Exploring Language Hierarchy for Video Grounding

Xinpeng Ding, Nannan Wang, *Member, IEEE*, Shiwei Zhang, Ziyuan Huang, Xiaomeng Li, *Member, IEEE*, Mingqian Tang, Tongliang Liu, *Senior Member, IEEE*, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— The understanding of language plays a key role in video grounding, where a target moment is localized according to a text query. From a biological point of view, language is naturally hierarchical, with the main clause (predicate phrase) providing coarse semantics and modifiers providing detailed descriptions. In video grounding, moments described by the main clause may exist in multiple clips of a long video, including both the ground-truth and background clips. Therefore, in order to correctly discriminate the ground-truth clip from the background ones, this co-existence leads to the negligence of the main clause, and concentrate the model on the modifiers that provide discriminative information on distinguishing the target proposal from the others. We first demonstrate this phenomenon empirically, and propose a Hierarchical Language Network (HLN) that exploits the language hierarchy, as well as a new learning approach called Multi-Instance Positive-Unlabelled Learning (MI-PUL) to alleviate the above problem. Specifically, in HLN, the localization is performed on various layers of the language hierarchy, so that the attention can be paid to different parts of the sentences, rather than only discriminative ones. Furthermore, MI-PUL allows the model to localize background clips that can be possibly described by the main clause, even without manual annotations. Therefore, the union of the two proposed components enhances the learning of the main clause, which is of critical importance in video grounding. Finally, we evaluate that our proposed HLN can plug into the current methods and improve their performance. Extensive experiments on challenging datasets show HLN significantly improve the state-of-the-art methods, especially achieving 6.15% gain in terms of *Recall*1@*IoU*0.5 on the TACoS dataset.**

***Index Terms*— Video and language, video understanding, language hierarchy.**

Manuscript received 18 September 2021; revised 22 April 2022; accepted 21 June 2022. Date of publication 6 July 2022; date of current version 12 July 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grant 61922066, Grant 61876142, Grant 62036007, and Grant 62176195; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; and in part by the Open Research Projects of Zhejiang Lab under Grant 2021KG0AB01. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ajmal S. Mian. *(Corresponding authors: Nannan Wang; Shiwei Zhang.)*

Xinpeng Ding is with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China, and also with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: xdingaf@connect.ust.hk).

Nannan Wang is with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: nnwang@xidian.edu.cn).

Shiwei Zhang and Mingqian Tang are with Alibaba Group, Hangzhou, Zhejiang 311100, China (e-mail: zhangjin.zsw@alibaba-inc.com; mingqian.tmq@alibaba-inc.com).

Ziyuan Huang is with the Advanced Robotics Centre, National University of Singapore, Singapore 117543 (e-mail: ziyuan.huang@u.nus.edu).

Xiaomeng Li is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China, and also with The Hong Kong University of Science and Technology Shenzhen Research Institute, Shenzhen 518057, China (e-mail: eexmli@ust.hk).

Tongliang Liu is with the Trustworthy Machine Learning Laboratory, Faculty of Engineering, School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: tongliang.liu@sydney.edu.au).

Xinbo Gao is with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gaoxb@cqupt.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3187288

I. Introduction

VIDEO grounding is a popular yet challenging topic in the computer vision field [1]–[3]. Due to its various applications in video understanding [4], [5], video retrieval [6], [7], and human-computer interaction [8], [9], it has drawn growing attention from both industry and academia.

The purpose is to predict the start and end of the target moment in an untrimmed long video given a natural language query, in which the understanding of language is of crucial importance [10], [11]. From a biological perspective, human comprehends language in a hierarchical structure [12]–[14]. This reveals the two steps to understand language for humans: (i) the attention is first paid to the main clause [1] of the sentence to get a coarse yet key information; (ii) the rest of the sentence that provides further details is then considered for fine and precise information, *i.e.*, the modified phrases. This property allows easy and meaningful expression of language [14].

In video grounding, the main clauses and other modified phrases are complementary and indispensable. For instance, the main clauses provides a coarse but critical localization to acquire several proposals in a long video and then the modified phrases give the discriminative information to select the best matching moment from the proposals. Most of the current video grounding methods [1], [2], [16]–[18] are generally in a *comparison-and-selection* way, which compares the total text query with all candidate proposals and select the best matching proposal. For instance, Gao *et al.* [2] first generate proposals by a set of fixed sliding windows. After comparing the proposals with the text query, a ranking score for each proposal is obtained and the target moment is selected with the highest score. 2D-TAN [17] adopts a 2D temporal map to model temporal anchors, which can extract the temporal relations among candidate proposals. However, the visual

[1] In English grammar, main clause, also known as independent clause, is a group of words made up of a subject and a predicate that together express a complete concept [15]. In this task, we regard it as the predicate phrase.

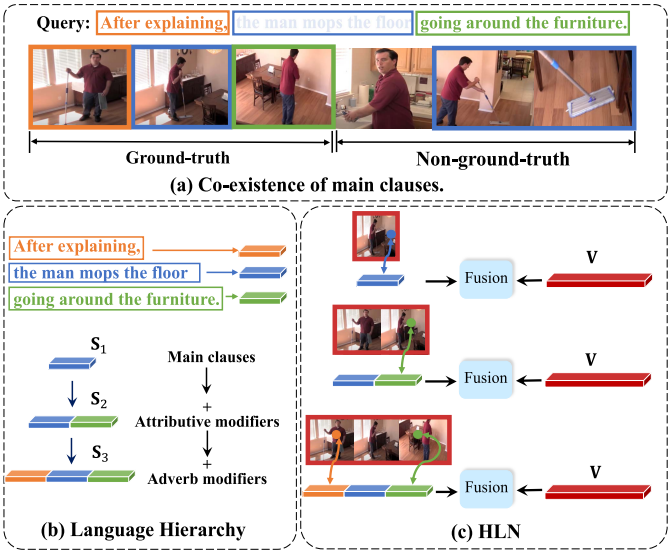


Fig. 1. (a) Co-existence of main clauses. The moments described by the main clause may simultaneously exist in ground-truth and background intervals, which may encourage the model to ignore the main clause; (b) Language hierarchy. We decompose the sentence into three top-down levels; (c) The proposed HLN, which improves language understanding for video grounding.

contents that can be described by the main clauses may occur in multiple clips (including the non ground truth clips) of a long real-world video, as shown in Fig. 1 (a). We regard the above phenomenon as the co-existence of main clauses. This leads to the negligence of the main clauses: to distinguish the target moment from the background clips, the discriminative parts (*i.e.*, modified phrases) in the text query is emphasized, while the main clause that only provides coarse semantics is neglected, as shown in Fig. 2.

We conduct the experiments using 2D-TAN [17] on TACoS [19]. In fact, the unique variable factor is input content, while the training and testing settings follow [17]. In Fig. 2 (a), the blue and green bars indicate we train and test with only MC and non-MC inputs respectively, which show the effect of MC. In Fig. 2 (b), the blue and green bars mean we test with MC and non-MC inputs, but the model are learned with whole sentence, which show the model tends to ignore the MC.

To address the negligence of main clauses, in this paper, we explore the hierarchical language for video grounding by devising a novel **H**ierarchical **L**anguage **N**etwork (HLN) composed of a language hierarchy and a new learning approach named **M**ulti-**I**nstance **P**ositive-**U**nlabeled **L**earning (MI-PUL). Through language hierarchy, the model is enforced to pay more attention to different phrases of the query and enhance the critical place of the main clause. Specifically, the whole sentence is first decomposed into separate phrases, *i.e.*, the main clause, the attributives and the adverbs, as shown in Fig. 1 (b). Then, the hierarchical textual information are constructed by sequentially connecting the modifiers to the main clause, which is similar to the procedure of human comprehension. As Fig. 1 (c) shows, we finally fuse the sentence at each level with the video information to extract the relations between text-video modalities for latter moment localization individually, so that different levels would focus on different phrases.

However, the sentences in the top two levels, *e.g.*, the main clause, may exist in both ground-truth and non ground-truth clips; see blue boxes in Fig. 1 (a). Hence, simply regarding ground-truth as supervision for top two levels may be not reasonable. Regarding the moments corresponding to main clauses as the positive proposals, the positive samples may exist in both ground-truth and non ground-truth ones, which is known as the Positive-Unlabeled Learning (PUL) [20]–[23] problems. Furthermore, the ground truth clips may consist of both positive and negative examples and contain at least one positive examples, which is corresponding to Multi-Instance Learning (MIL) [24], [25]. Hence, we propose a new learning approach, *i.e.*, MI-PUL, to allow the positive proposals in non ground truth clips, even without manual annotations of the moments. Concretely, based on the ground-truth (GT), we construct the bag of positive candidates which contains at least one positive samples and estimate the risk of this bag in a MIL manner. Then, we adopt importance re-weighting methods [21], [26], which treat unlabeled samples as weighted negative ones. With the combination of LH and MI-PUL, our proposed method can enhance the attention of different phrases to mitigate the ignoring of the main clauses in the training stage. As shown in Fig. 2 (b), our method (scatters) can avoid the negligence of main clauses and show similar trends as Fig. 2 (a). To demonstrate the effectiveness of the proposed HLN, we conduct experiments on challenging TACoS [19], Charades-STA [27] and ActivityNet [28] datasets. We also prove that HLN is a general training strategy, which can be easily plugged into the state-of-the-art methods and improve their performance.

In summary, our contributions are four-fold:

- To the best of our knowledge, we are the first to point out the negligence of the main clauses in video grounding.
- We design a novel HLN to enforce the model to understand language semantics in hierarchy and hence concentrate on different phrases in each hierarchical level.
- We propose the multi-instance positive-unlabeled learning in the first two levels of the language hierarchy to localize the moments described by MC, even without annotations.
- We perform extensive experiments to validate the proposed HLN can improve state-of-the-art methods on three public datasets, especially gain 6.15% in terms of $Recall1@IoU0.5$ on the TACoS dataset for 2D-TAN.

