BMRN: Boundary Matching and Refinement Network for Temporal Moment Localization with Natural Language

Muah Seol[1,2] Jonghee Kim[1] Jinyoung Moon[1,2*]
[1]Electronics and Telecommunications Research Institute (ETRI), South Korea
[2]University of Science and Technology (UST), South Korea
{seolmuah, jhkim27, jymoon}@etri.re.kr

Abstract

Temporal moment localization (TML) aims to retrieve the best moment in a video that matches a given sentence query. This task is challenging as it requires understanding the relationship between a video and a sentence, as well as the semantic meaning of both. TML methods using 2D temporal maps, which represent proposal features or scores on all moment proposals with the boundary of start and end times on the m and n axes, have shown performance improvements by modeling moment proposals in relation to each other. The methods, however, are limited by the coarsely pre-defined fixed boundaries of target moments, which depend on the length of training videos and the amount of memory available. To overcome this limitation, we propose a boundary matching and refinement network (BMRN) that generates 2D boundary matching and refinement maps along with a proposal feature map to obtain the final proposal score map. Our BMRN adjusts the fixed boundaries of moment proposals with predicted center and length offsets from boundary refinement maps. In addition, we introduce a length-aware proposal feature map that combines a cross-modal feature map and a similarity map between the predicted duration of the target moment and moment proposals. Our approach leads to improved TML performance on Charades-STA and ActivityNet Captions datasets, outperforming state-of-the-art methods by a large margin.

1. Introduction

Temporal moment localization with natural language (TML) methods has become increasingly important in recent years due to the growing demand for efficient and user-friendly methods to access specific moments in videos. TML aims to retrieve the temporal interval of a target moment that best matches a given sentence query within an input video. TML is one of the most challenging tasks, as it requires a comprehensive understanding of both every moment (*i.e*. local understanding) and the relationship between moments (*i.e*. global understanding). To attain the local understanding of a moment, TML methods must be capable of comprehending how every semantic information in a query sentence is visually expressed, as well as how each visual information in a video is linguistically expressed within the sentence through cross-modal interaction. Furthermore, TML methods must be able to grasp both a target moment and other moments if the sentence encompasses both local and global information, as shown in Fig. 1.

*corresponding author

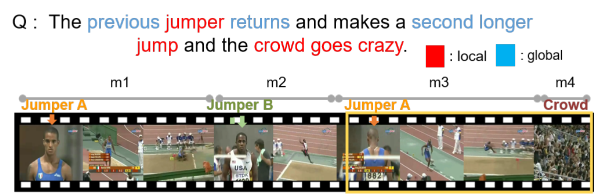
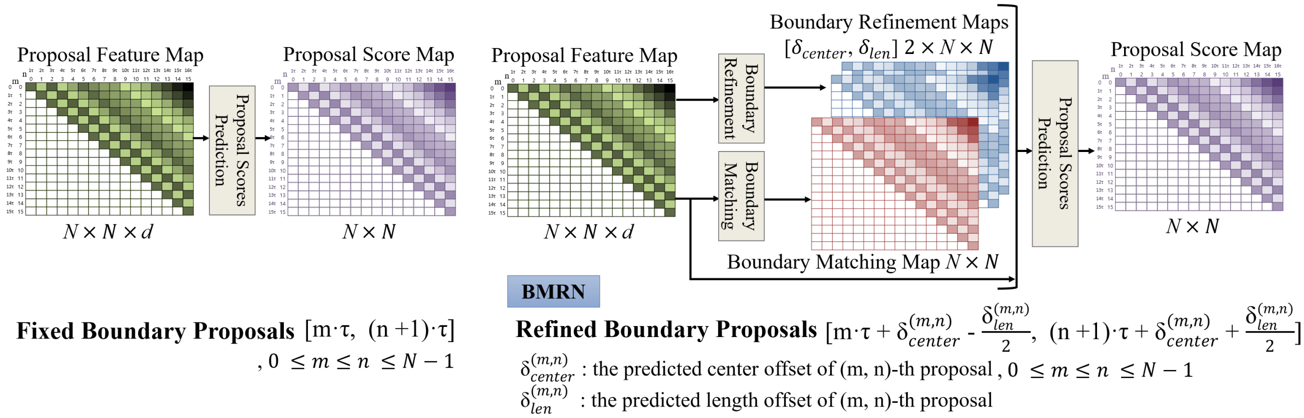


Figure 1. Example of temporal moment localization with an untrimmed input video and a sentence query. TML methods require considering both local and global information in a query distributed across both one target moment (*i.e*. the combined m3 and m4) and the other moments(*i.e*. m1, m2, m3, m4, or the combination of them, respectively.)

Depending on how the boundaries of a target moment are obtained, existing TML methods are largely divided into proposal-free [5, 11, 12, 15, 18–21, 23, 24, 28] and proposal-based methods [2, 4, 6, 13, 22, 25–27]. The proposal-free methods directly regress the start and end times of the target moment by using the cross-modal features for the entire input video. Alternatively, the proposal-free methods obtain the confidence score sequences for start and end times within the video duration and obtain the boundary of the target moment by choosing the start and end times with the highest confidence scores. In contrast, the proposal-based methods generate multiple moment proposals without prior information on the boundary of moment proposals. Using



