A Survey on Temporal Sentence Grounding in Videos

XIAOHAN LAN, Tsinghua Shenzhen International Graduate School, China
YITIAN YUAN[*], Meituan, China
XIN WANG[†], Tsinghua University, China
ZHI WANG, Tsinghua Shenzhen International Graduate School, China
WENWU ZHU[†], Tsinghua University, China

Temporal sentence grounding in videos (TSGV), which aims to localize one target segment from an untrimmed video with respect to a given sentence query, has drawn increasing attentions in the research community over the past few years. Different from the task of temporal action localization, TSGV is more flexible since it can locate complicated activities via natural languages, without restrictions from predefined action categories. Meanwhile, TSGV is more challenging since it requires both textual and visual understanding for semantic alignment between two modalities (*i.e.*, text and video). In this survey, we give a comprehensive overview for TSGV, which i) summarizes the taxonomy of existing methods, ii) provides a detailed description of the evaluation protocols (*i.e.*, datasets and metrics) to be used in TSGV, and iii) in-depth discusses potential problems of current benchmarking designs and research directions for further investigations. To the best of our knowledge, this is the first systematic survey on temporal sentence grounding. More specifically, we first discuss existing TSGV approaches by grouping them into four categories, *i.e.*, two-stage methods, single-stage methods, reinforcement learning-based methods, and weakly supervised methods. Then we present the benchmark datasets and evaluation metrics to assess current research progress. Finally, we discuss some limitations in TSGV through pointing out potential problems improperly resolved in the current evaluation protocols, which may push forwards more cutting edge research in TSGV. Besides, we also share our insights on several promising directions, including four typical tasks with new and practical settings based on TSGV.

CCS Concepts: • **Computing methodologies** → *Natural language processing*; *Computer vision*; • **Information systems** → **Video search**.

Additional Key Words and Phrases: video understanding, multi-modality, vision and language, cross-modal video retrieval

Just Accepted

1 INTRODUCTION

With the increasing development of multimedia technologies on mobile phones and other terminal devices, people have gained easier access to videos from all around the world. Compared with other mediums for information transmission and exchange like texts and images, videos contain more dynamic activities and are of richer semantics to convey complex while understandable information. Basically, one video is composed of a continuing sequence of frame images possibly accompanied by audio and subtitles. Moreover, the videos from online websites in the wild are also surrounded by multiple forms of natural language texts (*e.g.*, comments written by video

[*] This work started when Yitian Yuan was at Tsinghua university.
[†] Corresponding Authors

Authors' addresses: Xiaohan Lan, lanxh20@mails.tsinghua.edu.cn, Tsinghua Shenzhen International Graduate School, University Town of Shenzhen, Shenzhen, China; Yitian Yuan, Meituan, Beijing, China, yuanyitian@foxmail.com; Xin Wang, Tsinghua University, 30 Shuangqing Rd, Beijing, China, xin_wang@tsinghua.edu.cn; Zhi Wang, Tsinghua Shenzhen International Graduate School, University Town of Shenzhen, Shenzhen, China, wangzhi@sz.tsinghua.edu.cn; Wenwu Zhu, Tsinghua University, 30 Shuangqing Rd, Beijing, China, wwzhu@tsinghua.edu.cn.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2022 Copyright held by the owner/author(s).

1551-6857/2022/5-ART
https://doi.org/10.1145/3532626

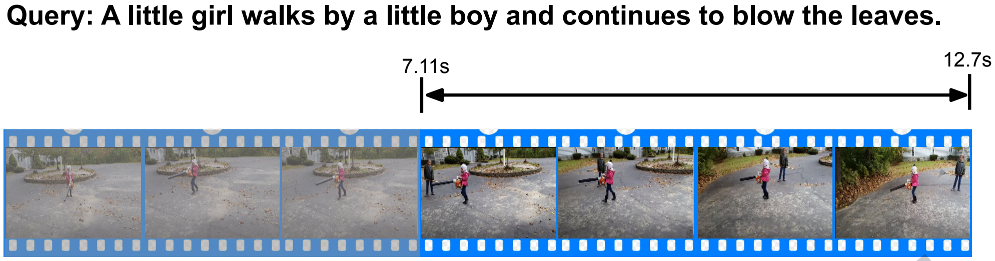

Query: A little girl walks by a little boy and continues to blow the leaves.
7.11s
12.7s

Fig. 1. An example of Temporal Sentence Grounding in Videos (TSGV), *i.e.*, to determine the start and end timestamps of the target video segment corresponding to the given sentence query.

viewers, video descriptions uploaded by creators, recommendation reasons edited by website editors). Thus, videos have natural advantages for multimedia intelligence exploration and research. However, the raw videos (*e.g.*, the user-generated video data [8] or surveillance videos [93]) are too redundant and of content sparsity against the user-specific retrieval demands. Furthermore, it is also challenging to maintain and manage these raw videos since they need to occupy a huge number of storage resources [22]. Therefore, the ability to quickly retrieve a specific video segment (*i.e.*, moment) from a long untrimmed video can allow users to locate highlighted moments of their interests conveniently and help information providers to optimize the storage fundamentally, thus being of great importance and interest in the research community.

Given the urgent need in both academia and industry, a vast number of studies attempt to automatically capture the key information within a video, *e.g.*, video summarization [65, 123, 131], video highlight detection [43, 114]. More fundamentally, some works [4, 44, 51, 64, 82, 84, 99, 117] treat the task of detecting a video segment that performs a specific action as a video classification problem, denominating this type of task as *action detection* or *temporal action localization* (TAL) [5]. Though TAL is able to extract effective information from the untrimmed videos, it is restricted by predefined action categories. Even the categorization is becoming more and more complicated, it is still not fully adequate to cover all kinds of interactive activities. Thus, it is natural to utilize natural language to describe those various and complex activities. Temporal Sentence Grounding in Videos (TSGV) is such a task to match a descriptive sentence with one segment (or moment) in an untrimmed video that is of the same semantics. As shown in Fig. 1, given the query "A little girl walks by a little boy and continues to blow the leaves" as input, the goal of TSGV is to predict the start and end points (*i.e.*, 7.11s to 12.7s) of the target segment within the whole video, and the predicted segment should contain the activities indicated by the input query. Like other visual-and-language tasks (*e.g.*, visual question answering [1, 2, 120], image/video captioning [18, 72, 73, 109, 115, 116], visual grounding [45, 98, 118] and vision-and-language pre-training [19, 24, 61, 90]), TSGV requires both understanding of visual and textual inputs. Moreover, it could also serve as an intermediate task for various downstream vision-and-language tasks such as video question answering [28, 48, 50, 107] and video summarization [23, 66, 81, 123, 140]. For example, related segments can be first grounded through the textual question and then analyzed for discovering the final answer to the input question. Also, by providing concise sentence summaries of videos, semantic coherent video segments can be grounded, retrieved and composed as the visual summaries of the original videos. Hence, it is worthwhile to go into a deep exploration in TSGV, which connects computer vision and natural language processing communities, as well as further promotes a variety of downstream applications. However, TSGV is much more challenging for the following reasons:

Just Accepted

- Both videos and sentence queries are in the form of temporal sequences with rich semantic meanings. Therefore, matching the relationships between videos and sentences is quite complicated and needs to be modeled in a fine-grained manner for accurate temporal grounding.
- The target segments corresponded to the provided sentence queries are quite flexible in terms of spatial and temporal scales in videos. It will be computationally expensive to fetch candidate video segments of different lengths in different locations via sliding windows, followed by individually matching them with the sentence query. Therefore, obtaining video segments with different temporal granularities to comprehensively cover the target segments efficiently also poses challenges for TSGV.
- Activities in a video often do not appear independently, instead they have internal semantic correlations and temporal dependencies on each other. Therefore, modelling the video context information, together with the inner logic relations among different video contents under the semantic guidance from sentence, becomes an important and challenging step to ensure the accuracy of temporal grounding approaches.

Despite the above challenges, there exist many promising research works which bring continuous improvement in TSGV in the past few years, ranging from early two-stage matching-based methods [29, 32, 38, 57, 103], single-stage methods [14, 122, 124, 128], RL-based methods [35, 36, 105], to the recent weakly supervised setting that draws people's attention [26, 67]. Therefore, a systematic review for TSGV which summarizes the current works, analyzes their strengths and weaknesses, as well as promotes the future research directions becomes a necessity for the community. Both Yang *et al.* [113] and Liu *et al.* [59] provide a method review on existing TSGV methods with a future direction discussion. Comparing to these previous ones, our survey covers more SOTA models that have been newly published and provides a clearer taxonomy of existing methods. The in-depth analysis of the limitations of current evaluation protocols is an additional advantage. In this survey, we summarize the taxonomy of existing methods, present the evaluation protocols, critically reveal the potential problems based on the current benchmarking designs, and further identify promising research directions to promote the development of this field.

The remainder of this article is organized as follows: Sec. 2 gives a detailed taxonomy and analysis on the existing approaches. Sec. 3 reviews benchmark datasets and evaluation metrics, summarizing the current research progress via comprehensive performance comparisons. Sec. 4 contains a discussion of the hidden risks behind current evaluation setting and point out promising research directions, followed by Sec. 5 that concludes the whole paper.

Just Accepted

2 METHODS OVERVIEW

We establish the taxonomy of existing approaches based on their characteristics (*c.f.*, Fig. 2). Early works adopt a two-stage architecture (*c.f.*, Fig. 3a), *i.e.*, they first scan the whole video and pre-cut various candidate segments (*i.e.*, proposals or moments) via sliding window strategy or proposal generation network, and then rank the candidates according to the ranking scores produced by the cross-modal matching module. However, such a *scan-and-localize* pipeline is time-consuming due to too much redundant computation of overlapping candidate segments, and the individual pairwise segment-query matching may also neglect the contextual video information.

Considering the above concerns, some researchers start to use single-stage methods to solve TSGV without the process of pre-cutting candidate moments (*c.f.*, Fig. 3b). Instead, multi-scale candidate moments ended at each time step are maintained by LSTM sequentially or convolutional neural networks hierarchically, and such single-stage methods are named anchor-based methods. Some other single-stage methods predict the probabilities for each video unit (*i.e.*, frame-level or clip-level) being the start and end point of the target segment, or straightforwardly regress the target start and end coordinates based on the multimodal feature of the providing video and sentence

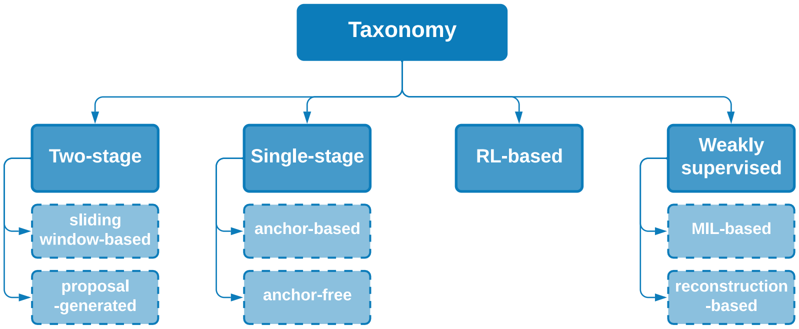


Fig. 2. The taxonomy of existing approaches, grouped into early two-stage methods, typical single-stage methods, reinforcement learning (RL)-based methods, and weakly supervised methods. According to the ways of proposal generation, the two-stage methods can be subsequently divided into sliding window-based and proposal-generated ones. Meanwhile, the single-stage methods can be divided into anchor-based and anchor-free ones, which depends on whether or not anchors (candidate moments) are produced for ranking. The weakly supervised methods could also be further grouped into MIL-based and reconstruction-based.

Just Accepted

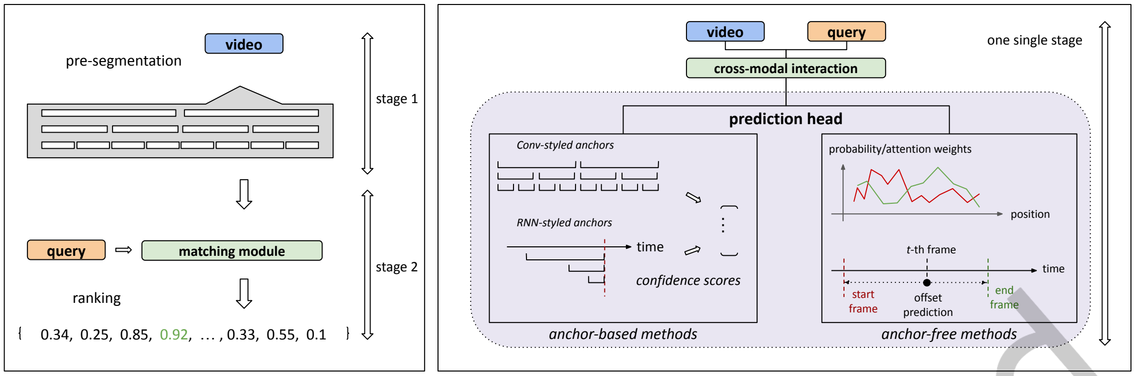
query. These methods do not depend on any candidate proposal generation process, and are named anchor-free methods.

Besides, it is worth noting that some works resort to deep reinforcement learning techniques to address TSGV, taking the sentence localization problem as a sequential decision process, which are also of anchor-free. To reduce intensive labor for annotating the boundaries of groundtruth moments, weakly supervised methods with only video-level annotated descriptions have also emerged, which can be either MIL-based or reconstruction-based. In the following, we will present all the approaches and perform a deep analysis of the characteristics for each type.

