Visual Co-Occurrence Alignment Learning for Weakly-Supervised Video Moment Retrieval

Zheng Wang, Jingjing Chen, Yu-Gang Jiang*
{zhengwang17,chenjingjing,ygj}@fudan.edu.cn
Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University

ABSTRACT

Video moment retrieval aims to localize the most relevant video moment given the text query. Weakly supervised approaches leverage video-text pairs only for training, without temporal annotations. Most current methods align the proposed video moment and the text in a joint embedding space. However, in lack of temporal annotations, the semantic gap between these two modalities makes it predominant to learn joint feature representation for most methods, with less emphasis on learning visual feature representation. This paper aims to improve the visual feature representation with supervisions in the visual domain, obtaining discriminative visual features for cross-modal learning. Based on the observation that relevant video moments (i.e., share similar activities) from different videos are commonly described by similar sentences; hence the visual features of these relevant video moments should also be similar despite that they come from different videos. Therefore, to obtain more discriminative and robust visual features for video moment retrieval, we propose to align the visual features of relevant video moments from different videos that co-occurred in the same training batch. Besides, a contrastive learning approach is introduced for learning the moment-level alignment of these videos. Through extensive experiments, we demonstrate that the proposed visual co-occurrence alignment learning method outperforms the cross-modal alignment learning counterpart and achieves promising results for video moment retrieval.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**.

KEYWORDS

video retrieval, cross-modal interaction, noise contrastive learning, weakly-supervised

ACM Reference Format:
Zheng Wang, Jingjing Chen, Yu-Gang Jiang. 2021. Visual Co-Occurrence Alignment Learning for Weakly-Supervised Video Moment Retrieval. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475278

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475278

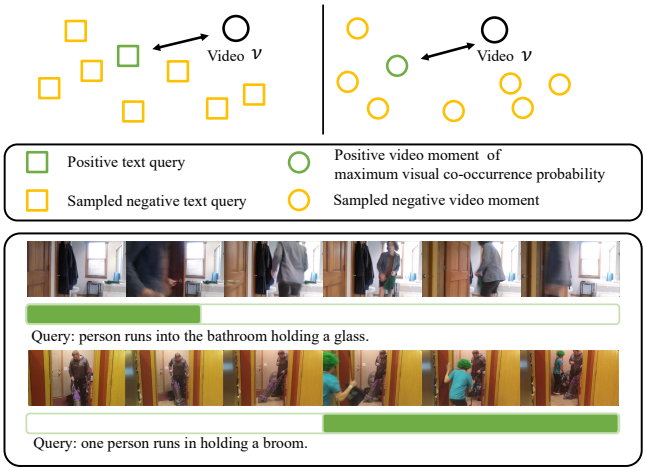

Figure 1: Given a video v and its associated text query for trainig a video moment retrieval model, most approaches aim at finding a moment that is most relevant to the query text (top left). We propose to select the moment that maximize the visual co-occurrence probability between video moments of similar query text within the same training batch(top right), in the light of activities in two different videos typically share semantically similar textual descriptions (bottom).

1 INTRODUCTION

Retrieving a video moment from a long video plays an important role in many real world scenarios such as video summarization [3, 31], video search [26, 48], video question answering [25, 42] and video object grounding [49]. Early works on localizing activities of interest from untrimmed videos [11, 38] mainly focus on pre-defined action categories, which are not able to deal with the situation when the queries are in different forms. Therefore, the problem of retrieving video moments [2, 13, 29] given free-form textual queries emerges and receives much attention recently.Nevertheless, early works on video moment retrieval focus on the fully-supervised scenarios [2, 13, 29], where the temporal boundaries of query sentences are given for the model training. Despite promising performances have been achieved, such fully-supervised methods requires explicit annotations for temporal boundaries, which are cumbersome and expensive. In addition, the temporal boundaries of activities are usually subjective and ambiguous, which may bring noises for model training. Considering that there are a bunch of narrated

and untrimmed videos available on the web and would benefit the learning eventually, weakly-supervised methods [14, 33, 41], which leverages the video-level textual annotations only and without the requirement of temporal annotations, are introduced for video moment retrieval.

The existing weakly-supervised video moment retrieval methods are mostly focusing on utilizing pairwise ranking loss for learning a joint embedding space between visual and textual embedding [12, 30, 33, 41]. Such a paradigm is widely adopted in cross-modal visual-language retrieval. As for the video moment retrieval task, previous approaches mainly focus on maximizing relevant video frames' similarity with the corresponding text. Due to the fact of lacking strong supervision information, the performance of such weakly-supervised methods are usually unsatisfactory. As a result, to obtain stronger supervision information, recent works focus on training the model with two ranking losses, i.e., an intra-video ranking loss and an inter-video ranking loss. The intra-video ranking loss aims to select the most relevant video moment in the video that is aligned with the query sentence. In contrast, the inter-video ranking loss maximizes the similarity between a video and the corresponding text while minimizes the similarity to other irrelevant texts. During the training, the frame features are typically pre-computed by action classification models, which results in negligible semantic discrepancy among different frames within videos. Therefore, distinguishing a set of frames relevant to the query sentence under weakly-supervised setting is inherently a challenge problem.

This paper studies the problem of weakly-supervised video moment retrieval. Specifically, different from previous works using paired sentences as supervision information, we propose to learn more discriminative and robust visual features by mining visual supervision information. Figure 1 shows the motivation of our method. As shown in the figure, visually relevant activities may occur in different videos and people are likely to describe them with semantic similar sentences. To achieve better video moment retrieval performances, the visual features of these relevant video moments should also be similar despite that they come from different videos. Therefore, to find such similar video moments across different videos and align their visual features, we propose the visual co-occurrence alignment (VCA) method. By mining and aligning the visual similar video moments co-occurred in a batch during the training process, our method is able to learn better visual features for video moment retrieval task. Besides, as a video moment could be coupled with different semantic relevant video moments in different training batches, the proposed method can mine multiple positive video moments during the training process. To this end, we adopt the noise-contrastive estimation (NCE) loss [5, 15] for visual feature representation learning. Different from the triplet loss that only compares the positive example with one negative example during each update, the NCE loss contrasts the positive example with all negative examples in the training batch, which provides a more robust way for learning feature representations [5].