(a) Existing methods providing fixed boundary proposals

(b) Our BMRN providing variable boundary proposals

Figure 2. Comparison of the existing methods and our proposed BMRN. (a) Existing methods using 2D proposal maps, which are based on 2D-TAN [26], and (b) our BMRN providing variable boundary proposals through boundary refinement

the extracted proposal features through cross-modal interaction between a video and a sentence, the proposal-based methods rank the generated proposals and then choose top-K proposals with the highest proposal scores. The initial proposal-based methods [2, 4, 22, 27] rank the individual proposals independently without taking into account the inter-relationship between them. To overcome this limitation, Zhang et al. [26] introduced 2D temporal proposal maps for proposal features and scores, as shown in Fig. 2a. These maps incorporate two dimensions, one indicating the start time and the other indicating the end time, to better represent temporally adjacent moments together in a map. Using the 2D proposal feature map, the proposal-based methods densely predict the 2D proposal score map by considering the relationship between different proposals. However, the proposal-based methods using the 2D proposal maps [6, 13, 25, 26] originally have a weakness in that the retrieved moments have fixed boundaries with coarsely predefined start and end times. The intervals between the predefined start and end times are determined by the duration of the training videos and the available amount of GPU memory.

To obtain a more precise boundary of a target moment using 2D proposal maps, we propose an end-to-end boundary matching and refinement network (BMRN). Our BMRN adjusts the fixed boundaries of moment proposals with the predicted center and length offsets, as shown in Fig. 2b. To this end, we first create a length-aware proposal feature map by combining an intermediate feature map, which is generated by cross-modal interaction between a video and a sentence, with a similarity map between the proposal length and the duration of the target moment, which is predicted from the intermediate feature sequences. In our BMRN, we consider the length similarity between a proposal and the target moment as an auxiliary proposal confidence score, inspired by [6]. Next, we use the length-aware proposal feature map as input to obtain the boundary matching map and boundary refinement maps for center and length offsets. We predict the final proposal score map using the proposal feature map, boundary matching map, and boundary refinement maps. Our BMRN adjusts the fixed boundaries of all moment proposals with center and length offsets from the boundary refinement maps. Finally, our BMRN generates a selection of top-K moment proposals with variable boundaries based on the best proposal scores from the final proposal score map. Figure 2 provides a comparison of our proposed BMRN with existing 2D proposal map-based methods [6, 13, 25, 26].

Our key contributions can be summarized as follows:

- We propose a novel boundary matching and refinement mechanism that adjusts the temporal boundaries of moment proposals with the highest scores from the proposal score map. This is the first attempt to obtain variable boundaries from the 2D temporal proposal map.
- We introduce a length-aware proposal feature map extraction method that combines cross-modal proposal feature maps with the similarity map between the proposal length and the predicted duration of the target moment before proposal interaction in order to generate a proposal feature map.
- Our BMRN outperforms state-of-the-art TML methods by a large margin on the TML benchmark datasets, including Charades-STA and ActivityNet Captions.

2. Related Work

The field of action recognition plays a vital role in video understanding, as it aims to determine the action class of an action instance in a well-trimmed video. To extract unit-level video features, popular backbone models, such as C3D [16] and I3D [1], have been pre-trained on Sports-1M [16] and Kinetics [1] datasets, respectively. These pre-trained models are widely used in various video understanding tasks, including video classification, video question and answering, temporal action localization and detection, and temporal moment localization.

The objective of temporal action localization (TAL) is to accurately determine the temporal intervals that represent action instances within an untrimmed video. TAL methods need to effectively differentiate between numerous background frames and significant action instances that may belong to multiple action classes. To address this, BSN [10], proposed by Lin et al., generates action proposals using predicted score sequences for the start and end times of action instances. This method was subsequently extended to BMN [9], which introduced a 2D temporal proposal map into the TAL task.

TML, also known as temporal sentence grounding in videos, can be broadly categorized into proposal-free and proposal-based methods based on how the boundaries of a target moment are obtained. Using the cross-modal feature, the proposal-free methods [5, 11, 12, 15, 18–21, 23, 24, 28] obtain a single proposal of the target moment by directly regressing the start and end times or by selecting the start and end times with the highest scores from the predicted start and end score sequences. In contrast, the proposal-based methods [2,4,6,22,25–27] generate multiple moment proposals without any prior cues on the proposal boundar, rank the proposals, and then obtain the best proposals from the predicted proposal scores. 2D temporal maps were introduced to the TML task by Zhang et al. [26] for dense prediction of proposal features and scores between temporally adjacent proposals, and subsequently, several methods [6, 13, 25] have been proposed that utilize 2D temporal maps. Although existing methods achieved performance gains through dense prediction based on 2D proposal maps, they have inherent shortcomings in that retrieved moments have fixed boundaries with coarsely pre-defined start and end times.

3. Proposed Method

The proposed BMRN network largely consists of uni-modal and multi-modal feature encoding, proposal feature map extraction, boundary matching and refinement map extraction, and proposal score map prediction, as shown in Figure 3.

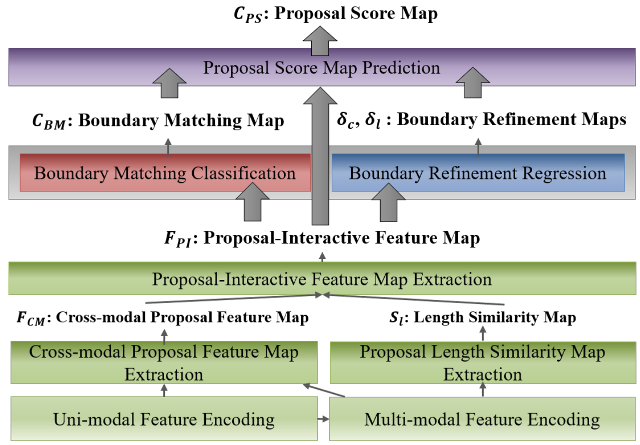


Figure 3. Overall architecture of the proposed BMRN

3.1. Problem Formulation

Given an untrimmed video V and a sentence query S, the goal of temporal moment localization is to localize the temporal boundaries (τ_{start}, τ_{end}) of the target moment, which is described by the sentence query within the video.