II. Related Work

A. Video Grounding

Early approaches [1], [2], [29], [30] usually use a two-stage visual-textual matching strategy to tackle video grounding. Specifically, proposals are first generated by sliding windows with different fixed scales, then each proposal and the query sentence are interacted by the cross-modal module. Finally, a ranking score is generated for each proposal. Due to these proposals are query-irrelevant, large numbers of proposals need to be generated for best matching. To generate proposals

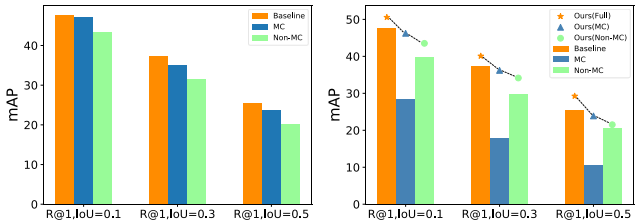


Fig. 2. Results demonstrating the negligence of main clauses. We split each query into two parts, main clauses (MC) and non main clauses (Non-MC). (a) The model is trained with MC/Non-MC and tested with MC. It shows that main clause carry crucially important information. (b) The model is trained with full sentences and tested with MC and Non-MC respectively. It indicates that the main clause is neglected by training with full sentence. In contrast, our approach avoids the negligence.

of high quality, SCDM [11] incorporates the query text into the visual feature for correlating and composing the sentence-related video contents over time. 2D-TAN [17] adopts a 2D temporal map to model temporal anchors, which can extract the temporal relations between video moments. To process more efficiently, recently, many one-stage methods [3], [10], [31]–[33] are proposed to predict starting and ending times directly. He *et al.* [32] formulate this task as a problem of sequential decision making by reinforcement learning, which learns an agent to regulates the temporal grounding boundaries. ABLR [10] proposes a cross-modal co-attention mechanism to generate both video and sentence attentions to highlight the sentence details for temporal localization. Zeng *et al.* [3] avoid the imbalance training by leveraging much more positive training samples, which improves the grounding performance. CBLN [34] adopts a biaffine mechanism to incorporate both local and global contexts into features. To reduce the query uncertainty and label uncertainty, Zhou *et al.* [35] propose a de-bias method to predict more diverse segments. Compared with the detection or regression pipelines, Wang *et al.* [36] propose a mutual matching network to compute the similarity between textual queries and video moments in a joint embedding space. However, existing comparison-selection methods [3], [17] tend to focus too much discriminative phrases, and ignore the main clauses which exists in both ground-truth and non ground-truth segments. In this paper, we propose a novel hierarchical language network to extract different semantics of the query and make the model focus on different phrases.

B. Language Modeling

Language Modeling is critical for video grounding. Previous works [37] mostly focus on decompose the sentence into word-level, phrase-level and sentence-level. Besides conducting the decomposition on sentences, Liu *et al.* [31] adopt different windows to encoded world-level embedding to obtain language features with different receptive fields. Recent works [10], [38]–[40] use attention mechanism to encourage the model to focus on different parts of sentences. Different from these methods, in this paper, our proposed HLN focuses on sentence structure from coarse to fine.

C. Positive-Unlabeled Learning

The standard supervised learning requires both positive and negative labels for training. However, in real world, only few positive samples would be obtained, while other samples are difficult to be collected [23]. For example, only diagnosed patient are be considered as the 'positive' (healthy), the much larger undiagnosed people are generally mixed with both 'positive' and 'negative' (patient) examples [41]. Directly setting the undiagnosed people as the negative ones would lead to biased classifiers. To tackle those practical problems, Positive-Unlabeled Learning (PUL) has achieved more and more attention in recent years. Early works [42], [43] tried to leverage semi-supervised or hand-crafted examining methods identify reliable negative examples from the unlabeled samples. Recently, importance re-weighting methods [21], [26] treating unlabeled data as weighted negative ones have achieved the state-of-the-art. Self-PU [23] seamlessly integrates PUL and self-training to learn capability of the model itself, which could have provided reliable supervision. Importance re-weighting methods requires the prior probability of the positive samples, which cannot be obtained in our task. In this paper, we estimate this prior probability via measuring the similarity between the ground-truth/non ground truth moments and text queries.

D. Multi-Instance Learning for Video Understanding

In multi-instance learning [24], samples are divided into two bags, named positive bags and negative bags respectively. In positive bags, there are at least one positive sample, while any negative bag contains no positive samples. The algorithm for multi-instance learning aims to classify each sample to be positive or negative, based on these bags. The multi-instance learning has been introduced to many video understanding tasks, such as anomaly detection [44], weakly supervised action classification [45], [46], temporal action localization [47]–[51], and video-text pre-training [52]. In this paper, we introduce a novel learning approach MIL-PUL, which allows allow the positive proposals in non ground truth clips, even without manual annotations of the moments.

III. Proposed Method

In this section, we first introduce the basic definition of video grounding in Section III-A. Then, the video and text encoding procedures are presented in Section III-B. After that, a hierarchical language network and a learning approach, termed MI-PUL, are proposed to explore the language hierarchy in Section III-C.

A. Problem Definition

Video grounding is a challenging cross-modal task, which involves in the visual and textual modalities. Specifically, given an untrimmed video with L frames and a query sentence with N words, the purpose is to localize the best matching moment (t_s, t_e) in the video corresponding to the query sentence, where t_s and t_e denote the starting and ending times respectively.

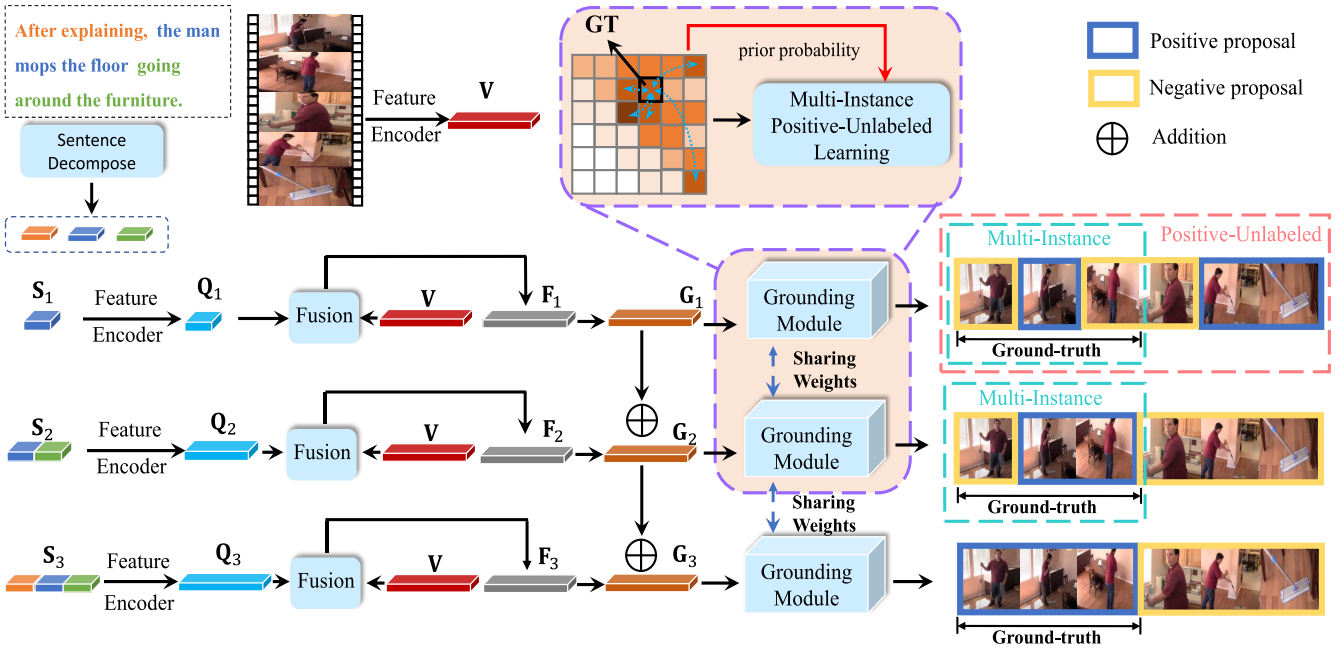


Fig. 3. Architecture of the proposed HLN. We first decompose the sentence into three top-down levels, defined as $\{\mathbf{S}_i\}_{i=1}^3$. After feeding them into the feature extractor, we obtain a set of textual features defined as $\{\mathbf{Q}_i\}_{i=1}^3$. Then, we fuse the textual features $\{\mathbf{Q}_i\}_{i=1}^3$ and visual features $\mathbf{V}$, and output the fused features $\{\mathbf{F}_i\}_{i=1}^3$. We adopt a FPN architecture to calculate $\{\mathbf{G}_i\}_{i=1}^3$, based on which we can perform prediction. During the train phase, we leverage the MI-PUL to find the moments occurring in both ground truth and background clips.

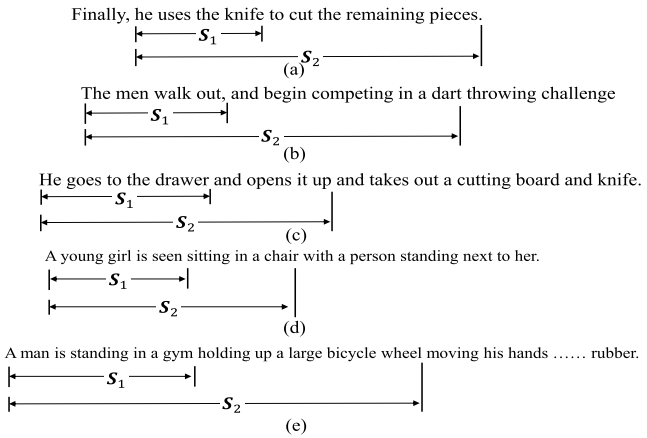


Fig. 4. Illustration of the decomposition of the complex sentences. The sentences are selected from TACoS [2] and ActivityNet-Captions [28]. For clarity, we just show $\mathbf{S}_1$ and $\mathbf{S}_2$. $\mathbf{S}_3$ is the whole sentence.