Figure 2 shows the proposed framework, which contains three modules, including a visual-language representation learning module, a temporal proposal generation module and a visual co-occurrence alignment learning module. The visual-language representation learning module adopts bi-directional LSTMs (Bi-LSTMs) for both visual stream and textual stream to learn frame-/word-wise feature representation. On top of the Bi-LSTM in the text stream, an extra single-modal multi-head attention is introduced to enhance the textual representation. Then a cross-modal multi-head attention mechanism is introduced to learn the cross-modal representations. With the learned cross-modal representation, the temporal proposal generation module generates and selects the temporal proposals of maximum agreement with the given text queries. Then the visual co-occurrence alignment learning module utilizes the similarity among sentences to mine the positive pairs of relevant video moments from different videos as well as negatives pairs for contrastive learning.

The main contributions of this paper are summarized as follows:

- The proposed Visual Co-occurrence Alignment method alleviates the semantic gap problem, learning from both cross-modal supervision and alignment information within the visual embedding space.
- We conduct extensive experiments on two datasets: Charades-STA and ActivityNet Caption and the results demonstrate the effectiveness of our method.

2 RELATED WORK

Cross-modal Video Retrieval. The target of cross-modal learning methods is to learn a common representation space, where the similarity between samples from different modalities can be measured directly. Cross-modal video retrieval takes the text as the query to retrieve relevant videos from a set of video candidates, the challenge is how to pull the embedding of the relevant video closer and repel the others. Video moment retrieval selects the most relevant part of a video given the text, which focuses on dissecting video frames and seeking subtle differences that pull some consequent frames of the relevant video much closer to the text. Variety methods for cross-modal retrieval have been proposed to learn the common representation space. MCN [2] model aggregates features over the video moment by mean pooling and compares them to the feature of the text query to minimize the distance to the text. CTRL [13] model fuses the features from a video moment and the feature of text query using operations like element-wise multiplication and vector concatenation, followed by separated fully connected layers for scoring and location regression. These aggression methods are efficient but have limited exploration of the correlation between the two modalities. Some works focus on leveraging one modal to guide the feature generated from one another. A dynamic filter is adopted in [37] to generate temporal attention over the video for regression of the starting and ending points. MAN [51] encodes the video features with dynamic filtered language features, and constructs an iterative graph adjustment network for learning moment-wise relations. ROLE [29] adopts a word attention network relied on the video moment context, which concentrates on these important words for cross-modal processing. Method proposed in LoGAN [41] performs cross-modal interaction from both side.

Weakly-supervised Video Moment Retrieval. A variety of approaches has been proposed to train models in a weakly-supervised setting, mainly consisting of three parts: 1) cross-modal feature learning, 2) proposal generation, and 3) optimization.

To learn the textual representation, sequence learning methods such as GRU [7] and LSTM [20] are cooperated. Features from videos are learned interactively within modal and cross modal. [40, 53] represent all video moments as a moments map and apply convolutions on it for moments-level interaction. Cross-modal interactions includes single-hop interaction [27, 53] and multi-hop interactions [30, 41].

The learning of sliding window is a typical way for generating candidate video moments [27, 30, 33, 40, 41], which preset a set of candidate video moments for retrieval. Moments boundaries can be obtained by distinguishing the described interval from the background region [28].

Optimization methods are roughly divided into three categories: 1) cross-modal similarity comparing, 2) foreground/background score ranking, and 3) sentence reconstruction. The cross-modal similarity comparing methods are inspired by the success of video retrieval methods, the joint video-text embedding is learned with ranking loss that pulls the embedding of the relevant video closer to the query text and repels the others. Instead of aligning the whole video with the text, TGA [33] model obtains a probability distribution over candidate video moments aligning with the text, and all the video moments are aggregated by their probabilities as a video-level representation for cross-modal similarity learning. LoGan [41] model compares the similarity between frame embedding within the video moment and frame-specific sentence embeddings, which is optimized by the margin-based ranking loss that penalizes both sampled negative videos and sampled negative sentences. The foreground/background score ranking methods have been adopted for weakly-supervised temporal action localization [24, 28], which contains two branches to separate the action localization from the action classification. A recent weakly-supervised moment retrieval work adopts the two-branch architecture and treats the background frames as negative samples [53]. Sentence reconstruction methods utilize the proposed video moments to generate the query sentence [40] or predict masked words in the query sentence [27]. In this paper, we propose a novel training scheme for weakly-supervised video moment retrieval, aligning the co-occurrence moments from two different videos accompanied with similar query sentences. Moments from different videos are directly compared to each other in the visual world, alleviating the problem of inconsistent semantics between visual data and textual data. Different from [12], which maximizing the similarity of the foreground of video pairs while minimizing the background, and only videos of similar text descriptions are coupled for learning, we regard all other videos within the mini-batch as negatives. Our model does not consider the activities in the background clips as the only negative but take all other irrelevant activities as negatives, which is a more robust way to learn the visual feature representation.

Self-supervised Representation. Self-supervised learning is a feature representation learning method, where supervision is created out of signals such as clip orders [23, 46], motion [8, 17], predicting the future [16], or temporal coherence [22, 45]. TCC [10] learns video representations by matching frames to their nearest neighbors in another video of the same category. Other than learning from the videos only, multi-modal self-supervised representation focus on learning the feature representation through the alignment of different modalities, for instance, contrastive loss [5, 18, 19] is adopted to learn the correspondence between video and narrations [32], or between frames and audio [1, 34]. We also make use of contrastive loss as an objective for weakly-supervised video moment retrieval.

