World Wide Web (2023) 26:1525–1544
https://doi.org/10.1007/s11280-022-01105-3

Spatiotemporal contrastive modeling for video moment retrieval

Yi Wang[1,2] · Kun Li[1,2] · Guoliang Chen[1,2] · Yan Zhang[1,2] · Dan Guo[1,2] · Meng Wang[1,2]

Received: 7 August 2022 / Revised: 8 September 2022 / Accepted: 19 September 2022 /
Published online: 26 September 2022

Abstract
With the rapid development of social networks, video data has been growing explosively. As one of the important social mediums, spatiotemporal characteristics of videos have attracted considerable attention in recommendation system and video understanding. In this paper, we discuss the video moment retrieval (VMR) task, which locates moments in a video based on different textual queries. Existing methods are of two pipelines: 1) proposal-free approaches are mainly in modifying multi-modal interaction strategy; 2) proposal-based methods are dedicated to designing advanced proposal generation paradigm. Recently, contrastive representation learning has been successfully applied to the field of video understanding. From a new perspective, we propose a new VMR framework, named spatiotemporal contrastive network (STCNet), to learn discriminative boundary features of video grounding by contrast learning. To be specific, we propose a boundary matching sampling module for dense negative sample sampling. The contrast learning can refine the feature representations in the training phase without any additional cost in inference. On three public datasets, Charades-STA, ActivityNet Captions and TACoS, our proposed method performs competitive performance.

Keywords Video moment retrieval · Spatiotemporal modeling · Contrastive learning · Language query · Temporal localization

1 Introduction

With the booming of the Internet, the number of videos on the web is growing at an unprecedented rate. The analysis of video spatiotemporal data allows us to utilize useful information and knowledge in a timely manner, which may further improve the

This article belongs to the Topical Collection: *Special Issue on Spatiotemporal Data Management and Analytics for Recommend* Guest Editors: Shuo Shang, Xiangliang Zhang and Panos Kalnis

✉ Kun Li
kunli.hfut@gmail.com

✉ Guoliang Chen
chenguoliang_hfut@126.com

Extended author information available on the last page of the article

effectiveness, reliability and efficiency of various tasks of video understanding [1–4]. Over the past few years, a lot of work has been conducted for video content recommendation applications with action recognition [5, 6] or action retrieval techniques [7]. Temporal action or event discovery aims to detect or retrieval a potential variation from untrimmed video, yet the variation belongs to a pre-defined action set or a specific query. Recently, with the development of computer vision and natural language processing, video moment retrieval is emerging and becoming a hot topic. Given a textual query, the goal of video moment retrieval is to locate a video segment (with starting and ending timestamps) that corresponds to the semantics of the query. Compared with temporal action detection, the task of video moment retrieval requires the models to understand the video and query in holistically and simultaneously. It has to process multi-modal spatiotemporal data and build cross-modal interaction models efficiently, and it has many challenges and broad application scenarios.

Many existing works [1, 8–10] have been proposed to tackle the task of video moment retrieval, including *proposal-based* and *proposal-free* methods. In video moment retrieval, a proposal is a segment candidate that may correspond to the target ground-truth. In the early years, video moment retrieval was typically treated as a matching problem, some proposal-based approaches tackle it in a propose-then-rank manner. These methods usually generate proposals with pre-defined sliding windows or anchors, then compute the semantic similarity between the query and each proposal. The proposal with the highest score is considered as the query results. Liu et al. [8] proposed attentive model to emphasize the importance of query. However, these proposal-based approaches are sensitive to the set of sliding windows or anchors. In addition, the proposal evaluation requires a lot of memory and computation resources. Inspired by the fast of proposal generation in temporal action detection [11], Yuan et al. [12] proposed a stacked convolution block to build dense proposal with the help of semantic conditioned dynamic modulation.

To address the consumption of enumerating calculation in proposal-based methods, many proposal-free based methods have emerged and developed. Proposal-free methods typically attempt to predict the start and end timestamps directly, without any proposal generation and ranking. The temporal modeling of video moments proposed by Yuan et al. [13] has become increasingly popular in recent years. The main difference from the existing work lies in the design of the multi-modal fusion module. For example, Mun et al. [9] proposed a local-global video-text interaction approach for deeply modelling the semantics of phrases and video clips along the timeline. However, the challenges of spatiotemporal modeling and video context understanding are still to be explored.

In this paper, we still focus on the spatiotemporal modeling and context understanding of videos. As shown in Figure 1a, given a language query, video moment retrieval aims to locate the action corresponding to the query. As shown in Figure 1b, we target to investigate the effectiveness of contrastive learning for video moment retrieval. Given a sentence query as the anchor, we first perform multi-modal interaction, then sample the positive and negative features from the set of multi-modal features. Finally, we apply the contrastive learning to narrow the semantic distance between the positive sample and the anchor, and enlarge the semantic distance between the negative samples and the anchor. As a result, the target temporal features become more distinguishable. The main contributions of our method are as follows:

- We propose a novel framework spatiotemporal contrastive network (STCNet) for video moment retrieval, which aims to enhance the feature representation of target temporal locations.

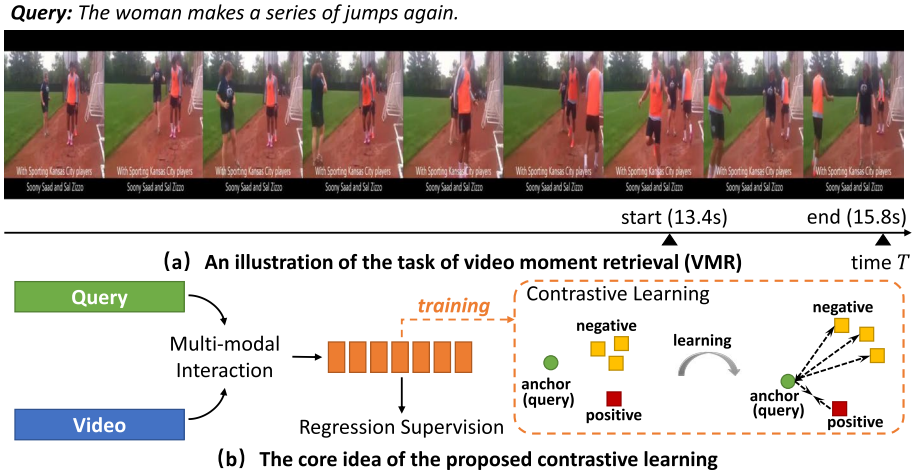

Figure 1 (**a**) An illustration of the video moment retrieval task. Given a video and an language query, the target is to locate a moment, which semantically corresponds to the query. (**b**) The core idea of the proposed contrastive learning

- We propose a Boundary Matching (BM) sampling module for dense negative sample sampling. Given a query, we deem the temporal region of ground-truth as positive and sample the adjacent regions but aligned to query as negative samples. Moreover, we use the Gaussian filter to calculate the sampling mask. We use contrastive learning to refine the discrimination of target temporal features, and it does not need any cost in inference.
- We propose the Local-Global Temporal Context Module (LGTCM) to perform local-global context modeling of multi-modal features. Specifically, we use 1D convolutional layer to model local context, and the non-local network to build the long temporal dependencies of global context.
- Extensive experiments are conducted on three benchmark datasets, Charades-STA, ActivityNet Captions, TACoS, and demonstrate the effectiveness of the proposed method. Ablation studies and qualitative visualizations also verify each component.

