

DAHYUN KIM (Member, IEEE) received the B.S. degree from Inha University, in 2019, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, in 2022. His research interests include vision-language reasoning, video question answering, and video moment retrieval.

JUNYEONG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from KAIST, South Korea, in 2015, 2017, and 2021, respectively. He was a Postdoctoral Researcher Associate at the Artificial Intelligence and Machine Learning Laboratory, School of Electrical Engineering, KAIST. He has been a Faculty Member with the Chung-Ang University (CAU), South Korea, where he is an Assistant Professor with the Department of Artificial Intelligence. His research interests include visual-language reasoning, visual question answering, and various video-based problem.

CHANG D. YOO (Senior Member, IEEE) received the B.S. degree in engineering and applied science from the California Institute of Technology, the M.S. degree in electrical engineering from Cornell University, and the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology. From January 1997 to March 1999, he was Senior Researcher at Korea Telecom (KT). Since 1999, he has been a Faculty Member with the Korea Advance Institute of Science and Technology (KAIST), where he is currently a tenured Full Professor with the School of Electrical Engineering and an Adjunct Professor with the Department of Computer Science. He also served as the Dean of the Office of Special Projects and the Office of International Relations.