

SCANet: Scene Complexity Aware Network for Weakly-Supervised Video Moment Retrieval

Sunjae Yoon Gwanhyeong Koo Dahyun Kim Chang D. Yoo

Korea Advanced Institute of Science and Technology (KAIST)

{sunjae.yoon, kookie, dahyun.kim, cd_yoo}@kaist.ac.kr

Abstract

Video moment retrieval aims to localize moments in video corresponding to a given language query. To avoid the expensive cost of annotating the temporal moments, weakly-supervised VMR (wsVMR) systems have been studied. For such systems, generating a number of proposals as moment candidates and then selecting the most appropriate proposal has been a popular approach. These proposals are assumed to contain many distinguishable scenes in a video as candidates. However, existing proposals of wsVMR systems do not respect the varying numbers of scenes in each video, where the proposals are heuristically determined irrespective of the video. We argue that the retrieval system should be able to counter the complexities caused by varying numbers of scenes in each video. To this end, we present a novel concept of a retrieval system referred to as Scene Complexity Aware Network (SCANet), which measures the ‘scene complexity’ of multiple scenes in each video and generates adaptive proposals responding to variable complexities of scenes in each video. Experimental results on three retrieval benchmarks (i.e. Charades-STA, ActivityNet, TVR) achieve state-of-the-art performances and demonstrate the effectiveness of incorporating the scene complexity. The project is available at: github.com/dbstjsw0505/SCANet

intensive annotating process (*i.e.* Annotators should find the start-end time of moments corresponding to query descriptions), and sometimes the annotated moments are ambiguous. To cope with this problem, many weakly-supervised VMR (wsVMR) methods [21, 5, 42, 44] have been proposed by only utilizing the video-query pairs, which are less laborious to annotate.

To perform the weak supervision using video-query pairs, if one query is paired (*i.e.* annotated) with multiple videos, we can identify the common scene among these videos and determine the alignment between the query and the scene. To implement this, all videos are divided into multiple segments, and the retrieval system maximizes the similarity scores between each query and paired segments while suppressing the scores between the query and unpaired segments in other videos. During the inference, the system selects a segment with the highest score as a moment prediction for a given query. For the wsVMR systems to accurately classify the best segment in a video, numerous video-language joint representation learning methods [16, 24, 44, 32] have been proposed.

Recently, researchers also have another focus on a study of how to generate video segments to capture many scenes in a video [20, 44]. These segments are referred to as ‘candidate moment proposals’, which is crucial, as they directly affect the retrieval performances by regulating the proposal quantities. Unfortunately, as supervision is not available in generating proposals, wsVMR systems [43, 44] use a fixed number of proposals for all input videos under heuristic optimization of a specific dataset, which is not reasonable to deal with varying numbers of scenes in a video. While some methods [20, 10] consider varying numbers of proposals for each video, they still rely on spurious correlations, such as generating proposals proportionally to the video length or using sliding window. Therefore, the current proposal generation method could not accurately respond to the diverse number of scenes in each video. We refer to this situation as a ‘scene-proposal mismatch’. For instance, in Figure 1(a), the systems produce an unnecessarily large number of proposals by referring to the long length of the video,

1. Introduction

Video search has the core building block of recently growing video streaming services (*e.g.* YouTube, Netflix). To enhance the capability of video search, video moment retrieval (VMR) aims to localize the start and end time of the moment pertinent to a given language query in an untrimmed video. The success of the VMR provides us with accurate video contextual information in less time and effort. Until recently, these remarkable search performances have been dependent on the size and quality of labeled training datasets. However, these datasets cost a labor-

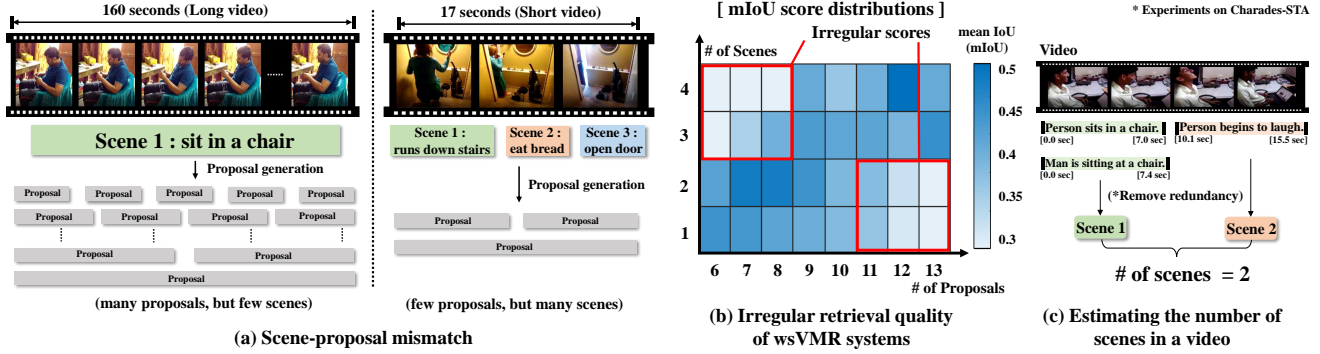


Figure 1: Scene-proposal mismatch in current wsVMR systems: (a) shows an unnecessary many proposals on a video containing few scenes and few proposals on a video containing many scenes, (b) shows mIoU scores of the current model’s predictions according to the number of scenes and the number of generated proposals and (c) shows a method for estimating the number of scenes, where redundant scenes are removed from the counts.

but the video only contains a single scene (*i.e.* scene of sitting still in a chair throughout the video), which should be handled by small amounts of proposals. They also show scene-proposal mismatch by producing a small number of proposals for the video containing many scenes, such that those scarce proposals seem not to work correctly.

Our experimental evidence in Figure 1(b) validates the current wsVMR systems’ incorrectness due to the scene-proposal mismatch. We plot performances (*i.e.* mean Intersection over Union (mIoU) scores) over predictions along the number of scenes in videos and the number of proposals generated, which shows irregularities in the scores. The scores are low for videos with many scenes but few proposals and also low for videos with few scenes but many proposals. To estimate the number of scenes in a video, as shown in Figure 1(c), we counted the number of paired queries for each video as a discrete approximation of the scene. Here, we found that some queries describing the same scene led to redundancy in the counting. Thus, we remove the redundancy of those queries via calculating their IoUs¹ between temporal boundary annotations². Our study further showed that the scene-proposal mismatch affects about 41% of videos in VMR benchmarks (*i.e.* Charades-STA [7], ActivityNet-Caption [13]).