2 Related work

In this section, we review the related works about video moment retrieval and focus on the proposal-free methods that our proposed method is mainly compared. Subsequently, we introduce contrastive learning works and how it can potentially be used in the task of video moment retrieval.

2.1 Video moment retrieval

Video moment retrieval (VMR) [7, 14], is also called video grounding [9, 10], which aims to retrieve the temporal moments in the untrimmed video that semantically

correspond to the linguistic query. It plays a crucial role in the field of video understanding [10, 15–17]. The main solutions can be divided into the following two types.

Proposal-based methods Gao et al. [1] first put forward a novel task formulation of temporal activity localization via natural language query. And then this task is evolved into solving the grounding actions and objects by language in videos. Early works scan videos with various sliding windows to generate candidate proposals and compare the proposals with the sentence query, then get the best matched proposal as the optimal result [1, 8, 18]. Based on this manner, the researchers consider the semantic correlation of the video and query is a vital part of the task; thus, they make use of various attention mechanisms to strengthen the interaction learning between the video and sentence query. For example, the moment alignment network (MAN) model [19] explicitly models moment-wise temporal relations as a structured graph and designs an iterative reasoning diagram to learn the relationship among candidate proposals. The Temporal GroundNet (TGN) model [20] captures more fine-grained frame-by-word interaction between the video and sentence query and generates the final grounding result by integrating the contextual information of temporal sequence. In addition, a novel 2D Temporal Map [21] has been proposed to describe all possible proposals. It tackles the problem well since it takes time dependence into account rather than considering temporal moments individually. In other words, the 2D Temporal Map has the ability to anticipate the score prediction for all possible proposals.

Proposal-free methods However, as we all know, the above-mentioned proposal-based methods are restricted by computational expense and limited efficiency. Yuan et al. [13] first propose a proposal-free approach to solve the VMR problem. This method is performed based on direct temporal boundary regression by achieving multi-modal interactions through simple "splicing", which solves the temporal sentence localization problem from a global perspective through an attention-based location regression (ABLR) approach.

To eliminate the imbalance of training samples in boundary regression, Lu et al. [22] achieve dense positive samples by predicting offsets from the ground-truth moment center. Another direction of the proposal-free methods is the boundary probability regression, which aims to predict the probability curves of start and end positions over each. To consider that the variation between consecutive video frames is small and the words in query may have different meanings between adjacent ones, the video span localizing network (VSLNet) model [23] uses contextual query attention to perform fine-grained multi-modal interactions and performs two conditional span predictors to predict the start and end boundaries of answer spans.

In addition to the above two basic proposal-free frameworks, more and more improvements in task can be achieved by modifying the multi-modal interaction module or introducing other advanced semantic understanding modules. For example, Gao et al. [7] replace the cross-modal interaction module with a cross-modal common space to achieve fast video moment retrieval. Adversarial bi-directional interaction network (ABIN) [24] designs an auxiliary adversarial discriminator network to generate coordinates and frame-dependent distributions for moment boundary refinement. By comparison, our work aims to enhance the discrimination of target temporal features in moment retrieval by using contrastive learning. As a result, the model will locate the target segment more accurately.

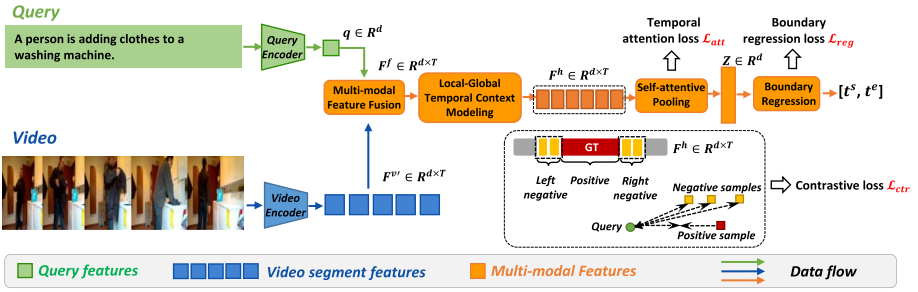

Figure 2 Overview of the proposed framework, which consists of feature encoders, multi-modal feature fusion, local-global temporal context modeling, contrastive learning and self-attentive regression

2.2 Contrastive learning

Contrastive learning has been successfully applied to unsupervised representation learning tasks in recent years. Contrastive learning gives the backbone network the ability to distinguish relevant and irrelevant samples by maximizing the difference between positive samples (the pair of samples in the input modality and the corresponding sample in the target modality) and negative samples (randomly selected samples in the input modality and the target modality). Most of the current contrastive learning methods are conducted based on the paradigm of self-supervised feature learning. Lorre et al. [25] propose a self-supervised video representation learning method based on Contrast Predictive Coding (CPC), which learns the long-term relationships behind the original signal sequence and predicts the potential representations of future clips in the video. He et al. [26] propose Momentum Contrast (MoCo), arguing that using more negative samples is what makes the contrast learning method work better. This is because more negative samples are needed to effectively cover the underlying data distribution, and MoCo trains the model by increasing the proportion of negative samples, surpassing even supervised target detection methods based on ImageNet initialization in the field of visual representation learning. Most of the current video comparison learning methods have similar loss functions. The difference is how to make positive and negative pairs. Sun et al. [27] propose Contrastive Bidirectional Transformer (CBT), which uses the video clip and its masked version as a positive pair. Han et al. [28] propose dense predictive coding (DPC), which uses the predicted autoregressive features and the ground-truth features at the same spatio-temporal location as positive pairs.

3 Proposed method

The overall architecture of the STCNet is shown in Figure 2. The STCNet consists of feature encoding, multi-modal fusion, temporal learning and attentive regression. In this section, we first describe the problem of video grounding in Section 3.1. Then, we use two feature encoders to obtain the video and text features in Section 3.2. Next, the outputs of the feature encoders are fused into the contextual semantics by Local-Global Context Modeling in Section 3.3. The core contrastive learning is then introduced in Section 3.4. Followed by a regular regression module in Section 3.5. The obtained multi-modal features

are turned into vectors to predict the query corresponding to the starting and ending times (t^s, t^e). Finally, the loss function used in STCNet is presented in Section 3.6.