3 METHODS

Given a pair of video and associated text query (V_i, T_i), where $V \in \mathcal{V}$ stands for a video and $T \in \mathcal{T}$ for a text query, the model targets to find the most relevant video moments $V_i^{[st,ed]}$, where st, ed are indices of start frame and end frame respectively. Our proposed model consists of three modules, including a visual-language representation learning module, a temporal proposal generation module and a visual co-occurrence alignment learning module. We detail our model in the following.

3.1 Visual-language Representation Module

We first extract the video features using a pre-trained feature extractor such as C3D [43] and then apply a fully connected layer on the feature vector to reduce the feature dimension. For the query sentence, the word features are represented with a pre-trained Glove embedding [35]. The word features are also encoded with a fully connected layer to obtain the same feature dimension as the video features. To model the features with contextual information, we feed both the frame feature sequence and the word feature sequence into a Bi-LSTM [20].

To generate more discriminate feature embeddings, we adopt multi-head attention function [44] for learning cross-modal feature correlation. The attention function maps a query $Q(\cdot)$ and a set of key-value pairs $(K(\cdot), V(\cdot))$ to an output, where the query, keys, values, and output are all vectors. The output is a weighted sum of the values, where the weight is computed by a similarity function between the query and the corresponding key. For the word sequence, we apply a single-modal multi-head attention module to model the inter-modal correlation among words:

$$Attn(Q(T_i), K(T_i), V(T_i)) = softmax(\frac{Q(T_i)K(T_i)^T}{\sqrt{d_k}})V(T_i), \quad (1)$$

where $Q(\cdot), K(\cdot), V(\cdot)$ are three linear projections of different parameters, d_k is the feature dimension of $K(\cdot)$. For the video sequence, we apply a cross-modal multi-head attention module to model both the correlation among features across modals. The cross-modal multi-head attention explores the compatibility of frames to the text. The frame sequence is regarded as the query and the word sequence makes up the key-value pairs:

$$Attn(Q(V_i), K(T_i), V(T_i)) = softmax(\frac{Q(V_i)K(T_i)^T}{\sqrt{d_k}})V(T_i). \quad (2)$$

Multiple outputs $Attn(\cdot)$ learned with different parameters are concatenated as multi-head attention. We adopt standard configuration for both multi-head attention and feed-forward network, where residual connection and layer normalization are added.

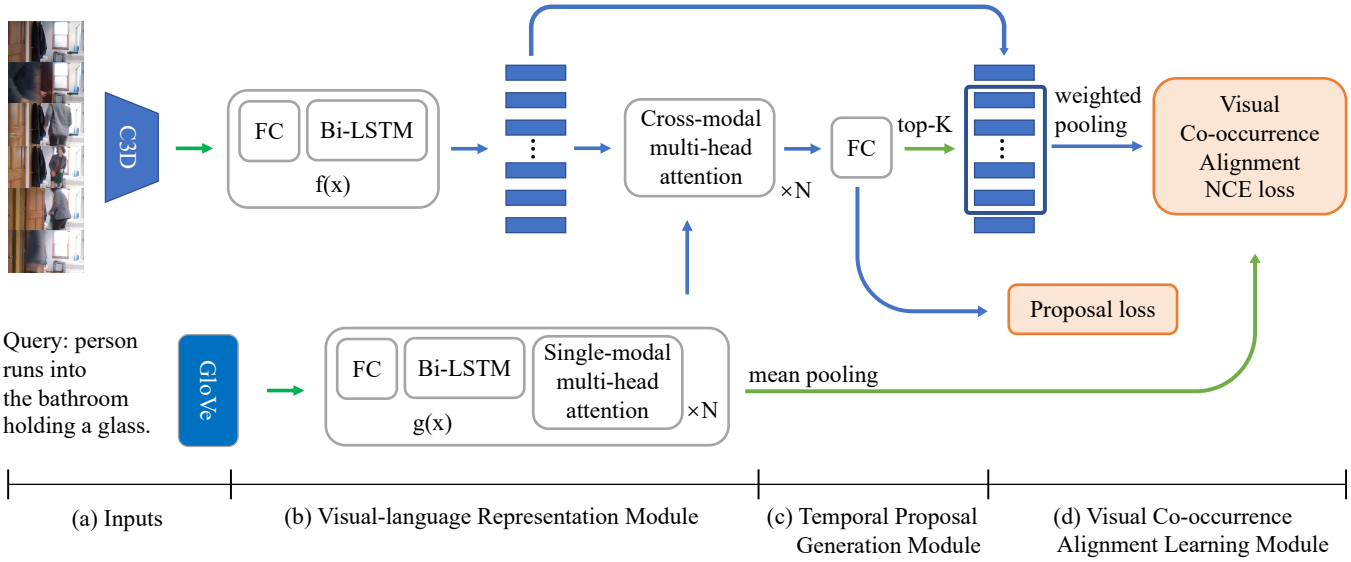

Figure 2: The network architecture of the proposed Visual Semantic Co-occurrence Alignment Learning method for weakly-supervised video moment retrieval. The arrows indicate the data flowing between modules, where gradients from the loss functions can be backpropagated through blue arrows but not through the green arrows.

3.2 Temporal Proposal Generation Module

Similar to one-stage detection methods [36] for object detection, both the candidate sliding window and the corresponding confidence score are generated in a single pass for video moment retrieval. To generate video moments proposals, each frame predicts N_c temporal windows as well as confidence scores for those windows. These confident scores reflect how confident the model is about whether the candidate window contains the activities described by the query sentence. If no activity happened within the window, the confidence score should be close to zero. If the described activity has a large intersection over union (IoU) with the window, the confidence score of the proposed window should be higher than that of other candidate windows.

