Support-Set Based Cross-Supervision for Video Grounding

Xinpeng Ding[1,2], Nannan Wang[1*], Shiwei Zhang[2*], De Cheng[1], Xiaomeng Li[3],
Ziyuan Huang [4], Mingqian Tang[2], Xinbo Gao[5]
[1]Xidian University, [2]Alibaba Group, [3]The Hong Kong University of Science and Technology,
[4]National University of Singapore, [5]Chongqing University of Posts and Telecommunications
xpding.xidian@gmail.com, {nnwang,dcheng}@xidian.edu.cn, eexmli@ust.hk
{zhangjin.zsw, mingqian.tmq}@alibaba-inc.com, ziyuan.huang@u.nus.edu, gaoxb@cqupt.edu.cn

Abstract

Current approaches for video grounding propose kinds of complex architectures to capture the video-text relations, and have achieved impressive improvements. However, it is hard to learn the complicated multi-modal relations by only architecture designing in fact. In this paper, we introduce a novel Support-set Based Cross-Supervision (Sscs) module which can improve existing methods during training phase without extra inference cost. The proposed Sscs module contains two main components, i.e.*, discriminative contrastive objective and generative caption objective. The contrastive objective aims to learn effective representations by contrastive learning, while the caption objective can train a powerful video encoder supervised by texts. Due to the co-existence of some visual entities in both ground-truth and background intervals,* i.e. *mutual exclusion, naively contrastive learning is unsuitable to video grounding. We address the problem by boosting the cross-supervision with the support-set concept, which collects visual information from the whole video and eliminates the mutual exclusion of entities. Combined with the original objectives, Sscs can enhance the abilities of multi-modal relation modeling for existing approaches. We extensively evaluate Sscs on three challenging datasets, and show that our method can improve current state-of-the-art methods by large margins, especially 6.35% in terms of R1@0.5 on Charades-STA.*

1. Introduction

Video grounding aims to localize the target time intervals in an untrimmed video by a text query. As illustrated in Fig. 1 (a), given a sentence 'The person pours some water into the glass.' and a paired video, the target is to localize the best matching segment, *i.e.*, from 7.3s to 17.3s. There are various methods [51, 49, 12] have been proposed for this task, and they have made significant progresses. These methods can reach an agreement that video-text relation modeling is one of the crucial roles. An effective relation should be that semantically related videos and texts must have high responses, and vice versa.

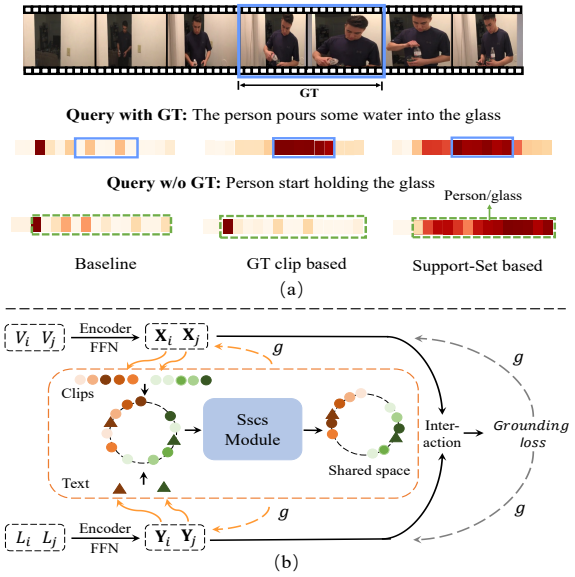


Figure 1. (a) Comparison of the attention map of the similarity between video clips and text queries. The darker the color, the higher the similarity. 'GT' indicates the ground-truth. (b) The proposed Support-Set based Cross-Supervision (Sscs) Module. Sscs makes the embedding of semantic-related clip-text pairs (dark circles and triangles) to be close in the shared feature space.

To achieve this goal, existing methods focus on carefully designing complex video-text interaction modules. For example, Zeng *et al.* [49] propose a pyramid neural network to consider multi-scale information. Local-global strategy [30] and self-modal graph attention [26] are applied as the interaction operations to learning the multi-modal relations. After that, they use the interacted features to perform

*Nannan Wang and Shiwei Zhang are the corresponding authors.

video grounding straightway. However, the multi-modal relations are complicated because the video and text have unequal semantics, *e.g.*, 'person' is just one word but may last a whole video. Hence, existing methods based on the architecture improvements have limited capacities to learn video-caption relations; see Fig. 1 (a) (Please see 'Baseline').

Motivated by the advances of multi-modal pre-training [28, 33, 29], we propose a Support-Set Based Cross-Supervision, termed Sscs, to improve multi-modal relation learning for video grounding in a supervision way compared with the hand-designed architectures. As shown in Fig. 1, the Sscs module is an independent branch that can be easily embedded into other approaches in the training stage. The proposed Sscs includes two main components, *i.e.*, contrastive objective and caption objective. The contrastive objective is as typical discriminative loss function, that targets to learn multi-modal representations by applying infoNCE loss function [28, 33]. In contrast, the caption objective is a generative loss function, which can be used to train a powerful video encoder [15, 53]. For an untrimmed video, there are some vision entities appear in both ground-truth and background intervals, *e.g.*, the person and glass in Fig. 2, but the original contrastive learning may wipe away the same parts between the foreground and background, including the vision entities. These vision entities are also important for video grounding task, *e.g.*, thus it is unsuitable to directly apply the contrastive learning into the video grounding task directly. To solve this problem, we apply the support-set concept, which captures visual information from the the whole video, to eliminates the mutual exclusion of entities. By this means, we can improve the cross-supervision module naturally and further enhance the relation modeling. To prove the robustness, we choose two state-of-the-art approaches as our baselines, *i.e.*, 2D-TAN [51] and LGI [30], and the experimental results show that the proposed Sscs can achieve a remarkable improvement.