2.1 Two-stage method

For a two-stage method, the pre-segmenting of proposal candidates is conducted separately with the model computation. It takes the pre-segmented candidates and the sentence query as inputs of a cross-modal matching module for target segment localization. The two-stage methods can be grouped into two categories based on different ways to generate proposals.

2.1.1 sliding window-based. Early methods adopt multi-scale sliding window sampling strategy for the generation of candidate proposals. There are two pioneering works MCN [38] and CTRL [29] to define the TSGV task and construct benchmark datasets. Firstly, Hendricks *et al.* [38] propose MCN, which samples all the candidate moments (*i.e.* segments) via sliding window mechanism, and then projects the video moment representation and query representation into a common embedding space. The ℓ_2 distance between the sentence query and the corresponding target video moment in this space is minimized to supervise the model training (*c.f.*, Fig. 4(b)). Specifically, MCN encourages the sentence query to be closer to the target moment than negative moments in a shared embedding space. Since the negative moments either come from other segments within the same video (intra-video) or from different videos (inter-video), MCN devises two similar but different ranking loss



(a) Two-stage Method

(b) Single-stage Method

Fig. 3. Two-stage Methods *vs.* Single-stage Methods. (a): At stage 1, the entire video is pre-segmented into multi-scale candidate moments. At stage 2, the matching module takes query-moment pairs as inputs and outputs matching scores for ranking. (b): Single-stage methods can be either anchor-based or anchor-free. Anchor-based methods use different types of anchors (*e.g.*, Conv-styled, RNN-styled) as candidate moments, while anchor-free methods use prediction head to directly obtain moment boundaries (*e.g.*, generate probability/attention weights for each position, or predict the offsets from a certain frame to the start/end groundtruth boundaries).

functions:

$$\mathcal{L}_i^{intra}(\theta) = \sum_{n \in \Gamma \setminus \tau^i} \mathcal{L}^R(D_\theta(s^i, v^i, \tau^i), D_\theta(s^i, v^i, n)),$$
$$\mathcal{L}_i^{inter}(\theta) = \sum_{j \neq i} \mathcal{L}^R(D_\theta(s^i, v^i, \tau^i), D_\theta(s^i, v^j, \tau^i)), \tag{1}$$

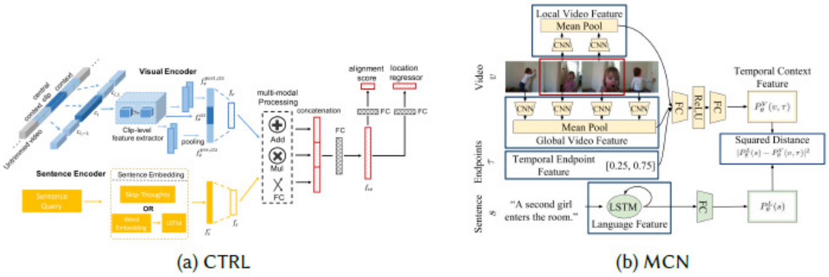
where $\mathcal{L}^R(x, y) = \max(0, x - y + b)$, b is a margin. As for training sample i, the intra-video ranking loss encourages sentence i to be closer to the target moment at the location τ^i than the negative moments from other possible locations within the same video, while the inter-video ranking loss encourages sentence i to be closer to the target one at location τ^i than the negative ones from other videos of the same location τ^i. The intra-video ranking loss is able to differentiate between subtle difference within a video while the inter-video ranking loss can differentiate between broad semantic concepts.

At the same time, Gao *et al.* [29] propose CTRL, which is the first one to adapt R-CNN [34] methodology from object detection to the TSGV domain. Particularly, CTRL also leverages sliding window to obtain candidate segments of various lengths, and as shown in Fig. 4(a), it exploits a multi-modal processing module to fuse the candidate segment representation with the sentence representation by three operators (*i.e.*, add, multiply, and fully-connected layer). Then, CTRL feeds the fused representation into another fully-connected layer to predict the alignment score and location offsets between the candidate segment and the target segment. CTRL designs a multi-task loss function to train the model, including visual-semantic alignment loss and location regression loss:

$$\mathcal{L}_{aln} = \frac{1}{N} \sum_{i=0}^{N} [\alpha_c \log(1 + \exp(-cs_{i,i})) + \sum_{j=0, j \neq i}^{N} \alpha_w \log(1 + \exp(cs_{i,j}))], \tag{2}$$

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{i=0}^{N} [R(t_{x,i}^* - t_{x,i}) + R(t_{y,i}^* - t_{y,i})], \tag{3}$$

Just Accepted



(a) CTRL

(b) MCN

Fig. 4. The Cross-modal Temporal Regression Localizer (CTRL) and Moment Context Network (MCN) frameworks, two pioneer works that firstly present TSGV task. CTRL uses joint language-segment representations to get the final alignment scores and refines the temporal boundaries by location regressor, while MCN tries to minimize the ℓ_2 distance between the language and video representations in a common space, figures from [29] and [38].

where $\mathcal{L}_{aln}$ is the visual-semantic alignment loss considering both aligned (video segment, query) pairs and misaligned pairs. $cs_{i,j}$ measures the alignment score between video segment c_i and sentence s_j. The location regression loss $\mathcal{L}_{reg}$ is only accounted for aligned pairs to predict the correct coordinates. R is a smooth-L1 function.

Compared to above CTRL that treats the query as a whole, Liu *et al.* [58] further make some improvements by decomposing the query and adaptively get the important textual components according to the temporal video context. Meanwhile, TMN [53] dynamically generates a modular neural network layout based on the semantic structure of the query to reason over the video.

Since CTRL overlooks the spatial-temporal information inside the moment and the query, Liu *et al.* [57] further propose an attentive cross-modal retrieval network (ACRN). With a memory attention network guided by the sentence query, ACRN adaptively assigns weights to the contextual moment representations for memorization to augment the moment representation. SLTA [42] also devises a spatial and language-temporal attention model to adaptively identify the relevant objects and interactions based on the query information. Considering that the inherent spatial-temporal structure of videos can not be fully captured by one-dimensional vectors in CTRL, Song *et al.* [86] propose to employ voxel- and channel-wise attention over the visual 3D feature maps to improve visual features and cross-modal correlation.

Wu and Han [103] propose a multi-modal circulant fusion (MCF) in contrast to the simple fusion ways employed in CTRL including element-wise product, element-wise sum, and concatenation. MCF extends the visual/textual vector to the circulant matrix, which can fully exploit the interactions of the visual and textual representations. By plugging MCF into CTRL, the grounding accuracy is further improved.

Previous works like CTRL, ACRN and MCF directly calculate the visual-semantic correlation without explicitly modelling the activity information within two modalities, and the candidate segments fairly sampled by sliding window may contain various meaningless noisy contents which do not contain any activity. Hence, Ge *et al.* [32] explicitly mine activity concepts from both visual and textual parts as prior knowledge to provide an actionness score for each candidate segment, reflecting how confident it contains activities, which enhances the localization accuracy. MMRG [126] employs a multi-modal relational graph explicitly considering the interactions among visual and textual objects. It also designs customized pre-training tasks to enhance the visual representations.

Despite the simplicity and effectiveness of such sliding window-based methods, they suffer from inefficient computation since there are too many overlapped areas re-computed due to the densely sampling process with predefined multi-scale sliding windows.

Just Accepted

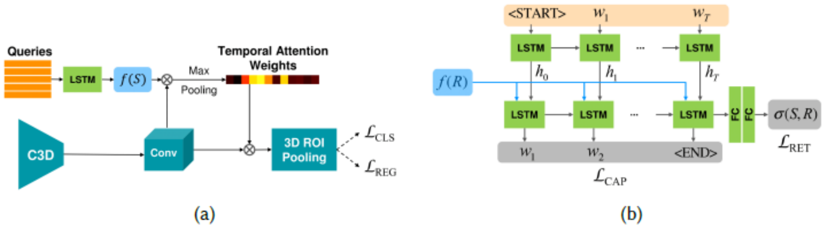


Fig. 5. The structure of Query-guided Segment Proposal Network (QSPN), including the query-guided segment proposal network (*c.f.*, 5(a)) and a fine-grained early-fused similarity model for retrieval (*c.f.*, 5(b)), figures from [108].

2.1.2 proposal-generated. Considering the inevitable drawbacks of sliding window-based methods, some approaches devote to reduce the number of proposal candidates, namely proposal-generated method. Such proposal-generated methods still adopt a two-stage scheme but avoid densely sliding window sampling through different kinds of proposal networks.

QSPN [108] relieves such a computation burden by proposing temporal segments conditioned on the query so as to reduce the number of candidate segments (*c.f.*, Fig. 5). As shown in Fig. 5(a), the query-guided SPN first incorporates the query embeddings into the video features to get the attention weight for each temporal location, and further integrates the temporal attention weights into the convolutional process for video encoding to propose query-aware representations of candidate segments. Afterwards, the generated proposal visual feature from Fig. 5(a) is incorporated into the sentence embedding process at each time step of the second layer of the two-layer LSTM in an early fusion way (*c.f.*, Fig. 5(b)). QSPN devises a triplet-based retrieval loss which is similar to MCN:

$$\mathcal{L}_{RET} = \sum_{(S,R,R')} \max\{0,\ \eta + \sigma(S,R') - \sigma(S,R)\}\,, \tag{4}$$

where (S,R) is the positive (sentence, segment) pair while R' is the sampled negative segment. QSPN also devises an auxiliary captioning task which re-generate the query sentence from the retrieved video segment. The loss for captioning is as follows:

$$\mathcal{L}_{CAP} = -\frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T_k} \log P(w_t^k | f(R), h_{t-1}^{(2)}, w_1^k, \ldots, w_{t-1}^k)\,, \tag{5}$$

where a standard captioning loss is introduced to maximize the normalized log-likelihood of the words generated at all T unrolled time steps, over all K groundtruth matching query-moment pairs.

Similarly, SAP proposed by Chen and Jiang [15] integrates the semantic information of sentence queries into the generation process of activity proposals. Specifically, the visual concepts extracted from the query sentence and video frames are used to compute visual-semantic correlation score for every frame. Activity proposals are generated by grouping frames with high visual-semantic correlation scores.

Summary. Despite the intuitiveness and success of this two-stage matching-based paradigm, it also has some drawbacks. In order to achieve high localization accuracy (*i.e.*, the candidate pool should have at least one proposal that is close to the groundtruth moment), the duration and location distribution of the candidate moments should be diverse, thus inevitably increasing the number of candidates, which leads to inefficient computation of the subsequent matching process.

Just Accepted

2.2 Single-stage method

The single-stage model follows one single-pass pattern. We divide it into two types, *i.e.*, anchor-based and anchor-free, based on whether the method uses anchors (*i.e.*, proposals) to make predictions.

2.2.1 anchor-based. Anchor-based methods employ different types of anchors (*e.g.*, Conv-styled, RNN-styled) to yield candidate moments. TGN [11] adopts a typical single-stage deep architecture, which can localize the target moment in one single pass without handling heavily overlapped pre-segmented candidate moments. As shown in Fig. 6, TGN dynamically matches the sentence and video units via a sequential LSTM grounder with fine-grained frame-by-word interaction, and at each time step, the grounder would simultaneously score a group of candidate segments with different temporal scales ending at this time step.

CMIN [136] sequentially scores a set of candidate moments of multi-scale anchors like TGN but with a sequential BiGRU network, and refines the candidate moments with boundary regression. To further enhance the cross-modal matching, it devises a novel cross-modal interaction network, which first leverages a syntactic GCN to model the syntactic structure of queries, and captures long-range temporal dependencies of video context with a multi-head self-attention module.

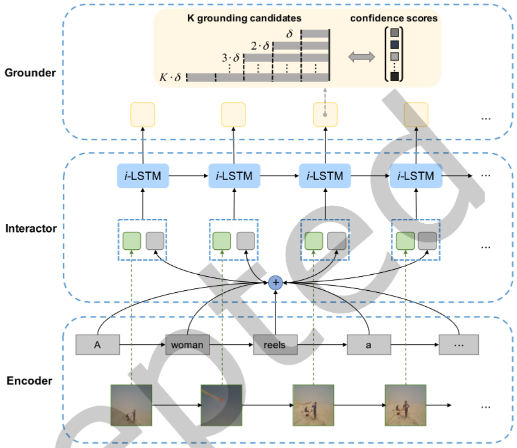


Fig. 6. The architecture of TGN, adopting a frame-by-word interaction single-stream framework. The grounder would generate multi-scale grounding candidates (anchors) that end at the same time step, figure from [11].

Likewise, CBP [97] builds a single-stream model with sequential LSTM, which jointly predicts temporal anchors and boundaries at each time step to yield precise localization. To better detect semantic boundaries, CBP devises a self-attention based module to collect contextual clues instead of simply concatenating the contextual features like [29, 32, 38].

CSMGAN [56] also adopts such a single-pass scheme. It builds a joint graph for modelling the cross-/self-modal relations via iterative message passing, to capture the high-order interactions between two modalities effectively. Each node of the graph aggregates the messages from its neighbor nodes in an edge-weighted manner and updates its state with both aggregated message and current state through ConvGRU. Qu *et al.* [74] present a fine-grained iterative attention network (FIAN), which devises a content-oriented strategy to generate candidate moments differing from the anchor-based methods with sequential RNNs mentioned above. FIAN employs a refined cross-modal guided attention block to capture the detailed cross-modal interactions, and further adopts a symmetrical iterative attention to generate both sentence-aware video and video-aware sentence representations.