3.1 Problem formulation

Given a query sentence, the task of video grounding aims to localize temporal moments with the starting and ending times (t^s,t^e) in the queried videos, which semantically corresponds to the query. We denote the visual features of a video with T segments as $V = \{v_1, v_2, v_3, \dots, v_T\} \in \mathbb{R}^{d_v \times T}$, where v_i is the i −th segment feature with the dimension of d_v. The textual features are denoted as $Q = \{q_1, q_2, q_3, \dots, q_L\} \in \mathbb{R}^{d_q \times L}$, where L is the length of query and d_q is the dimension of textual feature. Give video V and query Q, we aim to learn a deep learning model as $\mathcal{F}$, and the corresponding queried video segment (*i.e.*, starting time t^s, ending time t^e) can be predicted by:

$$(t^s, t^e) = \mathcal{F}(V, Q, \Theta), \tag{1}$$

where Θ is a set of parameters of the model $\mathcal{F}$.

3.2 Feature encoder

This task refers to a multi-modal understanding task, we have to handle the feature encoding of both video and query.

Video Encoder Given an untrimmed video V, we first equidistant sample and extract the segment-level features with a fixed length T by using the pre-trained 3D network, namely $F^v \in \mathbb{R}^{d_v \times T}$. Following the common practice in this field, we use the feature encoder in QANet [29] to further embed the visual features. Specifically, the feature encoder for videos is composed of Positional Encoding (PE), Multi-head Self-attention (MHA), Feed-forward Network (FFN) and LayerNorm (LN) operation. The calculation of the feature encoder are as follows:

$$\begin{aligned} F^{v'} &= \text{VideoEncoder}\,(F^v), F^v \in \mathbb{R}^{d_v \times T}, \\ &= \begin{cases} \widetilde{F^v} = \text{PE}(\text{FC}(F^v)); \\ \widehat{F^v} = \text{LN}(\text{MHA}(\widetilde{F^v}) + \widetilde{F^v}); \\ F^{v'} = \text{LN}(\text{FFN}(\widehat{F^v}) + \widehat{F^v}), \end{cases} \end{aligned} \tag{2}$$

where PE denotes the positional encoding function stated in [30], d denotes the dimension of feature encoding. Up to now, we get the advanced visual feature $F^{v'} \in \mathbb{R}^{d \times T}$.

Query Encoder For a query sentence with L words, we encode it by using GloVe embedding [31] and represent the word-level texture features as $Q = \{q_1, q_2, q_3, \dots, q_L\} \in \mathbb{R}^{d_q \times L}$, where d_q denotes the word embedding dimension. Then, we use a Bi-directional LSTM [32] to encode Q into the sentence-level vector.

$$q = BiLSTM(FC(Q)) \in \mathbb{R}^d, \tag{3}$$

where d denotes the feature dimension.

Multi-modal Feature Fusion To integrate the above multi-modal features of video and query, we perform a segment-level modality fusion by using the Hadamard product. The whole process is summarized as follows:

$$F^f = W_f(W_v F^{v'} \odot W_t q) \in \mathbb{R}^{d\times T}, \tag{4}$$

where $W_v, W_t, W_f \in \mathbb{R}^{d\times d}$ are learnable embedding matrices for multi-modal feature fusion, and $\odot$ is the Hadamard product operator.

3.3 Local-global temporal context modeling

The task of video moment retrieval aims to localize a temporal segment along the temporal dimension. Based on the above obtained feature sequence $F^f \in \mathbb{R}^{d\times T}$, we make efforts to refine it for the final prediction. To be specific, we propose a Local-Global Temporal Context Module (LGTCM). The LGTCM first learns the local context information through 1D convolutional layer, then learns the global context via non-local block [3]. As shown in Figure 3, the local contextual modeling is formulated as follows:

$$\begin{aligned} \widetilde{F^f} &= ResBlock([F^f_1, \cdots, F^f_T]); \\ &= Conv(LN([F^f_1, \cdots, F^f_T])), \end{aligned} \tag{5}$$

where F^f_i denotes the i-th feature in F^f and *ResBlock* is a residual block [9] consisting of two temporal convolution layers in our work. $F^f_1, \cdots$, and F^f_T share the same model parameter of *ResBlock* in this calculation.

Then, the global contextual modeling is formulated as follows:

$$\begin{aligned} F^h &= NLBlock(\widetilde{F^f}); \widetilde{F^f} \in \mathbb{R}^{d\times T}; \\ &= \widetilde{F^f} + (W_{rv}\widetilde{F^f}) softmax(\frac{(W_{rq}\widetilde{F^f})^T (W_{rk}\widetilde{F^f})}{\sqrt{d}})^T, \end{aligned} \tag{6}$$

where $W_{rv}, W_{rq}, W_{rk} \in \mathbb{R}^{d\times d}$ are learnable matrices, and $NLBlock(\cdot)$ denotes the non-local neural networks [3]. Finally, $F^h \in \mathbb{R}^{d\times T}$ is the final feature sequence output by the LGTCM module.

To summarize, we first use 1D convolution to learn the relationship of adjacent moment points (local features) in the fused multi-modal features, and then use multi-head self-attention in non-local block to model the global information features. The parameter N represents that 1D convolution is conducted in an N-layers stack, and the parameter M denotes the LGTCM is stacked M times.

3.4 Spatiotemporal contrastive learning

In this section, we propose a spatiotemporal contrastive learning for the feature enhancement of target temporal location. We first introduce the boundary matching sampling, and then introduce how the contrastive learning can be used in the task of video moment retrieval.

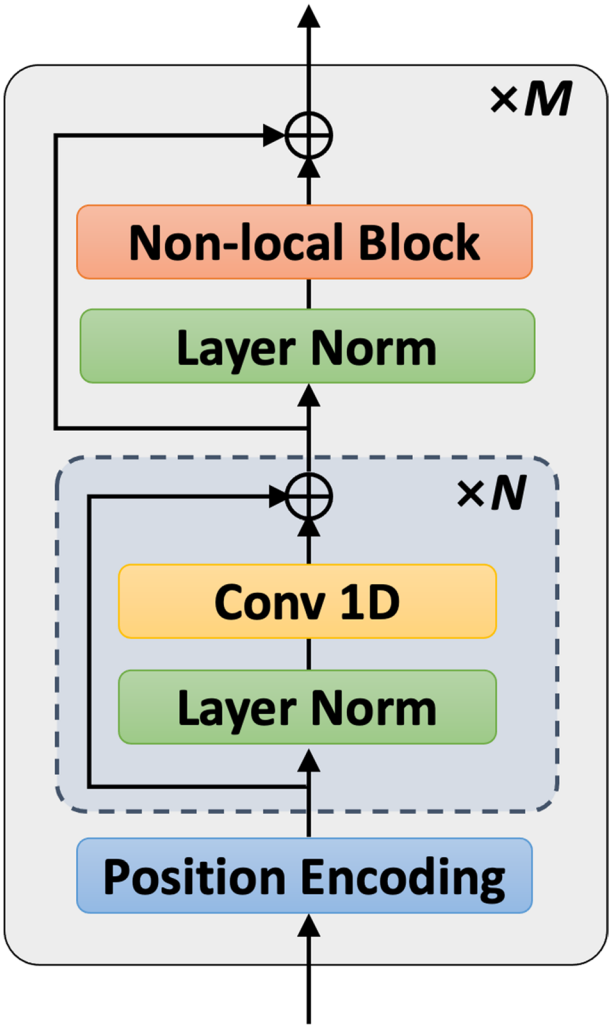

Figure 3 The network architecture of Local-Global Temporal Context Module (LGTCM)