Intrigued by the scene-proposal mismatch, this paper proposes a wsVMR system referred to as Scene Complexity Aware Network (SCANet), which allows the system to mitigate the scene-proposal mismatch problem and generate proposals adaptive to the complexity of the scenes contained in the video. For a given input video, SCANet first defines the scene complexity with a scalar, meaning how difficult for the system to find (*i.e.* retrieve) a specific scene among multiple distinguishable scenes in the video,

which can be effective prior knowledge of video by complementing weak supervision of VMR. On top of the scene complexity, SCANet adaptively generates proposals and enhances their representations. Therefore, SCANet incorporates (1) Complexity-Adaptive Proposal Generation (CPG) that generates adaptive proposals by leveraging the quantities of proposals under consideration of the complexity and (2) Complexity-Adaptive Proposal Enhancement (CPE) that enhances the proposals’ representations corresponding to the scene complexity. Furthermore, motivated by recent success [42, 44] of contrastive learning for wsVMR system, we introduce technical contributions to mine hard negatives in the input video and further video corpus together under our designed framework. Our extensive experiments show the effectiveness of the proposed SCANet, and qualitative results validate enhanced interpretability.

2. Related Works

2.1. Advancements in Video Moment Retrieval

Video Moment Retrieval (VMR) [7], as one of the high-level vision-language tasks, aims to localize video segments corresponding to scene descriptions automatically. Previous successes of multi-modal interaction [28, 19] have contributed to many respectful works [38, 39, 29, 40, 17] to boost retrieval performances by improving the joint representation of video and language to understand their semantic similarities. SMIN [30] and MPPG [25] show the recent state-of-the-art performances of video searching technologies. Researchers are also challenged to make these VMR systems to be more generalized and practical. For general usage of VMR, corpus-level retrieval systems [14, 15, 36, 35] have been proposed, and for practical usage, fast retrieval system [8] has been proposed. Moreover, the laborious annotating problem has been another historical issue of VMR systems. Annotating video segments cor-

¹We remove redundancy by scenes with IoU > 0.5.

²Temporal annotations are used only for identifying the proposal-scene mismatch problem and they are not involved in the wsVMR task

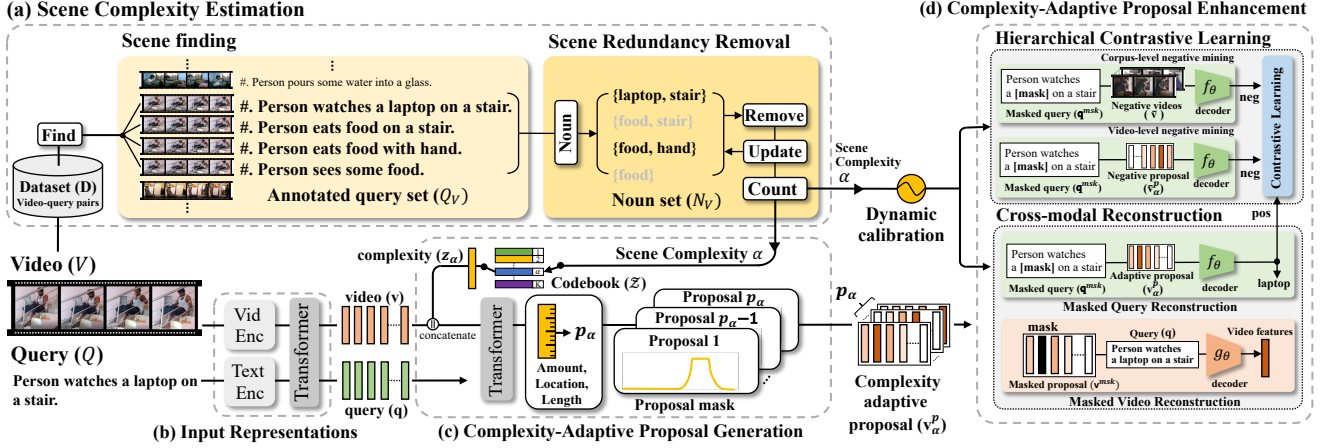


Figure 2: Illustration of proposed SCANet. (a) shows a scene complexity estimation which takes an input video and estimates a scene complexity using video-query pairs, (b) shows input representations, (c) shows a complexity-adaptive proposal generation which generates adaptive proposals according to the complexity, and (d) shows a complexity-adaptive proposal enhancement, which introduces multiple representation enhancements and calibrates them corresponding to the complexity.

responding to given scene descriptions is quite difficult and sometimes inaccurate due to temporal ambiguity. To overcome this, weakly-supervised learning methods have been considered, where it is assumed that systems are not given temporal annotations (start-end time). There has been much literature on weakly-supervised methods. We elaborate on this below in another section with detailed explanations.

2.2. Weakly-supervised Video Moment Retrieval

Weakly-supervised Video Moment Retrieval (wsVMR) shares the same goal as VMR and also aims to reduce the cost of annotation. Therefore, researchers have made an effort to train video-language alignment without temporal boundary annotations, where they considered introducing more affordable supervision. TGA [21] and WS-DEC [5] were the first weakly-supervised VMR systems that utilized the pairing information in video-language pairs³ as a weak-supervision for the alignment. Annotating video-language pairs is less laborious than moment-language pairs for fully-supervised learning. Thus many works have been performed in this weakly-supervised setting. To improve multi-modal interactions, attention-based models [20, 26] have been proposed. In addition, to achieve fine-grained retrieval, methods for refining predictions [31, 37, 4, 34, 10, 41, 33] have also been developed. These methods have made significant contributions to generating candidate moment proposals to predict. With the success of self-supervised learning, recent wsVMR systems [3, 16, 24] introduce the word reconstruction framework from the masked word in the query sentence. Henceforth, contrastive learning achieves large performance gains via

³Fully-supervised setting uses moment-language pairs for training.

mining hard negative retrievals [6, 42, 43, 44] and positive retrievals [32, 9]. However, current systems have never considered the scene-proposal mismatch problem and still suffer from this. Thus we first propose a method to mitigate the mismatch via scene complexity measurements.