B. Video and Sentence Encoding

1) Video Encoding: The videos usually last several minutes in the task, hence it is hard to train the model end-to-end limited by the GPU memory. Therefore, we encode the videos in a segment-based manner like existing methods [11], [17]. A long untrimmed video is first divided into T non-overlap segments with a fixed length, then the features of each segment are extracted by a pre-trained CNN model, such as C3D [53] and VGG [54] models. Then, a new fully connected layer is employed to calculate the final feature vector. We denote by $\mathbf{v}_t \in \mathbb{R}^d$ the features of the t-th segment, where d is the channel. All the segment features are then stacked as $\mathbf{V} \in \mathbb{R}^{T\times d}$.

2) Sentence Encoding: To explore the hierarchical semantics of the query sentences, we first decompose each sentence into three parts according to the natural sentence structure, *i.e.*, *main clause*, *the attributives* and *the adverbs*, by using an off-the-shelf semantic role parsing toolkit [55]. Based on these parts, we construct the language hierarchies by combining the attributives and adverbs to the main clause sequentially. As shown in Fig. 1(b), the hierarchical architecture has three top-down levels, including (i) main clause ($\mathbf{S}_1$), (ii) main clause + attributives ($\mathbf{S}_2$) and (iii) complete sentence ($\mathbf{S}_3$), which corresponds to the coarse-to-fine semantics of the sentences. For the complex sentences, a main verb and some parallel verbs will be found by Stanza [56]. $\mathbf{S}_1$ is generated by selecting the sub-sentence that contains the main verb. Adding sub-sentence containing the verb closes to the main verb to S1 to obtain $\mathbf{S}_2$. $\mathbf{S}_3$ is the whole sentence. More examples are shown in Fig. 4.

Following the hierarchy construction, we need to extract feature representation for each level of the architecture. We adopt a GloVe word2vec model [57] generate the word embedding $\mathbf{w} \in \mathbb{R}^{d_w}$, and combine them in each textual hierarchy as $\{\mathbf{S}_i \in \mathbb{R}^{N_i\times d_w}\}_{i=1}^3$, where d_w and N_i denote the dimension and word number of the i-th level. We then feed $\mathbf{S}_i$ into a three-layer bidirectional LSTM network with d channels [58], and use the last hidden state as the feature representation of the i-th hierarchy, denoted as $\{\mathbf{Q}_i \in \mathbb{R}^d\}_{i=1}^3$.

C. Architecture of Our Method

The proposed method mainly contains two components, *i.e.*, a hierarchical language network and multi-instance positive-unlabelled learning, as shown in Fig. 3. The hierarchical net-

work is a top-down architecture by decomposing the sentence into three levels. For the second and the third level, the positive samples may exist in both GT and Non GT, while GT may also contains positive ones, we design a novel learning approach named multi-instance positive-unlabelled learning to address the problem. Due to the content described by the main clause may simultaneously occur within and outside the ground-truth, simply taking the clips within ground-truth as positive samples and others as negative samples tends to encourage the model to ignore the main clause. Our proposed HLN can alleviate this ignorance by following two means: 1) feature enhance. We individually encode $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ to obtain $\mathbf{F}_1$, $\mathbf{F}_2$ and $\mathbf{F}_3$, and embed $\mathbf{F}_1$ and $\mathbf{F}_2$ onto $\mathbf{F}_3$. Thus, the $\mathbf{S}_3$ stream can be enhanced to focus on different parts of the query; 2) sharing weights. The Grounding Module shares weights among $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$. Meanwhile, we apply MI-PUL on $\mathbf{S}_1$ and $\mathbf{S}_2$ streams to avoids the shortcut of the Grounding Module, which can further improve the $\mathbf{S}_3$ stream. Following we will present the hierarchical language network and multi-instance positive-unlabelled learning in detail.

1) Hierarchical Language Network: As shown in Section III-B, we decompose sentences in hierarchy and extract the hierarchical language representations $\{\mathbf{Q}_i \in \mathbb{R}^d\}_{i=1}^3$, and extract the video features $\mathbf{V} \in \mathbb{R}^{T\times d}$. Then, we use the fusion module to interact the language representations with the video features in each hierarchy. After that, we combine the fused features from all hierarchies for following grounding. In this paper, the fusion module is conducted following 2D-TAN [17]. Specifically, we generate a 2D temporal map on $\mathbf{V}$, defined as $\mathbf{M} \in \mathbb{R}^{T\times T\times d}$, where the first dimension indicates the starting time and the second dimension indicates the end time. For example, the coordinate (i, j) of the temporal map represents the candidate moment starting at $i \cdot \tau$ and ending at $(j+1)\cdot\tau$, where τ is the time interval. Then, we fuse the 2D temporal feature map with sentence feature $\mathbf{Q}_i$ as follows:

$$\mathbf{F}_i = \|(\mathbf{w}^q \cdot \mathbf{Q}_i \cdot \mathbf{1}^\top) \odot (\mathbf{w}^m \cdot \mathbf{M})\|_F, \tag{1}$$

where $\mathbf{w}^q$ and $\mathbf{w}^m$ represent the learned parameters, $\mathbf{1}^\top$ is the transpose of an all-ones vector, $\odot$ indicates Hadamard product, and $\|\cdot\|$ is Frobenius normalization.

To explore the best configuration of HLN, we design different fusion and prediction ways in Fig. 5. One of the fusion method is indicated in Fig. 5 (a), which fuses video and sentence features in each level separately, terms as Individual Fusion (IF). The second fusion way is Joint Fusion (JF), which fuse hierarchical features $\mathbf{Q}_i$ with fused features from last level, as is shown Fig. 5 (b). The experimental results in Table III shows that each hierarchy should focus on the individual information. Four kinds of prediction modules are indicated in Fig. 5 (c)-(f), and we prove the effectiveness of multi-level grounding. See details in Section IV-C.3.

Based on the fused feature maps $\{\mathbf{F}_i\}_{i=3}^3$, we adopt FPN [59] to combine them to get $\{\mathbf{G}_i\}_{i=1}^3$, which are latter fed into the grounding module. The grounding module consists of several convolutional layers and the last layer predicts the matching scores for the 2D temporal map (See details in 2D-TAN [17]).

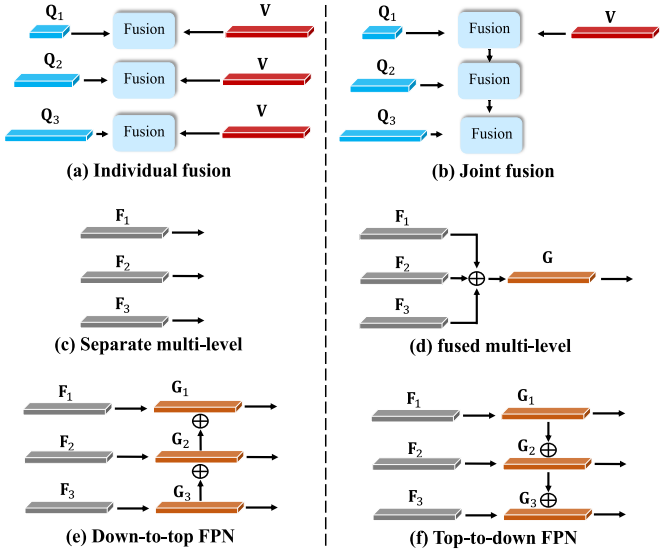


Fig. 5. Illustration of (a) individual fusion, (b) joint fusion, (c) separate multi-level, (d) fused multi-level, (e) down-to-top FPN and (f) top-to-down FPN. Note that $\mathbf{Q}_i$ and $\mathbf{V}$ mean the textual features at the i-th level and visual features, which have been defined in Section III-B. $\{\mathbf{F}_i\}_{i=1}^3$ and $\{\mathbf{G}_i\}_{i=1}^3$ are the fused feature maps and the features obtained from FPN, which are defined in Section III-C.

Formally, we denote the valid scores of the temporal map in i-th level of the language hierarchy as $X_i = \{x_j^i\}_{j=1}^{l_i}$, where l_i is the total number of moment candidates. Each x_j^i represents the confidence of the corresponding moment matching the queried sentence. The maximum value indicates the best matching moment. Further analysis and evaluation of these modules will be conducted in the experimental section.

In the training phase, we adopt a normalized IoU value as the ground-truth label rather than a hard binary score which is applied in 2D-TAN. We compute temporal IoU, defined as o_j^i, for each proposal with the ground truth. Then the IoU value o_j^i is min-max normalized by two hyperparameters t_{min} and t_{max}, which is formulated as:

$$y_j^i = \begin{cases} 0 & o_j^i < t_{\min} \\ \dfrac{o_j^i - t_{min}}{t_{\max} - o_j^i} & t_{\min} \le o_j^i \le t_{\max} \\ 1 & o_j^i > t_{\max} \end{cases} \tag{2}$$

Then, the objective function of the i-th levels is defined as:

$$\mathcal{L}_{vg}^i = -\frac{1}{l_i}\sum_{j=1}^{l_i} y_j^i \log x_j^i + (1-y_j^i)\log(1-x_j^i), \tag{3}$$

where $Y_i = \{y_j^i\}_{j=1}^{l_i}$. Finally, the overall objective of three levels can be $\mathcal{L}_{all} = \sum_{i=1}^3 \mathcal{L}_{vg}(X^i, Y^i)$. In inference stage, we select the highest score in all valid candidates as the matched moment given the query.