TGN establishes the temporal grounding architecture through a sequential LSTM network, while Yuan *et al.* [122] propose SCDM, which exploits a hierarchical temporal convolutional network to conduct target segment localization, and couples it with a semantics-conditioned dynamic modulation to fully leverage sentence semantics to compose the sentence-related video contents over time. As shown in Fig. 7, the multimodal fusion module fuses the entire sentence and each video clip in a fine-grained manner. The fused representation is formulated as:

$$\mathbf{f}_t = \mathrm{ReLU}(\mathbf{W}^f(\mathbf{v}_t||\bar{\mathbf{s}}) + \mathbf{b}^f)\,. \tag{6}$$

Just Accepted

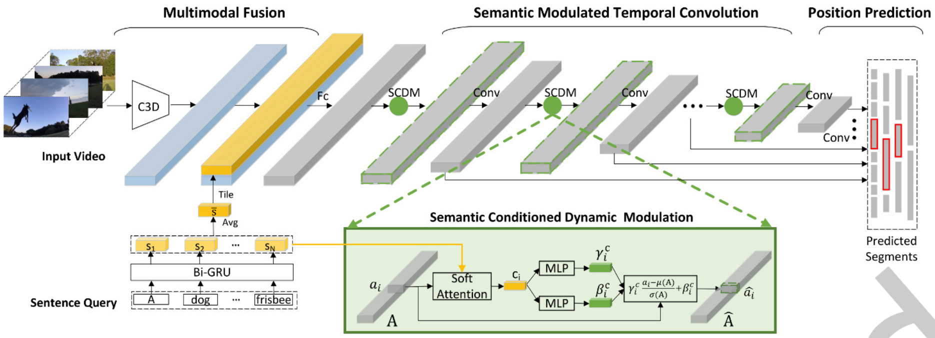


Fig. 7. The architecture of SCDM, which couples semantics-conditioned dynamic modulation into the temporal convolutional network to correlate the sentence-aware video contents over time, figure from [122].

With such fused representations as inputs, the semantic modulated temporal convolution module further correlates sentence-related video contents in a temporal convolution procedure, dynamically modulating the temporal feature maps conditioned on the sentence. Specifically, for each temporal convolutional layer, the feature map is denoted as $\mathbf{A} = \{\mathbf{a}_i\}$. The feature unit $\mathbf{a}_i$ will be modulated based on the modulation vectors γ_i^c and β_i^c:

$$\hat{\mathbf{a}}_i = \gamma_i^c \cdot \frac{\mathbf{a}_i - \mu(\mathbf{A})}{\sigma(\mathbf{A})} + \beta_i^c\,, \tag{7}$$

where the modulation vectors are computed based on the sentence representation $\mathbf{S} = \{\mathbf{s}_n\}_{n=1}^N$:

$$\begin{aligned} \rho_i^n &= \mathrm{softmax}(\mathbf{w}^T \tanh(\mathbf{W}^s \mathbf{s}_n + \mathbf{W}^a \mathbf{a}_i + \mathbf{b}))\,, \mathbf{c}_i = \sum_{n=1}^{N} \rho_i^n \mathbf{s}_n\,, \\ \gamma_i^c &= \tanh(\mathbf{W}^\gamma \mathbf{c}_i + \mathbf{b}^\gamma)\,, \beta_i^c = \tanh(\mathbf{W}^\beta \mathbf{c}_i + \mathbf{b}^\beta)\,. \end{aligned} \tag{8}$$

Finally, the position prediction module outputs the location offsets and overlap scores of candidate video segments based on the modulated features. Similar to SCDM, RMN [54] also correlates video contents conditioned on the query semantics via a modulation module, and it further employs a cascade of several rectification-modulation layers for multi-step reasoning.

MAN [128] leverages temporal convolutional network to address the TSGV task as well, where the sentence query is integrated as dynamic filters into the convolutional process. Specifically, MAN encodes the entire video stream using a hierarchical convolutional network to produce multi-scale candidate moment representations. The textual features are encoded as dynamic filters and convolved with such visual representations. Additionally, MAN exploits the graph-structured moment relation modelling adapted from Graph Convolution Network (GCN) [46] for temporal reasoning to further improve the moment representations. Similar to MAN, Soldan *et al.* [85] also adopt GCN and present a video-language graph matching network (VLG-Net) for modelling the fine-grained inter-modal interaction.

Both SCDM and MAN only consider 1D temporal feature maps, while the 2D-TAN [134] network models the temporal relations of video segments via a two-dimensional map. As shown in Fig. 8, it firstly divides the video into evenly spaced video clips with duration τ. The (i, j)-th location on the 2D temporal map represents a candidate moment (or anchor) from the time $i\tau$ to $(j + 1)\tau$. This kind of 2D temporal map covers diverse video moments with different lengths, while representing their adjacent relations. The proposed temporal adjacent

Just Accepted

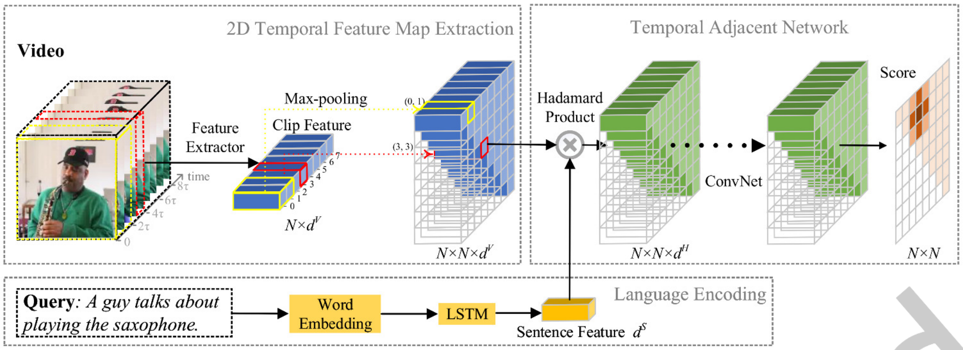


Fig. 8. The architecture of 2D temporal adjacent network (2D-TAN), which consists of a text encoder for language representation, a 2D temporal feature map extractor for video representation and a temporal adjacent network for moment localization, figure from [134].

network fuses the sentence representation with each of the candidate moment features and then leverages convolutional neural network to embed the video context information, and finally predicts the confidence score of each candidate to be the final target segment.

Some works [9, 30, 96, 133] also adopt the same proposal generation approach as that of 2D-TAN. Wang *et al.* [96] propose a structured multi-level interaction network (SMIN), which makes further modifications on the 2D temporal feature map as its proposal generation module. SMRN [9] adds a residual connection within the hierarchical convolution network of 2D-TAN and further utilizes the query semantics to modulate short connections within residual blocks. Zhang *et al.* [133] present a visual-language transformer backbone followed by a multi-stage aggregation module to get discriminative moment representations for more accurate localization. Gao *et al.* [30] design a fine-grained semantic distillation framework for retrieving desired moments with superiority in both accuracy and efficiency.

It is worth noting that Bao *et al.* [3] present an anchor-based dense events propagation network (DepNet) for a more challenging task namely dense events grounding, which aims to localize multiple moments given a paragraph. DepNet aggregates the visual-semantic information of dense events into a compact set and then propagates it to localize each single event, thus fully exploiting the temporal relationships between dense events.

Despite the superior performance anchor-based methods have achieved, the performance is sensitive with the heuristic rules manually designed (*i.e.*, the number and scales of anchors). As a result, such anchor-based methods can not adapt to the situation with variable video length. Meanwhile, although the pre-segmentation like two-stage methods is not required, it still essentially depends on the ranking of proposal candidates, which will also influence its efficiency.

2.2.2 anchor-free. Instead of ranking a vast number of proposal candidates, the anchor-free methods start from more fine-grained video units such as frames or clips, and aim to predict the probability for each frame/clip being the start and end point of the target segment, or directly regress the start and end points from the global view.

Yuan *et al.* [124] propose ABLR, which solves TSGV from a global perspective without generating anchors. Specifically, as shown in Fig. 9, to preserve the context information, ABLR first encodes both video and sentence via bidirectional LSTM networks. Then, a multi-modal co-attention mechanism is introduced to generate not only video attention which reflects the global video structure, but also sentence attention which highlights the crucial details for temporal localization. Finally, an attention-based coordinates prediction module is designed to regress the temporal coordinates (*i.e.* the starting timestamp t^s and the ending timestamp t^e) of sentence query from the

Just Accepted

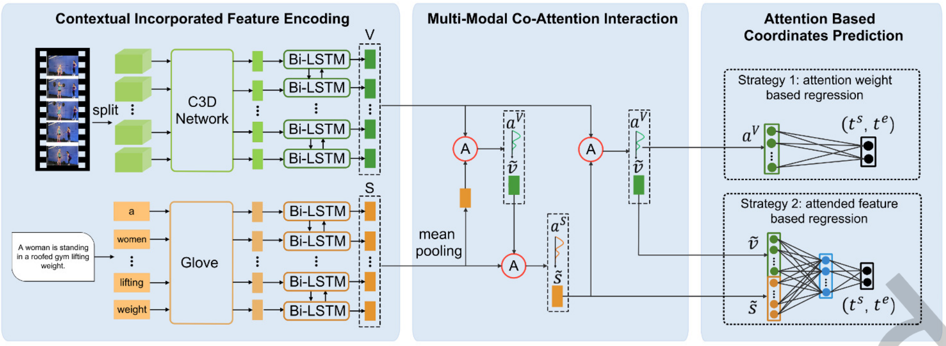
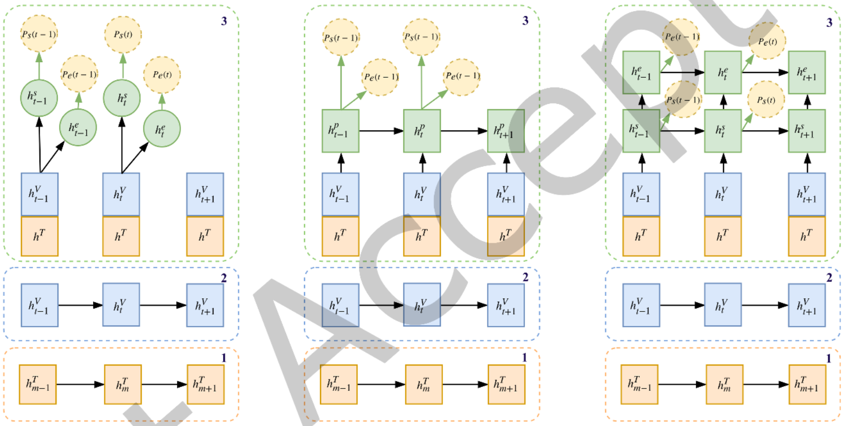

Contextual Incorporated Feature Encoding
Multi-Modal Co-Attention Interaction
Attention Based Coordinates Prediction
split
C3D Network
Bi-LSTM
V
S
A woman is standing in a roofed gym lifting weight.
a
women
lifting
weight
Glove
mean pooling
A
a^V
$\tilde{v}$
a^S
$\tilde{s}$
Strategy 1: attention weight based regression
(t^s, t^e)
Strategy 2: attended feature based regression

Fig. 9. The architecture of Attention Based Location Regression (ABLR) model, which regresses the target coordinates with a multi-modal co-attention mechanism, figure from [124].



(a) MLP predictor (b) Tied-LSTM predictor (c) Conditioned-LSTM predictor

Fig. 10. The architecture of ExCL. It consists of three modules: a sentence encoder (shown in orange squares), a video encoder (shown in blue squares) and three variants of frame predictor (*i.e.*, MLP, Tied-LSTM and Conditioned-LSTM). The frame predictor outputs the start and end probabilities for each frame, figure from [33].

former output attentions. Meanwhile, there are two different regression strategies (*i.e.*, attention weight-based regression and attended feature-based regression) with the location regression loss $L^{ablr}_{reg} = \sum_{i=1}^{K}[R(\tilde{t}^s_i - t^s_i) + R(\tilde{t}^e_i - t^e_i)]$, where R is a smooth L1 function. Besides the location regression loss that aims to minimize the distance between the temporal coordinates of the predicted and the groundtruth segments, ABLR also designs an attention calibration loss $\mathcal{L}_{cal}$ to get the video attentions more accurately:

$$\mathcal{L}_{cal} = -\sum_{i=1}^{K} \frac{\sum_{j=1}^{M} m_{i,j} \log(a^{V_i}_j)}{\sum_{j=1}^{M} m_{i,j}} . \tag{9}$$

Here, $\mathcal{L}_{cal}$ encourages the attention weights of the video clips within the groundtruth segment to be higher.

Just Accepted

LGI [68] formulates the TSGV task as the attention-based location regression like ABLR. It further presents a more effective local-global video-text interaction module, which models the multi-level interactions between semantic phrases and video segments. Chen *et al.* [14] propose pairwise modality interaction (PMI) via a channel-gated modality interaction model to explicitly model the channel-level and sequence-level interactions in a pairwise fashion. Specifically, a light-weight convolutional network is applied as the localization head to process the feature sequence and output the video-text relevance score and boundary prediction. HVTG [16] also computes the frame-level relevance scores and makes boundary prediction based on these scores. To perform the fine-grained interaction among the visual objects and between the visual object and the language query, HVTG devises a hierarchical visual-textual graph to encode the features.

Unlike ABLR that regresses the coordinates of target moment directly, ExCL [33] borrows the idea from the Reading Comprehension task [10] in natural language processing area. The process of retrieving a video segment from the video is analogous to extract a text span from the passage. Specifically, as shown in Fig. 10, ExCL employs three different variants of start-end frame predictor networks (*i.e.*, MLP, Tied-LSTM and Conditioned-LSTM) to predict start and end probabilities for each frame.

Likewise, VSLNet [130] employs a standard span-based Question Answering framework. VSLNet further distinguishes the differences between video sequence and text passage for better adaption to TSGV task. To address the differences, it designs a query-guided highlighting strategy to narrow down the search space to a smaller coarse highlight region. L-Net [12] introduces a boundary model to predict the start and end boundaries, semantically localizing the video segment given the language query. It devises a cross-gated attended recurrent network to emphasize the relevant video parts while the irrelevant ones are gated out, and a cross-modal interactor for fine-grained interactions between two modalities.

TMLGA [76] also predicts start and end probabilities for each video unit. It further models the uncertainty of boundary labels, using two Gaussian distributions as groundtruth probability distributions. CPN [141] devises a cascaded prediction network based on the segment-tree data structure. It performs two sub-tasks (*i.e.*, decision navigation and signal decomposition) on each level from top to down for final boundary prediction. PEARL [132] integrates the subtitles of videos and convolves the query filters into the visual and subtitle branches to locate the boundaries.

Lu *et al.* [60] propose a dense bottom-up grounding framework (DEBUG), which localizes the target segment by predicting the distances to bidirectional temporal boundaries for all frames inside the groundtruth segment. In this way, all frames inside the groundtruth segment can be seen as positive samples, alleviating the severe imbalance issue caused by only regarding the groundtruth segment boundaries as positive samples. As shown in Fig. 11, a typical dense anchor-free model usually contains a backbone framework for multimodal feature encoding and a head network for frame-level predictions. Specifically, DEBUG adopts QANet as its backbone network which models the interaction between videos and queries, and designs three branches as head networks which aim to separately predict the classification score, boundary distances, and confidence score for each frame.

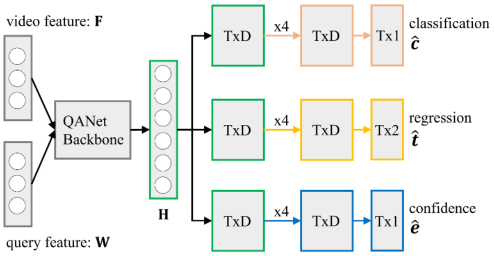


Fig. 11. The architecture of DEBUG, consisting of a backbone framework (QANet) to model the multimodal interaction and a head module with three branches for dense regression, figure from [60].

Similarly, DRN [125] and GDP [13] also adopt such a dense anchor-free framework. For backbone, DRN uses a video-query interaction module to obtain fused hierarchical feature maps. For head network, DRN densely predicts the distances to boundaries, matching score and estimated IoU for each frame within the groundtruth

Just Accepted

segment. Meanwhile, for backbone, GDP leverages a Graph-FPN layer which conducts graph convolution over all nodes in the scene space to enhance the integrated frame features. For head network, GDP predicts the distances from its location to the boundaries of target moment and a confidence score to rank its boundary prediction for each frame.

Another graph-based method DORi [77] utilizes a spatio-temporal graph to model the temporally-changing inter-object interactive relationships based on the language query, which can further improve the activity representations. Instead of adopting well-designed graph-based structures, Yu *et al.* [119] propose a simple yet effective approach that only conducts bi-directional cross-modal interaction via multi-head attention with multiple training objectives.

Compared with anchor-based methods, the anchor-free methods are obviously computation-efficient and robust to variable video duration. Despite these significant advantages, it is difficult for anchor-free methods to capture segment-level features for multimodal interactions.

Other Single-stage Methods. Different from the aforementioned single-stage methods which either samples from multi-scale anchors or directly regresses the final coordinates, some methods out of these patterns have emerged. The boundary proposal network (BPNet) [106] keeps the advantages of both anchor-based and anchor-free methods and avoids the defects, which generates proposals by anchor-free methods and then matches them with the sentence query in an anchor-based manner. Wang *et al.* [95] propose a dual path interaction network (DPIN) containing two branches (*i.e.*, a boundary prediction pathway for frame-level features and an alignment pathway for segment-level features) to complementarily localize the target moment. Inspired from the dependency tree parsing task in natural language processing community, a biaffine-based architecture named context-aware biaffine localizing network (CBLN) [55] has been proposed which can simultaneously score all possible pairs of start and end indices. Ding *et al.* [25] introduce a support-set cross-supervision (Sscs) module. The Sscs module can be a plug-in branch to enhance multi-modal relation modelling for both anchor-based and anchor-free methods.

2.3 Reinforcement learning-based method

As another kind of anchor-free approach, RL-based frameworks view such a task as a sequential decision process. The action space for each step is a set of handcraft-designed temporal transformations (*e.g.*, shifting, scaling).

He *et al.* [36] first introduce deep reinforcement learning techniques to address the task of TSGV, which formulates TSGV as a sequential decision making problem. As depicted in Fig. 12, at each time step, the observation network outputs the current state of the environment for the actor-critic module to generate an action policy (*i.e.*, the probabilistic distribution of all the actions predefined in the action space), based on which the agent will perform an action to adjust the temporal boundaries. This iterative process will be ended when encountering the STOP action or reaching the maximum number of steps (*i.e.*, T_{max}). Specifically, at each step, the current state vector is computed as $s^{(t)} = \Phi(E, V_G, V_L^{(t-1)}, L^{(t-1)})$, where $s^{(t)}$ is generated by one FC layer whose inputs are the concatenated features including the segment-specific features (*i.e.*, the normalized boundary pair $L^{(t-1)} = [l_s^{(t-1)}, l_e^{(t-1)}]$ and local segment C3D feature $V_L^{(t-1)}$) and global features (*i.e.*, the sentence embedding E and entire video C3D feature V_G). Then the actor-critic module employs GRU to model the sequential decision making process. At each time step, GRU takes $s^{(t)}$ as input and the hidden state is used for policy (denoted as $\pi(a_i^{(t)}|s^{(t)}, \theta_\pi)$) generation and state-value (denoted as $v(s^{(t)}|\theta_v)$) estimation. The reward for each step r_t is designed to encourage a higher tIoU compared to that of the last step. The accumulated reward function is then defined as (γ is a constant discount factor):

$$R_t = \begin{cases} r_t + \gamma * v(s^{(t)}|\theta_v), & t = T_{max} \\ r_t + \gamma * R_{t+1}, & t = 1, 2, \ldots, T_{max} - 1 \end{cases} . \tag{10}$$

Just Accepted

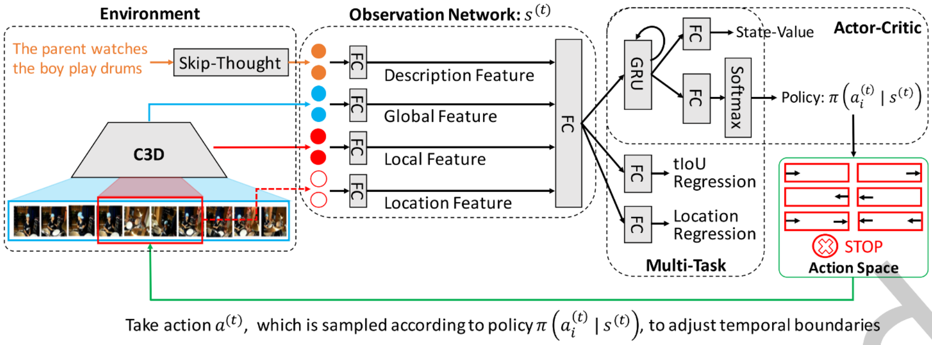

Environment
Observation Network: $s^{(t)}$
Actor-Critic
The parent watches the boy play drums
Skip-Thought
C3D
FC
Description Feature
Global Feature
Local Feature
Location Feature
GRU
State-Value
Softmax
Policy: $\pi\left(a_i^{(t)} \mid s^{(t)}\right)$
tIoU Regression
Location Regression
Multi-Task
STOP
Action Space
Take action $a^{(t)}$, which is sampled according to policy $\pi\left(a_i^{(t)} \mid s^{(t)}\right)$, to adjust temporal boundaries

Fig. 12. The architecture of R-W-M framework. The observation network takes the environment (*i.e.*, description feature, global video feature, local feature and location feature) as input to compute current state. Then one of those seven operators in the action space is determined to adjust the temporal boundaries of current segment. Two auxiliary supervised tasks including tIoU regression and location regression are also leveraged, figure from [36].

Then they introduce the advantage function as objective which is approximated by the Mente Carlo sampling to get the policy gradient:

$$\mathcal{L}'_A(\theta_\pi) = -\sum_t (\log \pi(a_i^{(t)}|s^{(t)}, \theta_\pi))(R_t - v(s^{(t)}|\theta_v))\,. \tag{11}$$

They further leverage two supervised tasks (*i.e.*, tIoU regression and location regression) so the parameters can be updated from both policy gradient and supervised gradient to help the agent obtain more accurate information about the environment.

Wang *et al.* [100] propose an RNN-based RL model which sequentially observes a selective set of video frames and finally obtains the temporal boundaries given the query. Cao *et al.* [6] firstly leverage the spatial scene tracking task, which utilizes a spatial-level RL for filtering out the information that is not relevant to the text query. The spatial-level RL can enhance the temporal-level RL for adjusting the temporal boundaries of the video. TripNet [35] uses gated attention to align textual and visual features, leading to improved accuracy. It incorporates a policy network for efficient search, which selects a fixed temporal bounding box moving around without watching the entire video.

TSP-PRL [105] adopts a tree-structured policy that is different from conventional RL-based methods, inspired by a human's coarse-to-fine decision-making paradigm. The agent receives the state from the environment (video clips) and estimates a primitive action via tree-structured policy, including root policy and leaf policy. The action selection is depicted by a switch over the interface in the tree-structured policy. The alignment network will predict a confidence score to determine when to stop. Meanwhile, AVMR [7] addresses TSGV under the adversarial learning paradigm, which designs an RL-based proposal generator to generate proposal candidates and employs Bayesian Personalized Ranking as a discriminator to rank these generated moment proposals in a pairwise manner.

2.4 Weakly supervised method

For the annotation of groundtruth data in TSGV, the annotators should read the query and watch the video first, and then determine the start and end points of the query-indicated segment in the video. Such a human-labored process is very time-consuming. Therefore, due to the labor-intensive groundtruth annotation procedure, some

Just Accepted

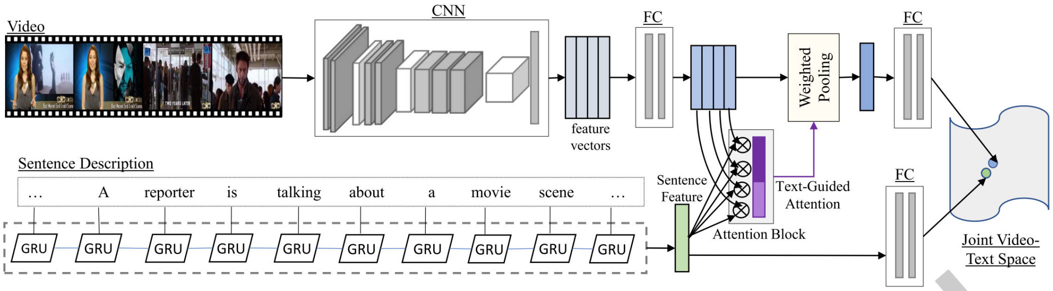


Fig. 13. The overall framework of TGA. It learns a joint embedding network to align the text and video features. The global video representation is generated by weighted pooling based on text-guided attentions, figure from [67].

works start to extend TSGV to a weakly supervised scenario where the locations of groundtruth segments (*i.e.*, the start and end timestamps) are unavailable in the training stage. This is formally named as weakly supervised TSGV. In general, weakly supervised methods for TSGV can be grouped into two categories (*i.e.*, MIL-based and reconstruction-based). One representative work will be illustrated in detail for each category, after which we will introduce the remaining.

Some works [20, 31, 67, 91] adopt multi-instance learning (MIL) to address the weakly TSGV task. When temporal annotations are not available, the whole video is treated as a bag of instances with bag-level annotations, and the predictions for instances (video segment proposals) are aggregated as the bag-level prediction.

TGA [67] is a typical MIL-based method which learns the visual-text alignment in the video level by maximizing the matching scores of the videos and their corresponding descriptions while minimizing the matching scores of the videos and the descriptions of others. It presents text-guided attention (TGA) to get text-specific global video representations, learning the joint representation of both the video and the video-level description. As illustrated in Fig. 13, TGA first employs a GRU for sentence embedding and a pretrained image encoder for extracting frame-level features. The similarity between j^{th} sentence and the k^{th} temporal feature within the i^{th} video denoted as s^i_{kj} is computed and a softmax opration is applied to get the text-guided attention weights for each temporal unit denoted as a^i_{kj}:

$$s^i_{kj} = \frac{\mathbf{w}^{iT}_j \mathbf{v}^i_k}{\left\|\mathbf{w}^i_j\right\|_2 \left\|\mathbf{v}^i_k\right\|_2}, \quad a^i_{kj} = \frac{\exp(s^i_{kj})}{\sum_{m=1}^{nv_i} \exp(s^i_{mj})}. \tag{12}$$

Thus we could get the sentence-wise global video feature $\mathbf{f}^i_j = \sum_{k=1}^{nv_i} a^i_{kj} \mathbf{v}^i_k$.

WSLLN [31] is another MIL-based end-to-end weakly supervised language localization network conducting clip-sentence alignment and segment selection simultaneously. Huang *et al.* [41] present a cross-sentence relations mining (CRM) method exploring the cross-sentence relations within paragraph-level scope to improve the per-sentence localization accuracy.

A video-language alignment network (VLANet) proposed by Ma *et al.* [63] prunes the irrelevant moment candidates with the Surrogate Proposal Module and utilizes multi-directional attention to get a sharper attention map for better multimodal alignment. Wu *et al.* [104] attempt to apply an RL-based model for weakly TSGV, which proposes a boundary adaptive refinement framework (BAR) for achieving boundary-flexible and content-aware grounding results. Chen *et al.* [20] propose a novel coarse-to-fine model (WSTG) based on MIL. First, the coarse stage selects a rough segment from a set of predefined sliding windows, which semantically corresponds to the

Just Accepted

given sentence. Afterwards, the fine stage mines the fine-grained matching relationship between each frame in the coarse segment and the sentence. It thereby refines the boundary of the coarse segment by grouping the frames and gets more precise grounding results.

Tan *et al.* [91] propose a Latent Graph Co-Attention Network (LoGAN), a novel co-attention model that performs fine-grained semantic reasoning over an entire video. LoGAN is also a MIL-based method, which performs a similar frame-by-word interaction with the supervised method TGN [11] and adapts the graph module from another supervised method MAN [128] for iterative frame representation update. Wang *et al.* [101] present a fine-grained semantic alignment network (FSAN), which enables iterative multi-head attention based cross-modal interaction to capture fine-grained video-language alignment. In order to learn more robust and discriminative moment features, VCA [102] devises a visual co-occurrence alignment NCE loss that maximizes the similarity between video moments from different videos with similar descriptions.

Since MIL-based methods typically learn the visual-text alignment with a triplet loss, these methods heavily depend on the quality of randomly-selected negative samples, which are often easy to distinguish from the positive ones and cannot provide strong supervision signals.

The reconstruction-based methods [17, 26, 52, 87] attempt to reconstruct the given sentence query based on the selected video segments and use the intermediate results for sentence localization. Unlike MIL-based methods, the reconstruction-based methods learn the visual-textual alignment in an indirect way. Lin *et al.* [52] propose a semantic completion network (SCN) to predict the masked important words within the query according to the visual context of generated and selected video proposals. Specifically, for each proposal G^k, denoted by $\hat{\mathbf{v}}^k = \{\mathbf{v}_i\}_{i=s_k}^{e_k}$, with the masked query representation $\hat{q}$, the energy word distribution $\mathbf{e}_i^k$ at i^{th} time step can be computed as $\mathbf{e}_i^k = \mathbf{W}_v\mathbf{f}_i^k + \mathbf{b}_v$, where $\mathbf{f}^k = \{\mathbf{f}_i^k\}_{i=1}^{n_q}$ are the cross-modal semantic representations, computed by $\mathbf{f}^k = \mathbf{Dec}_q(\hat{\mathbf{q}}, \mathbf{Enc}_v(\hat{\mathbf{v}}^k))$, $\mathbf{Dec}_q$ and $\mathbf{Enc}_v$ are respectively the textual decoder and visual encoder based on bi-directional Transformer [94]. Afterwards, the reconstruction loss can be computed by adding up all negative log-likelihood of masked words:

$$\mathcal{L}_{rec}^k = -\sum_{i=1}^{n_q-1}\log p(\mathbf{w}_{i+1}|\hat{\mathbf{w}}_{1:i}, \hat{\mathbf{v}}^k) = -\sum_{i=1}^{n_q-1}\log p(\mathbf{w}_{i+1}|\mathbf{e}_i^k)\,. \tag{13}$$

Song *et al.* [87] present a Multi-Level Attentional Reconstruction Network (MARN), which leverages the idea of attentional reconstruction. MARN uses proposal-level attentions to rank the segment candidates and refine them with clip-level attentions.

Duan *et al.* [26] formulate and address the problem of weakly supervised dense event captioning in videos (*i.e.*, to detect and describe all events of interest in a video), which is a dual problem of weakly supervised TSGV. It presents a cycle system to train the model which can solve such a pair of dual problems at the same time. In other words, weakly supervised TSGV can be regarded as an intermediate task in such a cycle system. Similar to [26], Chen and Jiang [17] also employ a loop system for dense event captioning. They adopt a concept learner to construct an induced set of concept features to enhance the information passing between the sentence localizer and event captioner.

Besides, instead of proposing a reconstruction-based or MIL-based method, Zhang *et al.* [138] design a counterfactual contrastive learning paradigm to improve the visual-and-language grounding tasks. A regularized two-branch proposal network (RTBPN) [137] is also presented to explore sufficient intra-sample confrontment with sharable two-branch proposal module for distinguishing the target moment from plausible negative moments.

3 DATASETS AND EVALUATIONS

In this section, we present benchmark datasets and evaluation metrics for TSGV, and provide detailed performance comparisons among the above mentioned approaches.

Just Accepted

Table 1. The statistics of videos and annotations of the benchmark datasets.

	Video Statistics				Annotation Statistics			
	# Videos	Aver. Video Duration(s)	Domain	Video Source	# Queries	# Moments	Aver. Moment Duration(s)	Aver. Query Length
DiDeMo	10,642	29.3	Open	Flickr	41,206	28,925	6.9	7.5
TACoS	127	286.6	Cooking	Lab Kitchen	18,227	7,069	27.9	9.4
Charades-STA	9,848	30.6	Indoor activity	Activity	16,124	11,767	8.1	6.2
ActivityNet Captions	14,926	117.6	Open	Activity	71,957	71,718	37.1	14.4

3.1 Datasets

Several datasets for TSGV from different scenarios with their distinct characteristics have been proposed in the past few years. There is no doubt that the effort of creating these datasets and designing corresponding evaluation metrics do promote the development of TSGV. Table 1 and Table 2 provide an overview about the statistics of public datasets. Table 1 gives an overall introduction about the videos and annotated query-moment pairs. As we can see that some datasets (*i.e.*, TACoS, Charades-STA) are constrained in a narrow and specific scene (*e.g.*, kitchen or indoor activity), while others (*i.e.*, DiDeMo, ActivityNet Captions) involve more complicated activities in open domains. Since each query refers to exactly one moment but multiple queries may refer to the same moment (a moment here means a video segment which can be identified by a {video id, start timestamp, end timestamp} triplet), the number of queries would be equal to the number of all samples (query-moment pairs) while the number of moments would be less than that, which is actually the number of unique {video id, start timestamp, end timestamp} triplets. Moreover, the detailed language statistics are reported in Table 2. Larger vocabulary size and average number of verbs/adjectives/nouns tokens indicate greater challenges in textual semantic understanding. Obviously, the sentences in ActivityNet Captions are the most difficult and those of Charades-STA are relatively simple with the smallest action (verb) set. We will introduce these four datasets more concretely as follows.

Table 2. Language statistics of the benchmark datasets.

	Vocabulary Statistics			Sentence Statistics		
	(Number of Used Unique Tokens			(Aver. Number per Query)		
	Verb	Adjectives	Nouns	Verbs	Adjectives	Nouns
DiDeMo	1.50K	1.40K	4.30K	1.20	0.58	2.64
TACoS	0.58K	0.42K	0.98K	1.48	0.23	2.64
Charades-STA	0.25K	0.17K	0.63K	1.26	0.06	2.40
ActivityNet Captions	2.60K	2.90K	8.90K	2.56	0.66	3.73

DiDeMo [38]. This dataset is collected from Flickr, and consists of various human activities uploaded by personal users. Hendricks *et al.* [38] split and label video segments from original untrimmed videos by aggregating five-second clip units, which means the lengths of groundtruth segments are times of five seconds. They claim that this trick is for avoiding ambiguity of labeling and accelerating the validation process. However, such a length-fixed issue makes the retrieval task easier since it compresses the searching space into a set with limited candidates. The data split is also provided by [38], with 33,005/4,180/4,021 video-sentence pairs for training/validation/test, respectively. Besides, a new dataset TEMPO [37] involving more temporally-related events is collected based on the DiDeMo, which is explored by some works as well [88, 135].

TACoS [75]. TACoS is built based on MPII-Compositive dataset [78]. It contains 127 complex videos featuring cooking activities, and each video has several segments being annotated by sentence descriptions illustrating people's cooking actions. The average length of videos in TACoS is around 300s, which is much longer than that of other benchmark datasets. The total amount of query-moment pairs is 18,227 in this dataset, and 50%, 25%, 25% of which are used for training, validation, and test, respectively.

Charades-STA [29]. Charades-STA is built upon Charades [83], which is originally collected for video activity recognition, and consists of 9,848 videos depicting human daily indoor activities. Specifically, Charades contains 157 activity categories and 27,847 video-level sentence descriptions. Based on Charades, Gao *et al.* [29] construct

Just Accepted

Charades-STA with a semi-automatic pipeline, which parses the activity label out of the video description first and aligns the description with the original label-indicated temporal intervals. As such, the yielded (description, interval) pairs can be seen as the (sentence query, target segment) pairs for TSGV. Since the length of original description in Charades-STA is quite short, Gao *et al.* [29] further enhance the complexity of the description by combining consecutive descriptions into a more complex sentence for test.

ActivityNet Captions [47]. ActivityNet Captions is originally proposed for dense video captioning upon ActivityNet dataset [5], and the query-moment pairs in this dataset can naturally be utilized for TSGV. ActivityNet Captions contains the largest amount of videos, and it aligns videos with a series of temporally annotated sentence descriptions. On average, each of the 20k videos contains 3.65 temporally localized sentences, resulting in a total of 100k sentences. Each sentence has an average length of 13.48 words. The sentence length is also normally distributed. Since the official test set is withheld for competitions, most TSGV works merge the two available validation subsets "val1" and "val2" as the test set. In summary, there are 10,009 videos and 37,421 query-moment pairs in the training set, and 4,917 videos and 34,536 query-moment pairs in the test set.

3.2 Metrics

There are two types of metrics for TSGV, *i.e.*, R@*n*,IoU=*m* and mIoU, both of which are first introduced for TSGV in [29]. Since IoU (Intersection over Union) is widely used in object detection to measure the similarity between two bounding boxes, similarly for TSGV, as illustrated in Fig. 14, many TSGV methods adopt temporal IoU to measure the similarity between the groundtruth moment and the predicted one. The ratio of intersection area over union area ranges from 0 to 1, and it will be equal to 1 when these two moments are totally overlapped.

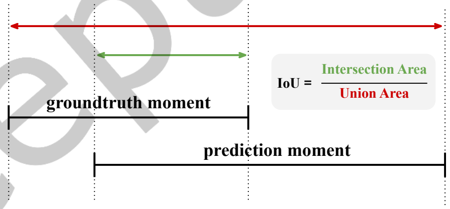


Fig. 14. The illustration of Temporal IoU (Intersection over Union).

Thereby, one of the metrics is mIoU (*i.e.*, mean IoU), a simple way to evaluate the results through averaging temporal IoUs of all samples. The other commonly-used metric is R@*n*, IoU=*m* [40]. As for sample *i*, it is accounted as positive when there exists one segment out of top *n* retrieved segments whose temporal IoU with the groundtruth segment is over *m*, which can be denoted as $r(n, m, q_i) = 1$. Otherwise, $r(n, m, q_i) = 0$. R@*n*, IoU=*m* is the percentage of positive samples over all samples (*i.e.*, $\frac{1}{N_q} \sum_i r(n, m, q_i)$).

The community is accustomed to setting $n \in \{1, 5, 10\}$ and $m \in \{0.3, 0.5, 0.7\}$. Usually, $n = 1$ when the method adopts a proposal-free manner (*i.e.*, belongs to either anchor-free or RL-based frameworks). Moreover, it is worth noting that MCN [38] adopts a particular metric with the IoU threshold $m = 1.0$ since the groundtruth segments in DiDeMo is generated by aggregating the clip units of five seconds, and MCN also employs a matching-based method thus the predicted moment has chance to fully coincide with the target moment, satisfying such a high IoU threshold.

3.3 Performance Comparison

In this section, we give a thorough performance comparison of the aforementioned approaches based on four benchmark datasets. For convenience and fairness, we uniformly adopt $n = 1$ and $m \in \{0.3, 0.5, 0.7\}$ for the metric of R@*n*,IoU=*m*. Though different types of extracted visual features may influence the grounding accuracy, we uniformly report the best results for each method as reported in literature. Table 3 lists all the experimental results grouped by their categories (*i.e.*, belonging to two-stage, single-stage, RL-based or weakly supervised

Just Accepted

Table 3. The performance comparison of all TSGV methods grouped by their categories (SW:sliding window-based, PG:proposal-generated, AB:anchor-based, AF:anchor-free, OT:other single-stage methods, RL:RL-based, WS:weakly supervised).