3. Method

Figure 2 presents an overall pipeline of the proposed Scene Complexity Aware Network (SCANet) for retrieval systems. SCANet first takes a video and measures scene complexity by estimating how many different scenes are in the video. The scene complexity determines the difficulty of selecting (*i.e.* retrieving) a specific scene among multiple scenes in a given video, which is effective prior knowledge that can be incorporated into weak supervision. Founded on the scene complexity, SCANet builds (1) Complexity-Adaptive Proposal Generation (CPG) and (2) Complexity-Adaptive Proposal Enhancement (CPE). The CPG adaptively leverages proposal generation, which mitigates the scene-proposal mismatch, and the CPE enhances the proposals' representations and dynamically calibrates enhancements according to the scene complexity.

3.1. Scene Complexity

Videos contain a varying number of scenes, and if we can know about the quantities of scenes existing in each video, it should be an effective prior knowledge by giving specified search space to perform retrieval in the video (*i.e.* especially effective in the method of generating retrieval candidates like moment proposals). In that sense, our proposed scene complexity aims to make the retrieval system identify how many scenes exist in the search space of a given video. To

this end, we propose a scene complexity estimation algorithm, which takes inputs from a video V and the dataset D composed of video-query pairs and produces the number of scenes contained in the video as given below:

$$\alpha = f_{sc}(V, D) \in \mathbb{R}^1, \quad (1)$$

where α is the number of different scenes in the video and we define it as the scene complexity. Following, we present a detailed process of the f_{sc} , which includes two procedures: (1) scene finding and (2) scene redundancy removal.

Scene finding. Scene finding is to specify all the candidate scenes in a given video. To this, we utilized the annotated queries sharing the same video as a discrete approximation of the scenes. For obtaining the annotated queries, we investigate video ID⁴ and collect the queries that share the same video ID among the video-query pairs dataset D :

$$Q_V = \text{Find}(V_{id}, D), \quad (2)$$

where V_{id} is video ID, Q_V is annotated query set and $\text{Find}(\cdot, \cdot)$ is a function to collect queries sharing the same video ID. Figure 2(a) also gives examples of annotated queries, but we also find a semantic redundancy⁵ among the queries (e.g. queries meaning “eating food”). It is required to remove the redundancy for the accurate counting of the number of different scenes. Therefore we devise a redundancy removal for our scene complexity estimation.

Scene Redundancy Removal. We utilize the word overlaps⁶ to sense the semantic redundancy among the queries in the annotated query set Q_V . In detail, we identify a part of speech of all words in the queries and sample nouns using the natural language toolkit [18] and then remove the redundancy from the queries that have overlaps by a noun. An example of this process is shown in Figure 2(a), where the annotated query set is provided as $Q_V = \{\text{‘Person watches a laptop on a stair’}, \text{‘Person eats food on a stair’}, \text{‘Person eats food with hand’}, \text{‘Person sees some food’}\}$, we should exclude the redundant queries that share similar meaning. To estimate which queries are semantically redundant, we identify nouns in each query as $N_V = \{\{\text{laptop, stair}\}, \{\text{food, stair}\}, \{\text{food, hand}\}, \{\text{food}\}\}$, where N_V is defined as annotated noun set. Here, nouns meaning human (e.g. person) are excluded from the findings. If there is word overlap between elements of N_V , the element with the most overlapping is removed first. This process is continued until there is no overlap. Thus, the redundancy removal process prunes out the elements in N_V and updates

upto $N_V = \{\{\text{laptop, stair}\}, \{\text{food, hand}\}\}$ or $\{\{\text{laptop, stair}\}, \{\text{food}\}\}$ after that, we count the number of elements in N_V as the final number of scenes in the given video V . To summarize the scene complexity estimation process, we formally define an algorithm about f_{sc} below:

Algorithm 1 Scene complexity estimation algorithm f_{sc}

- 1: **Input:** Video V , video-query pairs dataset D
 - 2: **Output:** Scene complexity α
 - 3: Find annotated query set: $Q_V = \text{Find}(V_{id}, D)$
 - 4: Find nouns: $N_V = \text{Noun}(Q_V)$
 - 5: **while** \exists word overlap among elements in N_V **do**
 - 6: Remove the most overlapping: $N_V \leftarrow \text{Remove}(N_V)$
 - 7: **end**
 - 8: **return** The number of elements of N_V
-

$\text{Noun}(\cdot)$ is a function to filter out noun⁷ and $\text{Remove}(\cdot)$ is a function to find the element with the most overlap and remove it. The final number of elements in N_V is defined as scene complexity α of an integer scalar. Following, our proposed SCANet utilizes α to adapt the retrieval in terms of proposal generation and proposal enhancement.

3.2. Scene Complexity Aware Network

Input Representations. We first give formal definitions of the input video V and query Q used in SCANet. Frame-level video features are obtained from a pre-trained video encoder [2, 27] and word-level query features are obtained from text encoder [23]. Both features are embedded into d -dimensional joint space. After adding positional encoding [28] and applying layer normalization [1], we get the final video features $\mathbf{v} \in \mathbb{R}^{N_v \times d}$ and the query features $\mathbf{q} \in \mathbb{R}^{N_q \times d}$, where N_v is the number of video frames and N_q is the number of words in the query.

Multi-Modal Interaction. To give multi-modal interactions between the query and video, we use Transformer Attention [28]. The video features \mathbf{v} and query features \mathbf{q} are concatenated and prepared for the Attention inputs below:

$$[\mathbf{v}||\mathbf{q}] = \text{Attention}([\mathbf{v}||\mathbf{q}]) \in \mathbb{R}^{(N_v+N_q) \times d}, \quad (3)$$

where $[\cdot||\cdot]$ denotes the concatenation and we get attended video features $\mathbf{v} \in \mathbb{R}^{N_v \times d}$ and query features $\mathbf{q} \in \mathbb{R}^{N_q \times d}$.

3.3. Complexity-Adaptive Proposal Generation

Complexity-Adaptive Proposal Generation (CPG) in Figure 2(c) is designed to generate candidate moment proposals adapting the scene complexity α of a given video, where α accounts for three aspects of proposals: amount, location, and length. To implement this, we devise a ‘complexity vector’ corresponding to the complexity level, such

⁴Previously, video ID (e.g. ‘ID: 0BH84’) is just used for accessing the feature data, but we further utilize the ID to build Q_V .