2) Multi-Instance Positive-Unlabeled Learning: As shown in the top in Fig. 6, the sentences (*e.g.,* "The man mops the floor") in the top two levels are only a sub-sentence of the original text query, their corresponding moments (blue boxes) may exist in both ground-truth and non ground-truth,

which can be regarded as a positive-unlabeled learning problem. Furthermore, the ground-truth may contain both positive and negative candidates, *i.e.*, multi-instance learning. Hence, we proposed MI-PUL to address the above problems. In the this section, we first review the problem settings and risk estimators in PUL. Then we introduce the proposed multi-instance positive-unlabeled learning method.

a) Problem settings: In PUL, the training samples $\mathbf{M}$, which corresponds to the 2D map in our method, consists of a positive set $\mathbf{M}_p$ and an unlabeled set $\mathbf{M}_u$, where we have $\mathbf{M} = \mathbf{M}_p \cup \mathbf{M}_u$. $\mathbf{M}_p$ contains n_p positive examples m_p sampled form $P(m|Y = +1)$ and $\mathbf{M}_u$ contains n_u unlabeled examples m_u sampled form $P(m)$, where $Y \in \{+1, -1\}$ is the output random variables. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be the convolutional layers in the grounding module, *i.e.*, $X = g(\mathbf{M})$ and $\mathcal{L} : \mathbb{R} \times \{+1, -1\} \rightarrow \mathbb{R}$ be the loss function. Hence we have $X_p = g(\mathbf{M_p})$ and $X_n = g(\mathbf{M_n})$.

b) Risk estimators: The risk of g can be defined as $R(g) = \pi_p R_p^+(g) + \pi_n R_n^-(g)$, where $\pi_p = P(Y = +1)$ is the class-prior probability and $\pi_n = P(Y = -1) = 1 - \pi_p$. In this paper, we regard the number of proposals with high similarity to GT as the prior for the number of positive samples (see red square in Fig. 3). In positive and negative learning, since the availability of $\mathbf{M}_p$ and $\mathbf{M}_n$, $R(g)$ can be approximated by:

$$\begin{aligned}\widehat{R}_{pn}(g) &= \pi_p \widehat{R}_p^+(g) + \pi_n \widehat{R}_n^-(g) \\ &= \frac{\pi_p}{n_p} \sum_{x \in X_p} \mathcal{L}(x, +1) + \frac{\pi_n}{n_n} \sum_{x \in X_n} \mathcal{L}(x, -1), \quad (4)\end{aligned}$$

Due to the $\mathbf{M}_n$ is unavailable in PUL, we approximate $R_n^-(g)$ directly [21]. Since $\pi_n P_n(x) = P(x) - \pi_p P_p(x)$, $\pi_{\mathrm{n}} \widehat{R}_{\mathrm{n}}^-(g)$ can be obtained as follows:

$$\begin{aligned}\pi_{\mathrm{n}} \widehat{R}_{\mathrm{n}}^-(g) &= \widehat{R}_{\mathrm{u}}^-(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^-(g) \\ &= \frac{1}{n_u} \sum_{x \in X_u} \mathcal{L}(x, -1) - \frac{\pi_p}{n_p} \sum_{x \in X_p} \mathcal{L}(x, -1). \quad (5)\end{aligned}$$

Then, $\widehat{R}_{pu}(g)$ can be approximated indirectly by:

$$\begin{aligned}\widehat{R}_{pu}(g) = &\frac{\pi_p}{n_p} \sum_{x \in X_p} \mathcal{L}(x, +1) \\ &+ \frac{1}{n_u} \sum_{x \in X_u} \mathcal{L}(x, -1) - \frac{\pi_p}{n_p} \sum_{x \in X_p} \mathcal{L}(x, -1). \quad (6)\end{aligned}$$

Eq. 6 is known as the unbiased risk estimator [20], [22], [26]. Specifically, for any g, $\widehat{R}(g) \geq 0$ should be fixed. However, $\widehat{R}_{\mathrm{u}}^-(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^-(g) \geq 0$ is not always true. Hence, $\pi_{\mathrm{n}} \widehat{R}_{\mathrm{n}}^-(g)$ may be negative, which leads to the model to overfit [21]. To avoid this drawback, the non-negative version [21] of Eq. 6 is proposed to make sure $\pi_{\mathrm{n}} \widehat{R}_{\mathrm{n}}^-(g)$ to be non-negative, as follows:

$$\begin{aligned}&\{\pi_{\mathrm{n}} \widehat{R}_{\mathrm{n}}^-(g)\}^+ \\ &= \max\{0, \widehat{R}_{\mathrm{u}}^-(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^-(g)\} \\ &= \max\{0, \frac{1}{n_u} \sum_{x \in X_u} \mathcal{L}(x, -1) - \frac{\pi_p}{n_p} \sum_{x \in X_p} \mathcal{L}(x, -1)\}. \quad (7)\end{aligned}$$

Finally, the non-negative PUL can be formulated as:

$$\begin{aligned}&\widehat{R}_{nnpu}(g) \\ &= \pi_p \widehat{R}_p^+(g) + \{\pi_{\mathrm{n}} \widehat{R}_{\mathrm{n}}^-(g)\}^+ \\ &= \frac{\pi_p}{n_p} \sum_{x \in X_p} \mathcal{L}(x, +1) \\ &\quad + \max\{0, \frac{1}{n_u} \sum_{x \in X_u} \mathcal{L}(x, -1) - \frac{\pi_p}{n_p} \sum_{x \in X_p} \mathcal{L}(x, -1)\}. \quad (8)\end{aligned}$$

c) MI-PUL: Different from the standard PUL that all labeled samples are positive, labeled samples within the ground-truth may contain both positive and negative proposals in our task, as shown in Fig. 6. That is, we have no prior that where the main clause is in the video, but just know it must exist in the ground-truth moment. To tackle this problem, we propose MI-PUL to construct a bag of positive candidates for each ground-truth moment and its corresponding query pair. Formally, let t_s and t_e denote the starting and ending time of ground-truth. Then the bag can be defined as $B = \{(l_s, l_e) | l_s, l_e \in [t_s, t_e]; l_s \leq l_e\}$. For instance, as shown in Fig. 6, the coordinate of ground-truth on 2D map is $(1, 3)$ and the bag can be formed as $B = \{(1, 1), (1, 2), (1, 3), (2, 2), (2, 3), (3, 3)\}$. Compared to minimizing the first line in Eq. 8, multi-instance PUL aims to minimize the following risk:

$$\widehat{R}_g^+(g) = \frac{1}{|B|} \sum_{(i,j) \subseteq B} a_{i,j} \mathcal{L}(g(\mathbf{M}_{i,j}), +1), \quad (9)$$

where $M_{i,j}$ is the coordinate (i, j) on M and $a_{i,j} =< Q, M_{i,j} >$ is the similarity between Q and $M_{i,j}$. Based on B and $a_{i,j}$, we define the set of the score and similarity as $A = \{(x = g(\mathbf{M}_{i,j}), a_{i,j}) | (i, j) \subseteq B\}$. Based on Eq. 3 and Eq. 8, the objective function of the top two levels of language hierarchy can be rewrote as:

$$\begin{aligned}\mathcal{L}_{mipu}^i = &\frac{\pi_p^i}{|A^i|} \sum_{(x,a) \subseteq A^i} a \mathcal{L}_{ce}(x, 1) \\ &+ \max\{0, \frac{1}{n_u^i} \sum_{x \in X_n^i} \mathcal{L}_{ce}(x, 0) - \frac{\pi_p^i}{n_p^i} \sum_{x \in X_p^i} \mathcal{L}_{ce}(x, 0)\}, \quad (10)\end{aligned}$$

where $\mathcal{L}_{ce}(x, y) = ylogx + (1 - y)log(1 - x)$ is the cross entropy loss and π_p^i, A^i, n_u^i, X_n^i, X_p^i, X_n^i is the i-th level of π_p, A, n_u, X_n, X_p, X_n. Finally, the overall objective of our HLN is as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{vg}^3 + \lambda_1 \mathcal{L}_{mipu}^1 + \lambda_2 \mathcal{L}_{mipu}^2, \quad (11)$$

where λ_1 and λ_2 are hyperparameters to control the weight of MI-PUL objective in the first and second level respectively. In this paper, we set both of them to 1.0.

IV. Experiments

In this section, we evaluate our proposed HLN on three challenge datasets, including TACoS [19], Charades-STA [27] and ActivityNet Captions [60].

In this section, we first introduce these datasets and our implementation details, and then compare the performance

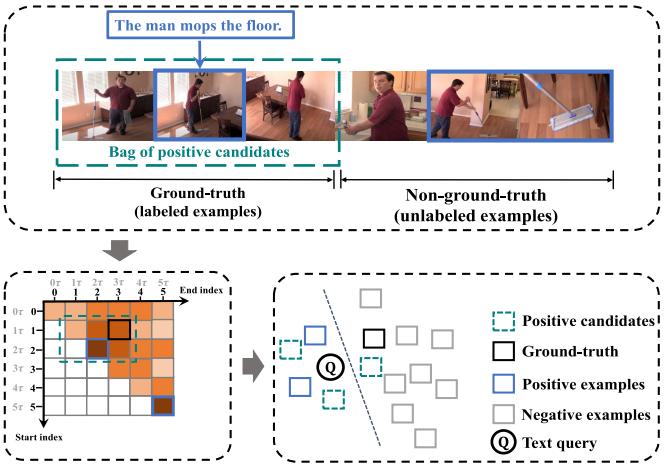


Fig. 6. Illustration of MI-PUL. Since the GT may contain both positive and negative proposals, we combine all proposals within the GT to construct the bag of positive candidates that may include positive and negative ones rather than only set GT as positive samples.