3.4.1 Boundary matching (BM) sampling

Since the multi-modal features F^h contain rich information for target moment prediction, we attempt to impose an effective contrastive restriction on the features to refine them. First, we have to select positive and negative samples. In our work, given a query q as the anchor, we deem the ground-truth temporal region $[t_s^{gt}, t_e^{gt}]$ as the positive sampling range. Here, we mainly discuss the negative sampling strategy.

Inspired by the boundary matching network [33], we enlarge possible candidate proposals around $[t_s^{gt}, t_e^{gt}]$ and propose an Boundary Matching (BM) sampling module for splitting these possible proposals. We give an instance visualization of BM sampling in Figure 4. To be specific, given a query q and its ground-truth temporal region $[t_s^{gt}, t_e^{gt}]$ in Figure 4(a), there is a temporal interval $d_q = t_e^{gt} - t_s^{gt}$. We use a sampling hyperparameter α to set closely similar but negative boundary windows $[t_s^{gt} - \alpha \cdot d_q, t_s^{gt}]$ and $[t_e^{gt}, t_e^{gt} + \alpha \cdot d_q]$ unaligned with query q.

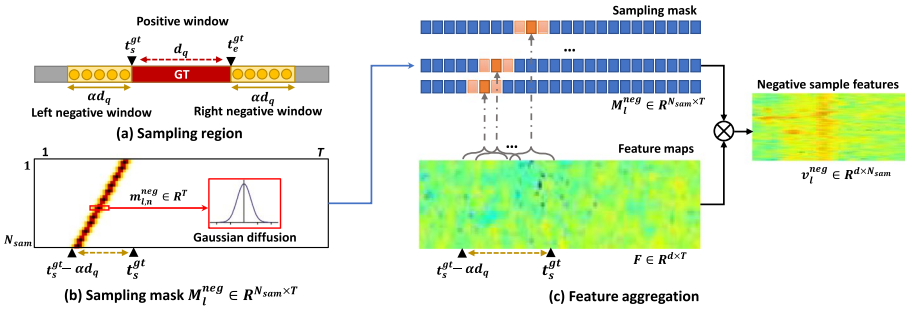

Figure 4 (**a**) Illustration of the positive and negative sampling strategy in contrastive learning. (**b**) Illustration of sampling mask (weight mask) $M_l^{neg} \in \mathbb{R}^{N_{sam} \times T}$. (**c**) Illustration of the Boundary Matching (BM) sampling module

Next, we sample N_{sam} points uniformly in $[t_s^{gt} - \alpha \cdot d_q, t_s^{gt}]$ and do the same sampling operation in $[t_e^{gt}, t_e^{gt} + \alpha \cdot d_q]$ too. Thus, there are $2N_{sam}$ negative sampling points in total.

As shown in Figure 4(b), taking the n-th sampling as example, we get the sampling timestamp t_n and create a mask vector $m_n^{neg} \in \mathbb{R}^T$. We use classical Gaussian filter to diffuse the mask and get the weight mask $m_n^{neg} \in \mathbb{R}^T$, which is formulated as follows:

$$m_n^{neg} = Gaussian(t_n, \sigma) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(1-dec(t))^2}{2\sigma^2}} & ift = floor(t_n) \\ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{dec(t)^2}{2\sigma^2}} & ift = floor(t_n) + 1 \\ \frac{1}{\sigma\sqrt{2\pi}} & ift = others \end{cases}, \tag{7}$$

where σ denotes the standard deviation of Gaussian kernel and its default value is 2.0, $n \in [1, N_{sam}]$. Thus, by the BM negative sampling, we can obtain respective weight mask $M_l^{neg} = [m_{l,1}^{neg}, \cdots, m_{l,N_{sam}}^{neg}] \in \mathbb{R}^{N_{sam} \times T}$ for the left boundary window $[t_s^{gt} - \alpha \cdot d_q, t_s^{gt}]$ and $M_r^{neg} = [m_{r,1}^{neg}, \cdots, m_{r,N_{sam}}^{neg}] \in \mathbb{R}^{N_{sam} \times T}$ for the right boundary window $[t_e^{gt}, t_e^{gt} + \alpha \cdot d_q]$.

3.4.2 Contrastive learning

In this work, we perform the contrastive learning at positive and negative region levels. To facilitate the calculation, we implement the feature aggregation for each positive and negative region.

For the positive sample, given the target segment label $[t_s^{gt}, t_e^{gt}]$, we directly aggregate the multi-modal features of $F^h \in \mathbb{R}^{d \times T}$ in region $[t_s^{gt}, t_e^{gt}]$ by element-wise sum operation. Thus, for each query, we get the feature of sole positive sample $v^{pos} \in \mathbb{R}^d$.

For the negative sample, as shown in Figure 4(c), there are two negative sampling windows $[t_s^{gt} - \alpha \cdot d_q, t_s^{gt}]$ and $[t_e^{gt}, t_e^{gt} + \alpha \cdot d_q]$ that are semantically unaligned with the query q_i. For each window, we sample $2N_{sam}$ times. Thus, we can obtain $2N_{sam}$ negative features. For the left negative sampling window $[t_s^{gt} - \alpha \cdot d_q, t_s^{gt}]$, the negative features are calculated as follows:

$$M_l^{neg} = BMSampling([t_s^{gt} - \alpha \cdot d_q, t_s^{gt}]) \in \mathbb{R}^{N_{sam} \times T} \tag{8a}$$

$$V_l^{neg} = M_l^{neg} \cdot F^h \in \mathbb{R}^{d \times N_{sam}}, \tag{8b}$$

where N_{sam} is the sampling number and α is the control parameter of sampling area.

Same as (8a) and (8b), the right negative features are calculated as follows:

$$M_r^{neg} = BMSampling([t_e^{gt}, t_e^{gt} + \alpha \cdot d_q]) \in \mathbb{R}^{N_{sam} \times T} \tag{9a}$$

$$V_r^{neg} = M_r^{neg} \cdot F^h \in \mathbb{R}^{d \times N_{sam}}, \tag{9b}$$

In a nutshell, we get the sole positive sample and $2N_{sam}$ negative samples, and the loss optimization is introduced in Section 3.5.