Type	Method	DiDeMo			TACoS			Charades-STA			ActivityNet Captions		
		0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
SW	MCN [38]	-	-	-	-	-	-	13.57	4.05	-	-	-	-
	CTRL [29]	-	-	-	18.32	13.3	-	-	21.42	7.15	-	-	-
	MCF [103]	-	-	-	18.64	12.53	-	-	-	-	-	-	-
	ROLE [58]	29.4	15.68	-	-	-	-	25.26	12.12	-	-	-	-
	ACRN [57]	-	-	-	19.52	14.62	-	-	-	-	-	-	-
	SLTA [42]	-	30.92	17.16	17.07	11.92	-	38.96	22.81	8.25	-	-	-
	VAL [86]	-	-	-	19.76	14.74	-	-	23.12	9.16	-	-	-
	ACL-K [32]	-	-	-	24.17	20.01	-	-	30.48	12.2	-	-	-
	MMRG [126]	-	-	-	57.83	39.28	-	71.6	44.25	-	-	-	-
PG	QSPN [108]	-	-	-	-	-	-	54.7	35.6	15.8	45.3	27.7	13.6
	SAP [15]	-	-	-	-	18.24	-	-	27.42	13.36	-	-	-
AB	TGN [11]	-	-	-	21.77	18.9	-	-	-	-	45.51	28.47	-
	MAN [128]	-	-	-	-	-	-	-	46.53	22.72	-	-	-
	CMIN [136]	-		-	24.64	18.05	-	-	-	-	63.61	43.4	23.88
	SCDM [122]	-	-	-	26.11	21.17	-	-	54.44	33.43	54.8	36.75	19.86
	CBP [97]	-	-	-	27.31	24.79	19.1	-	36.8	18.87	54.3	35.76	17.8
	2D-TAN [134]	-	-	-	37.29	25.32	-	-	39.7	23.31	59.45	44.51	26.54
	FVMR [30]	-	-	-	41.48	29.12	-	-	55.01	33.74	60.63	45	26.85
	SMRN [9]	-	-	-	42.49	32.07	-	-	43.58	25.22	-	42.97	26.79
	RMN [54]	-	-	-	32.21	25.61	-	-	59.13	36.98	67.01	47.41	27.21
	FIAN [74]	-	-	-	33.87	28.58	-	-	58.55	37.72	64.1	47.9	29.81
	CSMGAN [56]	-	-	-	33.9	27.09	-	-	-	-	68.52	49.11	29.15
	SMIN [96]	-	-	-	48.01	35.24	-	-	64.06	40.75	-	48.46	30.34
	Zhang *et al.* [133]	-	-	-	48.79	37.57	-	-	-	-	-	48.02	31.78
	VLG-Net [85]	25.57	71.65	-	45.46	34.19	-	-	-	-	-	46.32	29.82
AF	ABLR [124]	-	-	-	18.9	9.3	-	-	-	-	55.67	36.79	-
	DEBUG [60]	-	-	-	23.45	-	-	54.95	37.39	17.69	55.91	39.72	-
	GDP [13]	-	-	-	24.14	-	-	54.54	39.47	18.49	56.17	39.27	-
	PMI [14]	-	-	-		-	-	55.48	39.73	19.27	59.69	38.28	17.83
	ExCL [33]	-	-	-	44.4	27.8	14.6	61.4	41.2	21.3	62.1	41.6	23.9
	DRN [125]	-	-	-	-	23.17	-	-	45.4	26.4	-	42.49	22.25
	HVTG [16]	-	-	-	-	-	-	61.37	47.27	23.3	57.6	40.15	18.27
	TMLGA [76]	-	-	-	24.54	21.65	16.46	67.53	52.02	33.74	51.28	33.04	19.26
	LGI [68]	-	-	-	-	-	-	72.96	59.46	35.48	58.52	41.51	23.07
	VSLNet [130]	-	-	-	29.61	24.27	20.03	70.46	54.19	35.22	63.16	43.22	26.16
	CPN [141]	-	-	-	47.69	36.33	21.58	75.53	59.77	36.67	62.81	45.1	28.1
	DORi [77]	-	-	-	31.8	28.69	24.91	72.72	59.65	40.56	57.89	41.35	26.41
	CI-MHA [119]	-	-	-	-	-	-	69.87	54.68	35.27	61.49	43.97	25.13
	PEARL [132]	-	-	-	42.94	32.07	18.37	71.9	53.5	35.4	-	-	-
OT	BPNet [106]	-	-	-	25.96	20.96	14.08	55.46	38.25	20.51	58.98	42.07	24.69
	DPIN [95]	-	-	-	46.74	32.92	-	-	47.98	26.96	62.4	47.27	28.31
	CBLN [55]	-	-	-	38.98	27.65	-	-	61.13	38.22	66.34	48.12	27.6
RL	R-W-M [36]	-	-	-	-	-	-	-	36.7	-	-	36.9	-
	SM-RL [100]	-	-	-	20.25	15.95	-	-	24.36	11.17	-	-	-
	TripNet [35]	-	-	-	-	-	-	51.33	36.61	14.5	48.42	32.19	13.93
	TSP-PRL [105]	-	-	-	-	-	-	-	45.45	24.75	56.02	38.82	-
	STRONG [6]	-	-	-	72.14	49.73	18.29	78.1	50.14	19.3	-	-	-
	AVMR [7]	-	-	-	72.16	49.13	-	77.72	54.59	-	-	-	-
WS	WSDEC [26]	-	-	-	-	-	-	-	-	-	41.98	23.34	-
	TGA [67]	-	-	-	-	-	-	32.14	19.94	8.84	-	-	-
	WSLLN [31]	-	-	-	-	-	-	-	-	-	42.8	22.7	-
	EC-SL [17]	-	-	-	-	-	-	-	-	-	44.29	24.16	-
	SCN [52]	-	-	-	-	-	-	42.96	23.58	9.97	47.23	29.22	-
	WSTG *et al.* [20]	-	-	-	-	-	-	39.8	27.3	12.9	44.3	23.6	-
	VLANet [63]	-	-	-	-	-	-	45.24	31.83	14.17	-	-	-
	FSAN [101]	-	-	-	-	-	-	-	-	-	55.11	29.43	
	MARN [87]	-	-	-	-	-	-	48.55	31.94	14.81	47.01	29.95	-
	RTBPN [137]	-	-	-	-	-	-	60.04	32.36	13.24	49.77	29.63	-
	BAR [104]	-	-	-	-	-	-	44.97	27.04	12.23	49.03	30.73	-
	CCL [138]	-	-	-	-	-	-	-	33.21	15.68	50.12	31.07	-
	VCA [102]	-	-	-	-	-	-	58.58	38.13	19.57	50.45	31	-
	LoGAN [91]	-	-	-	-	-	-	51.67	34.68	14.54	-	-	-
	CRM [41]	-	-	-	-	-	-	53.66	34.76	16.37	55.26	32.19	-

Just Accepted

methods) which are segmented by double horizontal lines. Table 4 separately reports the experimental results on DiDeMo dataset with metrics of R@$\{1, 5\}$,IoU=1.0 and mIoU.

Two-stage method. As shown in Table 3, the overall performance of two-stage methods seems poorer than other approaches. The possible reasons lie in three folds: (1) Firstly, most of the two-stage methods combine video and sentence features coarsely, and neglect the fine-grained visual and textual interactions for accurate temporal sentence grounding in videos. (2) Secondly, separating the candidate segment generation and query-moment matching procedures will make the model unable to be globally optimized, which can also influence the overall performance. (3) Thirdly, establishing matching relationships between sentence queries and individual segments will make the local video content separate with the global video context, which may also hurt the temporal grounding accuracy.

Specifically, for the sliding window (SW)-based methods, all of them achieve the lowest grounding accuracy on the TACoS compared to the other three datasets with the same metrics. The reason is that the cooking activities in TACoS take place in the same kitchen scene with only some slightly varied cooking objects (*e.g.*, chopping board, knife, and bread). Thus, it is hard to do temporal location predictions for such fine-grained activities. Meanwhile, the lengths of videos in TACoS are also longer, which will greatly increase the target segment searching space and bring more difficulties. Obviously, MMRG outperform other SW-based methods with great gains on both TACoS and Charades-STA. Despite using the same moment sampling strategies with CTRL, the multi-modal relational graph MMRG employs can capture subtle differences of candidate moments from the same video and the customized self-supervised pre-training tasks further improve the visual features. Regardless of MMRG, ACL-K also significantly outperforms the remaining SW-based methods on TACoS and Charades-STA, proving the effectiveness of aligning the activity concepts mined from both textual and visual parts. MCN gets the most inferior results on the Charades-STA, which shows that its simple multimodal matching and ranking strategy for candidate segments cannot deal well with the segments of various and flexible locations. However, CTRL, ACRN, ROLE, SLTA, VAL, ACL-K and MMRG can adjust the candidate segment boundaries based on the model location offsets prediction, which can therefore improve the performances. All of the sliding window-based methods have not conducted experiments on the large-scale ActivityNet Captions dataset, which may due to the costly computation for multi-scale sliding window sampling.

The proposal-generated (PG) methods achieve even better performance than the SW-based methods though the number of proposal candidates decreases. QSPN with query-guided segment proposal network and auxiliary captioning loss significantly outperforms other two-stage methods (except MMRG) on the Charades-STA, which demonstrates that the presented query-guided proposal network is able to provide more effective candidate moments with finer temporal granularity without dealing with redundant sliding window sampled moments. QSPN also conducts experiments on ActivityNet Captions that is comprised of richer scenes, and it even achieves competitive results with single-stage anchor-based methods, which further proves the effectiveness of captioning supervision and query-guided proposals. Since the videos in Charades-STA are of shorter lengths and contain less diverse activities, it is necessary to focus more on the metrics with higher IoU thresholds. SAP consistently outperforms those SW-based methods on Charades-STA with a higher IoU threshold, which attributes to its discriminative generated proposals and additional refinement process.

Single-stage method. For anchor-based (AB) methods, TGN achieves the lowest performance on TACoS and ActivityNet Captions. CMIN also performs poorly on TACoS. The common inferior accuracy achieved by TGN, CMIN and CBP may attribute to their single-stream anchor-based localization frameworks. With sequential RNNs, they fail to reason complex cross-modal relations on datasets (*i.e.*, TACoS and ActivityNet Captions) of longer video lengths. Instead of employing RNN-styled anchors, both SCDM and MAN use convolutional neural networks to better capture fine-grained interactions and diverse video contents of different temporal granularities, which can achieve better performance (*e.g.*, SCDM performs better than TGN/CMIN and TGN/CBP on TACoS and ActivityNet Captions, respectively). To make further improvement, 2D-TAN extends it to 2D feature maps

Just Accepted

to model the adjacent relations of various candidate moments of multi-anchors. SMIN and Zhang *et al.* [133] that adopt such a similar 2D structure modelling the relationships of candidate moments, also achieve superior results out of AB methods on TACoS, Charades-STA and ActivityNet Captions. Specifically, the model presented by Zhang *et al.* [133] performs the best on TACoS while SMIN has surpassed other methods on Charades-STA, which also prove the effectiveness of 2D moment relationship modelling. Furthermore, CSMGAN, SMIN and Zhang *et al.* [133] all achieve superior results on ActivityNet Captions. It is noted that although CSMGAN adopts the similar sequential RNN like TGN but it builds a joint graph for modelling the cross-/self-modal relations which can capture the high-order interactions between two modalities effectively.

For anchor-free (AF) methods, the overall performance is slightly behind that of AB methods especially with the challenging ActivityNet Caption dataset. More specifically, reading comprehension-inspired methods (ExCL, VSLNet, TMLGA, CPN, DORi, CI-MHA and PEARL) outperform other types of anchor-free methods with a significant gap. However, TMLGA achieves the lowest performance with the metrics of R@1,IoU=$\{0.3, 0.5\}$ on ActivityNet Captions. One possible reason is that the subjectivity of annotation is the hardest to model for this challenging dataset. The dense AF methods including DRN, GDP and DEBUG outperform the early sparse regression network ABLR, justifying the importance of increasing the number of positive training samples. However, the additional regression-based methods including PMI, HVTG and LGI achieve superior performance on ActivityNet Captions, which may result from more effective interaction between visual and textual contents. An obvious observation is that DORi achieves the highest grounding accuracy (with IoU=0.7) among all single-stage methods on TACoS and Charades-STA. Since its spatio-temporal graph is able to model more fine-grained object interactions that change over time. It is noted that L-Net has not been included in the table since the original paper [12] did not report the specific experimental values.

Additionally, other single-stage methods (BPNet, DPIN and CBLN) which can not be grouped into either anchor-based or anchor-free method achieve comparable results on three datasets (except DiDeMo). Specifically, CBLN achieves superior performance among all single-stage methods on Charades-STA and ActivityNet Captions, which quite highlights the advantages of combining anchor-based and anchor-free schemes and its special biaffine-based architecture.

RL-based method. Although the overall performance of RL-based methods can not reach that of traditional single-stage SOTA methods, they provide brand-new thoughts to address the TSGV task and the sequential decision-making process can also enhance the ability of interpretability. Particularly, TSP-PRL outperforms TropNet and R-W-M on ActivityNet Captions and Charades-STA, which may contribute to its tree-structured policy design inspired by the coarse-to-fine human-decision-making process. STRONG and AVMR achieve the best performance out of the RL-based frameworks on TACoS due to the effectiveness of spatial RL for scene tracking and the employment of adversarial learning, respectively. R-W-M, TripNet and SM-RL achieve relative inferior performance. Specifically, SM-RL achieves lowest performance on Charades-STA and TACoS while TripNet keeps the lowest performance on ActivityNet Captions.

Weakly supervised method. Due to the great challenge that temporal annotations of groundtruth moments are not available at training stage for weakly supervised (WS) methods. The experimental results on Charades-STA and ActivityNet Captions are apparently not as good as above fully-supervised ones. We cannot tell which framework (*i.e.*, MIL-based or reconstruction-based) has absolute advances according to their overall performances. But among all WS methods, CCL, VCA and CRM achieve superior performance on both Charades-STA and ActivityNet Captions. The results are also competitive compared with those of other fully supervised methods. To investigate the reasons, one finding is that they all design special training objectives that help in better visual-semantic alignment even without the annotated boundary information. Specifically, CCL is able to construct fine-grained supervision signals from counterfactual results for the contrastive training. Meanwhile, VCA re-defines the TSGV problem and design a new loss for visual co-occurrence alignment learning. Moreover, CRM

Just Accepted

minimizes the mismatched sentence-moment pairs during training by expanding the scope to paragraph level that can further consider the temporal ordering between sentences.