TABLE I

ABLATION STUDY OF LANGUAGE HIERARCHY ON THE TACOS DATASET. THE SCORES AT $Rank1@0.5$ ARE PRESENTED

Models	$\mathbf{F}_1$	$\mathbf{F}_2$	$\mathbf{F}_3$	$Rank1@0.5$
Baseline_1	$\mathbf{V}\&\mathbf{Q}_3$	✗	✗	25.55
Baseline_2	$\mathbf{V}\&\mathbf{Q}_3$	$\mathbf{V}\&\mathbf{Q}_3$	$\mathbf{V}\&\mathbf{Q}_3$	25.43
HLN_1	$\mathbf{V}\&\mathbf{Q}_1$	✗	✗	23.62
HLN_2	$\mathbf{V}\&\mathbf{Q}_2$	✗	✗	24.55
HLN_3	$\mathbf{V}\&\mathbf{Q}_1$	$\mathbf{V}\&\mathbf{Q}_2$	✗	27.29
HLN_4	$\mathbf{V}\&\mathbf{Q}_1$	$\mathbf{V}\&\mathbf{Q}_3$	✗	27.99
HLN_5	$\mathbf{V}\&\mathbf{Q}_2$	$\mathbf{V}\&\mathbf{Q}_3$	✗	27.56
HLN_6	$\mathbf{V}\&\mathbf{Q}_1$	$\mathbf{V}\&\mathbf{Q}_2$	$\mathbf{V}\&\mathbf{Q}_3$	**29.26**

of our method with other state-of-the-art approaches. Finally, we investigate the impact of different components via a set of ablation studies.

A. Datasets

1) TACoS: TACoS is collected by Regneri *et al.* [19] which consists of 127 videos on cooking activities with an average length of 4.79 minutes for video grounding and dense video captioning tasks. There are 18,818 video-query pairs for video grounding task, and each video contains corresponding 148 queries on average. Due to temporal duration of moments over only a few seconds even a few frames, TACoS dataset is very challenging. We follow the same split of the dataset as Gao *et al.* [2] for fair comparisons, which has 10,146, 4,589, and 4,083 video-query pairs for training, validation, and testing respectively.

2) Charades-STA: Charades is originally collected for daily indoor activity recognition and localization [27], which consists of 9,848 videos. Gao *et al.* [2] build the Charades-STA by annotating the temporal boundary and sentence description of Charades [27], where there are 12,408 video-query pairs in the training set, and 3,720 video-query pairs in the test set.

TABLE II

ABLATION STUDY OF DIFFERENT SENTENCE DECOMPOSITION ON THE TACOS DATASET IN TERMS OF $Rank1@0.5$

$\mathbf{F}_1$			$\mathbf{F}_2$			$\mathbf{F}_3$	$Rank1@0.5$
MC	ADP	ATP	MC	ADP	ATP	Full	
✓			✓	✓		✓	**29.26**
✓			✓		✓	✓	29.10
✓				✓	✓	✓	28.54
	✓		✓	✓		✓	27.97
	✓			✓	✓	✓	26.59
	✓		✓		✓	✓	26.45
		✓	✓		✓	✓	26.83
		✓		✓	✓	✓	26.17
		✓	✓	✓		✓	26.08

3) ActivityNet-Captions: ActivityNet [28] is a large-scale dataset which is collected for video recognition and temporal action localization [47], [50], [62]–[66], which are associated with 200 activity classes, where the content is more diverse compared to Charades-STA. Krishna *et al.* [60] extend ActivityNet to ActivityNet-Captions for the dense video captioning task. ActivityNet-Captions consists of $20K$ videos with $100K$ queries, and each video, with around 2-minute duration. it contains 3.65 queries on average and each query has an average length of 13.48 words. The ActivityNet-Captions dataset is split into the training set, validation set, testing set with a 2:1:1 ratio, including 3,7421, 1,7505 and 1,7031 video-query pairs separately.

B. Implementation Details

1) Evaluation Metric: For fair comparisons, following the setting as previous work [2], we evaluate our model by computing *Rank* $n@m$. Specifically, it is defined as the percentage of sentence queries having at least one correct grounding prediction in the top-n predictions, and the grounding prediction is correct when its IoU with the ground truth is larger than m. Similar to [17], we evaluate our method with specific settings of n and m for different datasets. Specifically, we report the results at $n \in \{1, 5\}$ with $m \in \{0, 1, 0.3, 0.5\}$ for TACoS dataset, $n \in \{1, 5\}$ with $m \in \{0.5, 0.7\}$ Charades-STA dataset, $n \in \{1, 5\}$ with $m \in \{0.3, 0.5, 0.7\}$ for ActivityNet Captions dataset.

2) Feature Extractor: For fair comparison, we extract video features following previous works [3], [17]. Specifically, We use the C3D [53] network pre-trained on Sports-1M [67] as the feature extractor. For Charades-STA, we also use VGG feature [54] to compare out results with [2], [17]. We divided the video into segments contains several frames. The input of C3D network is a segment with 16 frames for three datasets. When using VGG feature for Charades-STA, the number of frames in a segment is set to 4. Non maximum suppression (NMS) with a threshold of 0.5 is applied during the inference. The dimension d_v, d_w and d are 512, 300 and 512 respectively. To generate 2D temporal map, we adopt an 8-layer 2D convolution network with kernel size of 5 for Charades-STA and TACoS, and a 4-layer 2D convolution

TABLE III
ABLATION STUDY OF VARIOUS FUSION AND PREDICTION METHODS ON THE TACOS AND ACTIVITYNET DATASETS

Fusion	Prediction	TACoS			ActivityNet-Captions		
		0.1	0.3	0.5	0.3	0.5	0.7
JF	Separate multi-level	47.62	37.52	25.37	59.55	44.89	26.67
	Fused multi-level	47.97	38.47	26.38	59.63	45.03	26.73
	DT-FPN($\mathbf{G}_1$)	47.78	38.06	26.15	59.70	45.09	26.81
	DT-FPN($\mathbf{G}_{1-3}$)	47.99	38.16	26.20	59.82	45.17	26.93
	TD-FPN($\mathbf{G}_3$)	48.01	38.27	26.29	59.67	45.25	27.05
	TD-FPN($\mathbf{G}_{1-3}$)	48.04	38.52	26.86	59.82	45.59	27.23
IF	Separate multi-level	47.15	35.10	23.45	59.80	45.46	26.82
	Fused multi-level	48.18	38.02	27.23	59.95	45.55	26.97
	DT-FPN($\mathbf{G}_1$)	49.24	38.74	28.45	59.97	45.62	27.18
	DT-FPN($\mathbf{G}_{1-3}$)	50.18	39.48	28.23	60.05	46.07	28.03
	TD-FPN($\mathbf{G}_3$)	50.15	39.91	28.87	59.98	45.83	27.86
	TD-FPN($\mathbf{G}_{1-3}$)	**50.56**	**40.09**	**29.26**	**60.15**	**46.35**	**28.13**

network with kernel size of 9 for ActivityNet Captions. The normalized thresholds t_{min} and t_{max} defined in Eq. 2 are set to 0.5 and 0.1 for Charades-STA and ActivityNet-Capstions, and 0.3 and 0.7 for TACoS.

3) Training Settings: We use Adam [68] with learning rate of 1×10^{-4} and batch size of 32 for optimization. We decay the learning rate with ReduceLROnPlateau in Pytorch [69]. All of our models are implemented by PyTorch and trained under the environment of Python 3.6 on Ubuntu 16.04.

C. Ablation Studies

We ablate the proposed HLN to evaluate the effectiveness of each component. Our baseline method is 2D-TAN [17], which fuses the text and video information by using the whole sentence and the complete video features with a single level structure. It corresponds to setting that only $\mathbf{V}$ and $\mathbf{Q}_3$ are used in our method.

1) The Effect of the Language Hierarchy: We conduct experiments on the TACoS dataset to evaluate the effect of the Language Hierarchy (LH). In these experiments, we fuse the two modalities with the Individual Fusion (IF) at each level rather than the Joint Fusion (JF). We will discuss the effectiveness of IF and JF in Section IV-C.3. In Table I, we report the performance of video grounding with different LH strategy. $\{\mathbf{F}_i\}_{i=1}^3$ indicate the fused features, as defined in Section III-C, and $\mathbf{V}\&\mathbf{Q}_1$ indicates fusion operation between $\mathbf{V}$ and $\mathbf{Q}_1$. Comparing HLN_1 and HLN_2 with the baseline, we can find that the main clauses are critical for video grounding. Adding the main clause (HLN_4) to the baseline method brings a significant improvement by 2.44%. Comparison between HLN_3 and HLN_4 indicates the whole sentences $\mathbf{Q}_3$ are more critical than $\mathbf{Q}_2$, because they can provide more details. Finally, applying all levels of LH can achieve best performance proves the effectiveness of the hierarchical architecture in the video grounding.

2) Exploration of the Sentence Decomposition: We conduct experiments to evaluate the effect of different sentence decomposition on TACoS dataset, and present the results in Table II. $\mathbf{F}_1$, $\mathbf{F}_2$ and $\mathbf{F}_3$ represent the fused features from local to global respectively. We can also see that Main Clause (MC) is the most important part than Attribute Phrases (ATP) and Adverbial Phrases (ADP). For example, the model with the setting of $\mathbf{F}_1$=MC outperforms $\mathbf{F}_1$=ADP and $\mathbf{F}_1$=ATP by 1.29% and 2.43% respectively. Then we also have the conclusion that ATP achieves better performance than ADP, such as 26.59% *vs* 26.17%. Moreover, when given a specific $\mathbf{F}_1$, *e.g.* MC, $\mathbf{F}_2$=MC+ATP and $\mathbf{F}_2$=MC+ADP gain 0.72% and 0.56% than $\mathbf{F}_2$=ADP+ATP, which means the top-to-down fusion procedure is more suitable for video grounding task.