3.2. Uni-modal and Multi-modal Feature Encoding

1) Uni-modal Feature Encoding: To encode video features, we first divide a long untrimmed video into T_v non-overlapping segments of fixed length, and extract video unit features using pre-trained CNN models such as C3D [16] and I3D [1]. We then feed the video unit features into a fully connected layer for dimensionality reduction, resulting in video features $F_v \in \mathbb{R}^{T_v \times d}$.

For sentence encoding, we use the pre-trained BERT [3] model. First, sentences are tokenized using the BERT tokenizer, which adds the special tokens [CLS] at the beginning and [SEP] at the end. Each token is then mapped to learned embeddings and summed with learned positional encodings. The input vectors are passed through transformer encoder blocks that include multi-head self-attentions. Finally, we feed the last hidden BERT features into a fully connected layer to obtain the sentence features $F_s \in \mathbb{R}^{T_s \times d}$, where T_s is the length of input tokens.

2) Multi-modal Feature Encoding: To capture long-range dependencies among video features and interaction between video and sentence unit features, we use a multi-head self-attention module(MHSA) of Transformer [17]. For this, the video features F_v added 1D sinusoidal position embeddings (PE) and sentence features F_s are concatenated and then fed into a MHSA module. We then get the transformed video features $F_{tv} \in \mathbb{R}^{T_v \times d}$ and the transformed sentence features $F_{ts} \in \mathbb{R}^{T_s \times d}$ from the encoder output. This can be expressed as:

$$[F_{tv}, F_{ts}] = \mathtt{MHSA}([F_v + PE, F_s]), \tag{1}$$

where [ ] denotes concatenation.

In addition, we use a multi-head cross-attention module [17] by feeding the video segment features F_v as queries and the sentence features F_s as keys and values. This enables us to extract guided sentence features $F_{gs} \in \mathbb{R}^{T_v \times d}$. This can be expressed as:

$$F_{gs} = \texttt{MHCA}(F_v,\ F_s,\ F_s), \tag{2}$$

where $\texttt{MHCA}(Q, K, V)$ is a multi-head cross-attention module having query Q, key K, and value V.

3.3. Cross-modal Proposal Feature Map Extraction

Using the encoded video and sentence features, we extract the cross-modal proposal feature map. To this end, we first partition transformed video features $F_{tv} \in \mathbb{R}^{T_v \times d}$ into N non-overlapping clips, each of which consists of T_v/N transformed video features. To extract statistical information from each clip, we apply to mean pooling and standard deviation pooling to some of the transformed video features F_{tv} that fall within each clip. And we feed the concatenated pooled mean and standard deviation of features into a linear layer, to get the transformed video clip features $F_{tvc} \in \mathbb{R}^{N \times d}$.

And then, to generate each (m, n)-th proposal feature in the $N \times N$ proposal feature map, we sample a segment of the transformed video clip features F_{tvc} from m-th clip to n-th clip, where $0 \leq m \leq n \leq N-1$. Since the lengths of the sampled proposal features may differ, we propose a scale-aware feature extraction approach that considers the common properties shared by features within the same scale of proposal length.

For each scale of the proposal lengths in [1, N], we first select all proposal features having the same length of s, $F^s_{tvc} \in \mathbb{R}^{N_s \times s \times d}$, where N_s is the number of proposals for each scale s, which are diagonally distributed in the 2D proposal feature map. We apply multi-head cross-attention with a learnable query for each scale s, q^s, and the proposal features at scale s, F^s_{tvc}, as key and value, as expressed in Eq. (3)

$$F^s_{vp'} = \texttt{MHCA}(q^s,\ F^s_{tvc},\ F^s_{tvc}). \tag{3}$$

In addition, to employ common statistical information in the same scale, we perform mean pooling and standard deviation pooling on the proposal features at scale s, F^s_{tvc}, followed by concatenation, which is expressed in Eq. (4)

$$F^s_{rvp'} = [\texttt{mean}(F^s_{tvc}), \texttt{std}(F^s_{tvc})], \tag{4}$$

where $\texttt{mean}$ and $\texttt{std}$ represent mean pooling and standard deviation pooling, respectively, and [ ] denotes concatenation. Then, the concatenated features are employed to key and value in multi-head cross-attention with the learnable query for each scale s as follows:

$$F^s_{rvp} = \texttt{MHCA}(q^s,\ F^s_{rvp'},\ F^s_{rvp'}). \tag{5}$$

Finally, the video proposal features at scale s, $F^s_{vp} \in \mathbb{R}^{N_s \times d}$, are obtained by Eq. (6) and Eq. (7).

$$F^s_{vp''} = F^s_{vp'} + F^s_{rvp}, \tag{6}$$

$$F^s_{vp} = F^s_{vp''} + \texttt{FFN}(F^s_{vp''}), \tag{7}$$

where $\texttt{FFN}$ represents a sequence of a fully connected layer, an activation function and a normalization layer. Note that we obtain the 2D video proposal feature map $F_{vp} \in \mathbb{R}^{N \times N \times d}$ by combining F^s_{vp} for all scales s.