Each frame at time step responses for predicting confident scores of several pre-defined temporal windows $C^t = (t, t + c^{n_t} * L_V)_{n_t=1}^{N_c}$, where $t, t + c^t * L_V)$ represent the start and end frame index of the candidate temporal window at time step t, $c^{n_t} \in (0, 1)$ is the n_t-th window size relative to the video of L_V frames. c^{n_t} is fixed for each time step. Finally the confidence scores at all time steps are predicted by a fully connected layer

$$S(C^t) = \sigma(W_s f_t + b_s), \tag{3}$$

where $S(C^t) \in R^{L_V \times N_c}$ is the confidence scores for N_c temporal windows of the t-th frame, and σ is a sigmoid function that makes the score between 0 and 1.

From all candidate temporal windows $\{C^t\}_{t=1}^{L_V}$, the temporal windows are ranked according to their corresponding confidence scores. We select the top-K moment proposals as $C_{top} = \{C_{top}^k\}_{k=1}^K$ of confidence scores $S_{top} = \{S_{top}^k\}_{k=1}^K$, where C_{top}^k is the video moments of the k-th proposal of confidence score S_{top}^k. To obtain diverse candidates for training, we use non-maximum suppression to eliminate proposals having high IoU with the chosen proposal. To cooperate with our training scheme, we alternately choose a moment C_{top}^k from the top-K proposals or a random candidate C_{rand} of confidence score S_{rand} with a sampling possibility p as [27], which is given by

$$p = \lambda_1 * \exp(-n_{update}/\lambda_2), \tag{4}$$

where λ_1, λ_2 are the hyper-parameters to control the increase rate of choose top-k proposal, n_{update} is the times of parameter updates. As the training proceeds, the top-K proposals are selected more frequently.

3.3 Contrastive Learning from Visual Co-occurrence Alignment

In this section, we first review a discriminative self-supervised learning method, namely the noise-contrastive estimation (NCE), which is effective for feature representation learning [15]. Then we introduce a cross-modal NCE where the two instances are generated from different modals (CM-NCE). Finally, we introduce the key idea of visual co-occurrence alignment for mining positive pairs of instance within the visual embedding space (VCA-NCE).

NCE. Given a dataset of N video instances, the objective for self-supervised feature representation learning is to obtain a function $f(\cdot)$ that can generate discriminative video embeddings for following tasks such as action recognition. An augmentation operation $\phi(\cdot)$ is performed to obtain an augmented video $x_i^p = \phi(V_i)$. For a specific video V_i, the positive instances and negative instances are defined as $\mathcal{P}_i = \{x_i^p\}$, and $\mathcal{N}_i = \{x_n), \forall n \neq i\}$, and the NCE loss is:

$$\mathcal{L}_{\text{NCE}} = -\log\left(\frac{e^{f(x_i)^\top f(x_i^p)}}{e^{f(x_i)^\top f(x_i^p)} + \sum\limits_{n \in \mathcal{N}_i} e^{f(x_i)^\top f(x_n)}}\right), \tag{5}$$

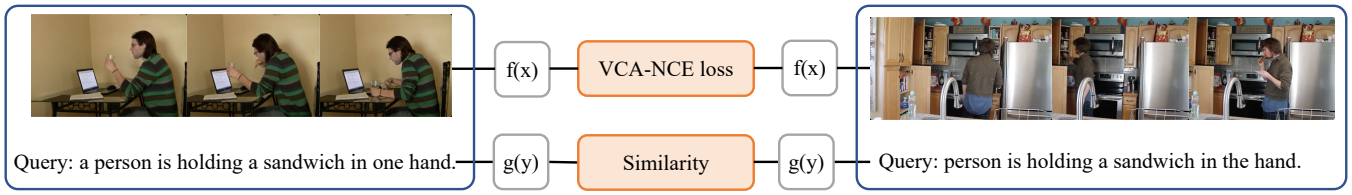

Figure 3: Two video-text pairs of similar textual descriptions. The query sentences are semantically similar between different videos despite the obvious visual difference in the whole video stream. This observation motivates the idea of visual co-occurrence alignment learning. VCA-NCE loss makes it possible to mining hard positives from each other and discriminates the most relevant video moment from other irrelevant video moments.

where the $f(x_i)^\top f(x_i^p)$ denotes the dot product between two video feature embeddings. Assuming the augmented instances are video moments, x_i, x_i^p are different moments at any time steps of any temporal window sizes, that could be relevant to the query text. These moments are represented as moment embeddings aggregated by a weighted pooling over all frame feature embeddings. The weighted pooling consists of a fully-connected outputs a single dimension vector, which is normalized by a softmax layer and is used as the weight for pooling. Reminding that our target is to select the video moments relevant to the query text, maximizing the similarity between random pairs of video moments of one video only pull the feature embeddings of video moments at different time steps closer to each other and repeal the moments from other videos, and the model isn't capable of generating feature representation that can discriminate the relevant moments from the other moments within the same video.