3) Evaluation of the Fusion and Prediction Ways: Table III evaluate different ways of fusion and prediction described in Fig. 5 (a)-(f) on TACoS and ActivityNet-Captions. The scores at *Rank*1 are presented. 'DT-FPN' represents down-to-top FPN, while 'TD-FPN' represents top-to-down FPN. $\mathbf{G}_1$ means that only one level $\mathbf{G}_1$ is applied when perform prediction, while $\mathbf{G}_{1-3}$ indicates all three levels are applied. The results show that Individual Fusion (IF) can obtain better performance that Joint Fusion (JF) under all the settings. It shows that each hierarchy should be trained individually and should not bring information from other levels. About the way of prediction, the FPN [59] architecture can achieve significant improvements. Moreover, we can also find that TD-FPN slightly outperforms DT-FPN, which can achieve best performance when combining all the levels.

4) Effect of MI-PUL: Table IV evaluate the effectiveness of the proposed MI-PUL on TACoS and ActivityNet-Captions datasets. In these experiments, we compare MI-PUL with GT that directly uses ground-truth segments as supervision, Label Smoothing (LS) [61] and Positive Unlabeled Learning (PUL) [21]. These results show that MI-PUL outperforms other methods based on both HLN_4 and HLN_6 models. The improvements of MI-PUL tend to increase with temporal IoU growing, which means our method can predict more precise boundaries. Moreover, we further explore MI-PUL from the false positive (FP) and false negative (FN) points, as shown in

TABLE IV

ABLATION STUDY OF THE MI-PUL ON THE TACOS AND ACTIVITYNET-CAPTIONS (ANET-CAP) DATASETS

Methods	Models		TACoS			ActivityNet-Captions		
	HLN_4	HLN_6	0.1	0.3	0.5	0.3	0.5	0.7
GT	✓		49.81	38.67	27.99	59.03	44.78	25.22
		✓	50.56	40.09	29.26	59.68	45.69	26.85
LS [61]	✓		50.79	40.19	28.64	59.23	44.94	25.25
		✓	51.69	40.41	29.78	**60.33**	45.28	26.89
PUL [21]	✓		51.12	40.15	29.95	59.02	45.08	27.08
		✓	51.93	40.63	30.55	59.51	45.98	27.64
MI-PUL	✓		51.66	40.23	30.25	59.57	45.24	27.70
		✓	**52.05**	**41.03**	**31.47**	60.15	**46.35**	**28.13**

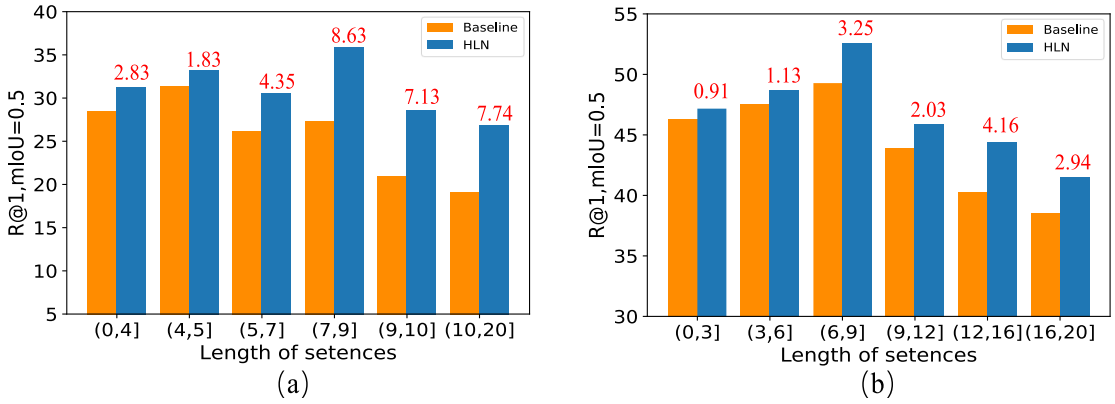


Fig. 7. Illustration of the performance of different length of sentences on (a) TACoS and (b) ActivityNet-Captions datasets. The red number over the bins is the relative performance improvement of HLN over the baseline.

TABLE V

COMPARISON OF FP, FN OF DIFFERENT METHODS ON THE TACOS AND ACTIVITYNET-CAPTIONS (ANET-CAP) DATASETS

Methods	Level	TACoS		ANet-Cap	
		FP	FN	FP	FN
GT	$\mathbf{F}_1$	5.82%	4.98%	11.35%	5.85%
	$\mathbf{F}_2$	4.79%	5.24%	7.61%	8.13%
	$\mathbf{F}_3$	3.35%	6.10%	4.22%	15.24%
PUL[21]	$\mathbf{F}_1$	10.18%	5.87%	13.15%	6.29%
	$\mathbf{F}_2$	6.04%	6.11%	9.22%	8.79%
	$\mathbf{F}_3$	4.28%	6.45%	5.60%	16.02%
MI-PUL	$\mathbf{F}_1$	**10.22%**	**4.52%**	**14.27%**	**5.78%**
	$\mathbf{F}_2$	**6.07%**	**5.05%**	**10.53%**	**8.12%**
	$\mathbf{F}_3$	**2.87%**	**5.73%**	**3.93%**	**14.28%**

TABLE VI

ABLATION STUDY OF *Rank*1 AT DIFFERENT IOUS OF DIFFERENT LEVELS IN LH ON THE TACOS AND ACTIVITYNET-CAPTIONS DATASETS

Level	TACoS			ANet-Cap		
	0.1	0.3	0.5	0.3	0.5	0.7
$\mathbf{G}_1$	16.20%	14.35%	7.07%	14.85%	8.55%	5.63%
$\mathbf{G}_2$	18.51%	15.24%	12.37%	20.71%	24.14%	21.27%
$\mathbf{G}_3$	65.29%	70.41%	80.56%	64.44%	67.31%	73.10%

Table V. Specifically, we first set a threshold value, i.e., 0.05. The predicted scores higher than 0.05 but outside of GT are FPs, while the ones lower than 0.05 but inside of GT are FNs. This aims to explore how does HLN work. For the top two levels, considering the more coarse semantics, FP should be higher, while in the bottom level with fine query, FP should be small. We find that PUL and MI-PUL can both obtain higher FP than GT, and MI-PUL can have smaller FP in the bottom level than PUL. Due to the sentence of the level from top to down being more and more restrict, FN is more and more higher. In all three methods, our MI-PUL shows the smallest FN, which indicates more accurate localization.

5) Evaluation of Different Levels and Lengths of the Sentences: Table VI shows the results of *Rank*1 at different levels on TACoS and ActivityNet-Captions datasets. It is clear that the deep layer can predict better moments with high IoUs. Meanwhile, we report the performance of different sentence length in terms of $Rank1@IoU = 0.5$ in Fig. 7. The improvements of the proposed HLN over the baseline mainly come form the long sentences, which means that language hierarchy can help to understand the languages.

TABLE VII
COMPARISONS WITH STATE-OF-THE-ARTS ON TACOS AND ACTIVITYNET-CAPTIONS

Methods	TACoS						ActivtiyNet-Captions					
	*Rank*1@			*Rank*5@			*Rank*1@			*Rank*5@		
	0.1	0.3	0.5	0.1	0.3	0.5	0.3	0.5	0.7	0.3	0.5	0.7
MCN [1]	14.42	-	5.58	37.35	-	10.33	39.35	21.36	6.43	68.12	53.23	29.70
CTRL [2]	24.32	18.32	13.30	48.73	36.69	25.42	47.43	29.01	10.34	75.32	59.17	37.54
MCF [70]	25.84	18.64	12.53	52.96	37.13	24.73	-	-	-	-	-	-
TGN [16]	41.87	21.77	18.9	53.40	39.06	31.02	43.81	27.93	-	4.56	44.20	
ACRN [71]	24.22	19.52	14.62	47.42	34.97	24.88	49.70	31.67	11.25	76.50	60.34	38.57
ROLE [72]	20.37	15.38	9.94	45.45	31.17	20.13	-	-	-	-	-	-
CMIN [73]	32.48	24.64	18.05	62.13	38.46	27.02	-	-	-	-	-	-
QSPN [29]	25.31	20.15	15.23	53.21	36.72	25.30	52.13	33.26	13.43	77.72	62.39	40.78
ABLR [10]	34.70	19.50	9.40	-	-	-	55.67	36.79	-	-	-	-
TripNet [74]	-	23.95	19.17	-	-	-	48.42	32.19	13.93	-	-	-
DRN [3]	-	-	23.17	-	-	33.36	-	45.45	24.36	-	77.97	50.30
HVTG [75]	-	-	-	-	-	-	57.60	40.15	18.27	-	-	-
CPNet [76]	-	42.61	28.29	-	-	-	-	40.56	21.63	-	-	-
Sscs [77]	50.78	41.33	29.56	72.53	60.65	48.01	61.35	46.67	27.56	86.89	78.37	63.78
2D-TAN [17]	47.59	37.29	25.32	70.31	57.81	45.04	59.45	44.51	26.54	85.53	77.13	61.96
+ Ours	52.05	41.03	31.47	73.21	63.57	49.18	60.15	46.35	28.13	87.25	78.87	64.08
above	**+5.46**	**+3.74**	**+6.15**	**+2.91**	**+5.67**	**+4.14**	**+0.70**	**+1.84**	**+1.95**	**+1.72**	**+1.74**	**+2.10**
CBLN [34]	49.16	38.98	27.65	73.12	59.96	46.24	66.34	48.12	27.60	88.91	79.32	63.41
+ Ours	**53.32**	**43.61**	**33.53**	**75.16**	**65.15**	**50.69**	**67.18**	**49.68**	**29.27**	**89.53**	**80.35**	**64.98**
above	**+4.16**	**+4.63**	**+5.88**	**+2.04**	**+4.09**	**+4.44**	**+0.84**	**+1.56**	**+1.67**	**+0.62**	**+1.03**	**+1.57**

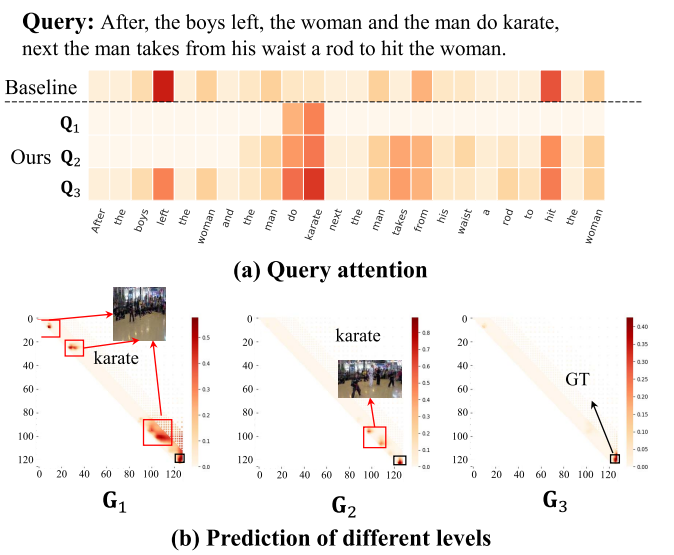


Fig. 8. Visualized results on ActivityNet Captions dataset: (a) attention of words and (b) prediction of different levels.