In the same way as above, we obtain the 2D sentence proposal feature map $F_{sp} \in \mathbb{R}^{N \times N \times d}$ by taking the guided sentence features F_{gs} as input instead of F_{tv}. And then we stack the feature maps F_{vp} and F_{sp} to obtain the proposal feature map $F_p \in \mathbb{R}^{2 \times N \times N \times d}$.

Furthermore, we modulate the proposal feature map to incorporate between the proposal feature map and the sentence features F_s. First, we feed the proposal feature map F_p as queries and the sentence feature $F_s \in \mathbb{R}^{T_s \times d}$ as keys and values into a multi-head cross-attention layer, as follows:

$$F_{mod'} = \texttt{MHCA}(F_p,\ F_s,\ F_s). \tag{8}$$

$$F_{mod''} = F_{mod'} + \texttt{mean}(F_s). \tag{9}$$

$$F_{mod} = F_{mod''} + \texttt{FFN}(F_{mod''}), \tag{10}$$

where $F_{mod} \in \mathbb{R}^{2 \times N \times N \times d}$.

Finally, we obtain the cross-modal proposal feature map $F_{CM} \in \mathbb{R}^{2 \times N \times N \times d}$ by using Hadamard product between the proposal feature map F_p and the modulating feature map F_{mod}, as expressed in Eq. (11).

$$F_{CM} = F_p \otimes F_{mod}, \tag{11}$$

where $\otimes$ denotes the Hadamard product.

3.4. Proposal Length Similarity Map Extraction

Inspired by [6], we predict the duration of a target moment t_s based on the transformed sentence features F_{ts}, in order to give a prior information on the duration by using the given sentence features. To this end, we feed the pooled transformed sentence features into a fully connected layer and sigmoid function, as follows:

$$t_s = \sigma(\texttt{FC}([\texttt{mean}(F_{ts}), \texttt{std}(F_{ts})])), \tag{12}$$

where $\texttt{FC}$ is a fully connected layer and σ is sigmoid function.

In addition, we predict the duration of a target moment t_v based on the transformed video features F_{tv}, in order to verify t_s. To this end, we define the moment score $M_{tv} \in \mathbb{R}^{T_v \times 1}$ which indicates 1 if each frame falls within the time interval of a target moment, otherwise 0, as expressed in Eq. (13)

$$M_{tv} = \sigma(\texttt{FC}(F_{tv})). \tag{13}$$

The frame-wise moment scores are averaged to predict video-based time duration t_v, as expressed in Eq. (14)

$$t_v = \mathtt{mean}(M_{tv}). \tag{14}$$

We assume that the duration of target moment t_s is reliable if it is similar to the video-based time duration t_v. Therefore, we obtain the confidence score of the predicted duration, cs_t, as expressed in Eq. (15)

$$cs_t = 1 - |t_s - t_v|. \tag{15}$$

Finally, we generate a proposal length similarity map $S_l \in \mathbb{R}^{N\times N}$, which is the similarity between the proposal lengths and the predicted duration of a target moment based on text t_t multiplied by the confidence score of the duration cs_t for all the proposals, as in Eq. (16)

$$S_l(m,n) = k_d^{1-|t(m,n)-t_s|} \cdot cs_t, \tag{16}$$

where $t(m,n)$ is a length of (m,n)-th proposal and k_d is a hyper-parameter greater than 1.

3.5. Proposal Interactive Feature Map Extraction

To effectively interact between moment proposals, we design a two-stream architecture with a CNN layer and a Transformer layer [17]. Specifically, we utilize the CNN layers and Transformer layers to capture local and global relationships among proposal features in the proposal feature map, respectively. First, we concatenate the cross-modal proposal feature map $F_{CM} \in \mathbb{R}^{2\times N\times N\times d}$ and proposal length similarity map $S_l \in \mathbb{R}^{\times N\times N}$ as input (i.e., $F'_{PI} \in \mathbb{R}^{2\times N\times N\times (d+1)}$). Note that we expand S_l to match the dimension between S_l and F_{CM}. We then reduce the dimension from $\mathbb{R}^{2\times N\times N\times (d+1)}$ to $\mathbb{R}^{N\times N\times (d/2)}$ through two different 3D convolution layers. Then, we feed the output into CNN and transformer layers, respectively, and then combine the output from each layer to obtain the final proposal feature map $F_{PI} \in \mathbb{R}^{N\times N\times d}$.

3.6. Boundary Matching and Refinement

The boundaries of proposals are determined by the unit length for proposals $\tau = 1/N$. A pair of boundaries for (m,n)-th proposal is represented by $[m\cdot\tau,\ (n+1)\cdot\tau]$, where $0 \le m \le n \le N-1$. As a result, the predicted boundary based on the proposal is roughly matched to the target moment if we employ the pre-defined boundaries. To overcome the constraint, we get the boundary matching score map and two boundary refinement maps for each proposal as follows. The boundary matching score map $C_{BM} \in \mathbb{R}^{N\times N}$ provides the boundary score of each proposal, as expressed in Eq. (17).

$$C_{BM} = \sigma(\mathtt{FFN}(F_{PI})), \tag{17}$$

where σ is the sigmoid function. The boundary refinement maps δ_c and δ_l generate each center and length offsets on each proposal, respectively, as expressed in Eq. (18) and Eq. (19)

$$\delta_c = \tanh\left(\mathtt{FFN}(F_{PI})\right), \tag{18}$$

$$\delta_l = \tanh\left(\mathtt{FFN}(F_{PI})\right). \tag{19}$$

3.7. Proposal Score Prediction

Finally, we obtain the final confidence scores of all proposals, $C_{PS} \in \mathbb{R}^{N\times N}$ by feeding the proposal features F_{PI}, boundary matching scores C_{BM}, and boundary refinement offsets for center δ_c and length offsets δ_l, through FC layers and sigmoid, as expressed in Eq. (20)

$$C_{PS} = \sigma(\mathtt{FFN}([F_{PI}, C_{BM}, \delta_c \otimes C_{BM}, \delta_l \otimes C_{BM}])), \tag{20}$$

where σ is the sigmoid function and $\otimes$ denotes the Hadamard product.