CM-NCE. Cross-model NCE method jointly learns the video representation and the text representation in a common representation space where relevant videos and texts are close to each other and far away otherwise. For moment retrieval tasks, given a set of N pairs of videos and associated text queries, the goal is to learn a visual feature representation function $f(\cdot)$ that maximizing the similarity between the most relevant video moment and the query text and minimizing the rest. For a specific video V_i and the paired text T_i, the only positive text instance for video x_i is y_i, which is represented by $g(y_i)$. Negative text instances are random text queries $\mathcal{N}_i = \{y_n), \forall n \neq i\}$, the CM-NCE loss is:

$$\mathcal{L}_{\text{CM-NCE}} = -\log\left(\frac{e^{f(x_i)^\top g(y_i)}}{e^{f(x_i)^\top g(y_i)} + \sum\limits_{n\in\mathcal{N}_i} e^{f(x_i)^\top g(y_n)}}\right), \tag{6}$$

where the $f(x_i)^\top g(y_p)$ denotes the dot product between the moment feature embedding and the text feature embedding. Assuming x_i is a moment proposed by the temporal proposal generation module, the aggregated moment embedding is compared with the text embedding. The frame-level feature representations of the selected moment are learned according to the compatibility between the moment embedding and the text embedding. The relevant frames will be encoded closer to the text embedding and it will be easier for the temporal proposal generation module to score these frames with higher score.

Note that MIL-NCE loss [32] introduced for cross-modal contrastive learning can be adopted for weakly-supervised moment retrieval as an untrimmed video typically contains multiple descriptions, but we are not going to discuss it here and we leave this for further exploration.

VCA-NCE. We now detailed the visual co-occurrence alignment method, and describe how to employ it for contrastive learning. Given two random pairs of video-text instances $\{(V_i, T_i), (V_j, T_j)\}$, we measure the similarity between their queries with a dot product of their text feature embedding. A large similarity indicates both the queries are describing analogous activities in two different videos, that is to say similar activities are co-occurring in the two videos (as shown in Figure 3). This observation enable the model to learn the alignment between two semantically relevant moments from two different videos. The VCA-NCE loss is:

$$\mathcal{L}_{\text{VCA-NCE}} = -\log\left(\frac{e^{f(x_i)^\top f(x_j)}}{e^{f(x_i)^\top f(x_j)} + \sum\limits_{n\in\mathcal{N}_i} e^{f(x_i)^\top f(x_n)}}\right), \tag{7}$$

where the $f(x_i)^\top f(x_j)$ denotes the visual similarity between the selected moment x_i and a positive video moment x_j, chosen from the videos that have the most similar queries to T_i:

$$\mathcal{P}_i = \{x_j | j \in \text{topK}(g(y_i) \cdot g(y_j)), \forall j \in [1, N], j \neq i\}. \tag{8}$$

Here, the $g(\cdot)$ is a mean pooled feature embedding of all words, $g(y_i) \cdot g(y_j)$ refers to the similarity between the i-th and j-th text query, and the topK$(\cdot)$ operator selects the indexes of topK instances over all available N samples. K is a dynamic parameter determined by a fixed similarity threshold. The negative set $\mathcal{N}_i$ contains moments selected from all videos other than the i-th and j-th video. It's worth mention that the supervision signal is directly back propagated to the Bi-LSTM, and it can be considered as a residual connection between the VCA-NCE loss and the Bi-LSTM that is to be optimized for learning more discriminative visual feature representation.

Comparing to these cross-modal alignment methods, there are two advantages for learning visual feature representation with the proposed method. First, because the most similar query is found within mini-batches, we can construct plenty of positive instances for learning. Second, the positive instances are text in these cross-modal alignment methods, causing the problem of semantic misalignment between different modals. While we directly select positive instances in the visual embedding space, making the embeddings more comparable.

Training with VCA-NCE. In this section, we describe the training scheme for weakly-supervised video moment retrieval. As shown in

Figure 2, The overall loss consist of a proposal loss and a VCA-NCE loss averaging over all N samples:

$$\mathcal{L}_{\text{overall}} = \frac{1}{N}\sum_{n=1}^{N}(\alpha\mathcal{L}_{\text{proposal}} + \mathcal{L}_{\text{VCA-NCE}}), \tag{9}$$

where $\mathcal{L}_{\text{VCA-NCE}}$ is given in Equation 7, aiming at learning discriminate feature representation for the temporal proposal generation module. The α term is a loss balancing hyper-parameter. The $\mathcal{L}_{\text{proposal}}$ is computed from the score of the chosen moment:

$$\mathcal{L}_{\text{proposal}} = \begin{cases} S^k_{top} & p_s \geq p \\ S_{rand} & p_s < p \end{cases}, \tag{10}$$

where $p_s \sim U(0,1)$ and p is defined in Equation 4, which maximizes the confidence score when a relevant video moment is proposed.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metric

We perform experiments on two public datasets to validate the effectiveness of our visual co-occurrence alignment learning method for weakly-supervised video moment retrieval.

Charades-STA. The Charades [39] dataset is originally introduced for in-door activity understanding, and it contains only temporal action annotation and multiple video-level descriptions for each video. The Charades-STA [13] dataset is modified from the Charades dataset for moment retrieval. Its temporal boundary annotations for sentences are generated using a semi-automatic way. The Charades-STA dataset comprises 12,408 pairs of descriptions and video moments for training, and 3,720 for testing. The average length of query sentences is 8.6 words, and the average duration of videos is 29.8 seconds.

ActivityNet Captions. The ActivityNet Captions [21] is a large-scale dataset for human activity understanding. This dataset contains 19,209 diverse untrimmed videos. Following the standard split, the Activitynet Captions dataset consists of 37,417 pairs of descriptions and video moments for training, and 17,031 for testing. The average duration of the video is about 2 minutes. Each video is described by multiple sentences with temporal annotations, and the average length of query sentences is 13.16 words.

Evaluation Metrics. To compare the performance of our Visual Co-occurrence Alignment Learning method with other recently proposed methods, we adopt the evaluation metric "Rank@k,IoU=γ" proposed by [13], and mIoU (*i.e.* the mean IoU) as well. If the temporal IoU between the proposal and the ground truth is bigger than γ, the query is counted as matched. R@k,IoU=γ calculate the percentage of matched queries over the total queries given k and γ. The mIoU is the mean IoU between the top-1 proposal and the ground truth.

4.2 Implementation Details

Data Pre-procession. The video feature is pre-extracted with C3D network [43] for fair comparisons. The C3D network has a temporal resolution of 16 frames and we set the temporal stride as 8 frames for frame feature extraction. The frame embedding is extracted from the second fully connected layer (fc7) of the network and is a 4096-dimension vector. The maximum length of the frame feature embedding is set to 200. The text query is tokenized to word sequence by NLTK and the word embeddings are initialized using the pre-trained Glove vector [35]. The word embeddings are of size 300. The maximum length of the words is set to 20. We keep the most common 8,000 vocabularies for ActivityNet Caption and all 1,016 vocabularies for Charades-STA.

Model Setting. At each time step of video, N_c temporal windows are pre-defined. We set N_c to 6 with ratios of [0.15, 0.3, 0.45, 0.6 0.75, 1] for ActivityNet Caption, and 5 with ratios of [0.1, 0.2, 0.3, 0.4, 0.5] for Charades-STA. We set the moment sampling hyper-parameters λ_1, λ_2 to 5, 2000. The loss balancing hyper-parameter α is set to 1. Both input features of frames and words are projected to the dimension of the hidden state using fully connected layers. The hidden state of Bi-LSTM, multi-head attention and all the fully-connected layers are set to 256. For both single-modal multi-head attention and cross-modal multi-head attention modules, the number of heads is set to 4 and are consists of 3 layers of multi-head attention. During training, we adopt the Adam optimizer with a learning rate of 0.0004 to optimize the model. After 400 warm-up steps, the learning rate increases to the maximum, then it decreases based on the number of updates [44]. The batch size is set to 512 for Charades-STA and 768 for ActivityNet Caption. The model is trained for 200 epochs for both Charades-STA and ActivityNet Caption.

Table 1: Comparison of feature representation module variants on Charades-STA.

Methods	Rank@1, IoU=		
	0.1	0.3	0.5
VCA w/o Bi-LSTM	44.07	26.08	10.01
VCA w/ Bi-LSTM	59.34	38.82	18.40

4.3 Ablations

To validate the effectiveness of modeling the weakly-supervised video moment retrieval problem as a visual co-occurrence alignment learning problem, we conduct a variety of ablation studies. All ablation studies are performed on the Charades-STA dataset.

Contrastive learning for frame-level feature representations. To validate the effectiveness of learning the frame-level feature representations with our proposed VCA-NCE loss, models with/without Bi-LSTM for the video stream are trained and the results are compared in Table 1. As gradients are directly backpropagated from VCA-NCE to the visual representation module, the Bi-LSTM responses for learning visual supervisions. Removing the Bi-LSTM, the visual feature representation function $f(x)$ is simply a fully connected layer reducing the dimension of the pre-extracted frame feature embedding to the hidden size. The results demonstrate that learning discriminative frame-level representations benefits the video moment retrieval task a lot.

Discriminative supervision with intra-modal alignment instead of cross-modal alignment. The CM-NCE (Equation 6) maximize the similarity between positive moment-text pairs, and the VCA-NCE (Equation 7) maximize the similarity between pairs of semantic relevant moments from different videos. The performance of CM-NCE and VCA-NCE is compared in Table 2, which indicates that VCA-NCE achieves higher performance than CM-NCE.

Table 2: Performance of model trained with different loss functions on Charades-STA.

Loss Function	Rank@1, IoU=		
	0.1	0.3	0.5
CM-NCE	44.24	25.81	10.13
VCA-NCE	59.34	38.82	18.40

Table 3: Performance of VCA under different batch size on Charades-STA.

Batch Size	Rank@1, IoU=		
	0.1	0.3	0.5
64	43.54	24.86	9.50
128	46.80	27.17	11.24
256	57.38	34.64	13.74
512	59.34	38.82	18.40

We suppose it's the network design that causes the performance gap, considering that augmented instances are encoded with the same representation function for NCE [15]. In detail, the visual embeddings and textual embeddings are encoded with different parameters for CM-NCE, while the moment embeddings share the same parameters for VCA-NCE. It makes the model much easier to learn with the given weak supervision.

Learning more similar moment pairs with larger batch size. Table 3 shows the impact of batch size when models are trained with VCA-NCE. We find that larger batch size has a significant advantage comparing to the smaller ones. During training, the moment associated with a text query is more like to be coupled with a moment described by a similar text as a result of batch size increasing. A larger batch size makes it easier to find similar text. As the training proceeds, co-occurred moments from videos with the most similar queries will be aligned to each other. Another advantage of using a larger batch size is that it provides more negative samples, facilitating the convergence of the model[5].

4.4 Comparison with Other Methods

We compare the proposed Visual Co-occurrence Alignment (VCA) learning method with fully-supervised methods and other state-of-the-art weakly-supervised methods, which are listed as follows.

Fully-supervised Methods. VSA-STV [13] projects the visual embeddings and textual embeddings to a common space and learns the alignments based on cosine similarity. CTRL [13] predicts an alignment score for the fused cross-modal embedding and regresses the temporal boundary with an adjustment. QSPN [47] performance multilevel visual-language interactions and introduces a captioning module as a side task. 2D-TAN [52] devises a 2D moment map modeling the temporal relations between nearby moments. MAN [51] considers the relations between proposed moments as a structured graph, which is updated with an iterative algorithm. ABLR [50] encodes both video and text stream with Bi-LSTM and applies cross-modal co-attention interactions for boundary regression.

Weakly-supervised Methods. TGA[33] represents the video as a weighted set of moments and aligns it with text in a joint embedding space for moment retrieval. SCN[27] takes the top-K proposed moments to predict the masked words in the query text. RTBPN[53] devises a language-aware filter to generate an enhanced and a suppressed video stream for modeling intra-sample confrontment. LoGan[41] introduces a two-stage cross-modal interaction mechanism for all frames and words. WSRA[12] leverages multiple queries accompanied with the single video for intra-video discrimination. WSLLN[14] simultaneously apply the alignment and detection module for learning pseudo labels. VLANet[30] groups moments according to their similarity with the query and reduces the search space. CTF[6] proposes a coarse-to-fine model to refine the prediction. MARN[40]. WS-DEC [9] alternates between dense caption generation and moment retrieval to explore the one-to-one alignment between multiple moments and captions.

Table 4: Performance Comparison on the Charades-STA Dataset.

Methods	Rank@1, Iou=			Rank@5, IoU=			mIoU
	0.3	0.5	0.7	0.3	0.5	0.7	
fully-supervised methods							
VSA-STV [13]	-	16.91	5.81	-	53.89	23.58	-
CTRL [13]	-	23.63	8.89	-	58.92	29.52	-
QSPN [47]	54.70	35.60	15.80	95.60	71.80	38.87	-
MAN [51]	-	46.53	22.72	-	86.23	33.09	-
2D-TAN [52]	-	39.81	23.25	-	79.33	52.15	-
weakly-supervised methods							
TGA [33]	32.14	19.94	8.84	86.58	65.52	33.51	-
SCN [27]	42.96	23.58	9.97	95.56	71.80	38.87	-
CTF [6]	39.80	27.30	12.90	-	-	-	27.30
WSRA [12]	50.13	31.20	11.01	86.75	70.50	39.02	31.00
VLANet [30]	45.24	31.83	14.17	95.72	**82.82**	33.33	-
MARN [40]	48.55	31.94	14.81	90.7	70.00	37.40	-
RTBPN [53]	**60.04**	32.26	13.24	97.48	71.85	41.18	-
LoGan [41]	51.67	34.68	14.54	92.74	74.30	**39.11**	-
VCA (Ours)	58.58	**38.13**	**19.57**	**98.08**	78.75	37.75	**38.49**

Table 5: Performance Comparison on the ActivityNet Caption Dataset.

Methods	Rank@1, IoU=			Rank@5, IoU=			mIoU
	0.1	0.3	0.5	0.1	0.3	0.5	
fully-supervised methods							
QSPN [47]	-	45.30	27.70	-	75.70	59.20	-
TGN [4]	-	43.81	27.93	-	54.56	44.20	-
CTRL [13]	49.10	28.70	14.00	-	58.92	29.52	20.5
ABLR [50]	73.30	55.67	36.79	-	-	-	36.99
2D-TAN [52]	-	59.45	44.51	-	85.53	77.13	-
weakly-supervised methods							
WS-DEC [9]	62.71	41.98	23.34	-	-	-	28.2
WSLLN [14]	**75.40**	42.80	22.70	-	-	-	32.2
CTF [6]	74.20	44.30	23.60	-	-	-	32.2
SCN [27]	71.48	47.23	29.22	90.88	71.45	55.69	-
RTBPN [53]	73.73	49.77	29.63	**93.89**	**79.89**	**60.56**	-
VCA (Ours)	67.96	**50.45**	**31.00**	92.14	71.79	53.83	**33.15**

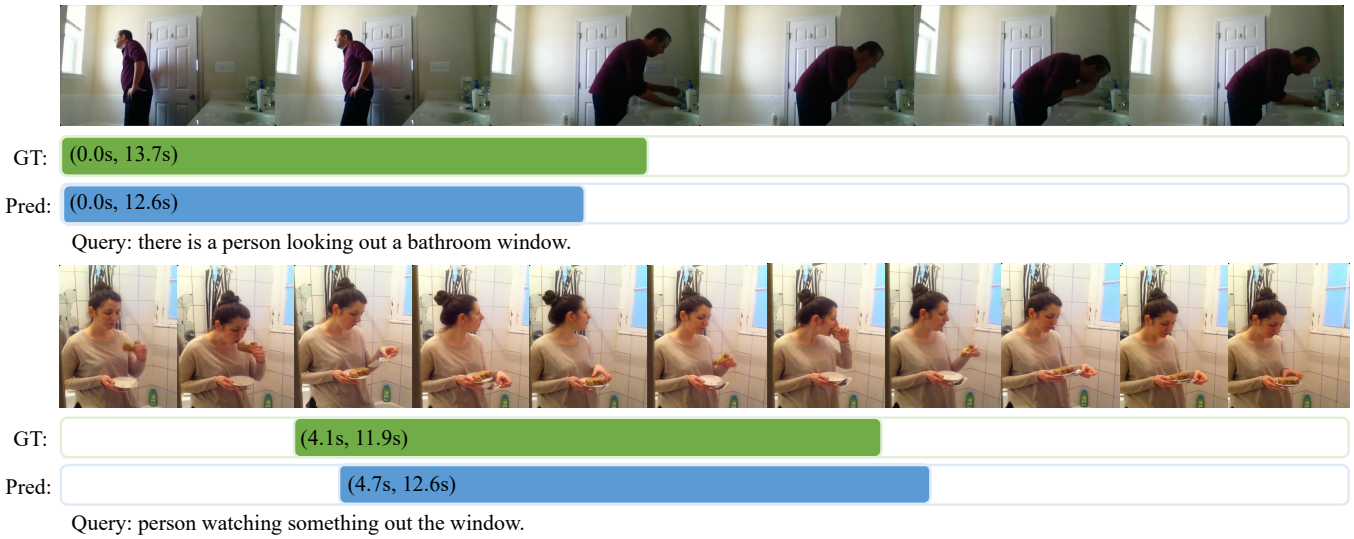

Figure 4: Examples of weakly-supervised video moments retrieval result on Charades-STA.