3.5 Self-attentive regression

Thanks to the above contrastive sampling paradigm, the multi-modal feature is further restricted and refined. In this part, we leverage a self-attentive regression module on the multi-modal feature F^h to predict the temporal boundaries. Specifically, a two-layer $MLP_{temporal}$ is used to compute the attention weight m of the cross-modal features over the temporal dimension. We use this m to obtain the fusion vector Z. After that, a two-layer MLP_{reg} is used to predict the starting and ending timestamps (t^s,t^e) corresponding to the query. The whole process is formulated as follows:

$$m = Softmax(MLP_{temporal}(F^h)) \in \mathbb{R}^T; \tag{10a}$$

$$Z = \sum_{i=0}^{T} m * F_i^h \in \mathbb{R}^d; \tag{10b}$$

$$(t^s, t^e) = MLP_{reg}(Z) \in \mathbb{R}^2. \tag{10c}$$

3.6 Loss optimization

To optimize the proposed model, we design a multi-task loss $\mathcal{L}$. The total objective function is:

$$\mathcal{L} = \mathcal{L}_{reg} + \mathcal{L}_{att} + \mathcal{L}_{ctr}, \tag{11}$$

where $\mathcal{L}_{reg}$ denotes the regression loss term, which directly evaluates the predicted starting and ending timestamps (t^s, t^e). $\mathcal{L}_{att}$ denotes the attention loss term, which is used to align the self-attention pooling vector m with the location label along the temporal dimension. $\mathcal{L}_{ctr}$ denotes the InfoNCE Loss function which is used for contrastive learning.

Concretely, the loss function of $\mathcal{L}_{reg}$ is calculated as follows:

$$\mathcal{L}_{reg} = SL_1(t_s^{gt} - t_s) + SL_1(t_e^{gt} - t_e), \tag{12}$$

where SL_1 is Smooth L1 loss, (t_s^{gt}, t_e^{gt}) denote the ground-truth starting and ending timestamps, and (t_s,t_e) denote the predicted starting and ending timestamps.

The loss function $\mathcal{L}_{att}$ is formulated as follows:

$$\mathcal{L}_{att} = -\frac{\sum_{i=1}^{T} \hat{m}_l \log(m_i)}{\sum_{i=1}^{T} \hat{m}_l}, \tag{13}$$

where m corresponds to m in (10a). Here, $\hat{m} \in \mathbb{R}^T$ is another representation of the ground-truth label over the temporal dimension with the value of 1 in the temporal interval and 0 otherwise.

The loss function $\mathcal{L}_{ctr}$ is formulated as follows:

$$\mathcal{L}_{ctr} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{s(q_i, v_i^{pos})/\tau}}{e^{s(q_i, v_i^{pos})/\tau} + \sum_{n=1}^{2N_{sam}} e^{s(q_i, v_{i,n}^{neg})/\tau}}, \tag{14}$$

where $s(\cdot,\cdot)$ denotes the calculation of cosine similarity, B denotes batch size, i denotes the query number in the batch and τ means temperature hyperparameter. q_i represents the sentence-level vector of i −th query. For the i −th query, v_i^{pos} represents the aggregated feature of positive samples in the target location interval, $v_{i,n}^{neg}$ denotes the n −th feature of negative samples coming from V_l^{neg} and V_r^{neg}.

4 Experiments

Extensive experiments have been conducted on three benchmark datasets to evaluate the proposed method. And we also test the effectiveness of contrastive learning in this task. In this section, we first introduce the experimental setup, including datasets, evaluation metrics and implementation details. Then, we make comparisons and analysis with state-of-the-art methods. We also present an ablation study to investigate the contribution of each component in the proposed framework.

4.1 Experimental setup

4.1.1 Datasets

Following previous works [1, 12, 13], we experiment on three public benchmark datasets: **Charades-STA** [1], **ActivityNet Captions** [34] and **TACoS** [35]. **Charades-STA** [1] contains 6,672 daily life videos with the duration of 30.59 seconds on average. Each video has around 2.4 annotated moments and the average duration of the moment is 8.2 seconds. The dataset contains 16,1248 query-clip pairs and is split into training and testing parts with 12,408 pairs and 3,720 pairs, respectively.

ActivityNet Captions (abbreviated as **ANet-Captions**) [34] is a benchmark dataset for the task of dense video captions, which is built upon the ActivityNet [36] dataset. ANet-Captions is originally proposed for dense video understanding. Compared with Charades-STA [1], ANet-Captions is more challenging due to its two properties: one is that ANet-Captions contains longer videos, and the other is that the queries are often complicated. The ANet-Captions dataset consists of 20K videos along with 100K language queries. On average, each video is annotated with 2.5 queries. Limited to the unreleased "test" set, in this paper, we adopt the setting of "train" for training, "val 1" for validation, and "val 2" for testing in [37]. Thus, the dataset is split into the training/validation/testing sets of 37,421, 17,505, and 17,031 query-clip pairs.

Table 1 The statistics of three public video grounding datasets. #Anns means denotes the number of query-moment pairs, L_{vid} denotes the average length of videos, L_{query} denotes the average length of query, L_{moment} denotes the average length of queried moment

Dataset	Domain	#Videos	#Anns Train	#Anns Val	#Anns Test	L_{Vid}	L_{Query}	L_{Moment}
Charades-Sta	Indoors	6,672	12,408	–	3,720	30.59S	7 Words	8.22S
Anet-Captions	Daily Life	14,926	37,421	17,505	17,031	117.61S	15 Words	36.18S
Tacos	Cooking	127	10,146	4,589	4,083	287.14S	10 Words	5.45S

TACoS [35] consists of 175 videos that are collected from the cooking room. The duration of each video is 4.79 minutes on average. Each video has 178 queries on average. Compared with Charades-STA and ANet-Captions datasets, TACoS has more dense queries on each video, causing more challenges. The TACoS dataset consists of 10,146, 4,589 and 4,083 query-clip pairs for training, validation, and test, respectively. The detailed statistics of these three datasets are listed in Table 1.

4.1.2 Evaluation metrics

Following previous works [1, 10] on video moment retrieval, we adopt "R@N, IoU@" as the evaluation metrics. The metric **"R@N, IoU@"** [1, 13] calculates the percentage of samples having larger temporal Intersection over Union (tIoU) than threshold in the top-N predicted segments. The higher the value of mIoU, the more accurate the prediction result of the model. Since the proposed method is proposal-free, we report all the results at R@1. We abbreviate it as **"IoU@"** in the following tables. Besides, **"mIoU"** denotes the average IoU for all the test queries. The pre-set thresholds can be set to {0.5,0.7} on Charades-STA and ANet-Captions, and {0.3,0.5} on TACoS.

$$R@N, IoU@\theta = \frac{1}{N} r(\theta, q_i), \tag{15}$$

where q_i is the $i-$th predicted segment, $r(\cdot)$ denotes the tIoU calculation [1], N is the number of predicted segments, and is a pre-set threshold.

4.1.3 Implementation details

For fair comparison, we use the C3D network [38] for ANet-Captions dataset [1], C3D [38] and I3D [5] networks for Charades-STA dataset, and VGG network [39] for TACoS dataset to extract visual features. To facilitate model training, we uniformly sample segments from each video with a fixed length $T = 128$. As for language features, we first transform all words in each query to lowercase and extract the GloVe word embedding [31] with the dimension of 300. Both the visual features and textual features are linearly mapped into 512-*dim* vectors. Finally, about training settings, we use Adam optimizer [40] to optimize the proposed network with a learning rate of $1e-4$. The batch size is set to 100. For contrastive learning, the sample point N_{sam} of BM

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

Table 2 Comparison results with state-of-the-art methods on Charades-STA dataset

Method	Venue	Feature	IoU@		mIoU
			0.7	0.5	
MCN [1]	ICCV'17	C3D	8.01	17.46	–
CTRL [41]	ICCV'17	C3D	8.89	23.63	–
ACRN [8]	SIGIR'18	C3D	7.64	20.26	–
MAC [42]	WACV'19	C3D	12.20	30.48	–
QSPN [18]	AAAI'19	C3D	15.80	35.60	–
ABLR [13]	AAAI'19	C3D	9.01	24.36	–
SAP [47]	AAAI'19	C3D	13.36	27.42	–
R-W-M [48]	AAAI'19	C3D	–	36.70	–
SM-RL [49]	CVPR'19	C3D	11.17	24.36	32.22
CBP [44]	AAAI'20	C3D	18.87	36.80	35.74
GDP [50]	AAAI'20	C3D	18.49	39.47	36.60
TSP-PRL [51]	AAAI'20	C3D	17.69	37.39	37.22
PMI [46]	ECCV'20	C3D	19.27	**39.73**	–
BPNet [45]	AAAI'21	C3D	20.51	38.25	**38.03**
Ours	–	C3D	**20.73**	38.44	37.33
TMLGA [52]	WACV'20	I3D	33.74	52.02	–
DRN [53]	CVPR'20	I3D	31.75	53.09	–
LGI [9]	CVPR'20	I3D	35.48	**59.46**	51.38
BPNet [45]	AAAI'21	I3D	31.64	50.75	46.34
SSMN [54]	TOMM'21	I3D	28.49	51.51	–
Ours	–	I3D	**38.68**	59.09	**52.18**

sampling is set to 32. For LGCTM, the kernel size of 1D convolution layer is set to 15, the parameters N and M are set to 2 and 2, respectively. The temperature parameter τ in contrastive learning loss term is set to $1e-7$. Note that, the contrastive learning is only engaged in the training process, and it performs without any memory consumed in the inference process.

4.2 Comparison with state-of-the-arts

We compare the proposed STCNet with the following state-of-the-art methods: 1) **Proposal-based methods:** CTRL [1], MCN [41], ACRN [8], MAC [42], CMIN [43]; SCDM [12], TGN [20], CBP [44]; 2) **Proposal-free methods:** ABLR [13], LGVTI [9], BPNet [45], CPNet [10], PMI-LOC [46].

The experimental results on Charades-STA [1], ANet-Captions [34] and TACoS [35] datasets are listed in Tables 2, 3, 4, where the best result in each column is highlighted in bold. From Table 2 on the charades-STA dataset, we can see that with C3D features, although the results of our method are not optimal, but the results are competitive with the current state-of-the-art methods, and the same are the results with the I3D features. The results of the ANet-Captions dataset are summarized in Table 3, our proposed method is higher than CPNet 0.22% and 0.38% at R@0.5 and mIoU, respectively. Our mIoU superiority is more obvious than IoU. Table 4 compares the performances on the

Table 3 Performance comparison with state-of-the-art methods on ANet-Captions dataset

Method	Venue	IoU@		mIoU
		0.7	0.5	
MCN [1]	ICCV'17	–	9.58	15.83
CTRL [41]	ICCV'17	–	14.00	20.54
ACRN [8]	SIGIR'18	–	16.17	24.16
TGN [20]	SIGIR'18	11.86	27.93	29.17
QSPN [18]	AAAI'19	13.60	27.70	–
ABLR [13]	AAAI'19	–	36.79	36.99
SCDM [12]	NeurIPS'19	19.86	36.75	–
TMLGA [52]	WACV'20	19.26	33.04	–
CBP [44]	AAAI'20	17.80	35.76	36.85
GDP [50]	AAAI'20	–	39.30	39.80
PMI [46]	ECCV'20	17.83	38.28	–
SSMN [54]	TOMM'21	20.03	35.38	–
CPNet [10]	AAAI'21	21.63	**40.56**	40.65
Ours	–	**21.85**	40.15	**41.03**

Table 4 Comparison results with state-of-the-art models on TACoS dataset

Method	Venue	IoU@		mIoU
		0.5	0.3	
MCN [1]	ICCV'17	5.58	–	–
CTRL [41]	ICCV'17	13.30	19.32	11.98
ACRN [8]	SIGIR'18	14.62	19.52	–
TGN [20]	SIGIR'18	20.21	25.13	17.93
CMIN [43]	SIGIR'19	18.05	24.64	–
ABLR [13]	AAAI'19	9.40	19.50	–
SAP [47]	AAAI'19	18.24	–	–
SM-RL [49]	AAAI'19	15.95	20.15	–
SCDM [12]	NeurIPS'19	21.17	26.11	–
CBP [44]	AAAI'20	24.79	27.31	21.59
GDP [50]	AAAI'20	13.50	24.14	16.18
2D-TAN [21]	AAAI'20	25.32	37.29	–
DRN [53]	CVPR'20	23.17	–	–
VSLNet [23]	ACL'20	24.03	29.61	24.11
ABIN [24]	TMM'21	20.16	23.63	–
BPNet [45]	AAAI'21	20.96	25.96	19.53
DCMH [55]	TIP'21	**25.58**	30.04	–
Ours	–	25.42	**38.84**	**26.25**

TACoS dataset, in which video samples are collected from cooking room. Compared with VSLNet [23], our method achieves the performance improvement 1.39%, 9.23% and 2.14% at R@0.5, R@0.3 and mIoU than it, respectively.

Table 5 Ablation studies of main components in our approach on ANet-Captions dataset

Method	IoU@			mIoU
	0.7	0.5	0.3	
w/o contrastive	18.92	36.39	54.46	37.87
w/o local	18.67	36.25	55.55	38.10
w/o global	19.48	36.50	54.75	37.83
STCNet (Ours)	**21.85**	**40.15**	**59.34**	**41.03**

4.3 Ablation study

4.3.1 Main components of STCNet

In this subsection, we experiment the ablation study of each component of STCNet. We test several variants of our model: 1) STCNet: our complete model based on both local-global context modeling and contrastive learning includes all the loss terms, 2) w/o contrastive: STCNet does not use contrastive learning module and contrastive loss term, 3) w/o local: STCNet does not use local contexts in the local-global context modeling module, 4) w/o global: STCNet does not use global contexts in the local-global context modeling module. The ablation studies are experimented on the ANet-Captions dataset.