3.8. Training of BMRN

Our BMRN is trained by the following four types of losses.

1) Moment Score Loss: we define the moment score indicating if each frame falls within the time interval of a target moment, its score is 1, otherwise is 0, we get M_{tv} and M_{gs} by feeding the transformed video features F_{tv} and guided sentence features F_{gs} into a fully connected layer and sigmoid function, as input, respectively. The moment score loss L_m is calculated by two binary cross-entropy losses from M_{tv} and M_{gs}, L_m are expressed as follows:

$$L_m = L_{m\_v} + L_{m\_s}, \tag{21}$$

$$\begin{aligned} L_{m\_v} = \sum_{i=1}^{T_v} & y_m(i)\log(M_{tv}(i)) \\ & + (1-y_m(i))\log(1-M_{tv}(i)), \end{aligned} \tag{22}$$

$$\begin{aligned} L_{m\_s} = \sum_{i=1}^{T_v} & y_m(i)\log(M_{gs}(i)) \\ & + (1-y_m(i))\log(1-M_{gs}(i)), \end{aligned} \tag{23}$$

where $y_m(i)$ is the label of the moment score of i-th frame.

2) Moment Duration Loss: For the moment duration loss L_d, we adopt a binary cross-entropy loss between the duration of the target moment y_d and the predicted duration t_s, as expressed in Eq. (24)

$$L_d = y_d\log(t_s) + (1-y_d)\log(1-t_s). \tag{24}$$

3) Proposal Score Loss: During training, we adopt a normalized IoU value as the supervision signal for proposal scores, which are related to boundary matching scores in Section 3.6 and proposal confidence scores in Section 3.7.

Note that there is a slight notation abuse for simplicity, *i.e.*, (c) represents (m,n). For each moment proposal $p(c)$, we compute its IoU with the GT moment $(\tau_{start},\tau_{end})$, $o(c) = \texttt{IoU}(p(c),(\tau_{start},\tau_{end}))$. Then, we divide the $o(c)$ by the maximum of all IoUs for all proposals for normalizing it to a value in $[0,1]$, as expressed in Eq. (25)

$$\tilde{o}(c) = o(c)/o_{max}, \tag{25}$$

where o_{max} is the maximum of IoUs for all proposals. Similar to the IoU score in [26], the IoU score $\tilde{o}(c)$ is then scaled with the threshold $\texttt{IoU}_{\texttt{min}}$, as expressed in Eq. (26)

$$y_s(c) = \begin{cases} \dfrac{\tilde{o}(c) - \texttt{IoU}_{\texttt{min}}}{1.0 - \texttt{IoU}_{\texttt{min}}}, & \text{if } \tilde{o}(c) > \texttt{IoU}_{\texttt{min}} \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

For boundary matching scores C_{BM}, we ramdomly samples a value of $\texttt{IoU}_{\texttt{min}}$ from the range of [0.5, 0.9]. Similar to dropout [14], the sampled $\texttt{IoU}_{\texttt{min}}$ also alleviate the risk of overfitting as changing $y_s(c)$. For final proposal scores C_{PS}, we set the value of $\texttt{IoU}_{\texttt{min}}$ to 0.5, as follows:

$$L_s = L_{bm} + L_{ps}, \tag{27}$$

$$\begin{aligned} L_{bm} = \frac{1}{C}\sum_{c=1}^{C} & y_{s\_rand}(c)\log(C_{BM}(c)) \\ &+(1 - y_{s\_rand}(c))\log(1 - C_{BM}(c)), \end{aligned} \tag{28}$$

$$\begin{aligned} L_{ps} = \frac{1}{C}\sum_{c=0}^{1} & y_{s\_0.5}(c)\log(C_{PS}(c)) \\ &+(1 - y_{s\_0.5}(c))\log(1 - C_{PS}(c)). \end{aligned} \tag{29}$$

where C is the total number of proposals. y_{s_rand} obtained by the uniformly sampled $\texttt{IoU}_{\texttt{min}}$ and $y_{s\_0.5}$ obtained by $\texttt{IoU}_{\texttt{min}} = 0.5$.

4) Proposal Refinement Loss: The refinement loss L_r consists of the center offset loss L_{co}, the length offset loss L_{lo}, and the refined IoU loss L_{rIoU}.

The center offset label $y_{o\_c}(c)$ and the length offset label $y_{o\_l}(c)$ are calculated between boundary of proposal $p(c)$ $(t_{start}(c), t_{end}(c)) \in [0,1]$ from 2D Map Proposals and the target moment $(\tau_{start}, \tau_{end}) \in [0,1]$, as follows:

$$\begin{aligned} y_{o\_c}(c) =& (\tau_{end} + \tau_{start})/2 \\ &- (t_{end}(c) + t_{start}(c))/2, \end{aligned} \tag{30}$$

$$\begin{aligned} y_{o\_l}(c) =& (\tau_{end} - \tau_{start}) \\ &- (t_{end}(c) - t_{start}(c)). \end{aligned} \tag{31}$$

And we calculate the refined boundary of proposal $(\tilde{t}_{start}(c), \tilde{t}_{end}(c))$ as follows:

$$\tilde{t}_{start}(c) = t_{start}(c) + \delta_c(c) - \delta_l(c)/2, \tag{32}$$

$$\tilde{t}_{end}(c) = t_{end}(c) + \delta_c(c) + \delta_l(c)/2. \tag{33}$$

We then, calculate refined IoU($\texttt{rIoU}$) between refined boundary($\tilde{t}_{start}(c)$, $\tilde{t}_{end}(c)$) and target moment (τ_{start}, τ_{end}), as follows:

$$\texttt{rIoU}(c) = \frac{\texttt{min}(\tau_{end}, \tilde{t}_{end}(c)) - \texttt{max}(\tau_{end}, \tilde{t}_{start}(c))}{\texttt{max}(\tau_{end}, \tilde{t}_{end}(c)) - \texttt{min}(\tau_{end}, \tilde{t}_{start}(c))}. \tag{34}$$

In order to refine only proposals with IoU greater than 0.5, we use the values of $y_{s\_0.5}$. The refinement loss L_r is defined as follows:

$$L_{o\_c} = \frac{1}{C}\sum_{c=1}^{C} |y_{o\_c}(c) - \delta_c(c)| \cdot y_{s\_0.5}(c) \tag{35}$$

$$L_{o\_l} = \frac{1}{C}\sum_{c=1}^{C} |y_{lo}(c) - \delta_l(c)| \cdot y_{s\_0.5}(c), \tag{36}$$

$$L_{rIoU} = \frac{1}{C}\sum_{c=1}^{C} -log(\texttt{rIou}(c)) \cdot y_{s\_0.5}(c), \tag{37}$$

$$L_r = L_{o\_c} + L_{o\_l} + L_{rIoU}. \tag{38}$$

The total loss is computed as follows:

$$L = \lambda_1 \cdot L_m + \lambda_2 \cdot L_d + \lambda_3 \cdot L_s + \lambda_4 \cdot L_r, \tag{39}$$

where is λ_i for i=1,2,3, and 4 are balancing parameters for the total loss.

3.9. Inference of BMRN

To obtain the final boundary of a target moment consisting of start and end times, we calculate the refined start and end times by using the fixed boundary of proposals from the proposal score map and the center and length offsets from the two boundary refinement maps in Eq. (32) and Eq. (33). We select the top-K moment proposals with the highest proposal scores, which are not highly intersected between them through NMS.

4. Experiments

4.1. Datasets

We use Charades-STA [4] and ActivityNet Captions [8] as TML benchmark datasets. Charades-STA contains 9,848 videos mainly involving indoor human actions. On average, a video is 30 second long. Charades-STA contains 12,408 and 3,720 moment annotations in the training and testing sets, respectively. ActivityNet Captions contains 19,209 untrimmed videos whose length is two minute long, on average. The whole dataset has 37,417, 17,505, and 17,031 moment annotations for training, validation, and testing, respectively.

4.2. Evaluation Metric

We evaluate our BMRN by using Rank $n@m$ (n is the number of top-K proposals and m is the threshold of IoU with GT moment). Rank $n@m$ is the percentage of queries with at least one correct moment in the top-n predicted moments. A predicted moment proposal is considered the correct proposal if its IoU with the GT moment is larger than m. On both Charades-STA and ActivityNet Captions, we report Rank $n@m$ with $n \in \{1, 5\}$ and $m \in \{0.5, 0.7\}$.

4.3. Implementation Details

We use Adam [7] with learning rate of 1×10^{-4} and batch size of 32 for optimization. We adopt pre-trained C3D [16] and I3D [1] models as a video unit feature extractor and pre-trained BERT model [3] for a sentence unit feature encoding. The number of video clips N, which determines the size of proposal maps, is set to 16 and 64 for Charades-STA and ActivityNet Captions, respectively. The 2D spare map strategy is the same in [26]. The non-maximum suppression (NMS) threshold is set to 0.5 during the inference. And we set k_d of the proposal length similarity map to 4. The balancing parameters for total loss L are set to $\lambda_1 = 0.5, \lambda_2 = \lambda_3 = \lambda_4 = 1$ on Charades-STA, and $\lambda_1 = 0.5, \lambda_2 = 0.5, \lambda_3 = 1, \lambda_4 = 2$ on ActivityNet Captions.

4.4. Performance Comparison

We evaluate our BMRN on ActivityNet Captions and Charades-STA and compare it with the recent state-of-the-arts including both proposal-free methods (MCN [5], ABLR [20], TMLGA [12], LGI [11], DRN [21], CPN [23], MSA [24], LPNet [18], ACRM [15], DTG [28], and HiSA [19]) and proposal-based methods (CTRL [4], SAP [2], MAN [22], CMIN [27], 2D-TAN [26], TACI [13], MS-2D-TAN [25], and STCM-Net [6]).

Table 1 presents the comparison of moment localization performance on Charades-STA, where our BMRN outperforms the state-of-the-art methods in all performance measures for both C3D and I3D features. Notably, our BMRN with I3D features achieves significantly better results than the state-of-the-art methods, with a large margin of 1.99%p, 2.76%p, 3.56%p, and 0.74%p in terms of R1@0.5, R1@0.7, R5@0.5, and R5@0.7, respectively.

Table 2 shows the comparison of moment localization performance on ActivityNet Captions, where our BMRN outperforms the state-of-the-art methods except for R1@IoU=0.7. Specifically, our method achieves R5 scores of 81.37% and 64.44% at IoU=0.5 and IoU=0.7, respectively, with a large margin of 2.57%p and 0.98%p.

4.5. Ablation Study

To demonstrate the effectiveness of the boundary matching and refinement maps and the length similarity map, we

Method	Rank1@		Rank5@	
	0.5	0.7	0.5	0.7
C3D video features				
CTRL [4]	23.63	8.89	58.92	29.52
ABLR [20]	24.36	9.01	-	-
DRN [21]	45.40	26.40	88.01	55.38
LPNet [18]	40.94	21.13	-	-
ACRM [15]	40.78	22.28	-	-
TACI [13]	36.60	18.33	-	-
MS-2D-TAN [25]	41.10	23.25	81.53	48.55
Our BMRN	**45.93**	**28.37**	**89.12**	**57.19**
I3D video features				
TMLGA [12]	33.04	19.26	-	-
LGI [11]	59.46	35.48	-	-
DRN [21]	53.09	31.75	89.06	60.05
CPN [23]	59.77	36.67	-	-
LPNet [18]	54.33	34.03	-	-
ACRM [15]	57.53	38.33	-	-
DTG [28]	60.19	39.38	87.53	66.91
HiSA [19]	61.10	39.70	-	-
TACI [13]	60.27	38.74	-	-
MS-2D-TAN [25]	60.08	37.39	89.06	59.17
Our BMRN	**63.09**	**42.46**	**92.62**	**67.65**

Table 1. Comparisons between our BMRN and the state-of-the-arts on Charades-STA.

Method	Rank1@		Rank5@	
	0.5	0.7	0.5	0.7
C3D video features				
CTRL [4]	29.01	10.34	59.17	37.54
MCN [5]	21.36	6.43	53.23	29.70
ABLR [20]	36.79	-	-	-
CMIN [27]	43.40	23.88	67.95	50.73
2D-TAN [26]	44.51	26.54	77.13	61.96
LGI [11]	41.51	23.07	-	-
DRN [21]	45.45	24.36	77.97	50.30
CPN [23]	45.10	28.10	-	-
MSA [24]	48.02	**31.78**	78.02	63.18
LPNet [18]	45.92	25.39	-	-
HiSA [19]	45.36	27.68	-	-
TACI [13]	45.50	27.23	-	-
MS-2D-TAN [25]	46.16	29.21	78.80	60.85
STCM-Net [6]	46.23	29.04	78.43	63.46
Our BMRN	**48.47**	31.15	**81.37**	**64.44**

Table 2. Comparisons between our BMRN and the state-of-the-arts on ActivityNet Captions.

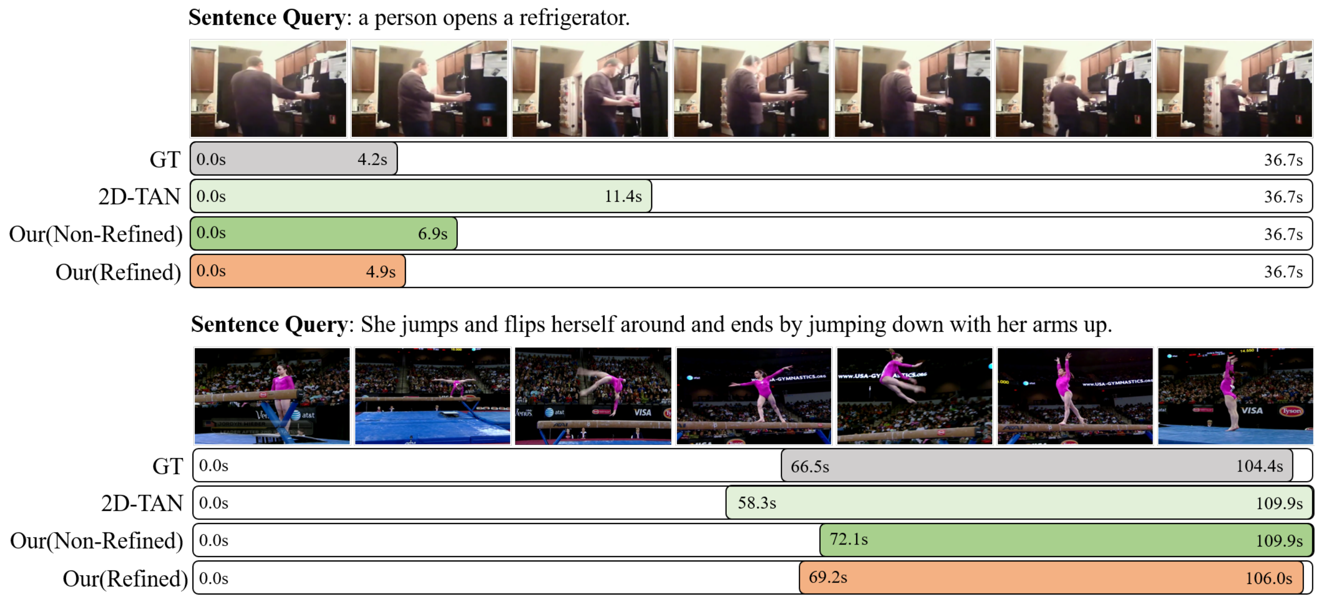


Figure 4. Qualitative evaluation of BMRN on Charade-STA [4] (top) and ActivityNet-Captions [8] (bottom). Each result shows ground-truth, 2D-TAN, BMRN without refinement, and full BMRN

Method	Rank1@		Rank5@	
	0.5	0.7	0.5	0.7
Full BMRN	63.09	42.46	92.62	67.65
w/o *BM and BR maps*	60.83 (-2.26)	40.54 (-1.92)	89.95 (-2.67)	67.89 (0.24)
w/o *Len Sim map*	62.23 (-0.86)	42.19 (-0.27)	90.54 (-2.08)	65.54 (-2.11)

Table 3. Ablation study of the effectiveness of the boundary matching and refinement maps and the length similarity map.

conducted ablation experiments on Charades-STA by comparing the performance of the full BMRN with two variants: one without the boundary matching and refinement maps and the other one without the length similarity map, as shown in Table 3.

4.6. Qualitative Evaluation

In Figure 4, we present qualitative results for two queries from Charades-STA and ActivityNet Captions datasets, comparing results obtained by ground-truth, 2D-TAN, BMRN without refinement, and the full BMRN. The results clearly demonstrate the significant performance improvement achieved by our proposed BMRN.

5. Conclusion

In this paper, we propose an end-to-end boundary matching and refinement network that adjusts the fixed boundaries of proposals from 2D proposal maps using the predicted center and length offsets from the boundary refinement maps. Our BMRN offers a selection of the top-K moment proposals with variable boundaries. Additionally, we introduce a length-aware proposal feature map by combining the proposal feature map with the similarity map between the proposal length and the duration of the target moment. The experimental results show performance improvements over the current state-of-the-arts on Charades-STA and ActivityNet Captions. As an ablation study, we demonstrate the effectiveness of the boundary matching and refinement maps and the length similarity map. As our future work, we plan to investigate the development of a TML method that can achieves highly accurate localization performance while using a significantly reduced number of proposals, which will be adjusted by a novel boundary refinement mechanism.

6. Acknowledgement

This work was supported by IITP grant funded by the Korea government(MSIT) (No.2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network).

References

[1] Jo˜ao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *ICCV*, pages 6299–6308, 2017. 3, 7

[2] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, pages 8199–98206, 2019. 1, 2, 3, 7

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, page 4171–4186, 2019. 3, 7

[4] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 1, 2, 3, 6, 7, 8

[5] Lisa A. Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. 1, 3, 7

[6] Zixi Jia, Minglin Dong, Jingyu Ru, Lele Xue, Sikai Yang, and Chunbo Li. Stcm-net: A symmetrical one-stage network for temporal language localization in videos. *Neurocomputing*, 471:194–207, 2022. 1, 2, 3, 4, 7

[7] Diederik P. Kingma and Jimmy Ba. A method for stochastic optimization. In *Proc. ICRL*, 2015. 7

[8] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 6, 8

[9] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: boundary-matching network for temporal action proposal generation. In *ECCV*, pages 1–17, 2018. 3

[10] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 1–17, 2018. 3

[11] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, pages 10810–10819, 2021. 1, 3, 7

[12] Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Fatemeh S. Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *IEEE WACV*, pages 2463–2473, 2020. 1, 3, 7

[13] Jungkyoo Shin and Jinyoung Moon. Learning to combine the modalities of language and video for temporal moment localization. *Computer Vision and Image Understanding*, 217:103375:1 – 103365:13, 2022. 1, 2, 3, 7

[14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958. 6

[15] Haoyu Tang, Jihua Zhu, Meng Liu, Zan Gao, and Zhiyong Cheng. Frame-wise cross-modal matching for video moment retrieval. *IEEE TMM*, 24:1338 – 1349, 2021. 1, 3, 7

[16] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 3, 7

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 4, 5

[18] Shaoning Xiaoy, Long Chenz, Jian Shaoy, Yueting Zhuangy, and Jun Xiao. Natural language video localization with learnable moment proposals. In *ENMLP*, pages 4008–4017, 2021. 1, 3, 7

[19] Zhe Xu, Da Chen, Kun Wei, Cheng Deng, and Hui Xue. Hisa: hierarchically semantic associating for video temporal grounding. *IEEE TIP*, 31:5178 – 5188, 2022. 1, 3, 7

[20] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: temporal sentence localization in video with attention based location regression. In *AAAI*, pages 9159–9166, 2019. 1, 3, 7

[21] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, pages 10287–10296, 2021. 1, 3, 7

[22] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. Man: moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, pages 1247–1257, 2019. 1, 2, 3, 7

[23] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng T. Shen. Cascaded prediction network via segment tree for temporal video grounding. In *CVPR*, pages 12669–12678, 2021. 1, 3, 7

[24] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng T. Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *CVPR*, pages 12669–12678, 2021. 1, 3, 7

[25] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *IEEE TPAMI*, 44(12):1247–1257, 2022. 1, 2, 3, 7

[26] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, pages 12870–12877, 2020. 1, 2, 3, 6, 7

[27] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*, pages 1247–1257, 2019. 1, 2, 3, 7

[28] Hao Zhou, Chongyang Zhang, Yan Luo, Chuanping Hu, and Wenjun Zhang. Thinking inside uncertainty: interest moment perception for diverse temporal grounding. *IEEE TCSVT*, 32(10):7190 – 7203, 2022. 1, 3, 7