Fig. 9. Qualitative results on three examples selected from Charades-STA and ActivityNet-Captions datasets.

D. Comparisons With State-of-the-Arts

Table VII and Table VIII compare the propose HLN with current state-of-the-art approaches on the three datasets. "above" in Table VII and Table VIII mean the improvement obtained by adding our method to the baseline. It is clear that the proposed HLN can improve all the methods by applying same C3D feature over almost all IoU thresholds on TACoS. Specifically, our method achieves 31.47% with the *Rank*1 metric at IoU 0.5, which significantly outperforms

TABLE VIII
COMPARISONS WITH STATE-OF-THE-ARTS ON CHARADES-STA

Methods	$Rank1@$		$Rank5@$	
	0.5	0.7	0.5	0.7
VGG				
MCN [1]	17.46	8.01	48.22	26.73
ACRN [71]	20.26	7.64	71.99	27.79
ROLE [72]	21.74	7.82	70.37	30.06
VAL [78]	23.12	9.16	61.26	27.98
ACL-K [30]	30.48	12.20	64.84	35.13
QSPN [29]	35.60	15.80	79.40	45.40
SM-RL [33]	24.36	11.17	61.25	32.08
SLTA [79]	22.81	8.25	72.39	31.46
DRN [3]	42.90	23.68	87.80	54.87
CBLN [34]	43.76	24.44	88.39	56.49
Sscs [77]	43.15	25.54	84.26	54.17
2D-TAN [17]	39.70	23.31	80.32	51.26
+ Ours	41.15	24.24	81.63	52.37
above	**+1.45**	**+0.93**	**+1.31**	**+1.11**
MS-2D-TAN [80]	46.65	27.20	87.72	56.42
+ Ours	**47.33**	**28.07**	**88.24**	**57.37**
above	**+0.68**	**+0.87**	**+0.52**	**+0.95**
C3D				
CTRL [2]	23.63	8.89	58.92	29.52
SAP [81]	27.42	13.36	66.37	38.15
DRN [3]	45.40	26.40	88.01	55.38
HVTG [75]	47.27	23.30	-	-
CPNet [76]	40.32	22.47	-	-
CBLN [34]	47.49	28.22	88.20	57.47
+ Ours	**48.26**	**29.31**	**88.82**	**58.52**
above	**+0.77**	**+1.09**	**+0.62**	**+1.05**

the baseline 2D-TAN by 6.15%. On ActivityNet-Captions, our method reaches competitive performance in terms of both $Rank1$ when $Rank5$ IoU= 0.5, outperforming the previous model 2D-TAN by 1.84% and 1.74% respectively. For a fair comparison, we employ VGG and C3D features to when evaluating on Charades-STA dataset, following DRN [3]. We can see that our method achieve the highest score at all IoUs with VGG feature, outperforming the previous best model 2D-TAN by 1.45% and 0.93% at $rank1@\{0.5, 0.7\}$ respectively. Actually, the queries of TACoS are more challenging than Charades-STA, while HLN obtains larger improvements on the TACoS, which can better demonstrate the effectiveness of the language hierarchy.

E. Qualitative Analysis

We present some visualization results in Fig. 8 to evaluate the proposed method. In Fig. 8 (a), we compare HLN with the baseline model in terms of the attention of the query. Darker color indicates a higher value. $\mathbf{Q}_1$, $\mathbf{Q}_2$ and $\mathbf{Q}_3$ indicate the

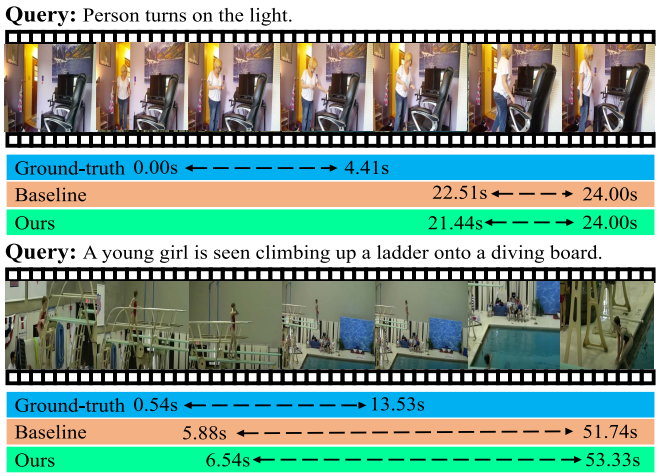


Fig. 10. Failure cases on two examples selected from Charades-STA and ActivityNet-Captions datasets.

sentence features at the first, second and third level respectively. Compared with the baseline model, HLN can focus on different parts of the query at different levels. Fig. 8 (b) presents the visualization of prediction different levels. It is clear that at top two levels, HLN can find the moments of 'do karate', even they are not within the ground truth. Fig. 9 shows some qualitative results from three query-video examples. Our methods can predict more precise segments. Specifically, for the second query, the baseline model may localize the segments which contain 'left' and 'hit', hence generating a more coarse prediction. Since our proposed HLN can focus more on other phrases, *i.e.*, 'do karate', the predicted segments are more precise.

F. Drawback Analysis

Since our model need decompose the sentence into different phrases, *i.e.*, main clauses, attribute phrases and adverbial phrases, there is only slight improvement for those sentences which only consists of main clauses. As Table VIII indicates, our proposed method only achieves a slight improvement than baseline (2D-TAN), *i.e.*, 24.24% vs 23.31% with VGG feature, which is only 0.93% over than the baseline model. That is due to that most of the sentences in the Charades-STA dataset are simple, such as 'person turns the light off'; see the top row in Fig. 10. Furthermore, the improvement on ActivityNet-Captions is also not as large as that on TACoS. This is due to that the sentences in ActivityNet-Captions contain too many verbs, such as 'is seen to do', which are not the real main clauses, as shown in Fig. 4 (d)-(e); see the bottom row in Fig. 10.

V. CONCLUSION

In this paper, we propose a novel HLN to explore the language hierarchy for the video grounding task. We discuss different strategies to construct the language hierarchy, and design a coarse-to-fine architecture. In this hierarchical architecture, several fusion and prediction approaches are well

evaluated, and the experiments show that the individual fusion and top-to-down FPN achieve the best performance. We also propose a learning approach, termed MI-PUL, which allows that it can contain the moments described by main clauses in the background intervals. Combined with MI-PUL, the proposed HLN improve the state-of-the-art methods on three challenging datasets. In future works, instead of decomposing sentences with manually defined rules (*i.e.*, main clause, the attributives and the adverbs), we would explore the automatic learned way to group sentences, which is more efficient and generalized.

REFERENCES

[1] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5803–5812.

[2] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: Temporal activity localization via language query," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5267–5275.

[3] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, and C. Gan, "Dense regression network for video grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10287–10296.

[4] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 20–36.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[6] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 471–487.

[7] J. Dong *et al.*, "Dual encoding for zero-example video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9346–9355.

[8] J. Singha, A. Roy, and R. H. Laskar, "Dynamic hand gesture recognition using vision-based approach for human–computer interaction," *Neural Comput. Appl.*, vol. 29, no. 4, pp. 1129–1141, 2016.

[9] F. Zhu, Y. Zhu, X. Chang, and X. Liang, "Vision-language navigation with self-supervised auxiliary reasoning tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10012–10022.

[10] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 9159–9166.

[11] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 536–546.

[12] S. L. Frank, R. Bod, and M. H. Christiansen, "How hierarchical is language use?" *Proc. Roy. Soc. B, Biol. Sci.*, vol. 279, no. 1747, pp. 4522–4531, Nov. 2012.

[13] S. Miyagawa, R. C. Berwick, and K. Okanoya, "The emergence of hierarchical structure in human language," *Frontiers Psychol.*, vol. 4, p. 71, Jan. 2013.

[14] E. B. Goldstein, *Cognitive Psychology: Connecting Mind, Research and Everyday Experience*. Camden, NJ, USA: Nelson, 2014.

[15] L. Rozakis, *The Complete Idiot's Guide to Grammar and Style*. Baltimore, MD, USA: Penguin, 2003.

[16] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 162–171.

[17] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2D temporal adjacent networks for moment localization with natural language," 2019, *arXiv:1912.03590*.

[18] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 8199–8206.

[19] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 25–36, Dec. 2013.

[20] Y. Xu, C. Xu, C. Xu, and D. Tao, "Multi-positive and unlabeled learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3182–3188.

[21] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," 2017, *arXiv:1703.00593*.

[22] Y. Xu *et al.*, "Positive-unlabeled compression on the cloud," 2019, *arXiv:1909.09757*.

[23] X. Chen *et al.*, "Self-PU: Self boosted and calibrated positive-unlabeled training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1510–1519.

[24] Z.-H. Zhou, "Multi-instance learning: A survey," Dept. Comput. Sci. Technol., Nanjing Univ., Nanjing, China, Tech. Rep. 2, 2004.