⁵video-query pair datasets usually contain many redundant queries

⁶Unlike Figure 1(c), labels (start-end times) are unavailable in wsVMR.

⁷Table 4 gives ablation studies (e.g. verb) to identify the redundancy.

that we build a codebook $\mathcal{Z} = \{\mathbf{z}_k\}_{k=1}^K \in \mathbb{R}^{K \times d}$ composed of d -dimensional learnable K vectors and select a single vector by indexing α as $\mathbf{z}_\alpha \in \mathbb{R}^d$, where the K (*e.g.* $K = 8$) is the maximum number of α . Thus the \mathbf{z}_α is our defined complexity vector that has sensibility according to the α . After providing semantics of input modalities as $[\mathbf{z}_\alpha || \mathbf{v} || \mathbf{q}] = \text{Attention}([\mathbf{z}_\alpha || \mathbf{v} || \mathbf{q}])$, in the following, the \mathbf{z}_α generates adaptive proposals deciding proposal properties in terms of amount, location and length.

To decide the amount of proposals, we first build an integer set $I = \{p_{\min}, p_{\min} + 1 \cdots, p_{\max}\} \in \mathbb{R}^n$ regarding the number of proposals, where p_{\min} and p_{\max} are the minimum (*e.g.* 5) and the maximum number (*e.g.* 10) of proposals, and the $n = p_{\max} - p_{\min} + 1$ is the number of elements in the I . The complexity vector \mathbf{z}_α decides a single number $p_\alpha \in I$ from the integer set I and generates p_α proposals. For the detailed implementation of this, we utilize a Multi-Layer Perceptron (MLP) that takes the input of \mathbf{z}_α and produces an n -dimensional selection vector as $\mathbf{a} = \text{MLP}(\mathbf{z}_\alpha) \in \mathbb{R}^n$, where the n denotes the same dimension of the integer set I . The selection is performed by using an output of $\text{argmax}(\mathbf{a})$ as index to select a single integer in I and the selected integer defines the number of proposals p_α . To make selection trainable, we introduce a Gumbel-Softmax [12], which provides functional n -dimensional one-hot vector \mathbf{g} for the deciding the number of proposals over integer set I below:

$$\begin{aligned} \mathbf{g} &= \text{Gumbel_Softmax}(\mathbf{a}) \in \mathbb{R}^n, \\ p_\alpha &= \sum_{i=1}^n \mathbf{g}_i \cdot I_i \in \mathbb{R}^1, \end{aligned} \quad (4)$$

where the number of proposals (*i.e.* amount of proposals) is finally determined by the number $p_{\min} \leq p_\alpha \leq p_{\max}$.

To decide the locations and lengths of proposals, we first build a proposal mask, which remains the video features in the region of the mask as the proposal features. A center point and width of the operating region of the mask correspond to the location and length of the proposal, thus complexity features $\mathbf{z}_\alpha \in \mathbb{R}^{1 \times d}$ regresse the center and width as $[\mathbf{c}_\alpha, \mathbf{w}_\alpha] = \sigma(\mathbf{z}_\alpha W_p) \in \mathbb{R}^2$, where $W_p \in \mathbb{R}^{d \times 2}$ is learnable weights for regression $\sigma(\cdot)$ is the sigmoid function, and $\mathbf{c}_\alpha \in \mathbb{R}^1$, $\mathbf{w}_\alpha \in \mathbb{R}^1$ are the center and width of the proposal.

Founded on the \mathbf{c}_α and \mathbf{w}_α , we design the proposal mask, and here, Figure 3(a) shows a popular example of the proposal mask using Gaussian curve [43, 44]. We consider that the Gaussian curve may not be reasonable because video features are unevenly attended inside the proposal. Therefore, as shown in Figure 3(b), we design Flatten Gaussian mask, which is simple, yet more reasonable by evenly remaining features inside the proposal⁸. To give a formal definition of our mask, we first construct the base mask using the Gaussian curve using the \mathbf{c}_α and \mathbf{w}_α as given below:

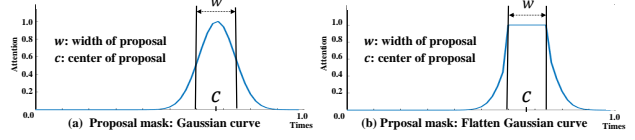


Figure 3: Illustration of (a) gaussian mask [43, 44] and (b) our proposed flatten gaussian mask.

$$\mathbf{m}_\alpha^p[i] = \frac{1}{\sqrt{2\pi}(\mathbf{w}_\alpha^p/\sigma)} \exp\left(-\frac{(i/N_v - \mathbf{c}_\alpha^p)^2}{2(\mathbf{w}_\alpha^p/\sigma)^2}\right), \quad (5)$$

where the $\mathbf{m}_\alpha^p \in \mathbb{R}^{N_v}$ is the base mask and \mathbf{c}_α^p , \mathbf{w}_α^p denote the center point and width, where a superscript $p \in [1, \dots, p_\alpha]$ denotes p -th proposal. The σ is a hyperparameter. The $i \in [1, \dots, N_v]$ denotes the i -th index in the length of video. We flatten the base mask in the region of proposals by substituting the attention values of it as mean values of them: $\mathbf{m}_\alpha^p[\text{st: ed}] = \text{Mean}(\mathbf{m}_\alpha^p[\text{st: ed}])$, where $\text{st} = \mathbf{c}_\alpha^p - (\frac{\mathbf{w}_\alpha^p}{2})$, $\text{ed} = \mathbf{c}_\alpha^p + (\frac{\mathbf{w}_\alpha^p}{2})$, and $\text{Mean}(\cdot)$ is a mean-pooling while keeping the input dimension. Thus, $\mathbf{m}_\alpha^p[i]$ is updated as i -th value of the Flatten Gaussian function corresponding to p -th proposal. Using \mathbf{m}_α^p , the video features $\mathbf{v} \in \mathbb{R}^{N_v \times d}$ are attended to produce p -th proposal features:

$$\mathbf{v}_\alpha^p = \mathbf{v} \circ \mathbf{m}_\alpha^p \in \mathbb{R}^{N_v \times d}, \quad (6)$$

where \circ is column-wise multiplication. Therefore \mathbf{v}_α^p is our final complexity adaptive proposal features. In the following, we use the \mathbf{v}_α^p to learn proposal-language alignment by our designed proposal enhancement framework.

3.4. Complexity-Adaptive Proposal Enhancement

Due to the unavailability of supervision, wsVMR systems rely on several training objectives. As shown in Figure 2(d), to enhance the adaptive proposal features \mathbf{v}_α^p , SCANet contains the multiple representation enhancements: (1) Cross-modal Reconstruction and (2) Hierarchical Contrastive Learning, where they are dynamically calibrated according to the scene complexity α .

Cross-modal Reconstruction. Cross-modal reconstruction aims to learn connections of common semantics among the modalities (*i.e.* video, query), such that we mask a part of the features in one modality and restore that part by referring to the other modality. Depending on which modality is masked, it is referred to as follows: (1) masked query reconstruction (MQR), and (2) masked video reconstruction (MVR). For the MQR, we randomly sample verb or noun tokens to mask in query $Q = \{w_1 \cdots w_{N_q}\}$ (*i.e.* w is a word token). Before masking, we define these tokens as target tokens w^{tgt} to predict, and they are replaced by [mask]

⁸Table 5 validate also the effectiveness of Flatten Gaussian curve.

tokens in the query, which makes masked query features $\mathbf{q}^{msk} \in \mathbb{R}^{N_q \times d}$ throughout text encoder. Thus a masked query reconstruction loss \mathcal{L}_{mqr} is defined by cross-entropy loss to predict the target words w^{tgt} in the mask from the \mathbf{q}^{msk} and proposal features \mathbf{v}_α^p as given below:

$$\mathcal{L}_{mqr}(\theta) = -\frac{1}{p_\alpha} \sum_{p=1}^{p_\alpha} \log f_\theta(w^{tgt} | \mathbf{q}^{msk}, \mathbf{v}_\alpha^p), \quad (7)$$

where θ is learnable weight and f_θ is a decoder to reconstruct the target words. For the MVR, we sample video frame features with a probability of 10%⁹, where they are defined as \mathbf{v}^{tgt} , and then replaced by zeros to make masked proposal features $(\mathbf{v}_\alpha^p)^{msk}$. As the \mathbf{v}^{tgt} is the d -dimensional video features, we introduce a regressor g_θ to regress the features. Thus, video reconstruction loss \mathcal{L}_{mvr} is defined by L2 loss between target and regressed features as below:

$$\mathcal{L}_{mvr}(\theta) = \frac{1}{p_\alpha} \sum_{p=1}^{p_\alpha} \|\mathbf{v}^{tgt} - g_\theta(\mathbf{q}, (\mathbf{v}_\alpha^p)^{msk})\|_2^2. \quad (8)$$

Hierarchical Contrastive Learning. Contrastive learning aims to enhance the representations of positive features (*i.e.* proposal region) via comparing negative features (*i.e.* non-proposal region). It is crucial to find a hard negative case (*i.e.* a scene similar to a positive but not a positive). Therefore, we build hierarchical contrastive learning to explore the hard negatives at the video-level and corpus-level. For the video-level, we use the input video V and mine the negative in the region of V excluding the adaptive proposal \mathbf{v}_α^p . We first make the negative mask from the positive mask as $1 - \mathbf{m}_\alpha^p$, and get the negative proposal features $\bar{\mathbf{v}}_\alpha^p = \mathbf{v} \circ (1 - \mathbf{m}_\alpha^p) \in \mathbb{R}^{N_v \times d}$. We then, compare the masked query reconstruction losses between positive proposals (\mathcal{L}_{mqr}) and negative proposals (\mathcal{L}_{mqr}^*), which defines video-level contrastive loss \mathcal{L}_{vid} with margin of δ_1 as:

$$\begin{aligned} \mathcal{L}_{vid}(\theta) &= \max(\mathcal{L}_{mqr}(\theta) - \mathcal{L}_{mqr}^*(\theta) + \delta_1, 0), \\ \mathcal{L}_{mqr}^*(\theta) &= -\frac{1}{p_\alpha} \sum_{p=1}^{p_\alpha} \log f_\theta(w^{tgt} | \mathbf{q}^{msk}, \bar{\mathbf{v}}_\alpha^p). \end{aligned} \quad (9)$$

For the corpus-level, we use a video corpus D_V composed of all videos in the dataset and mine the hard negative videos for contrastive learning. To find the hard negatives, we perform video retrieval on the corpus, which takes the query Q and video corpus D_V as inputs and predicts the top-k videos as outputs $V_k = \text{SCANet}^k(Q, D_V)$ ¹⁰, where V_k is the top-k videos with the lowest \mathcal{L}_{mqr} for the input query. As the V_k is utilized for negative videos, the ground-truth video is removed from it. Similar to video-level, we

⁹Masking is performed in a range of $[\mathbf{c}_\alpha^p - (\mathbf{w}_\alpha^p/2), \mathbf{c}_\alpha^p + (\mathbf{w}_\alpha^p/2)]$ in video frames assuming effective region of the proposal features \mathbf{v}_α^p .

¹⁰See details about video retrieval of SCANet^k(\cdot, \cdot) in Section 5.

compare the masked query loss between positive proposals (\mathcal{L}_{mqr}) and negative videos ($\mathcal{L}_{mqr}^\dagger$), which defines the corpus-level contrastive loss \mathcal{L}_{cps} with a margin δ_2 below:

$$\begin{aligned} \mathcal{L}_{cps}(\theta) &= \max(\mathcal{L}_{mqr}(\theta) - \mathcal{L}_{mqr}^\dagger(\theta) + \delta_2, 0), \\ \mathcal{L}_{mqr}^\dagger(\theta) &= -\frac{1}{k} \sum_{i=1}^k \log f_\theta(w^{tgt} | \mathbf{q}^{msk}, \bar{\mathbf{v}}_i), \end{aligned} \quad (10)$$

where $\bar{\mathbf{v}}_i \in \mathbb{R}^{N_v \times d}$ is i -th negative video features from V_k .

Dynamic calibration. The videos with high complexity are usually more difficult to train than videos with lower complexity, as the moment predictions in the videos with high complexity are performed under many proposals. Thus we dynamically calibrate the loss to place different weights according to the complexity α as below:

$$\mathcal{L} = \frac{\gamma}{1 + e^{-\alpha}} (\mathcal{L}_{mqr} + \mathcal{L}_{mvr} + \mathcal{L}_{vid} + \mathcal{L}_{cps}), \quad (11)$$

where γ is a hyperparameter. For the inference, SCANet predicts the proposal that generates the lowest $\mathcal{L}_{mqr} + \mathcal{L}_{mvr}$ among the proposals. For the best proposal, the start-end times ([st,ed]) are inferred from the corresponding proposal mask's width \mathbf{w}_α^p and center \mathbf{c}_α^p and scaled by the video duration as [st,ed] = $[\mathbf{c}_\alpha^p - \frac{\mathbf{w}_\alpha^p}{2}, \mathbf{c}_\alpha^p + \frac{\mathbf{w}_\alpha^p}{2}] * \text{duration}$

4. Experiments

4.1. Dataset

Our proposed SCANet is validated on three moment retrieval benchmark datasets, where the wsVMR system uses temporal annotations only for evaluation.

Charades-STA. Charades-STA includes about 30 seconds of videos for human behaviors and their language queries. Average length is about 29.8 seconds, and dataset contains 12,408 video-query pairs and 3,720 for testing.

ActivityNet Captions. ActivityNet Captions is a large-scale dataset including about 117 seconds videos of human actions and their language query. The dataset contains 19,290 videos with 37,417/17,505/17,031 samples for train/val.1/val.2 splits. SCANet is validated on the val.2.

TV show Retrieval. TV show Retrieval (TVR) [14] comprises 6 TV shows about diverse genres, including 109K queries from 21.8K multi-character videos with subtitles. Each video is about 60-90 seconds. The TVR is split into 80% train, 10% val, 10% test-public. The test-public is prepared for the challenge. As test-public is currently unavailable, SCANet is validated on the val.

Table 1: Performances of weakly-supervised video moment retrieval on the Charades-STA dataset.

Method	R@1,IoU=m			R@5,IoU=m		
	m=0.3	m=0.5	m=0.7	m=0.3	m=0.5	m=0.7
TGA [21]	32.14	19.94	8.84	86.58	65.52	33.51
CTF [4]	39.80	27.30	12.90	-	-	-
SCN [16]	42.96	23.58	9.97	95.56	71.80	38.87
WSTAN [31]	43.39	29.35	12.28	93.04	76.13	41.53
BAR [33]	44.97	27.04	12.23	-	-	-
LoGAN [26]	48.04	31.74	13.71	89.01	72.17	37.58
MARN [24]	48.55	31.94	14.81	90.70	70.00	37.40
WSRA [6]	50.13	31.20	11.01	86.75	70.50	39.02
CCL [42]	-	33.21	15.68	-	73.50	41.87
CRM [10]	53.66	34.76	16.37	-	-	-
VCA [32]	58.58	38.13	19.57	98.08	78.75	37.75
LCNet [34]	59.60	39.19	18.87	94.78	80.56	45.24
RTBPN [41]	60.04	32.36	13.24	97.48	71.85	41.18
CNM [43]	60.39	35.43	15.45	-	-	-
CPL [44]	65.99	49.05	22.61	96.99	84.71	52.37
SCANet (ours)	68.04	50.85	24.07	98.24	86.32	53.28

Table 2: Performances of weakly-supervised video moment retrieval on the ActivityNet Captions dataset.

Method	R@1,IoU=m			R@5,IoU=m		
	m=0.1	m=0.3	m=0.5	m=0.1	m=0.3	m=0.5
WS-DEC [5]	62.71	41.98	23.34	-	-	-
EC-SL [3]	68.48	44.29	24.16	-	-	-
MARN [24]	-	47.01	29.95	-	72.02	57.49
SCN [16]	71.48	47.23	29.22	90.88	71.56	55.69
BAR [33]	-	49.03	30.73	-	-	-
RTBPN [41]	73.73	49.77	29.63	93.89	79.89	60.56
CTF [4]	74.20	44.30	23.60	-	-	-
WSLLN [9]	75.40	42.80	22.70	-	-	-
LCNet [34]	78.58	48.49	26.33	93.95	82.51	62.66
CCL [42]	-	50.12	31.07	-	77.36	61.29
WSTAN [31]	79.78	52.45	30.01	93.15	79.38	63.42
CRM [10]	81.61	55.26	32.19	-	-	-
CNM [43]	78.13	55.68	33.33	-	-	-
CPL [44]	82.55	55.73	31.37	87.24	63.05	43.13
SCANet (ours)	83.62	56.07	31.52	94.36	82.34	64.09

4.2. Evaluation Metric

To evaluate the moment retrieval, we compute the average recall (R@n) over all queries, where temporal Intersection over Union (IoU=m) measures the overlap between prediction and ground-truth. The n denotes the recall rate of top-n predictions, and m is the predefined IoU threshold, thus quantifying the percentage of predicted moments with the IoU value larger than m among top-n predictions.

4.3. Experimental Results

Table 1 and Table 2 summarize the results on Charades-STA (C-STA) and ActivityNet Captions (ANC) datasets. SCANet is compared to previous works (Please, refer to Related Works for their detailed descriptions). SCANet shows

Table 3: Performances of weakly-supervised video moment retrieval on the TVR dataset (validation) (* reproduced).

Method	R@1,IoU=m			R@5,IoU=m		
	m=0.1	m=0.3	m=0.5	m=0.1	m=0.3	m=0.5
TGA* [21]	17.61	2.38	0.97	48.63	11.54	5.32
CPL* [44]	33.16	7.28	2.11	64.41	17.93	8.56
SCANet (ours)	37.51	10.76	4.24	67.47	20.32	10.21

Table 4: Ablation study of redundancy removal for scene complexity estimation (f_{sc}) along the types to find the redundancy among queries in the annotated query set Q_V .

Types	R@1,IoU=m			R@5,IoU=m
	m=0.1	m=0.3	m=0.5	m=0.1
none	77.43	49.78	27.32	88.32
noun	83.54	55.97	31.82	94.50
verb	81.26	52.32	30.91	93.11
noun & verb	80.52	51.31	30.42	92.52

the best performances of all metrics on C-STA, which is especially effective in metrics (R@1) and also the effectiveness in four metrics on ANC. Table 3 firstly summarizes the results of wsVMR systems on TVR. We reproduce the baseline and most recent model from their public codes. The videos in TVR are quite challenging because they include relatively more similar actions and backgrounds, which are difficult for the models to distinguish. Overall, SCANet shows improvements in retrieval quality, but it is also notable that those improvements are mainly from rectifying the video samples suffering scene-proposal mismatch problems, which can be confirmed in Figure 4(c).

4.4. Ablation Study

Ablation studies are performed on the ActivityNet Captions (validation)¹¹. To show performance variances in various metrics, we validate SCANet on the most challenging metric (R@1,IoU=0.5) and the easiest one (R@5,IoU=0.1). Table 4 summarizes the studies about redundancy removal in the annotated query set Q_V for scene complexity estimation (f_{sc}). The first section is the results of complexity estimation without redundancy removal, where the complexity equals the number of queries in Q_V . The below sections are the results with redundancy removal, where the noun or verb is used to find the redundant queries. Redundancy removal with the noun or verb shows effectiveness. However, using both together decreases the performance. We consider that finding redundant queries using noun and verb together can make sure to find almost the same descriptions (e.g. ‘person *drinks* a cup of *coffee*’ and ‘person *drinks coffee* in table’), but the redundancy shows variance for the

¹¹The validation set of Charades-STA is not available.

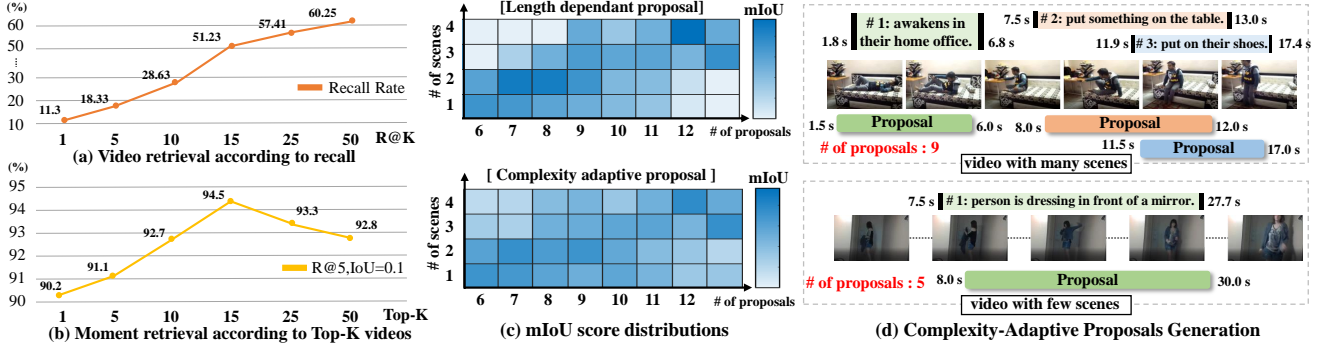


Figure 4: Qualitative results of SCANet: (a) shows the video retrieval performance of SCANet according to the R@k, (b) shows the moment retrieval performances according to involving top-K retrieved videos for hierarchical contrastive learning, (c) shows the IoU scores distributions according to the number of scenes and proposals (upper: length dependant proposals, below: complexity adaptive proposals), and (d) illustrates the proposals in SCANet according to videos with diverse scenes.

Table 5: Ablation study of generating proposals. (n: number of proposals), (w, s: frame width and stride of window), (G: Gaussian mask, FG: Flatten Gaussian mask)

Method	R@1, IoU=m		R@5, IoU=m
	m=0.1	m=0.5	m=0.1
fixed proposal (n:6)	77.43	27.32	88.32
fixed proposal (n:8)	78.23	27.98	90.12
sliding window (w:{20,40},s:5)	68.54	26.51	84.32
sliding window (w:{20,40,60},s:10)	69.14	24.33	86.72
complexity adaptive proposal (G)	81.32	29.43	92.41
complexity adaptive proposal (FG)	83.54	31.82	94.50

same scene by changing noun or verb (e.g. ‘one *holds* a cup of coffee’). Table 5 shows the ablation studies of proposal generations in SCANet. Our baseline was to generate proposals with fixed numbers, and here, to determine the locations and lengths of the proposals, we use a learnable d -dimensional single vector instead of complexity features $\mathbf{Z}_\alpha \in \mathbb{R}^d$. We modify the baseline with the sliding window and complexity adaptive proposals, where the adaptive method was the most effective with our designed Flatten Gaussian mask. Table 6 shows the ablation studies with two proposal enhancements and their calibrating method. Incremental improvements are shown by adding two enhancements and calibrating the enhancements with complexity.

Figure 4(a) shows video retrieval performances of SCANet that predicts recall rate with top-k videos (R@K), and (b) shows the moment retrieval performances according to using top-k videos for \mathcal{L}_{cps} . As K increases, the recall increases in (a), and the moment retrieval performance in (b) is also improved using the predicted videos as negative (i.e. ground-truth videos are removed). Meanwhile, over K=15, the performance degrades, where we presume that the top-k videos start to contain not hard negative videos.

Table 6: Ablation study in Complexity-Adaptive Proposal Enhancement. (CMR: cross-modal reconstruction, HCL: hierarchical contrastive learning, Calibration: calibrating training loss according to scene complexity α).

Proposal Enhancement			R@1, IoU=m	R@5, IoU=m
CMR	HCL	Calibration	m=0.5	m=0.1
✓			28.45	91.52
✓	✓		30.32	92.71
✓	✓	✓	31.82	94.50

4.5. Qualitative Results

Figure 4(c) shows the retrieval performances according to the number of proposals and the number of scenes for each video. Compared to the length-dependant proposals (e.g. sliding window), complexity adaptive proposals mitigates the scene-proposal mismatch by regularizing the retrieval quality along the number of proposals and scenes. Figure 4(d) illustrates the proposals of SCANet on (a) short-length video with many scenes and (b) long-length video with few scene. The adaptive proposals accurately capture the scenes in both cases. In the bottom left, we also add the number of generated proposals for each video, which shows that proposals do not depend on the length of the video.

5. Implementation Details

Data Settings. For the video encoder, I3D [2] model is used to get the Charades-STA video features, and C3D [27] model is used for the ActivityNet-Caption video features. Both video features are extracted by every 8 frames. For the word token embedding, we use word2vec from GloVe [23]. The size of the vocabulary is fixed as 8000 with maximum 20 word-length of sentence.

Query: Person tries to fix a loose doorknob.

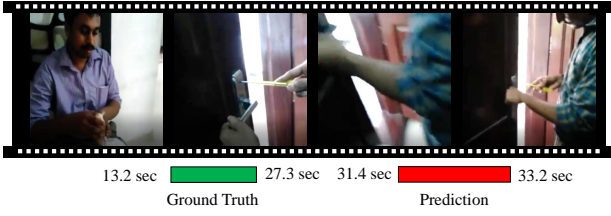


Figure 5: Illustration of failure case of moment prediction.

Model Settings. Hyperparameters in SCANet are as follows: $K = 12$ for the maximum number of scene complexity, the minimum number of proposal $p_{\min} = 5$, the maximum number of proposals $p_{\max} = 14$, the hyperparameter of calibration is $\gamma = 0.5$. The σ for Gaussian function is 8, the margins δ_1 for contrastive loss \mathcal{L}_{vid} and \mathcal{L}_{cps} are $\delta_1 = 0.1$, $\delta_2 = 0.5$, where the higher margin of δ_2 is designed for promoting to distinguish the positive video from negative videos with similar scenes.

Video Retrieval with SCANet. To prepare the ranked top-k videos used in $V_k = \text{SCANet}^k(Q, D_V)$, they are retrieved by SCANet trained from reconstruction losses (*i.e.* $\mathcal{L}_{mgr}, \mathcal{L}_{mvr}$) and video-level contrastive loss (*i.e.* \mathcal{L}_{vid}). SCANet retrieves top-k (*e.g.* $k=15$) videos from the video dataset for each query that have the lowest reconstruction losses. To train wsVMR, the top-k video IDs are utilized to provide hard negative videos, which makes \mathcal{L}_{cps} . The retrieval performances (*i.e.* R@K, prediction recall according to top-K videos) are presented in Figure 4.

6. Failure cases

Figure 5 presents the failure case of our proposed SCANet. For the query ‘Person tries to fix a doorknob’, SCANet predicts the moment of person opening the door. We consider the action of fixing something is not frequently paired with queries, and it may be unavailable for wsVMR systems to directly learn about the uncommon video-query pairs as a long-tailed action recognition problem. Therefore, wsVMR systems are more vulnerable to these actions with categories in the long tail and we believe that overcoming the long-tail problem should be a contribution to many tasks including moment retrieval. Our future works also include mitigating this long-tail problem.

7. Limitations

Our proposed method is based on the scene complexity of video by referring to the number of annotated queries to the video, which can also have more flexibility by applying other systems [22, 11] under weak supervision. How-

ever, in the real environment, it may not be available to get scene complexity of video from referring to other annotated queries (*i.e.* In real environment, we may not access to other annotated query sets for one video). We feel this is our current SCANet’s limitation, and to overcome this, we further made another effort to learn scene complexity via the neural network from the input of video, where we refer to this method as ‘Scene Complexity Neural Estimator’. In our supplementary materials, we elaborate on this with our current studies as another our experimental contributions.

8. Conclusion

SCANet is presented to consider a scene-proposal mismatch problem in the wsVMR. SCANet measures a scene complexity of multiple scenes in each video. Founded on the complexity, SCANet builds complexity-adaptive proposal generation to mitigate the scene-proposal mismatch and complexity-adaptive proposal enhancement to enhance the representation by calibrating with the complexity.

Acknowledgment

This work was supported by Institute for Information & communications Technology Planning & Evaluation (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. 2022R1A2C2012706).

References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [3] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8435, 2021.
- [4] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*, 2020.
- [5] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems*, 31, 2018.
- [6] Zhiyuan Fang, Shu Kong, Zhe Wang, Charless Fowlkes, and Yezhou Yang. Weak supervision and referring atten-

- tion for temporal-textual association learning. *arXiv preprint arXiv:2006.11747*, 2020.
- [7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.
 - [8] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1523–1532, 2021.
 - [9] Mingfei Gao, Richard Socher, and Caiming Xiong. Weakly supervised natural language localization networks, Nov. 26 2020. US Patent App. 16/531,343.
 - [10] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208, 2021.
 - [11] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2022.
 - [12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
 - [13] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.
 - [14] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer, 2020.
 - [15] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.
 - [16] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546, 2020.
 - [17] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022.
 - [18] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.
 - [19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
 - [20] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *European conference on computer vision*, pages 156–171. Springer, 2020.
 - [21] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019.
 - [22] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. Wtalc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018.
 - [23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
 - [24] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020.
 - [25] Xin Sun, Xuan Wang, Jialin Gao, Qiong Liu, and Xi Zhou. You need to read again: Multi-granularity perception network for moment retrieval in videos. *arXiv preprint arXiv:2205.12886*, 2022.
 - [26] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2083–2092, 2021.
 - [27] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
 - [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
 - [29] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. Dual path interaction network for video moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4116–4124, 2020.
 - [30] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2021.
 - [31] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. Weakly supervised temporal adjacent network for language grounding. *IEEE Transactions on Multimedia*, 2021.
 - [32] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1459–1468, 2021.
 - [33] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1283–1291, 2020.

- [34] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 30:3252–3262, 2021.
- [35] Sunjae Yoon, Ji Woo Hong, Soohwan Eom, Hee Suk Yoon, Eunseop Yoon, Daehyeok Kim, Junyeong Kim, Chanwoo Kim, and Chang D Yoo. Counterfactual two-stage debiasing for video corpus moment retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [36] Sunjae Yoon, Ji Woo Hong, Eunseop Yoon, Dahyun Kim, Junyeong Kim, Hee Suk Yoon, and Chang D Yoo. Selective query-guided debiasing for video corpus moment retrieval. In *European Conference on Computer Vision*, pages 185–200. Springer, 2022.
- [37] Sunjae Yoon, Dahyun Kim, Ji Woo Hong, Junyeong Kim, Kookhoi Kim, and Chang D Yoo. Weakly-supervised moment retrieval network for video corpus moment retrieval. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 534–538. IEEE, 2021.
- [38] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019.
- [39] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020.
- [40] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020.
- [41] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4098–4106, 2020.
- [42] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134, 2020.
- [43] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, page 3, 2022.
- [44] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15555–15564, 2022.