Our contributions are summarized as three-folds: **(a)** We introduce a novel cross-supervision module for video grounding, which can enhance the correlation modeling between videos and texts but not bring in the extra inference cost. **(b)** We propose to apply support-set concept to address the mutual exclusion of video entities, which make it is more suitable to apply contrastive learning for video grounding. **(c)** Extensive experiments illustrate the effectiveness of Sscs on three public datasets, and the results show that our method can significantly improve the performance of the state-of-the-art approaches.

2. Related Work

Video grounding. Early approaches [12, 1, 46, 14] for video grounding use a two-stage visual-textual matching strategy to tackle this problem, which require a large number of proposals. It is important for these methods to improve the quality of the proposals. SCDM [47] incorporates the query text into the visual feature for correlating and composing the sentence-related video contents over time. 2D-TAN [51] 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 [49, 26, 48, 17, 44, 45] are proposed to predict starting and ending times directly. Zeng *et al.* [49] avoid the imbalance training by leveraging much more positive training samples, which improves the grounding performance. LGI [30] improves the performance of localization by exploiting contextual information from local to global during bi-modal interactions.

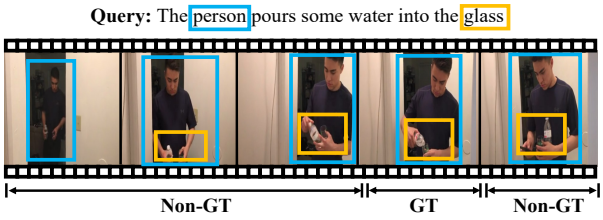


Figure 2. Mutual exclusion of entities. The 'person' and 'glass' entities appear in both ground-truth (GT) clips and non-ground-truth clips (Non-GT). Although there are no 'pour water' action happening in Non-GT clips, the semantics of the Non-GT video clips are also similar with those of GT ones, due to the common entities.

Multi-modal Representation Learning. A mass of self-supervised methods [9, 3, 10] have been proposed to pre-train models on large-scale multi-modal data, such as images [37], videos [5] and text [54]. To learn video-text representations, a large-scale instructional video dataset, HowTo100M [29], is released. Some works use the contrastive loss to improve video-text representations based on HowTo100M for tasks such as video caption [53], video retrieval [2] and video question answering [24]. MIL-NCE [28] brings the multi instance learning into the contrastive learning framework to address the misalignment between video content and narrations. Patrick *et al.* [33] combine both discriminative and generative objectives to push related video and text instance together. Compared with these approaches, our method targets to improve video grounding via multi-modal training without extra inference cost.

3. Proposed Method

3.1. Problem Formulation

Let's define a set of video-text pairs as $\mathcal{C} = \{(V_i, L_i)\}_{i=1}^{N}$, where N is the number of video-text pairs, V_i and L_i are the i-th untrimmed video and sentence respectively. Given a query sentence L_i, the purpose of video

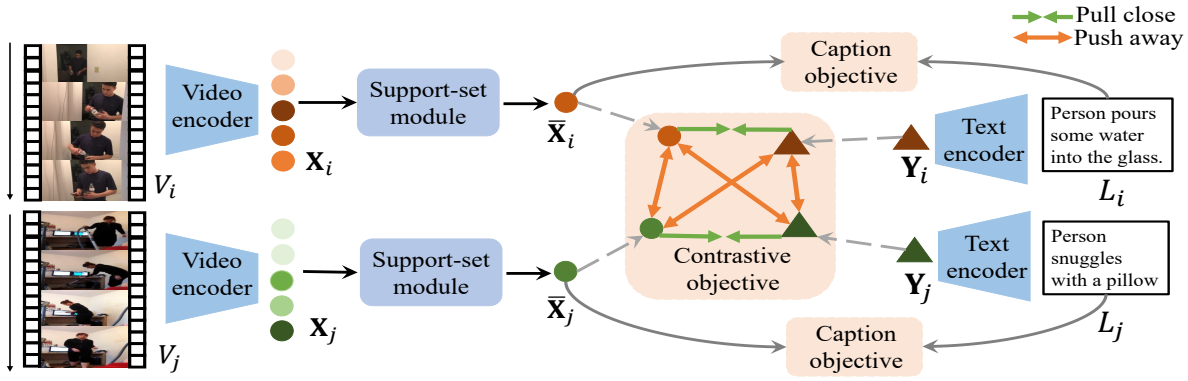


Figure 3. Illustration of our proposed Support-set Based Cross-Supervision Module. For clarity, we only present two video-text pairs $\{V_i, L_i\}$, $\{V_j, L_j\}$ in the batch. After feeding them into a video and text encoder, the clip-level and sentence-level embedding ($\{\mathbf{X}_i, \mathbf{Y}_i\}$ and $\{\mathbf{X}_j, \mathbf{Y}_j\}$) in a shared space are acquired. Base on the support-set module (see details in Fig. 4 (b)), we compute the weighted average of $\mathbf{X}_i$ and $\mathbf{X}_j$ to obtain $\overline{\mathbf{X}}_i$, $\overline{\mathbf{X}}_j$ respectively. Finally, we combine the contrastive and caption objectives to pull close the representations of the clips and text from the same samples and push away those from other pairs.

grounding is to localize a target time interval $\mathcal{A}_i = (t_s^i, t_e^i)$ in V_i, where t_s^i and t_e^i denote the starting and ending time respectively.

3.2. Video and Sentence Encoding

Video encoding. We first divide a long untrimmed video V_i into T clips, defined as $V_i = \{v_t^i\}_{t=1}^T$. Each clip consists of a fixed number of frames. Then, T clips are fed into a pre-trained 3D CNN model to extract the video features $\mathbf{F}_i = \{\mathbf{f}_t^i\} \in \mathbb{R}^{T \times D_v}$, where D_v denotes the dimension of the clip-based video features.

Sentence encoding. For m-th word l_m^i in a sentence L_i, we feed it into the GloVe word2vec model [35] to obtain the corresponding word embedding w_m^i. Then, the word embeddings are sequentially fed into one three layer bidirectional LSTM network [19], and we use its last hidden state as the features of the sentence L_i, *i.e.*, $\mathbf{G}_i \in \mathbb{R}^{D_l}$.

3.3. Cross-Supervised Video Grounding

In this section, we first outline the overall framework in Section 3.3.1. Then, in Section 3.3.2, we introduce the concept of support-set for video grounding in details. Finally, we introduce several kinds of support-set in Section 3.4.

3.3.1 The Overall Framework

The key of video grounding is to capture the relations between videos and texts. That is, it should have a high similarity between V_i and S_j if they are semantically-related and vice versa. For this purpose, most existing methods design multitudinous architectures to capture the relations by modeling the video-text interactions [51, 30]. Typically, they first fuse the visual and textual embeddings $\mathbf{X}_i$ and $\mathbf{Y}_i$, and then predict target time intervals $\hat{\mathcal{A}} = (\hat{t}_s^i, \hat{t}_e^i)$ directly. At the training stage, the loss function $\mathcal{L}^{\text{vg}}$ is applied on the fused features to optimize the models. The function $\mathcal{L}^{\text{vg}}$ may have different form in different method, *e.g.*, binary cross entropy loss function applied in 2D-TAN [51].

Unlike these methods, we introduce two cross-supervised training objectives that can improve existing methods just during training phase. The two objectives contain a contrastive objective and a caption objective, and can be insert existing methods directly. Thus the overall framework mainly contains two components, *i.e.*, the commonly used video grounding framework and the proposed cross-supervised objectives. Hence, the overall objective of our method is:

$$\mathcal{L} = \mathcal{L}^{\text{vg}} + \lambda_1 \mathcal{L}^{\text{contrast}} + \lambda_2 \mathcal{L}^{\text{caption}}, \tag{1}$$

where $\mathcal{L}^{\text{contrast}}$ and $\mathcal{L}^{\text{caption}}$ denote the contrastive objective and caption objective respectively. The hyperparameters λ_1 and λ_2 control the weight of two objectives.

3.3.2 Cross-Supervised Objectives

The target of the cross supervised objectives is to learn effective video-text relations, as illustrated in Fig. 3. To make it clear, we first introduce the GT clip-based learning, based on which we present the details of the proposed cross-supervised objectives. After that, we discus existing problems, *i.e.*, the mutual exclusion between the visual and textual modalities. Finally, we provide the solution by support set based learning.

GT Clip-Based Learning. In video grounding, a sentence usually corresponds to multiple clips, which are all contained in a ground-truth interval. An intuitive method to learn a powerful representation is to set the clips in ground-truth (GT) intervals as the positive samples, while others are negatives, *i.e.*, clips in Non-GT intervals and other videos.

Formally, we denote A *mini-batch* of samples from $\mathcal{C}$ with $\mathcal{B}$, hence $\mathcal{B} \subset \mathcal{C}$. Then the samples in the *mini-batch* can be defined as $\mathcal{B} = \{(V_i, L_i)\}_{i=1}^{B}$, where B is the size of the *mini-batch*. After fed into the video and text encoders, we can obtain base embeddings $\{(\mathbf{F}_i, \mathbf{G}_i)\}_{i=1}^{B}$. Then the embeddings can be mapped into a same space with equal dimension by $\mathbf{X}_i = \Psi(\mathbf{F}_i)$ and $\mathbf{Y}_i = \Phi(\mathbf{G}_i)$. For a pair of the video and text embeddings $(\mathbf{X}_i, \mathbf{Y}_i)$ in $\mathcal{B}$, we define the set of ground-truth clips as $\mathcal{M}_i = \{\mathbf{x}_t^i \,|\, t \in [t_s^i, t_e^i]\}$, where t_s^i and t_e^i denote the starting and ending time of the ground-truth, $\mathbf{x}_t^i$ is the t-th vector in $\mathbf{X}_i$. The set of background clips for V_i can be denoted as $\overline{\mathcal{M}}_i = \{\mathbf{x}_t^i \,|\, t \notin [t_s^i, t_e^i]\}$. Then, the positive pairs in $\mathcal{B}$ can be constructed by considering the ground-truth clips together with the corresponding text, defined as $\mathcal{P}_i = \{(\mathbf{x}, \mathbf{Y}_i) \,|\, \mathbf{x} \in \mathcal{M}_i\}$. The non GT clips and clips in other videos can be regarded as the negative samples of the text L_i, i.e, $\mathcal{N}_i = \{(\mathbf{x}, \mathbf{Y}_i) \,|\, \mathbf{x} \in \overline{\mathcal{M}}_i \cup \mathbf{X}_j,\, i \neq j\}$.

Contrastive objective. Based on the above definitions, we detail the contrastive objective here. The purpose of the contrastive objective is to learn effective video-text representations, for which we use a contrastive loss to increase the similarities of positive pairs in $\mathcal{P}$ and push away those from the negative pairs in $\mathcal{N}$. Specifically, we minimize the softmax version of MIL-NCE [28] as follows:

$$-\sum_{i=1}^{B} \log \left(\frac{\sum\limits_{(\mathbf{x},\mathbf{y}) \in \mathcal{P}_i} e^{\mathbf{x}^\top \mathbf{y}/\tau}}{\sum\limits_{(\mathbf{x},\mathbf{y}) \in \mathcal{P}_i} e^{\mathbf{x}^\top \mathbf{y}/\tau} + \sum\limits_{(\mathbf{x}',\mathbf{y}') \in \mathcal{N}_i} e^{\mathbf{x}'^\top \mathbf{y}'/\tau}} \right), \tag{2}$$

where τ is the temperature weight to control the concentration level of the sample distribution [18]. Therefore, the constrastive objective is a typical kind of discriminative loss function.

Caption objective. Besides the contrastive objective, we also introduce the caption objective [33, 22] to further improve the video-text representation. The caption objective can be formulized as:

$$\mathcal{L}^{\text{caption}} = -\frac{1}{B} \sum_{i=1}^{B} \log p\left(l_i \mid \mathbf{w}_i\right), \tag{3}$$

where l_i is the i-th word of L_i, $\mathbf{w}_i \in \mathbb{R}^D$ is the embedding for generating the sentence, which is obtained by $\mathbf{w}_i = \Phi'\left(\mathbf{X}_i^{gt}\right)$. $\mathbf{X}_i^{gt}$ is the concatenated features in ground-truth clips, *i.e.*, $\mathbf{X}_i^{gt} = [\mathbf{x}_{t_s^i}^i, .., \mathbf{x}_{t_e^i}^i]$. Φ' is the transformation layer which can be convolutional layers [40] or self-attention [43].

We name the model training with Eq. 2 and Eq. 3 as the GT clip-based Learning. The model will push the sentence feature $\mathbf{Y}_i$ and its corresponding GT clip features to be close, while push $\mathbf{Y}_i$ away from Non-GT clip features. Non-GT clip features of $\mathbf{Y}_i$ contain non-ground-truth clips and clips in other videos. However, in the same video, the entities may appear in both the GT and Non-GT clips, rather than only GT clips, as shown in Fig. 4 (a). By simply attracting $\mathbf{Y}_i$ and GT clip features and repulsing $\mathbf{Y}_i$ and Non-GT clip features, the GT clip-based learning would make the same entity (yellow cube in Fig. 4) in background clips also be far away from that in ground-truth clips. Hence, this method is too strict and the learned representations of video clips may be far away even they have similar semantics.

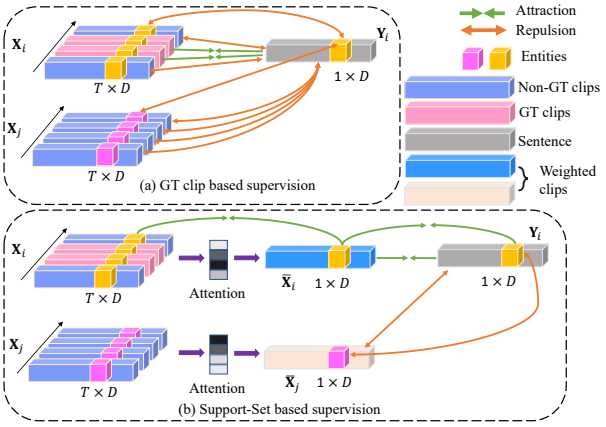


Figure 4. (a) GT clips based supervision. The GT clips based learning aims to encourage the GT clip features to be close with $\mathbf{Y}_i$ and push away the Non-GT clip features. (b) Support-set based supervision. Considering there are also entities from the query in Non-GT clips, *i.e.*, the yellow cube, we maximize the similarity between the weighted feature ($\overline{\mathbf{X}}_i$) and $\mathbf{Y}_i$.

Support Set-Based Supervision. To address the mutual exclusion between the videos and texts as analyzed above, we propose a support-set based supervision approach. Our motivation is that different clips in a video may share same semantic entities. For example, given a sentence query 'The person pours some water into the glass' and its corresponding video, the person entity and glass entity appear throughout the video, as shown in Fig. 2, and only in GT clips, the action of 'pour water' occurs. Although there are no 'pour water' happening in Non-GT clips, the semantics of them are also similar with those of ground-truth ones, *e.g.*, the semantics of 'The person pours some water into the glass' is much close to that of 'The person hold a glass', rather than that of 'Person put a notebook in a bag'. If we strictly push away the representations of the Non-GT clips, the model would only extract 'pour water' in the video and the text, while ignoring 'person' and 'glass'.

In order to make the learned representations of Non-GT clips with the same entities also have a certain degree of similarity with the corresponding text, we introduce a *support-set*, defined as S_i, for each text L_i. The clips in S_i normally have the same entities. In this work, we set all clips in a video as the support-set of its corresponding text, *i.e.*, $S_i = \{\mathbf{x}_t^i\}_{t=1}^{T}$, where $\mathbf{x}_t^i \in \mathbb{R}^D$ is the embed-

ding of v_t^i. This is because in video grounding, clips in the same video usually belong to the same scene, and most of the people and things in those clips are similar or even the same. Based on the support-set S_i, we first compute the similarity between all clips in S_i and L_i and then the clip-wise attention can be obtained as a softmax distribution over clip indices:

$$a_t^i = \frac{e^{\langle \mathbf{x}_t^i, \mathbf{Y}_i \rangle / \tau}}{\sum_{\mathbf{x} \in \mathcal{S}_i} e^{\langle \mathbf{x}, \mathbf{Y}_i \rangle / \tau}}, \tag{4}$$

where, $\langle \mathbf{x}_t^i, \mathbf{Y}_i \rangle$ is the cosine similarity between $\mathbf{x}_t^i$ and $\mathbf{Y}_i$. Then, we compute the weighted average of the embeddings in S_i as follows:

$$\overline{\mathbf{w}}_i = \sum_{t=1}^{T} a_t^i \cdot \mathbf{x}_t^i. \tag{5}$$

After acquiring $\overline{\mathbf{w}}_i$, we can redefine the positive samples and the negative samples in $\mathcal{B}$. Concretely, we set the $\{(\overline{\mathbf{w}}_i, \mathbf{Y}_i)\}_{i=1}^B$ as positive samples, and other pairs in the batch as negative ones, i.e, $\overline{\mathcal{N}}_i = \{(\overline{\mathbf{w}}_i, \mathbf{Y}_j)\}_{i \neq j}$. Then, the contrastive objective can be defined as follows:

$$\mathcal{L}^{\text{contrast}} = -\sum_{i=1}^{B} \log \left(\frac{e^{\overline{\mathbf{w}}_i^\top \mathbf{Y}_i / \tau}}{e^{\overline{\mathbf{w}}_i^\top \mathbf{Y}_i / \tau} + \sum\limits_{(\mathbf{x}', \mathbf{y}') \in \overline{\mathcal{N}}_i} e^{\mathbf{x}'^\top \mathbf{y}' / \tau}} \right), \tag{6}$$

and the caption objective is:

$$\mathcal{L}^{\text{caption}} = -\frac{1}{B} \sum_{i=1}^{B} \log p\left(l_i \mid \overline{\mathbf{w}}_i\right). \tag{7}$$

We name the model training with Eq. 6 and Eq. 7 as the support-set based supervision. As Fig. 4 (b) shows, besides pushing the sentence feature $\mathbf{Y}_i$ and its corresponding GT clip feature to be close, the representations of the same entity (yellow cube) in both Non-GT clip features and the sentence feature are also be attracted.

Comparison between [33] and Ours. The main differences with SS are two-fold: i) Motivations. Our goal is to apply the cross-supervision to capture the relations between the visual semantics and textual concepts. While [33] aims to improve video-text representations by relaxing the contrastive objective; ii) Solutions. In SS, the cross-captioning objective is to relax the strict contrastive objective, hence it's an adversarial relationship actually. While in Scsc, our two objectives are in a cooperative relationship because they both aim to learn the video-text relations. Furthermore, our contrastive objective is build on global video features encoded by support-set, while [33] applies a triplet ranking loss based on local clip features.

Table 1. Ablation study of different supervision methods on the Charades-STA dataset.

Model	$\mathcal{L}^{\text{contrast}}$	$\mathcal{L}^{\text{caption}}$	$Rank1@$		$Rank5@$	
			0.5	0.7	0.5	0.7
2D-TAN [51]			50.62	28.71	79.92	48.52
2D-TAN+GTC	✓		54.77	31.63	86.28	55.07
		✓	51.72	29.35	83.66	52.12
	✓	✓	55.40	32.15	87.07	55.62
2D-TAN+SS	✓		56.19	32.03	87.95	56.05
		✓	53.12	30.05	85.19	53.28
	✓	✓	**56.97**	**32.74**	**88.65**	**56.91**
LGI [30]			59.46	35.48	-	-
LGI+GTC	✓		59.63	35.71	-	-
		✓	59.88	35.92	-	-
	✓	✓	60.02	36.11	-	-
LGI+SS	✓		60.09	36.32	-	-
		✓	60.53	36.75	-	-
	✓	✓	**60.75**	**37.29**	-	-

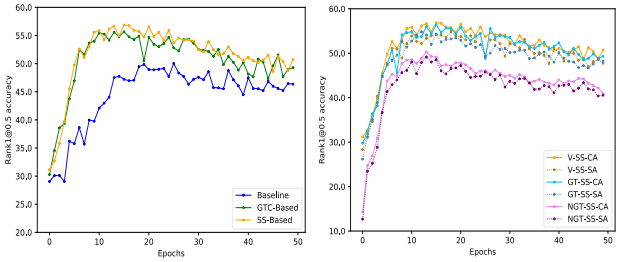
3.4. Several Kinds of Support-Set

The support-set based supervision contains two basic operations: (a) *the construction of the support-set* $\mathcal{S}_i$; and (b) *the function to map the support-set to a weighted embedding* $\overline{\mathbf{w}}_i$. In this section, we explore three kinds of functions to construct the support-set: **(a) video-level support set (V-SS)**: we set all clips in a video as the support, *i.e.*, $\mathcal{S}_i = \{\mathbf{x}_t^i\}_{t=1}^T$; **(b) ground-truth-level support set (GT-SS)**, which only contains the GT clips, *i.e.*, $\mathcal{S}_i = \mathcal{M}_i$; **(c) Non-GT level support set (Non-GT-SS)**: which only contains the Non-GT clips, *i.e.*, $\mathcal{S}_i = \overline{\mathcal{M}}_i$.

By these functions, we compare six ways as follows: **(a) cross-attention (CA).** The function is computed by Eq. 4 and Eq. 5; **(b) self-attention (SA).** We first concatenate the clips in $\mathcal{S}_i$ along clip indices to obtain $\mathbf{S}_i$, then we compute the similarity matrix of $\mathbf{S}_i$ by $\mathbf{Q}_i = \mathbf{S}_i^\top \mathbf{S}_i / \tau$. The t-th vector of $\mathbf{Q}_i$ is $\mathbf{q}_t^i \in \mathbf{R}^D$. Sum all the elements of $\mathbf{q}_t^i$ to obtain the summed scalar z_t^i. Then we obtain the clip-wise attention as follows:

$$a_t^i = \frac{e^{z_t^i}}{\sum_{z \in \mathcal{Z}_i} e^z}, \tag{8}$$

where $\mathcal{Z}_i$ is the set of all z_t^i for $\mathbf{Q}_i$. Finally, $\overline{\mathbf{w}}_i$ can be obtained by Eq. 5. **(c) fully-connected layer (FC).** In this way, after concatenating the clips in $\mathcal{S}_i$ along clip indices, the concatenated feature $\mathbf{S}_i$ is converted to $\overline{\mathbf{w}}_i$ by a fully-connected layer. **(d) Convolutional layer (Conv).** Similar to FC, we fed $\mathbf{S}_i$ into a convolutional layer to acquire $\overline{\mathbf{w}}_i$. **(e) Max-pooling (MP).** In this way, after concatenating the clips in $\mathcal{S}_i$ along clip indices, the concatenated feature $\mathbf{S}_i$ is fed into a max-pooling layer to acquire $\overline{\mathbf{w}}_i$. **(f) Average-pooling (AP).** Similar to MP, we fed $\mathbf{S}_i$ into a average-pooling layer to acquire $\overline{\mathbf{w}}_i$.



(a) Comparison of different learning methods (b) Comparison of different kinds of support-set

Figure 5. (a) Comparison of the accuracy curve of different learning methods. (b) Comparison of the accuracy curve of different kinds of support-set.

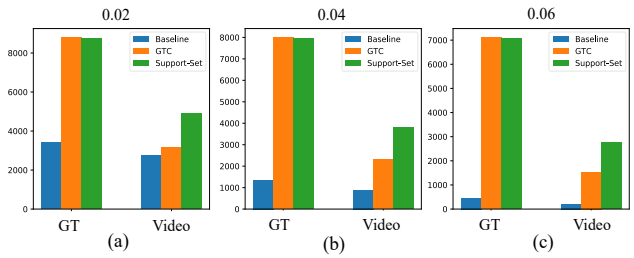


Figure 6. Comparison of recalls of high-related video-text pairs under different thresholds.

4. Experiments

4.1. Datasets

TACoS. TACoS is collected by Regneri *et al.* [36] which consists of 127 videos on cooking activities with an average length of 4.79 minutes for video grounding and dense video captioning tasks. We follow the same split of the dataset as Gao *et al.* [12] for fair comparisons.

Charades-STA. Charades is originally collected for daily indoor activity recognition and localization [39], which consists of $9,848$ videos. Gao *et al.* [12] build the Charades-STA by annotating the temporal boundaries and sentence descriptions of Charades[39].

ActivityNet-Captions. ActivityNet [4] is a large-scale dataset which is collected for video recognition and temporal action localization [25, 6, 13, 11, 38, 31, 34]. Krishna *et al.* [23] extend ActivityNet to ActivityNet-Captions for the dense video captioning task.

4.2. Implementation details

Evaluation metric. For fair comparisons, we follow the setting as previous work [12] and evaluate our model by computing $Rank\ n@m$. Specifically, it is defined as the percentage of 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 [51], we evaluate our method with specific settings of n and m for different datasets.

Feature Extractor. For a fair comparison, we extract video features following previous works [51, 49]. Specifically, We use the C3D [42] network pre-trained on Sports-1M [20] as the feature extractor. For Charades-STA, we also use VGG [40], C3D [42] and I3D [5] feature to compare out results with [12, 51]. We divided the video into segments and each contains fixed number 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. τ is set to 0.1. For Charades-STA, λ_1 and λ_2 are set to 0.1, and for TACoS, λ_1 and λ_2 are set to 0.001.

Baseline Model. Our work is built on two current state-of-the-art models for video grounding, 2D temporal adjacent network (2D-TAN) [51] and local-global video-text interactions (LGI) [30].

Training settings. We use Adam [21] with learning rate of 1.6×10^{-2} and batch size of 256 for optimization. We decay the learning rate with ReduceLROnPlateau function in Pytorch [32]. All of our models are implemented by PyTorch and trained under the environment of Python 3.6 on Ubuntu 16.04.

4.3. Ablation Study

In this section, all presented results are on Charades-STA [12] with I3D [5] features. For convenience, we use 'GTC' and 'SS' to refer to GT clip and support-set based supervision in following experiments.

Comparison of different supervision methods. In this ablation study (Table 1), we compare different learning methods, proposed in Section 3.3.2, including GT clip based supervision and support-set based supervision. It is clear from Table 1, the performance of SS outperforms that of GTC with large margins. What's more, the contrastive objective $\mathcal{L}^{\text{contrast}}$ bring larger performance improvement than the caption one $\mathcal{L}^{\text{caption}}$. Combining both the contrastive objective and the caption objective, our model obtain the best performance. The interactions of videos and text in 2D-TAN [51] is Hadamard product, and those in LGI [30] is in a coarse-to-fine manner which is more fine-grained than 2D-TAN. For 2D-TAN, the interaction and grounding modules are compute the similarity between video clips and text to predict target intervals, which is very similar to the objective of cross-supervision. Hence, our method achieves larger improvement in 2D-TAN. As Fig. 5 (a) indicates, with an extra Cscs branch, besides the higher performance, the model converges faster than the baseline method.

Comparison with different kinds of the support-set. In this ablation study, we compare different kinds of construction methods and function methods of the support-set. Table 2 presents the performance of different kinds of the support-set on the Charades-STA dataset. Specifically, we compare

Table 2. Ablation study of different kinds of construction methods and function methods on the Charades-STA dataset.

Construction method	Function method						$Rank1@$		$Rank5@$	
	CA	SA	FC	Conv	MP	AP	0.5	0.7	0.5	0.7
V-SS	✓						**56.97**	**32.74**	**88.65**	**56.91**
		✓					54.88	30.98	86.56	54.92
			✓				54.91	31.25	86.75	55.01
				✓			54.89	31.08	86.75	54.73
					✓		53.35	30.64	86.13	54.35
						✓	53.14	30.36	86.10	54.13
GT-SS	✓						55.91	32.03	88.12	55.25
		✓					54.89	31.23	87.11	54.40
			✓				54.90	31.17	87.10	54.85
				✓			54.85	31.10	87.52	54.88
					✓		53.62	30.80	86.54	54.79
						✓	53.70	30.91	86.78	54.88
Non-GT-SS	✓						50.12	28.96	85.82	52.78
		✓					48.55	26.64	83.27	50.62
			✓				48.52	26.56	83.31	50.64
				✓			48.29	26.44	81.13	50.44
					✓		48.87	26.57	83.40	50.60
						✓	48.33	26.48	83.34	50.52

Table 3. Comparisons with state-of-the-arts on Charades-STA.

Methods	Feature	$Rank1@$		$Rank5@$	
		0.5	0.7	0.5	0.7
VAL [41]	VGG	23.12	9.16	61.26	27.98
ACL-K [14]	VGG	30.48	12.20	64.84	35.13
TripNet [16]	VGG	36.61	14.50	-	-
DRN [49]	VGG	42.90	23.68	**87.80**	**54.87**
2D-TAN [51]	VGG	39.70	23.31	80.32	51.26
2D-TAN + ours	VGG	**43.15**	**25.54**	84.26	54.17
LGI [30]	VGG	41.72	21.48	-	-
LGI + ours	VGG	43.68	23.22	-	-
MAN [50]	I3D	46.53	22.72	86.23	53.72
DRN [49]	I3D	53.09	31.75	**89.06**	**60.05**
2D-TAN [51]	I3D	50.62	28.71	79.92	48.52
2D-TAN + Ours	I3D	56.97	32.74	88.65	56.91
LGI [30]	I3D	59.46	35.48	-	-
LGI + Ours	I3D	**60.75**	**36.19**	-	-

three types of construction methods: (a) V-SS, (b) GT-SS, (c) Non-GT-SS and six ways of function methods: (a) CA, (b) SA, (c) FC, (d) Conv, (e) MP, (e) AP (See details in Section 3.4). Our proposed method (V-SS + CA) achieve the best performance. The V-SS way can make the learned representation explore more similar entities in non-ground-truth clips. CA aims to find the high similarity between videos and text, while other function methods (*e.g.*, SA, FC, etc.) only considering the single modality information (i.e, videos). Hence, CA is more effective in the support-set. Since Non-GT-SS only contains non-ground-truth clips, the learned representation of the videos and text would let the ground-truth clips have dissimilar semantics, resulting in poor performance in video grounding. The comparison of accuracy curves presented in Fig. 5 (b).

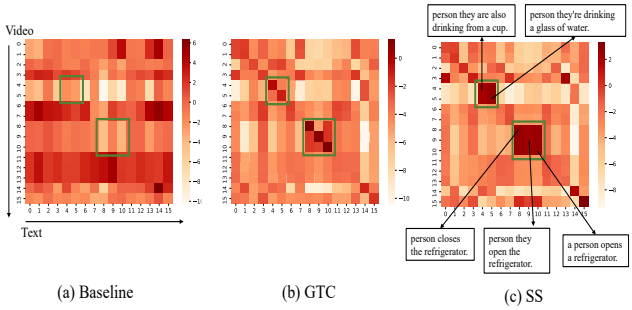
Recalls of high-related video-text pairs. In order to verify that the proposed approach can enhance the correlation between text and videos, we present the recalls of high-similar video-text pairs under different thresholds (0.02, 0.04 and 0.06) in Fig. 6. 'Video' indicates the average similarity between clips in the whole video and text, and 'GT' is the average similarity between GT clips and text. It is clear that, adding cross-supervision module can significantly improve the similarity between the video and text. Support-set based approach can have a more generalized representation, compared with the GT clip based learning.

4.4. Comparison with the State-of-the-Arts

We conduct experiments on TACoS, Charades-STA and ActivityNet-Captions datasets to compare with several State-Of-the-Art (SOTA) approaches. From Table 3 and Table 4, it clearly shows that the proposed method can largely improve the SOTA models, *i.e.*, 2D-TAN [51] and LGI [30], almost without any extra inference cost. We can also see that Sscs achieve smaller gain with LGI. The reasons may be that LGI is a *regression* based method that directly regress the boundaries, while 2D-TAN is a *comparison and selection* based method that compares text with dense proposals and selects the best one. In Scsc, SS is built on a contrastive objective, which has a similar spirit with 2D-TAN, hence it achieves larger gains on 2D-TAN. Furthermore, with 2D-TAN, SS obtains larger gains by 6.35% and 4.24% on Charades and TACoS than that 2.16% on Activities. We think it because that Charades and TACoS have static and smooth backgrounds and simple actions, while ActivityNet is more complex and diverse. Thus the improvement on ActivityNet is relatively small.

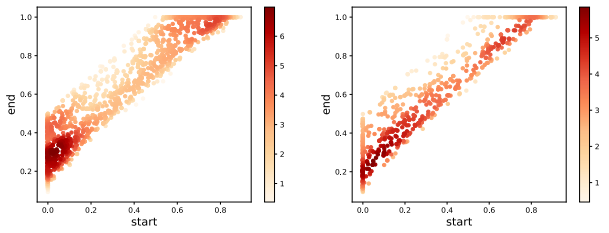
Table 4. Comparisons with state-of-the-arts on TACoS and ActivityNet-Captions.

Methods	TACoS						ActivtiyNet-Captions					
	Rank1@			*Rank5@*			*Rank1@*			*Rank5@*		
	0.1	0.3	0.5	0.1	0.3	0.5	0.3	0.5	0.7	0.3	0.5	0.7
TGN [7]	41.87	21.77	18.9	53.40	39.06	31.02	43.81	27.93	-	4.56	44.20	
ACRN [27]	24.22	19.52	14.62	47.42	34.97	24.88	49.70	31.67	11.25	76.50	60.34	38.57
CMIN [52]	32.48	24.64	18.05	62.13	38.46	27.02	-	-	-	-	-	-
QSPN [46]	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 [48]	34.70	19.50	9.40	-	-	-	55.67	36.79	-	-	-	-
DRN [49]	-	-	23.17	-	-	33.36	-	45.45	24.36	-	77.97	50.30
HVTG [8]	-	-	-	-	-	-	57.60	40.15	18.27	-	-	-
2D-TAN [16]	47.59	37.29	25.32	70.31	57.81	45.04	59.45	44.51	26.54	85.53	77.13	61.96
2D-TAN + Ours	**50.78**	**41.33**	**29.56**	**72.53**	**60.65**	**48.01**	**61.35**	**46.67**	**27.56**	**86.89**	**78.37**	**63.78**
LGI [30]	-	-	-	-	-	-	58.52	41.51	23.07	-	-	-
LGI + Ours	-	-	-	-	-	-	59.75	43.62	25.52	-	-	-



(a) Baseline (b) GTC (c) SS

Figure 7. similarity matrix of (a) the baseline, (b) GTC and (c) SS. We present 16 video-text sample pairs.



(a) Baseline (b) Ours

Figure 8. (a) The distributions of successfully predicted time intervals by the baseline. (b) The distributions of our model additionally success-predicted time intervals, compared with the baseline model.

4.5. Qualitative analysis

In this section, we present some qualitative results on Charades-STA. We present the similarity matrix of video-text pairs in Fig. 7. It is clear that the baseline model can not capture the semantic similarity of video-text pairs even they come from the same sample (see Fig. 7 (a)). On the contrary, the similarity score of videos and text from the same sample would be higher than others. Compared with GTC, SS can also capture the related semantics pairs, even they are not from the same sample. As Fig. 7 shows, the text in 4-th and 5-th samples have similar semantics, and the similarity of the corresponding videos are also high, which are not found in the baseline model and GTC.

Fig. 8 shows the successfully predicted time interval distributions. It is clear that most of the baseline model predicted time intervals are generally concentrated at the beginning of the video, and the duration of the fragments are about $20\% - 40\%$ of the total length of the video, as shown in Fig. 8 (a). Compared with the time intervals predicted by the baseline model, the proposed method can find more time intervals occurring in the middle of the videos, also the duration of the time intervals are shorter, which is indicated in Fig. 8 (b). This reason is that the proposed method can learned better video-text representations, thus we can find more time intervals that are difficult to locate.

5. Conclusion

In this paper, we introduce a Support-Set Cross-Supervision (Sscs) Module as an extra branch to video grounding to extract the correlation between videos and text. By conducting the contrastive and caption objective to the clip-level and sentence features in a shared space, the learned two-modality features are enforced to become similar, only if the semantics of them are related. To address the mutual exclusion of entities, we improve the cross-supervision with the support-set to collect all important visual clues from the whole video. The experimental results shows the proposed method can greatly improve the performance of the state-of-the-art backbones almost without any extra inference cost, and the ablation study verify the effective of the support-set.

Acknowledgements. 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 Grants 62036007, 61922066, 61876142, 61772402, and 6205017; in part by the Fundamental Research Funds for the Central Universities.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[2] Andre Araujo and Bernd Girod. Large-scale video retrieval using image queries. *IEEE transactions on circuits and systems for video technology*, 28(6):1406–1420, 2017.

[3] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *CVPR*, pages 9922–9931, 2020.

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[6] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*, pages 1130–1139, 2018.

[7] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *ICEMLP*, pages 162–171, 2018.

[8] Shaoxiang Chen and Yu-Gang Jiang. Hierarchical visual-textual graph for temporal activity localization via language. 2020.

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Xinpeng Ding, Nannan Wang, Xinbo Gao, Jie Li, Xiaoyu Wang, and Tongliang Liu. Weakly supervised temporal action localization with segment-level labels. *arXiv preprint arXiv:2007.01598*, 2020.

[12] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017.

[13] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. *arXiv preprint arXiv:1705.01180*, 2017.

[14] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *WACV*, pages 245–253. IEEE, 2019.

[15] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[16] Meera Hahn, Asim Kadav, James M Rehg, and Hans Peter Graf. Tripping through time: Efficient localization of activities in videos. *arXiv preprint arXiv:1904.09936*, 2019.

[17] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, volume 33, pages 8393–8400, 2019.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Bruno Korbar, Fabio Petroni, Rohit Girdhar, and Lorenzo Torresani. Video understanding as machine translation. *arXiv preprint arXiv:2006.07203*, 2020.

[23] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.

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

[25] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 3–19, 2018.

[26] Daizong Liu, Xiaoye Qu, Xiaoyang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. *arXiv preprint arXiv:2008.01403*, 2020.

[27] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *ACM SIGIR*, pages 15–24, 2018.

[28] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020.

[29] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.

[30] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, pages 10810–10819, 2020.

[31] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *ICCV*, pages 8679–8687, 2019.

[32] A. Paszke, S. Gross, and S. Chintala. Pytorch deep learning framework. Web page, 2017.

[33] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020.

[34] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*, pages 563–579, 2018.

[35] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[36] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[38] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, pages 1049–1058, 2016.

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

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] Xiaomeng Song and Yahong Han. Val: Visual-attention action localizer. In *Pacific Rim Conference on Multimedia*, pages 340–350. Springer, 2018.

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

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

[44] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, pages 334–343, 2019.

[45] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems*, 32:6838–6849, 2019.

[46] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, volume 33, pages 9062–9069, 2019.

[47] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NeurIPS*, pages 536–546, 2019.

[48] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, volume 33, pages 9159–9166, 2019.

[49] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, pages 10287–10296, 2020.

[50] Da Zhang, Xiyang Dai, Xin Wang, Yuan Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2018.

[51] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. *arXiv preprint arXiv:1912.03590*, 2019.

[52] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *ACM SIGIR*, pages 655–664, 2019.

[53] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, pages 8739–8748, 2018.

[54] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, pages 19–27, 2015.