DiDeMo evaluation results with particular metrics. As aforementioned, MCN [38] reports the results on DiDeMo dataset under the IoU threshold m=1.0. Some works [11, 63, 128] following MCN also adopt such metrics to assess their models. We supplementally list the evaluation results (*i.e.*, R@{1, 5},IoU=1.0 and mIoU) on DiDeMo shown in Table 4, which are grouped by the supervision manner. Specifically, LoGAN as a WS method achieves the best performance among both fully and weakly supervised methods, which is also due to the effective visual-semantic representation learning via a latent graph co-attention network. Another observation is that the top-1 recall values for all fully supervised methods are constrained into a certain small range (22%–27%). It demonstrates that DiDeMo can not greatly differentiate the performance of methods, which may result from its limitation of taking pre-defined segments as groundtruth.

Table 4. The evaluation results on DiDeMo (The IoU threshold m = 1.0).

Type	Method	R@1	R@5	mIoU
Fully supervised	TMN [53]	22.92	76.08	35.17
	TGN [11]	24.28	71.43	38.62
	VLG-Net [85]	25.57	71.65	-
	MCN [38]	28.1	78.21	41.08
	MAN [128]	27.02	81.7	41.16
Weakly supervised	TGA [67]	12.19	39.74	24.92
	VLANet [63]	19.32	65.68	25.33
	WSLLN [31]	19.4	53.1	25.4
	FSAN [101]	19.4	57.85	31.92
	RTBPN [137]	20.79	60.26	29.81
	LoGAN [91]	39.2	64.04	38.28

4 DISCUSSIONS

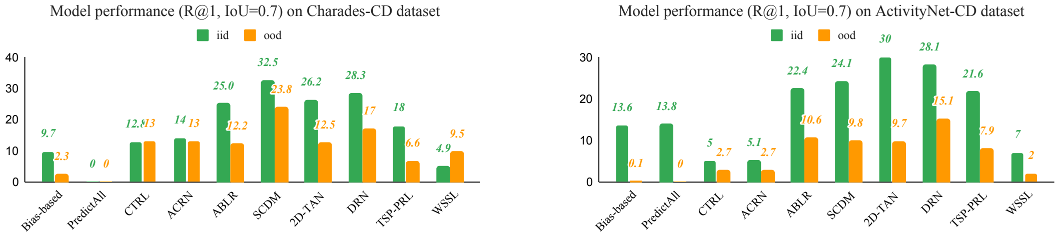
In this section, we discuss the limitations of current benchmarks and point out several promising research directions for TSGV. Firstly, we comprehensively divide these limitations into three categories, *i.e.*, the temporal annotation biases and ambiguous groundtruth annotations in public datasets, and the problematic evaluation metrics. These limitations may heavily mislead the TSGV approaches since each proposed method should be evaluated with these benchmarks. Meanwhile, we also present a couple of recent efforts to address these issues with proposing new datasets/metrics or proposing new methods. Then, we point out some promising research directions of TSGV including four typical tasks, *i.e.*, large-scale video corpus moment retrieval, spatio-temporal localization, audio-enhanced localization and video-language pre-training. We hope these research advances can provide more insights for future TSGV explorations, and thus further promote the development in this area.

4.1 Limitations of Current Benchmarks

Despite the promising results which have been made in TSGV, there are also some recent works [70, 121] doubting the quality of current datasets and metrics: (1) The joint distributions of starting and ending timestamps of target video segments are strongly biased and almost identical in the training and test splits of current datasets. Without truly modelling the video and sentence data, and just fitting such distribution biases in the training set, some baselines can still achieve good results and even outperform some well-designed methods. (2) The annotation of groundtruth segment location for TSGV is ambiguous and subjective, and may influence the model evaluation. (3) Current evaluation metrics are easily deceived by the above annotation biases in current datasets, and cannot measure the model performance effectively. Since TSGV is heavily driven by these datasets and evaluation metrics, such problematic benchmarks will influence the research progress of TSGV, and further mislead this research direction. In the following, we will detail the limitations on existing datasets and evaluation metrics, and present some recent solutions to address these issues.

Annotation distribution biases in datasets. Some recent studies [70, 121] attempt to visualize the temporal location distribution of groundtruth segments, finding joint distributions of starting and ending timestamps

Just Accepted



(a) (b)

Fig. 15. Performance of SOTA TSGV methods on re-organized data splits, figure adapted from [121].

of groundtruth segments identical in training and test sets with obvious distribution biases. They design some simple model-free methods, for example, a bias-based method [121], which samples locations from the observed training distribution and takes them as predicted locations of target segments at inference stage. This bias-based method can achieve good performance even surpassing some well-designed deep models, without any valid visual and textual inputs. Further, Yuan *et al.* [121] re-organize two benchmark datasets for out-of-distribution test. They create two different test sets: one test set follows the identical temporal location distribution with the training set, namely test-iid, and the other test set that has quite different distribution with the training set, namely test-ood. After comparing the experimental results of various baseline methods on these two test sets, they find that for almost all methods, the performance on test-ood drops significantly (*c.f.*, Fig. 15), which indicates that existing methods are heavily influenced by temporal annotation biases and do not truly model the semantic matching relationship between videos and texts. Thus, it is crucial for future works to construct de-biased datasets and build robust models unaffected by biases. Recently, there have been some attempts to address this issue. For example, Yang *et al.* [112] design a causal-inspired framework based on CTRL and 2D-TAN, which attempts to eliminate the spurious correlation between the input and prediction caused by hidden confounder (*i.e.*, the temporal location of moments).

Moreover, it is worth noting that there are some de-biased works [69, 142] that concentrate on other kinds of biases in TSGV instead of the moment annotation distribution biases. Zhou *et al.* [142] are devoted to dealing with the biases caused by single-style of annotations. The proposed DeNet with a debiasing mechanism can produce diverse yet plausible predictions. Nan *et al.* [69] propose an approach to approximate the latent confounder set distribution based on the theory of causal inference to deconfound selection biases introduced by datasets (*e.g.*, in datasets, it appears more often that a person is holding a vacuum cleaner than a person is repairing a vacuum cleaner).

Ambiguity of groundtruth annotation. One recent study [70] also mentions the ambiguous and inconsistent annotations among current TSGV datasets. Annotating the target segment location of the provided sentence query is a quite subjective task. In some cases, one query can be matched with multiple segments in videos, or different annotators will make different decisions on the grounded location of the sentence query. Therefore, only using one single groundtruth to evaluate the temporal grounding results is problematic. Otani *et al.* [70] suggest to re-annotate the benchmark datasets with multiple groundtruth moments for one given sentence query if exists, as shown in Fig. 16, they ask five annotators to re-annotate a video from ActivityNet Captions given the query “a woman is doing somersaults and big jumps alone”. These five re-annotated segments corresponding to the query are totally different and do not overlap with the groundtruth segment, justifying the ambiguity

Just Accepted

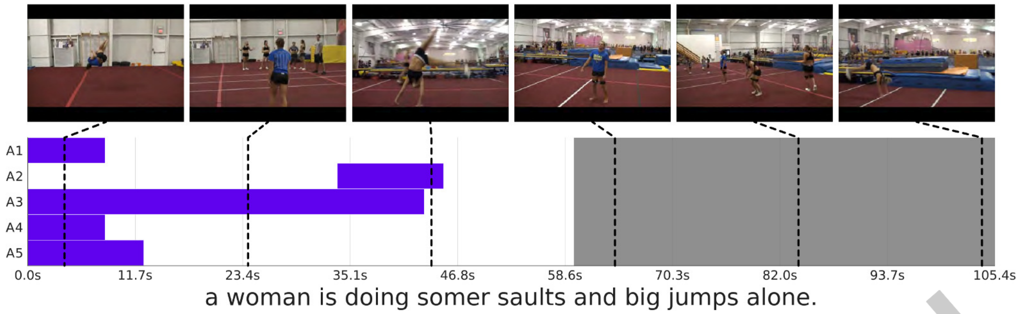


Fig. 16. The re-annotation example for ActivityNet Captions. The annotators annotate five different positive segments (shown as blue bars), all of which match the given query. While the original groundtruth segment is represented as grey bar, figure from [70].

Just Accepted

and subjectivity of groundtruth annotations. They further present two alternative evaluation metrics that take multiple annotated groundtruth moments into consideration.

Limitation of evaluation metrics. Besides the temporal annotation biases in current dataset, Yuan *et al.* [121] also find that some characteristics of the datasets may have negative effects on model evaluation. Most of previous TSGV methods [11, 58, 108, 124, 134] report their scores on some small IoU thresholds like $m \in \{0.1, 0.3, 0.5\}$. However, for ActivityNet Captions, a substantial proportion of groundtruth moments are of quite long lengths. Statistically, 40%, 20%, and 10% of sentence queries refer to a moment occupying over 30%, 50%, and 70% of the length of the whole video, respectively. Such annotation biases can obviously increase the chance of correct prediction under small IoU thresholds. Taking an extreme case as an example, if the groundtruth moment is the whole video, any predictions with duration longer than 0.3 can achieve R@1,IoU=0.3=1. Thus, the metric R@n, IoU=m with small m is unreliable for current biased annotated datasets. Therefore, to alleviate the above effects, they present a new metric namely discounted-R@n, IoU=m. This new metric considers that the hit score (*i.e.*, $r(n, m, q_i)$) for each positive sample i should not be limited to $\{0, 1\}$. It can be a real number $\in [0, 1]$ depending on the relative distances between the predicted and groundtruth boundaries. The formal definition for each sample i is as follows:

$$r(n, m, q_i) = (1 - \mathrm{nDis}(g_i^s, p_i^s)) \times (1 - \mathrm{nDis}(g_i^e, p_i^e)), \tag{14}$$

where the nDis operation calculates the distance between the groundtruth and predicted boundaries normalized to $[0, 1]$ by the video length. $(g_i^s, g_i^e)/(p_i^s, p_i^e)$ indicates the (start,end) timestamps of the groundtruth/predicted segment for sample i.

4.2 Promising Research Directions

We point out some promising research directions, including four TSGV-related tasks based on TSGV.

4.2.1 Large-scale video corpus moment retrieval. Large-scale video corpus moment retrieval (VCMR) is a research direction extended from TSGV that has been explored over the past few years [27, 49, 80, 127, 129]. It has more application value since it can retrieve the target segment semantically corresponding to a given text query from a large-scale video corpus (*i.e.*, a collection of untrimmed and unsegmented videos) rather than from a single video. As compared with TSGV, VCMR has higher efficiency requirements since it not only needs to retrieve a specific segment from one single video but also locates the target video from a video corpus.

Escorcia *et al.* [27] first extend TSGV to VCMR, introducing a model named Clip Alignment with Language (CAL) to align the query feature with a sequence of uniformly partitioned clips for moment composing. Lei *et al.* [49] introduce a new dataset for VCMR called TVR, which is comprised of videos and their associated subtitle texts. A Cross-modal Moment Localization (XML) network with a novel convolutional start-end detector module is also proposed to produce moment predictions in a late fusion manner. Zhang *et al.* [127] present a hierarchical multi-modal encoder (HAMMER) to capture both coarse- and fine-grained semantic information from the videos and train the model with three sub-tasks (*i.e.*, video retrieval, segment temporal localization, and masked language modelling). Zhang *et al.* [129] introduce contrastive learning for VCMR, designing a retrieval and localization network with contrastive learning (ReLoCLNet).

4.2.2 Spatio-temporal localization. Spatial-temporal sentence grounding in videos is another extension from TSGV which mainly localizes the referring object/instance as a continuing spatial-temporal tube (*i.e.*, a sequence of bounding boxes) extracted from an untrimmed video via a natural language description. Since fine-grained labeling process of localizing a tube (*i.e.*, annotate a spatial region for each frame in videos) for STSGV is labor-intensive and complicated, Chen *et al.* [21] propose to solve this task in a weakly-supervised manner which only needs video-level descriptions, with a newly-constructed VID-sentence dataset. Besides, VOGNet [79] commits to address the task of video object grounding, which grounds objects in videos referred to the natural language descriptions, and constructs a new dataset called ActivityNet-SRL. Zhang *et al.* [139] propose a spatio-temporal graph reasoning network (STGRN) for grounding multi-form sentences that depict an object and construct a new dataset VidSTG. Tang *et al.* [92] employ visual transformer to solve a similar task which aims to localize a spatio-temporal tube of the target person from an untrimmed video based on a given textural description with a newly-constructed HC-STVG dataset. Su *et al.* [89] further present a new STVGBert framework based on a visual-linguistic transformer to perform object tube predictions without any pre-trained object detectors.

4.2.3 Audio-enhanced localization. The current inputs for TSGV only contain the given sentence along with the untrimmed video. However, the audio signals are not effectively exploited, which may provide extra guidance for video localization, *e.g.*, the loud noise while using electronics in the kitchen or cheers from the audience when the football player kicks a goal. Such various forms of sounds do offer auxiliary but essential clues for more precise localization of the target moments, which has not been explored yet. Moreover, what people speak in videos can be converted into text with the Automated Speech Recognition (ASR) technique. The converted text also provides relevant information for the cross-modal alignment between video and the text query. Nowadays, there has been many works [39, 111] in visual-and-language area with audio-enhanced auxiliary proving its effectiveness for performance improvements. Thus, it is a promising future direction to embed the audio information for the TSGV task.