The results of the ablation experiments are shown in Table 5. Compared with the full STCNet, the metrics R@0.3, R@0.5, R@0.7, the mIoU of "w/o contrastive" decreases by 4.88%, 3.76%, 2.93%, and 3.16%, respectively. The severe performance degradation happens on "w/o contrastive", which demonstrates the superiority of contrastive learning module in STCNet. Therefore, the contrastive learning of positive and negative samples can better enhance the target temporal representation for answer prediction.

For the ablation study of local context modeling, the performance of "w/o local" is dropped by a large min (*e.g.,* mIoU from 41.03 to 38.10). These results show that it is not enough to merely rely on the global contexts in multi-modal features, but also need to model the fine-grained relationships between multi-modal features to achieve accurate moment retrieval. For the global context modeling, the performance of "w/o global" decreases significantly (*e.g.,* mIoU from 41.03 to 37.83). These results show that the model only relies on local context modeling and will pay more attention to local information, thus ignoring the overall semantics. In a nutshell, the results of "w/o local" and "w/o global" prove the effectiveness of the proposed LGTCM module. This module first uses 1D convolution to learn the relationship between adjacent moment points (local features) in multi-modal features and then uses the multi-head self-attention to model the global context relationship, which can effectively model multi-modal features.

4.3.2 Albation study for contrastive learning

We also analyze the role of two hyperparameters in the contrastive learning module - temperature hyperparameter (τ) and contrastive sampling hyperparameter (α). Tables 6 and 7 show their impacts on the ANet-Captions dataset. Table 6 lists the results of parameter $\tau \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$, from which we can see that as τ continues to increase, the trend of each metric value shows an upward trend. This is due to the fact that τ controls

Table 6 Ablation results of the temperature hyperparameter τ in the contrastive loss on ANet-Captions dataset

Method	IoU@			mIoU
	0.7	0.5	0.3	
τ=0.1	19.50	36.66	55.30	38.18
τ=0.2	18.43	35.48	54.03	37.02
τ=0.4	19.58	36.49	54.47	37.53
τ=0.6	20.31	36.64	55.38	37.75
τ=0.8	21.02	38.84	58.63	39.95
τ=1.0	**21.85**	**40.15**	**59.34**	**41.03**
τ=1.2	20.50	38.19	57.93	39.59

Table 7 Ablation results of the constrastive sampling hyperparameter α on ANet-Captions dataset

Method	IoU@			mIoU
	0.7	0.5	0.3	
α=0.5	19.62	36.95	54.95	38.07
α=0.6	18.70	36.06	54.03	37.31
α=0.7	18.71	36.35	55.68	38.30
α=0.8	19.11	36.35	54.83	37.76
α=0.9	19.22	37.01	55.60	38.19
α=1.0	**21.85**	**40.15**	**59.34**	**41.03**
α=1.1	18.92	36.45	54.63	37.77
α=1.2	17.25	35.21	53.81	36.73

the discrimination capability of model between positive sample and negative samples as stated in [56]. α denotes control parameters of the sampling area and its impact is shown in Table 7. It can be seen that there is not a linear trend, thus we set an optimal empirical value of α for each dataset. The utilization of positive and negative samples around the target temporal interval is achieved in a more reasonable way to further improve the performance of the model.

4.4 Qualitative results

Figure 5 shows two examples selected from the ANet-Captions dataset. Compared with several variants of our method, the full STCNet is more accurate in predicting the starting timestamp (t_s) and the ending time (t_e) in the video under the queried sentence. For example, in example $Q1$, "w/o constrastive" predicts 5.26 seconds and 2.39 seconds earlier than t_s and t_e predicted by STCNet, respectively. In example $Q2$, this case happens 11.05 and 18.17 seconds earlier, respectively. As mentioned previously, our contrastive strategy focuses on the representation learning between similar but different instances, and enlarges the representative difference between non-similar instances. In addition, the performance drops obviously without using local or global context modeling. If the model does not use local or global contexts, it will ignore the adjacent temporal changes and the whole storyline and cannot predict accurate boundaries. For example, the video in example $Q2$ displays a man with a blue shirt, and there are no obvious scene changes. Thus, it is hard for

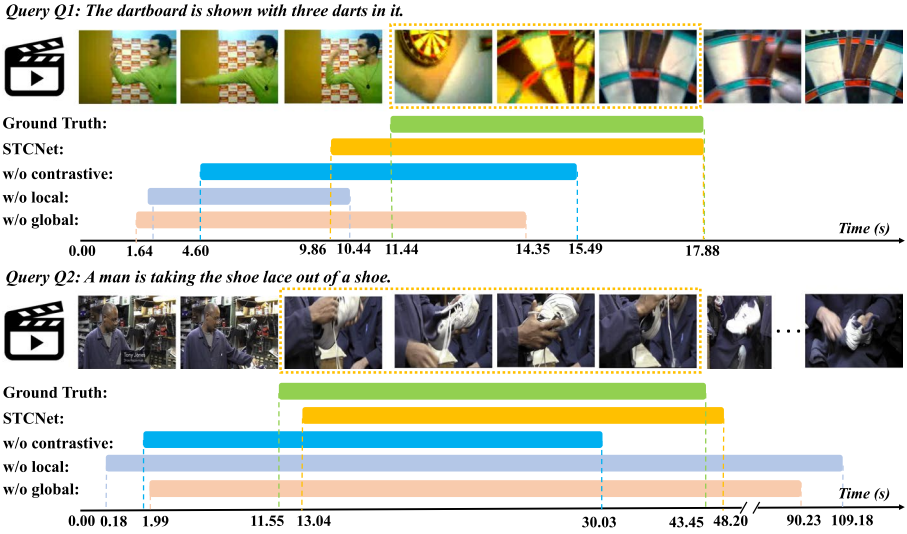

Figure 5 Qualitative results on ANet-Captions dataset

the model to identify where the beginning of queried action is; the model fails to locate the target and only predicts a result that is almost as long as the video.

5 Conclusion

In this paper, we propose a spatiotemporal constrastive learning approach named STCNet for video moment retrieval. Based on the feature encoding and fusion of video and query, we first perform the local-global contextual modeling of multi-modal features, and then use a spatiotemporal contrast learning module to enhance the target temporal feature representation. Experiments on Charades-STA, ActivityNet Captions and TACoS validate the effectiveness of our approach.

Acknowledgements Not applicable

Author Contributions Kun Li and Guoliang Chen designed the proposed method. Yi Wang and Kun Li wrote the main manuscript text. All authors reviewed the manuscript.