Results on Charades-STA. Table 4 summarizes the performance comparison results on the Charades-STA dataset. As seen in the results, our approach achieves the best result on all metrics except a slightly worse one, which verifies the effectiveness of our approach. It's worth mention that promising performance is achieved for large IoUs (e.g. 0.5, 0.7), demonstrating the advantage of visual co-occurrence alignment for video moment retrieval when the precise temporal boundary is not available. The results of our approach even outperform some fully supervised methods, indicating the importance of learning both discriminative intra-modal representations and compatible cross-modal representations.

Results on ActivityNet Caption. Table 5 shows the performance comparison results on the ActivityNet Caption dataset. It can be seen that results are inconsistent among different IoUs, methods that achieve high accuracy for small IoUs may obtain low accuracy on other metrics. From the results, our method obtains the highest performance except for Rank@1 IoU=0.1. As the queries in ActivityNet Caption are diverse and complicated, and the textual feature representation is only trained with the proposal loss, a simple mean polling operation can't well represent the meaning of the whole query text. When $g(y)$ fails to model the similarity between queries, these associated moments can't find suitable positive samples for training, which may cause unsatisfactory performance. Another possible reason is that videos in ActivityNet are long, frame-level feature representation can't cover the diverse activity conveyed within a longer moment.

4.5 Qualitative Results

To qualitatively show the performance of our visual co-occurrence alignment learning method for weakly-supervised video moment retrieval, two test samples of similar queries are provided in Figure 4. The green bar is the ground truth temporal boundary of the query sentence, the blue bar is the predicted temporal boundary that has the highest proposal score. Both proposed temporal boundaries are well localized around the ground truth. Both the shown video moments are described by semantic similar text and are localized well by our model regardless of the total time duration difference between them, showing the effectiveness of our proposed visual co-occurrence alignment method.

5 CONCLUSION

In this paper, we study the problem of weakly-supervised video moment retrieval. We propose a visual co-occurrence alignment learning method that cooperates with discriminative intra-modal representations and compatible cross-modal representations learned by the proposed VCA-NCE loss and proposal loss. In particular, the VCA-NCE loss maximizes the similarity between video moments from different videos of similar textual descriptions. The experiments on the Charades-STA dataset and ActivityNet Caption dataset demonstrate the effectiveness of our method.

There are several issues not being resolved in this paper. First, the current supervision of VCA-NCE is only involved in the frame-level feature representation, how to use it for learning moment-level feature representation is worth to be further explored. Second, the temporal proposal generation module is simple yet cannot model the relationship between different moments. Therefore, devise a more reliable temporal proposal generation module is another future direction.

ACKNOWLEDGMENTS

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. This work was supported by National Natural Science Foundation of Project (62072116) and Shanghai Pujiang Program (20PJ1401900).

REFERENCES

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2019. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667* (2019).

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*. 5803–5812.

[3] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video Summarization Using Deep Neural Networks: A Survey. *arXiv preprint arXiv:2101.06072* (2021).

[4] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 162–171.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[6] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. 2020. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308* (2020).

[7] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[8] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. 2019. Dynamonet: Dynamic action and motion network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6192–6201.

[9] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. *arXiv preprint arXiv:1812.03849* (2018).

[10] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2019. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1801–1810.

[11] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*. Springer, 768–784.

[12] Zhiyuan Fang, Shu Kong, Zhe Wang, Charless Fowlkes, and Yezhou Yang. 2020. Weak Supervision and Referring Attention for Temporal-Textual Association Learning. *arXiv preprint arXiv:2006.11747* (2020).

[13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*. 5267–5275.

[14] Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. 2019. Wslln: Weakly supervised natural language localization networks. *arXiv preprint arXiv:1909.00239* (2019).

[15] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.

[16] Tengda Han, Weidi Xie, and Andrew Zisserman. 2019. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[17] Tengda Han, Weidi Xie, and Andrew Zisserman. 2020. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709* (2020).

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

[19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* (2018).

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[21] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*. 706–715.

[22] Zihang Lai and Weidi Xie. 2019. Self-supervised learning for video correspondence flow. *arXiv preprint arXiv:1905.00875* (2019).

[23] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2017. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*. 667–676.

[24] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. 2020. Background modeling via uncertainty estimation for weakly-supervised action localization. *arXiv preprint arXiv:2006.07006* (2020).

[25] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696* (2018).

[26] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 447–463.

[27] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11539–11546.

[28] Daochang Liu, Tingting Jiang, and Yizhou Wang. 2019. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1298–1307.

[29] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*. 843–851.

[30] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. 2020. VLANet: Video-Language Alignment Network for Weakly-Supervised Video Moment Retrieval. In *European Conference on Computer Vision*. Springer, 156–171.

[31] Jingjing Meng, Suchen Wang, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. 2017. Video summarization via multi-view representative selection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1189–1198.

[32] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9879–9889.

[33] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11592–11601.

[34] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. 2020. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298* (2020).

[35] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[37] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2464–2473.

[38] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1049–1058.

[39] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*. Springer, 510–526.

[40] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048* (2020).

[41] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 2021. 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*. 2083–2092.

[42] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4631–4640.

[43] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[45] Xiaolong Wang, Allan Jabri, and Alexei A Efros. 2019. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2566–2576.

[46] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. 2019. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10334–10343.

[47] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9062–9069.

[48] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. [n.d.]. Tree-Augmented Cross-Modal Encoding for Complex-Query Video Retrieval. In *SIGIR, pages=1339–1348, year=2020.*

[49] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. 2020. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *ACM International Conference on Multimedia*. 1939–1947.

[50] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9159–9166.

[51] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1247–1257.

[52] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12870–12877.

[53] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. 2020. Regularized Two-Branch Proposal Networks for Weakly-Supervised Moment Retrieval in Videos. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4098–4106.