4.2.4 Video-language pre-training. Video-language pre-training [62, 71] has proven to improve many downstream text-based video understanding tasks, *e.g.*, video captioning, video question answering and video retrieval. Therefore, some pioneer works attempt to leverage video-language pre-training to benefit the TSGV task. Xu *et al.* [110] design boundary-aware proxy tasks to get boundary-sensitive video features for downstream localization, which can benefit many temporal localization tasks including TAL, TSGV, and step localization. Zeng *et al.* [126] introduce the graph pre-training upon their multi-modal relational graph to enhance the visual features with explicit relations. They design two node-level and graph-level self-supervised pre-training tasks (*i.e.*, attribute masking and cross-modal context prediction).

5 CONCLUSION

Temporal Sentence Grounding in Videos (TSGV) is a fundamental and challenging task connecting computer vision and natural language processing communities. It is also worth exploring since it can be seen as an

Just Accepted

intermediate task for some downstream video understanding applications such as video question answering, video summarization and video content retrieval.

In this survey, we take a systematic and insightful overview of the current research progress of the TSGV task, by categorizing existing approaches, benchmark datasets and evaluation metrics. The identified limitations of current benchmarks as well as our careful thoughts on promising research directions are also provided to researchers, aiming to further promote the development for TSGV. For future works, we suggest that i) more efforts should be made on proposing unbiased datasets and reliable metrics to better evaluate new methods for TSGV, and ii) models that are more robust and able to generalize well in dynamic scenarios should be paid with more attentions.

ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Program of China under grant No.2020AAA0106300 and National Natural Science Foundation of China No.62102222.

Just Accepted

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[3] Peijun Bao, Qian Zheng, and Yadong Mu. 2021. Dense events grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 920–928.

[4] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. 2017. End-to-End, Single-Stream Temporal Action Detection in Untrimmed Videos. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*.

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–970.

[6] Da Cao, Yawen Zeng, Meng Liu, Xiangnan He, Meng Wang, and Zheng Qin. 2020. STRONG: Spatio-Temporal Reinforcement Learning for Cross-Modal Video Moment Localization. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. 4162–4170.

[7] Da Cao, Yawen Zeng, Xiaochi Wei, Liqiang Nie, Richang Hong, and Zheng Qin. 2020. Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. 898–906.

[8] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 1–14.

[9] Cheng Chen and Xiaodong Gu. 2020. Semantic Modulation Based Residual Network for Temporal Language Queries Grounding in Video. In *International Symposium on Neural Networks*. Springer, 119–129.

[10] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1870–1879.

[11] 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.

[12] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8175–8182.

[13] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10551–10558.

[14] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. 2020. Learning modality interaction for temporal sentence localization and event captioning in videos. In *European Conference on Computer Vision*. Springer, 333–351.

[15] Shaoxiang Chen and Yu-Gang Jiang. 2019. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8199–8206.

[16] Shaoxiang Chen and Yu-Gang Jiang. 2020. Hierarchical Visual-Textual Graph for Temporal Activity Localization via Language. In *European Conference on Computer Vision*. Springer, 601–618.

[17] Shaoxiang Chen and Yu-Gang Jiang. 2021. Towards Bridging Event Captioner and Sentence Localizer for Weakly Supervised Dense Event Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8425–8435.

[18] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[19] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

[20] 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* abs/2001.09308 (2020).

[21] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. 2019. Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1884–1894.

[22] Howard Cheng and Xiaobo Li. 2000. Partial encryption of compressed images and videos. *IEEE Transactions on signal processing* 48, 8 (2000), 2439–2451.

[23] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3584–3592.

[24] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. 2021. ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross-and Intra-modal Knowledge Integration. In *Proceedings of the 29th ACM International Conference on Multimedia*. 797–806.

[25] Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. 2021. Support-set based cross-supervision for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11573–11582.

[26] Xuguang Duan, Wen-bing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly Supervised Dense Event Captioning in Videos. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. 3063–3073.

[27] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. 2019. Temporal localization of moments in video collections with natural language. *ArXiv preprint* abs/1907.12763 (2019).

[28] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1999–2007.

[29] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal Activity Localization via Language Query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 5277–5285.

[30] Junyu Gao and Changsheng Xu. 2021. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1523–1532.

[31] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. 2019. WSLLN:Weakly Supervised Natural Language Localization Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 1481–1487.

[32] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 245–253.

[33] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1984–1990.

[34] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. 580–587.

[35] Meera Hahn, Asim Kadav, James M. Rehg, and Hans Peter Graf. 2020. Tripping through time: Efficient Localization of Activities in Videos. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*.

[36] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. 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. 8393–8400.

[37] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337* (2018).

[38] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing Moments in Video with Natural Language. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 5804–5813.

[39] Chiori Hori, Huda AlAmri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Irfan Essa, Dhruv Batra, and Devi Parikh. 2019. End-to-end Audio Visual Scene-aware Dialog Using Multimodal Attention-based Video Features. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. 2352–2356.

Just Accepted

[40] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural Language Object Retrieval. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 4555–4564.

[41] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7199–7208.

[42] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. 2019. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the 2019 on international conference on multimedia retrieval*. 217–225.

[43] Yifan Jiao, Zhetao Li, Shucheng Huang, Xiaoshan Yang, Bin Liu, and Tianzhu Zhang. 2018. Three-dimensional attention-based deep ranking model for video highlight detection. *IEEE Transactions on Multimedia* 20, 10 (2018), 2693–2705.

[44] Svebor Karaman, Lorenzo Seidenari, and Alberto Del Bimbo. 2014. Fast saliency based pooling of fisher encoded dense trajectories. In *ECCV THUMOS Workshop*, Vol. 1. 5.

[45] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 787–798.

[46] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[47] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 706–715.

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

[49] 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.

[50] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8658–8665.

[51] Tianwei Lin, Xu Zhao, and Zheng Shou. 2017. Single Shot Temporal Action Detection. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. 988–996.

[52] 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.

[53] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2018. Temporal modular networks for retrieving complex compositional activities in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 552–568.

[54] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. 2020. Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network. In *Proceedings of the 28th International Conference on Computational Linguistics*. 1841–1851.

[55] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11235–11244.

[56] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly Cross- and Self-Modal Graph Attention Network for Query-Based Moment Localization. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. 4070–4078.

[57] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. 15–24.

[58] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. 843–851.

[59] Xinfang Liu, Xiushan Nie, Zhifang Tan, Jie Guo, and Yilong Yin. 2021. A survey on natural language video localization. *arXiv preprint arXiv:2104.00234* (2021).

[60] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. 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 (EMNLP-IJCNLP)*. 5144–5153.

[61] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[62] Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Hongyang Chao, and Tao Mei. 2021. CoCo-BERT: Improving Video-Language Pre-training with Contrastive Cross-modal Matching and Denoising. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5600–5608.

[63] 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.

Just Accepted

[64] Shugao Ma, Leonid Sigal, and Stan Sclaroff. 2016. Learning Activity Progression in LSTMs for Activity Detection and Early Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 1942–1950.

[65] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*. 533–542.

[66] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 202–211.

[67] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. 2019. Weakly Supervised Video Moment Retrieval From Text Queries. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 11592–11601.

[68] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. 10807–10816.

[69] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional Video Grounding with Dual Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2765–2775.

[70] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. 2020. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*.

[71] Yingwei Pan, Yehao Li, Jianjie Luo, Jun Xu, Ting Yao, and Tao Mei. 2020. Auto-captions on GIF: A Large-scale Video-sentence Dataset for Vision-language Pre-training. *arXiv preprint arXiv:2007.02375* (2020).

[72] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4594–4602.

[73] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6504–6512.

[74] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Fine-grained Iterative Attention Network for Temporal Language Localization in Videos. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. 4280–4288.

[75] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics* 1 (2013), 25–36.

[76] 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.

[77] Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, and Stephen Gould. 2021. DORi: discovering object relationships for moment localization of a natural language query in a video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1079–1088.

[78] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script data for attribute-based recognition of composite activities. In *European conference on computer vision*. Springer, 144–157.

[79] Arka Sadhu, Kan Chen, and Ram Nevatia. 2020. Video Object Grounding Using Semantic Roles in Language Description. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. 10414–10424.

[80] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. 2018. Find and focus: Retrieve and localize video events with natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 200–216.

[81] Aidean Sharghi, Jacob S Laurel, and Boqing Gong. 2017. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4788–4797.

[82] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 1049–1058.

[83] 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.

[84] Bharat Singh, Tim K. Marks, Michael J. Jones, Oncel Tuzel, and Ming Shao. 2016. A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 1961–1970.

[85] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. 2021. VLG-Net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3224–3234.

[86] Xiaomeng Song and Yahong Han. 2018. Val: Visual-attention action localizer. In *Pacific rim conference on multimedia*. Springer, 340–350.

[87] 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* abs/2003.07048 (2020).

[88] Jonathan C Stroud, Ryan McCaffrey, Rada Mihalcea, Jia Deng, and Olga Russakovsky. 2019. Compositional temporal visual grounding of natural language event descriptions. *arXiv preprint arXiv:1912.02256* (2019).

Just Accepted

[89] Rui Su, Qian Yu, and Dong Xu. 2021. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1533–1542.

[90] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

[91] 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.

[92] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2021. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology* (2021).

[93] Stefanie Tellex and Deb Roy. 2009. Towards surveillance video search by natural language query. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. 1–8.

[94] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 5998–6008.

[95] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. 2020. Dual Path Interaction Network for Video Moment Localization. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. 4116–4124.

[96] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 2021. Structured Multi-Level Interaction Network for Video Moment Localization via Language Query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7026–7035.

[97] Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12168–12175.

[98] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. 2021. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14090–14100.

[99] Limin Wang, Yu Qiao, and Xiaoou Tang. 2014. Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge* 1, 2 (2014), 2.

[100] Weining Wang, Yan Huang, and Liang Wang. 2019. Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 334–343.

[101] Yuechen Wang, Wengang Zhou, and Houqiang Li. 2021. Fine-grained semantic alignment network for weakly supervised temporal language grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 89–99.

[102] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. 2021. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1459–1468.

[103] Aming Wu and Yahong Han. 2018. Multi-modal Circulant Fusion for Video-to-Language and Backward. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. 1029–1035.

[104] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. 2020. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. 1283–1291.

[105] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. 2020. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12386–12393.

[106] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary Proposal Network for Two-Stage Natural Language Video Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2986–2994.

[107] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*. 1645–1653.

[108] 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.

[109] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[110] Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martinez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. 2021. Boundary-sensitive pre-training for temporal localization in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7220–7230.

[111] Yuecong Xu, Jianfei Yang, and Kezhi Mao. 2019. Semantic-filtered Soft-Split-Aware video captioning with audio-augmented feature. *Neurocomputing* 357 (2019), 24–35.

[112] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. 1–10.

Just Accepted

[113] Yulan Yang, Zhaohui Li, and Gangyan Zeng. 2020. A Survey of Temporal Activity Localization via Language in Untrimmed Videos. In *2020 International Conference on Culture-oriented Science & Technology (ICCST)*. IEEE, 596–601.

[114] Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 982–990.

[115] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*. 684–699.

[116] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE international conference on computer vision*. 4894–4902.

[117] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. 2016. End-to-End Learning of Action Detection from Frame Glimpses in Videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2678–2687.

[118] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1307–1315.

[119] Xinli Yu, Mohsen Malmir, Xin He, Jiangning Chen, Tong Wang, Yue Wu, Yue Liu, and Yang Liu. 2021. Cross interaction network for natural language guided video moment retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1860–1864.

[120] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6281–6290.

[121] Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. 2021. A closer look at temporal sentence grounding in videos: Dataset and metric. In *Proceedings of the 2nd International Workshop on Human-centric Multimedia Analysis*. 13–21.

[122] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. 534–544.

[123] Yitian Yuan, Tao Mei, Peng Cui, and Wenwu Zhu. 2017. Video summarization by learning deep side semantic embedding. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 1 (2017), 226–237.

[124] 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.

[125] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense Regression Network for Video Grounding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. 10284–10293.

[126] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. 2021. Multi-modal relational graph for cross-modal video moment retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2215–2224.

[127] Bowen Zhang, Hexiang Hu, Joonseok Lee, Ming Zhao, Sheide Chammas, Vihan Jain, Eugene Ie, and Fei Sha. 2020. A Hierarchical Multi-Modal Encoder for Moment Localization in Video Corpus. *ArXiv preprint* abs/2011.09046 (2020).

[128] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 1247–1257.

[129] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Video Corpus Moment Retrieval with Contrastive Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*.

[130] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based Localizing Network for Natural Language Video Localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6543–6554.

[131] Ke Zhang, Kristen Grauman, and Fei Sha. 2018. Retrospective encoders for video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 383–399.

[132] Lingyu Zhang and Richard J Radke. 2022. Natural language video moment localization through query-controlled temporal convolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 682–690.

[133] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. 2021. Multi-Stage Aggregated Transformer Network for Temporal Language Localization in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12669–12678.

[134] 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.

[135] Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1230–1238.

[136] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. 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, SIGIR 2019, Paris, France, July 21-25, 2019*. 655–664.

Just Accepted

[137] 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 *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020.* 4098–4106.

[138] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. *Advances in Neural Information Processing Systems* 33 (2020), 18123–18134.

[139] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020.* 10665–10674.

[140] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2018. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 7405–7414.

[141] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. 2021. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4197–4206.

[142] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. 2021. Embracing Uncertainty: Decoupling and De-bias for Robust Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8445–8454.

Just Accepted