Funding This research was supported by the National Natural Science Foundation of China (NSFC) under grants 61876058, 61725203, 62020106007, and U20A20183.

Declarations

Human and Animal Ethics Not applicable

Competing interests The authors declare no competing interests.

Ethics approval and consent to participate Not applicable

Consent for Publication Not applicable

References

1. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5267–5275 (2017)
2. Tian, H., Tao, Y., Pouyanfar, S., Chen, S.-C., Shyu, M.-L.: Multimodal deep representation learning for video classification. World Wide Web **22** (3), 1325–1341 (2019)
3. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
4. Guo, Y., Zhang, J., Gao, L.: Exploiting long-term temporal dynamics for video captioning. World Wide Web **22**(2), 735–749 (2019)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
6. Men, Q., Leung, H., Yang, Y.: Self-feeding frequency estimation and eating action recognition from skeletal representation using kinect. World Wide Web **22**(3), 1343–1358 (2019)
7. Gao, J., Xu, C.: Fast video moment retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1523–1532 (2021)
8. Liu, M., Wang, X., Nie, L., He, X., Chen, B., Chua, T.-S.: Attentive moment retrieval in videos. In: The 41St International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 15–24 (2018)
9. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10810–10819 (2020)
10. Li, K., Guo, D., Wang, M.: Proposal-free video grounding with contextual pyramid network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1902–1910 (2021)
11. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 988–996 (2017)
12. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In: Advances in Neural Information Processing Systems, pp. 536–546 (2019)
13. Yuan, Y., Mei, T., Zhu, W.: 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, pp. 9159–9166 (2019)
14. Wang, W., Gao, J., Yang, X., Xu, C.: Learning coarse-to-fine graph neural networks for video-text retrieval. IEEE Trans. Multimedia **23**, 2386–2397 (2020)
15. Jing, W., Nie, X., Cui, C., Xi, X., Yang, G., Yin, Y.: Global-view hashing: harnessing global relations in near-duplicate video retrieval. World Wide Web **22**(2), 771–789 (2019)
16. Li, X., Zhou, Z., Chen, L., Gao, L.: Residual attention-based lstm for video captioning. World Wide Web **22**(2), 621–636 (2019)
17. Liu, K., Liu, W., Ma, H., Huang, W., Dong, X.: Generalized zero-shot learning for action recognition with web-scale video data. World Wide Web **22**(2), 807–824 (2019)
18. Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9062–9069 (2019)
19. Zhang, D., Dai, X., Wang, X., Wang, Y.-F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1247–1257 (2019)
20. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.-S.: Temporally grounding natural sentence in video. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 162–171 (2018)
21. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12870–12877 (2020)
22. Lu, C., Chen, L., Tan, C., Li, X., Xiao, J.: Debug: a dense bottom-up grounding approach for natural language video localization. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 5144–5153 (2019)
23. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6543–6554 (2020)

24. Zhang, Z., Zhao, Z., Zhang, Z., Lin, Z., Wang, Q., Hong, R.: Temporal textual localization in video via adversarial bi-directional interaction networks. IEEE Trans. Multimedia **23**, 3306–3317 (2020)
25. Lorre, G., Rabarisoa, J., Orcesi, A., Ainouz, S., Canu, S.: Temporal contrastive pretraining for video action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 662–670 (2020)
26. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
27. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Learning video representations using contrastive bidirectional transformer. arXiv:1906.05743 (2019)
28. Han, T., Xie, W., Zisserman, A.: Video representation learning by dense predictive coding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
29. Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., Le, Q. V.: Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv:1804.09541 (2018)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
31. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543 (2014)
32. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
33. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: Bmn: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3889–3898 (2019)
34. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 706–715 (2017)
35. Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M., Schiele, B.: Script data for attribute-based recognition of composite activities. In: European Conference on Computer Vision, pp. 144–157. Springer (2012)
36. Fabian Caba Heilbron, B. G., Escorcia, V., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–970 (2015)
37. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6543–6554 (2020)
38. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
40. Kingma, D. P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2015)
41. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5803–5812 (2017)
42. Ge, R., Gao, J., Chen, K., Nevatia, R.: Mac: Mining activity concepts for language-based temporal localization. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 245–253. IEEE (2019)
43. Zhang, Z., Lin, Z., Zhao, Z., Xiao, Z.: Cross-modal interaction networks for query-based moment retrieval in videos. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 655–664 (2019)
44. Wang, J., Ma, L., Jiang, W.: Temporally grounding language queries in videos by contextual boundary-aware prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12168–12175 (2020)
45. Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., Xiao, J.: Boundary proposal network for two-stage natural language video localization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2986–2994 (2021)

46. Chen, S., Jiang, W., Liu, W., Jiang, Y.-G.: Learning modality interaction for temporal sentence localization and event captioning in videos. In: European Conference on Computer Vision, pp. 333–351. Springer (2020)
47. Chen, S., Jiang, Y.-G.: Semantic proposal for activity localization in videos via sentence query. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8199–8206 (2019)
48. He, D., Zhao, X., Huang, J., Li, F., Liu, X., Wen, S.: Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8393–8400 (2019)
49. Wang, W., Huang, Y., Wang, L.: Language-driven temporal activity localization: a semantic matching reinforcement learning model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 334–343 (2019)
50. Chen, L., Lu, C., Tang, S., Xiao, J., Zhang, D., Tan, C., Li, X.: Rethinking the bottom-up framework for query-based video localization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10551–10558 (2020)
51. Wu, J., Li, G., Liu, S., Lin, L.: Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12386–12393 (2020)
52. Rodriguez, C., Marrese-Taylor, E., Saleh, F. S., Li, H., Gould, S.: Proposal-Free Temporal Moment Localization of a Natural-Language Query in Video Using Guided Attention. In: The IEEE Winter Conference on Applications of Computer Vision, pp. 2464–2473 (2020)
53. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10287–10296 (2020)
54. Liu, X., Nie, X., Teng, J., Lian, L., Yin, Y.: Single-shot semantic matching network for moment localization in videos. ACM Trans. Multimedia Comput. Commun. Appl. **17**(3), 1–14 (2021)
55. Hu, Y., Liu, M., Su, X., Gao, Z., Nie, L.: Video moment localization via deep cross-modal hashing. IEEE Trans. Image Process. **30**, 4667–4677 (2021)
56. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. **2**(7) arXiv:1503.02531 (2015)

Authors and Affiliations

Yi Wang[1,2] · Kun Li[1,2] · Guoliang Chen[1,2] · Yan Zhang[1,2] · Dan Guo[1,2] · Meng Wang[1,2]

Yi Wang
wangyi_2018@mail.hfut.edu.cn

Yan Zhang
yanzhang.hfut@gmail.com

Dan Guo
guodan@hfut.edu.cn

Meng Wang
eric.mengwang@gmail.com

[1] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui 230601, China

[2] School of Artificial Intelligence, Hefei University of Technology, Hefei, Anhui 230601, China