[25] J. R. Foulds and E. Frank, "A review of multi-instance learning assumptions," *Knowl. Eng. Rev.*, vol. 25, no. 1, pp. 1–25, 2010.

[26] B. Zhang and W. Zuo, "Tri-training based learning from positive and unlabeled data," in *Proc. Int. Symposiums Inf. Process.*, vol. 27, May 2008, pp. 703–711.

[27] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 510–526.

[28] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[29] H. Xu, K. He, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 9062–9069.

[30] R. Ge, J. Gao, K. Chen, and R. Nevatia, "MAC: Mining activity concepts for language-based temporal localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 245–253.

[31] D. Liu, X. Qu, X.-Y. Liu, J. Dong, P. Zhou, and Z. Xu, "Jointly Cross- and self-modal graph attention network for query-based moment localization," 2020, *arXiv:2008.01403*.

[32] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8393–8400.

[33] W. Wang, Y. Huang, and L. Wang, "Language-driven temporal activity localization: A semantic matching reinforcement learning model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 334–343.

[34] D. Liu *et al.*, "Context-aware biaffine localizing network for temporal sentence grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11235–11244.

[35] H. Zhou, C. Zhang, Y. Luo, Y. Chen, and C. Hu, "Embracing uncertainty: Decoupling and de-bias for robust temporal grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8445–8454.

[36] Z. Wang, L. Wang, T. Wu, T. Li, and G. Wu, "Negative sample matters: A renaissance of metric learning for temporal grounding," 2021, *arXiv:2109.04872*.

[37] W. Yang, T. Zhang, Y. Zhang, and F. Wu, "Local correspondence network for weakly supervised temporal sentence grounding," *IEEE Trans. Image Process.*, vol. 30, pp. 3252–3262, 2021.

[38] X. Lan, Y. Yuan, X. Wang, Z. Wang, and W. Zhu, "A survey on temporal sentence grounding in videos," 2021, *arXiv:2109.08039*.

[39] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10810–10819.

[40] H. Zhang, A. Sun, W. Jing, and J. Tianyi Zhou, "The elements of temporal sentence grounding in videos: A survey and future directions," 2022, *arXiv:2201.08071*.

[41] H. K. Armenian and A. M. Lilienfeld, "The distribution of incubation periods of neoplastic diseases," *Amer. J. Epidemiol.*, vol. 99, no. 2, pp. 92–100, Feb. 1974.

[42] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 2, Jul. 2002, pp. 387–394.

[43] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 3, 2003, pp. 587–592.

[44] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[45] T. Leung, Y. Song, and J. Zhang, "Handling label noise in video classification via multiple instance learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2056–2063.

[46] N. Shapovalova, A. Vahdat, K. Cannons, T. Lan, and G. Mori, "Similarity constrained latent support vector machine: An application to weakly supervised action classification," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 55–68.

[47] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-TALC: Weakly-supervised temporal activity localization and classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 563–579.

[48] P. Nguyen, T. Liu, and G. Prasad, "Weakly supervised action localization by sparse temporal pooling network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6752–6761.

[49] P. Nguyen, D. Ramanan, and C. Fowlkes, "Weakly-supervised action localization with background modeling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5502–5511.

[50] X. Ding, N. Wang, X. Gao, J. Li, X. Wang, and T. Liu, "Weakly supervised temporal action localization with segment-level labels," 2020, *arXiv:2007.01598*.

[51] G. Li, J. Li, N. Wang, X. Ding, Z. Li, and X. Gao, "Multi-hierarchical category supervision for weakly-supervised temporal action localization," *IEEE Trans. Image Process.*, vol. 30, pp. 9332–9344, 2021.

[52] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9879–9889.

[53] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[55] P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," 2019, *arXiv:1904.05255*.

[56] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, 2020, pp. 1–8.

[57] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[59] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[60] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 706–715.

[61] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[62] S. Wang, Z. Miao, W. Xu, C. Ma, and M. Li, "Boundary sensitive and category sensitive network for temporal action proposal generation," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2019, pp. 3–19.

[63] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1130–1139.

[64] J. Gao, Z. Yang, and R. Nevatia, "Cascaded boundary regression for temporal action detection," 2017, *arXiv:1705.01180*.

[65] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1049–1058.

[66] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3C-Net: Category count and center loss for weakly-supervised action localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8679–8687.

[67] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[69] A. Paszke, S. Gross, and S. Chintala. (2017). *PyTorch Deep Learning Framework*. [Online]. Available: http://pytorch.org/

[70] A. Wu and Y. Han, "Multi-modal circulant fusion for video-to-language and backward," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, p. 8.

[71] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua, "Attentive moment retrieval in videos," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 15–24.

[72] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 843–851.

[73] Z. Zhang, Z. Lin, Z. Zhao, and Z. Xiao, "Cross-modal interaction networks for query-based moment retrieval in videos," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 655–664.

[74] M. Hahn, A. Kadav, J. M. Rehg, and H. P. Graf, "Tripping through time: Efficient localization of activities in videos," 2019, *arXiv:1904.09936*.

[75] S. Chen and Y.-G. Jiang, "Hierarchical visual-textual graph for temporal activity localization via language," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 601–618.

[76] K. Li, D. Guo, and M. Wang, "Proposal-free video grounding with contextual pyramid network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1902–1910.

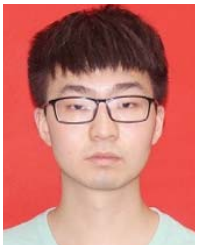
[77] X. Ding *et al.*, "Support-set based cross-supervision for video grounding," 2021, *arXiv:2108.10576*.

[78] X. Song and Y. Han, "VAL: Visual-attention action localizer," in *Proc. Pacific Rim Conf. Multimedia*. Cham, Switzerland: Springer, 2018, pp. 340–350.

[79] B. Jiang, X. Huang, C. Yang, and J. Yuan, "Cross-modal video moment retrieval with spatial and language-temporal attention," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 217–225.

[80] S. Zhang, H. Peng, J. Fu, Y. Lu, and J. Luo, "Multi-scale 2D temporal adjacent networks for moment localization with natural language," 2020, *arXiv:2012.02646*.

[81] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 8199–8206.



Xinpeng Ding received the B.Eng. degree in software engineering and the M.Sc. degree in information and telecommunications engineering from Xidian University, Xi'an, China, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree with the School of Engineering, The Hong Kong University of Science and Technology, Hong Kong. His current research interests include computer vision, pattern recognition, medical image analysis, and machine learning.

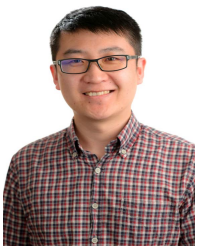
Nannan Wang (Member, IEEE) received the B.Sc. degree in information and computation science from the Xi'an University of Posts and Telecommunications in 2009 and the Ph.D. degree in information and telecommunications engineering from Xidian University, Xi'an, China, in 2015. He is currently a Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. He has published over 150 articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, CVPR, and ICCV. His current research interests include computer vision and machine learning.

Shiwei Zhang received the Ph.D. degree from the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, in 2019. He is currently a Researcher of computer vision at the DAMO Academy, Alibaba Group. His research interests include video understanding, video generation, multi-modal representation learning, and machine learning.

Ziyuan Huang received the B.Eng. degree in vehicle engineering from Tongji University in 2019. He is currently pursuing the Ph.D. degree with the Advanced Robotics Centre, National University of Singapore, supervised by Prof. Marcelo Ang. His main research interests are on video understanding, including action recognition and localization, video representation learning, multi-modal learning, and video-based scene understanding.

Xiaomeng Li (Member, IEEE) received the Ph.D. degree from The Chinese University of Hong Kong. She is currently an Assistant Professor of electronic and computer engineering at The Hong Kong University of Science and Technology. Her research lies in the interdisciplinary areas of artificial intelligence and medical image analysis, aiming at advancing healthcare with machine intelligence.

Mingqian Tang received the M.Sc. degree from the School of Computer Science and Technology, Xidian University, Xi'an, China, in 2014. She is currently a Senior Algorithm Expert at Alibaba Group and exploring the applications of artificial intelligence technology. Her research interests include video classification, video generation, video retrieval, and multi-modal learning.



Tongliang Liu (Senior Member, IEEE) is currently a Lecturer and the Director of the Trustworthy Machine Learning Laboratory, School of Computer Science, The University of Sydney. He has published papers on various top conferences and journals, such as NeurIPS, ICML, ICLR, CVPR, ECCV, KDD, IJCAI, AAAI, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON MULTIMEDIA. He is broadly interested in the fields of trustworthy machine learning and its interdisciplinary applications, with a particular emphasis on learning with noisy labels, transfer learning, adversarial learning, unsupervised learning, and statistical deep learning theory. He has received the ICME 2019 Best Paper Award and the PacificVis 2021 Best VisNotes Paper Award. He was a recipient of the Discovery Early Career Researcher Award (DECRA) from Australian Research Council (ARC) and the Cardiovascular Initiative Catalyst Award by the Cardiovascular Initiative. He was named in the Early Achievers Leaderboard of Engineering and Computer Science by the Australian in 2020. He is/was a Meta Reviewer for many conferences, such as NeurIPS, ICLR, AAAI, and IJCAI.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Postdoctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of the Ministry of Education of China, a Professor of pattern recognition and intelligent system with Xidian University, and a Professor of computer science and technology with the Chongqing University of Posts and Telecommunications. He has published six books and around 300 technical articles in refereed journals and proceedings. His current research interests include image processing, computer vision, multimedia analysis, machine learning, and pattern recognition. He is a fellow of the Institute of Engineering and Technology and the Chinese Institute of Electronics. He has served as the general chair/co-chair, the program committee chair/co-chair, or a PC member for around 30 major international conferences. He is on the editorial boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier).