Activity-driven Weakly-Supervised Spatio-Temporal Grounding from Untrimmed Videos

Junwen Chen, Wentao Bao, Yu Kong
{jc1088,wb6219,yu.kong}@rit.edu
Rochester Institute of Technology
Rochester, New York

ABSTRACT

In this paper, we study the problem of weakly-supervised spatio-temporal grounding from raw untrimmed video streams. Given a video and its descriptive sentence, spatio-temporal grounding aims at predicting the temporal occurrence and spatial locations of each query object across frames. Our goal is to learn a grounding model in a weakly-supervised fashion, without the supervision of both spatial bounding boxes and temporal occurrences during training. Existing methods have been addressed in trimmed videos, but their reliance on object tracking will easily fail due to frequent camera shot cut in untrimmed videos. To this end, we propose a novel spatio-temporal multiple instance learning framework for untrimmed video grounding. Spatial MIL and temporal MIL are mutually guided to ground each query to specific spatial regions and the occurring frames of a video. Furthermore, an activity described in the sentence is captured to use the informative contextual cues for region proposals refinement and text representation. We conduct extensive evaluation on YouCookII and RoboWatch datasets, and demonstrate our method outperforms state-of-the-art methods.

CCS CONCEPTS

• **Computational methodologies-Multimedia,Computer vision**;

KEYWORDS

Weakly-supervised object grounding, Untrimmed video, Activity, Multiple instance learning

ACM Reference Format:
Junwen Chen, Wentao Bao, Yu Kong. 2020. Activity-driven Weakly-Supervised Spatio-Temporal Grounding from Untrimmed Videos. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413614

1 INTRODUCTION

Grounding natural language in visual data is a fundamental task in the multimedia and computer vision communities with a variety of applications, including image/video retrieval [15], robotics [1] and human-computer interactions [26]. Given an image/video and its description sentence, for example "break the eggs", visual grounding aims at localizing the query objects described in the sentence on the given image or video. Recently, great progress has been made on image grounding [4, 15, 30, 31]. On the basis of this, researchers started to explore grounding in the video domain [5, 7, 14, 25, 35].

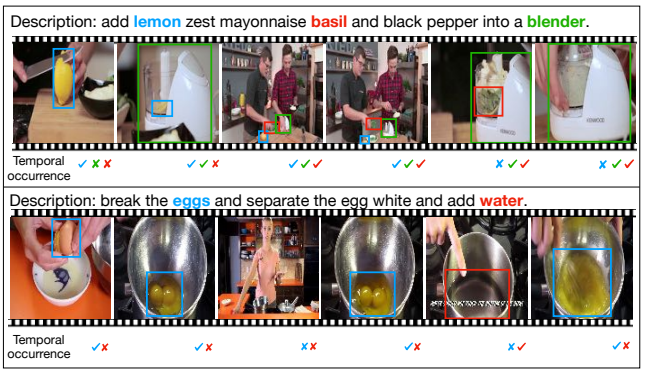

Figure 1: Given a video and its description sentence, our goal is to achieve the spatio-temporal grounding of the described queries on challenging untrimmed videos, where camera shot cuts frequently appear. Spatio-temporal grounding grounds each query to specific spatial regions and the frames of a video where the query object appears.

Nevertheless, in video grounding, it is labor-intensive to annotate a considerable number of bounding boxes for queries in videos. To address this challenge, multiple instance learning (MIL) methods [5, 7, 14, 25, 35] were proposed, which do not require bounding box annotations in the training videos. Video object grounding is achieved in a weakly-supervised fashion, where only a video and its description sentence are required during training. However, these methods are only able to infer the spatial occurrence of the query objects, and cannot tell the temporal occurrence of the objects. This problem was later addressed in [7] by generating region proposal tubes using object tracking methods. But their method is only applicable to trimmed videos without camera shot cuts.

We argue that a successful video grounding method should infer both the spatial and temporal occurrence of a query object without the need of expensive annotations. In addition, the method is expected to work in untrimmed videos, which can be of long duration and contain frequent visual inconsistency mainly caused by frame flickerings and camera shot cuts (see Fig. 1). A query object may appear discontinuously (it frequently appears and disappears) across frames in an untrimmed video. Existing video grounding

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413614

methods [7] that rely on visual trackers [28, 29] would undoubtedly fail as the trackers can be distracted if a camera shot cut appears.

We propose a novel multiple instance learning method for spatio-temporal grounding on untrimmed videos. Our method does not require extensive annotations of spatial and temporal occurrence[1] of the query object in training. At the spatial level, we assign each textual query to one of the region proposals in a frame, while at the temporal level, we represent each frame by query-specific region and ground each query to its relevant frames. We formulate spatio-temporal grounding as two MIL problems. The spatial MIL aims at selecting the best instance (top-ranked region) from a bag (frame). The temporal MIL aims at selecting the multiple instances (query occurring frames) from a bag (video). Two MILs are mutually guided to achieve the optimal spatio-temporal grounding results.

We also propose to model human activity operating on the query object. This allows us to capture the physical states of the object as well as the spatial relations between human and the object. Most of existing visual grounding methods [5, 15, 25, 35] simply compute the similarity between the visual and the textual features of the query object as a measurement for selecting candidate regions for the query. However, there is a granularity gap between the coarse textual and rich visual modalities. For example, the text-level query object "potatoes" might correspond to "potatoes" in different physical states: "mashed potatoes" means visually paste-like potatoes, while "peel potatoes and cut" corresponds to cube-shape potatoes. Directly computing the feature similarity as in [15, 25, 35] leads to a large discrepancy between the text and visual features. To address this, first, we propose to enrich the textual representation by incorporating the activity performed on the query to better align the text feature with the diverse visual features. Second, we propose an activity-driven region proposal refinement to find high-quality region proposals. Most of existing visual grounding methods [5, 15, 25, 35] build a candidate pool of top-N region proposals, in which query objects could be missed. Proposals with a high recall rate by increasing N typically lead to a large search space for grounding a query object. To tackle this dilemma, we exploit the intrinsic spatial relations between human and object using human activity to refine the search space for region proposal generation.

Our work is different from [25, 35], which also focus on grounding untrimmed videos. However, they ground query objects every frame even if the frame does not contain the query. This could result in a lot of false positives in the frames where the queried object does not appear due to its sparse existence in a long untrimmed video. On the contrary, we infer both the bounding box and the temporal occurrence of the query object. Therefore, our method can be used in more realistic scenarios.

Our main contribution can be summarized as follows: 1). We propose a spatio-temporal Multiple Instance Learning method to learn a spatio-temporal video grounding model for the challenging untrimmed videos in a weakly-supervised fashion; 2). We exploit the activity cues in the description sentence of the video, including enriching the query representation with activity effect and refine the object proposal generation; 3). Extensive results demonstrate that our method outperforms state-of-the-art weakly-supervised object grounding model in untrimmed videos by a large margin.

[1]Temporal occurrence of an object means the object appear in some frames of a video.

2 RELATED WORK

Weakly-supervised Visual Grounding. Weakly-supervised image grounding [4, 15, 23] has been extended to video domain [5, 7, 14, 25, 35], but they are only applicable to constrained scenarios. Early work [33] grounded sentences to objects in the constrained videos that are recorded in lab. A reference grounding model [14] extends proposal ranking [15] to video domain and further enhances the performance by modeling the reference relationships between video segments. Following [15], the work in [35] extends proposal ranking to video domain via a frame-wise weighting strategy. They also introduce an object grounding dataset based on YouCookII [36]. The work in [5, 25] follow the same problem setup as [35] and boost grounding performance by using contextual similarity and cross-modal context reasoning. However, during inference [5, 25, 35] only ground query in the frames where the objects occur without grounding frame occurrence in the temporal domain. Thus, the output of their methods contain a lot of false positives in the frames without the presence of query objects. The VID-sentence dataset is introduced in [7], which first grounds spatio-temporal tubes for a query. But their method and dataset are only for trimmed videos.

In this work, we aim at object grounding on untrimmed video streams by localizing the query objects in both spatial region and frame-level occurrence. Our method does not rely on tracking tubes due to frame flickering and camera shot cut in untrimmed videos.

Fully-supervised Spatio-Temporal Grounding has been developed by combining with other tasks, such as object tracking [32], video captioning [34] and visual question answering [2]. Yang et al. [32] add language description on an object tracking dataset [10] to make it for grounding task and propose a grounding and tracking integration model. But this dataset only contains single object in a video, which is much easier than our goal that multiple objects in a query need to be grounded. Zhou et al. [34] augment the challenging ActivityNet Captions dataset with 158K bounding boxes annotations and provide a framework to not only generate video captions but also link the sentence to the evidence in the video. However, these methods require dense spatio-temporal tube annotations for training, which are especially expensive to obtain. Our paper aims at solving spatio-temporal grounding in a weakly-supervised setting without bounding box annotations.

Weakly-supervised Video Object Localization localizes an object class or a video tag in the visual content. Object class or video tag comes from human labeling while the descriptive sentence in visual grounding can be accessed from the existing web video descriptions uploaded by users or the YouTube Automatic Speech Recognition scripts, which requires less human effort. Existing work of weakly-supervised object localization also formulate it as an MIL approach. Kwak et al. [17] integrated object tracking and frame-wise object detection together to achieve video object localization. Prest et al. [21] extract spatio-temporal tubes as proposals to be ranked and selected. However, similar to [7, 32], these methods heavily rely on tracking which is not applicable to long untrimmed video due to camera shot cut.

Weakly-supervised Temporal Grounding focuses on identifying relevant frames in a video from text descriptions without the annotations of temporal boundaries. Existing work [6, 9, 11, 19] extract a set of pre-defined temporal segment proposals and select

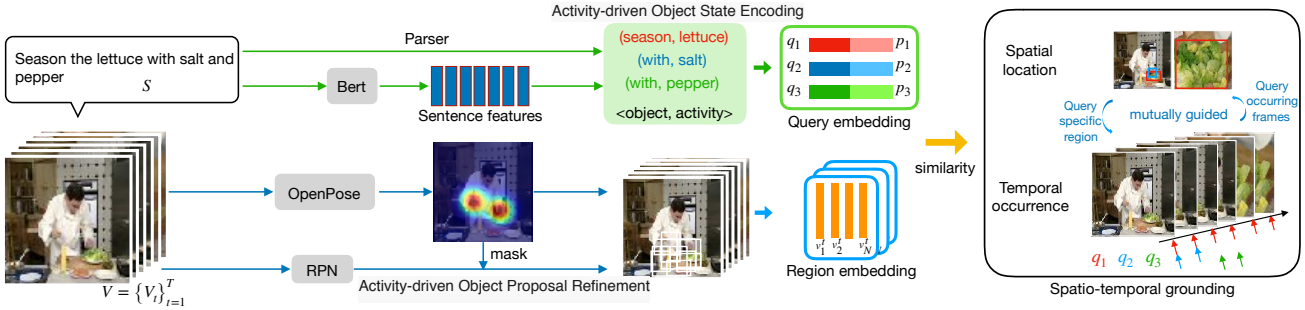

Figure 2: Overview of our framework. Given a video and its description sentence as input, we first extract the region features via a pre-trained region proposal network (RPN) and find the high-quality proposals by the proposed activity-driven object proposal refinement module. Given text data, we first encode the sentence by BERT and propose an activity-driven object state encoding module to enrich query representation by incorporating activity effect. Then, through a similarity alignment between the two modalities, spatio-temporal grounding is achieved by retrieving the frames that contain the queries and selecting the best-match spatial region in the frames where the queries appear. During training, the spatial and temporal levels are mutually guided. Training details can be seen in Section 3.3.

one of them that semantically best matches the description. Our task is more challenging as we ground the query not only to its occurring frames but also to specific locations. Moreover, query appears discontinuously in temporal domain due to the frequent camera shot cut, which may not be addressed by finding the best matched temporal proposal.

3 OUR APPROACH

3.1 Problem Setup

Given an untrimmed video and its description sentence, video grounding task grounds each query object described in the sentence to spatio-temporal visual regions in the video. The query can be either a noun, e.g. word "potato" or a pronoun with reference meaning, e.g. "they". Video grounding on untrimmed videos is of great significance while more challenging than trimmed video, since the untrimmed video contains large temporal incoherence caused by camera motion and camera shot cut[2].

We propose a spatio-temporal grounding model that can be applied to untrimmed videos with significant camera motion. Our model is trained in a weakly-supervised fashion, where only the video-description pairs are given during training; spatial and temporal occurrence of the query objects are not given. As shown in Fig. 2, our grounding model takes a video V and its description sentence S as a pairwise input, and predicts the temporal occurrence (i.e., what frames contain the object) of the query object across frames and the spatial location (using a bounding box) of the object on the frames where it appears. The video contains T frames and is denoted as $V = \{V_t\}_{t=1}^{T}$. Following [25, 35], each frame consists of N region proposals, denoted as $V_t = \{v_t^n\}_{n=1}^{N}$, where n indexes the proposals in the t-th frame. The description sentence S includes K queries, e.g. query "lettuce" and "pepper" in the description sentence "season the lettuce with salt and pepper". Each query s_k corresponds to a word or a phrase in S and all of queries in a sentence is denoted as $\{s_k\}_{k=1}^{K}$. The visual feature v_t^n and query feature Q_k of a query are encoded into a joint feature space, and their similarity is computed for the grounding purpose.

We formulate spatio-temporal grounding as a multiple instance learning (MIL) problem for untrimmed videos. We propose two ranking losses (one on spatial level and one on temporal level) mutually guiding each other to learn a shared metric space for grounding. We consider a weakly-supervised learning scenario without the annotations of bounding boxes and temporal occurrences. An activity-driven encoder is proposed to better align the visual and text modalities by considering the object state variations and spatial location prior of region proposals.

3.2 Activity-driven Encoding

Activity cues in both text and visual modalities are informative for grounding objects in an untrimmed video. For example, as shown in Fig. 3, the activity in the description sentence "mash the potatoes" results in paste-like potatoes in the visual data, while "peel potatoes and cut" results in cube-shaped potatoes. By modeling activities, various physical states of an object can be modeled at a fine-grained representation level, which allows us to accurately ground the object. In addition, the activity provides a spatial location prior for the object to be grounded. For example, "cut potatoes" indicates that query potatoes should appear close to human hand. The spatial location prior can be exploited to refine the candidate region proposals of visual data.

3.2.1 Activity-driven Object State Encoding. To encode the query into a representative feature, previous work [25, 35] extract each query word (e.g. "potatoes") from the description sentence and then represent it based on GloVe features [20]. This is ineffective because there is a semantic granularity gap between the text modality and visual modality (see Fig. 3). Existing methods simply attach the same textual representation to the diverse visual representation, which results in text-visual misalignment problem.

We propose to enrich the textual representation and align it with diverse visual objects. This allows us to capture rich cues of object

[2] A camera shot cut is the view change from one shot to another, e.g., from a distant view shot to a close view shot.

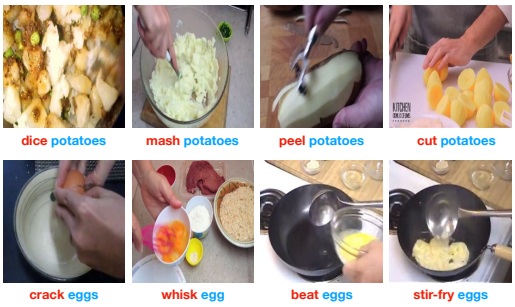

Figure 3: Examples of an object in different states. For example, "potatoes" in "mashed" and "peeled" are different in appearance. Blue word indicates query object and red word indicate the activity applied on the object. Best view in color.

physical states. Specifically, we introduce an activity-driven object state encoding module to enrich the query textual representation. We consider each query with the predicate performing on it and reformulate each query as an object-activity pair. We use Stanford CoreNLP parser [18] to parse each noun or pronoun and its predicate from a sentence S. Meanwhile, each sentence is encoded by a pre-trained BERT model [8]. Then, we crop the features of the k-th query and its predicate from the sentence representation as q_k and p_k, respectively. The textual representation of k-th query s_k in the sentence is enriched as (q_k, p_k), which is an object-activity pair. Query set $Q = \{(q_1, p_1), ..., (q_K, p_K)\}$ is denoted as the textual representations of every query in sentence. With activity-driven object states encoding, textual and visual modalities can be better aligned without the large granularity gap.

3.2.2 Activity-driven Object Proposal Refinement. Most of existing visual grounding methods [5, 15, 25, 35] are based on selecting the best matched proposal out of a candidate pool that contains top-N region proposals as a grounded object. However, query objects may not be included in these proposals and thus are unlikely to be grounded. A naive solution is to increase N but this will lead to a large search space especially for long untrimmed videos.

Human activity can provide spatial location prior to refine the region proposal generation. Intuitively, there is a spatial dependency between the activity performer and the activity receiver. For example, if a video is describing "peel a potato", "potato" tends to occur around "human" hands, which indicates the potential of spatial location prior between activity performer (human) and receivers (objects). We propose to model the spatial prior as a truncated normal distribution $\mathcal{N}(\mu_a, \nu_a)$ to mask out the irrelevant region proposals. μ_a is the pixel coordinate of activity-relevant joint and ν_a is the hyperparameter. We apply the normal distribution to activity-relevant key joints to form a mixture of Gaussian (see the heatmap in Fig. 2), and select top-N proposals by considering the densities. We use a pre-trained human detector [12] and OpenPose [3] to extract the human's joints. If there is no human detected in a frame like Fig. 3, most likely the frame is captured in a close view. Then, we simply keep the top-N region proposals without refinement. Using this method, we can include more query-related proposals and make the query more likely to be grounded.

3.3 Spatio-Temporal MIL for Video Grounding

We consider a weakly-supervised learning scenario, where the only supervision is the sentence description of the video. Existing work [7] addresses the weakly-supervised spatio-temporal grounding for trimmed videos using object tracking to generate proposal tubes. Different from [7], our goal is to achieve *untrimmed* video grounding. This is more challenging as an object does not necessarily appear on every frame and usually occur discontinuously due to the frequent camera shot cut in untrimmed videos. In this case, the tracking-based grounding methods would undoubtedly fail.

To address this problem, we resort to MIL and propose a novel spatio-temporal MIL framework. At the spatial level, we aim at grounding each textual query to one of N object proposals v_t^n ($n \in [1, N]$) extracted on a frame. At the temporal level, we aim at grounding each textual query to the frames where it occurs on a video. Both the temporal and spatial MILs are formulated by pair-wise ranking losses. The losses encourage the correct matching of an aligned video-sentence pair and discourages the matching of an unaligned pair, i.e., the sentence does not belong to the video[3]. The spatial and the temporal MILs are mutually guided to learn a spatio-temporal grounding model.

3.3.1 Spatial level MIL. The goal of spatial grounding is to ground each referred query to one of the top-N region proposals on a frame. In order to obtain the query specific region, we use normalized cosine distance as the metric of region-query similarity $a(v_t^n, Q_k)$:

$$a(v_t^n, Q_k) = \frac{{v_t^n}^{\mathrm{T}} \cdot (q_k + p_k)}{\|v_t^n\| \|(q_k + p_k)\|}, \tag{1}$$

where $(q_k + p_k)$ and v_t^n are feature embeddings of the textual object-activity pair of k-th query and the region proposal, respectively, in a joint d-dimensional feature space. k, t, n index queries, frame, and region proposals, respectively. T is a transpose.

The spatial-level MIL regards each frame as a bag and all region proposals in the frame as instances in the bag. Instance score *w.r.t.* query Q_k is the region-query similarity $a(v_t^n, Q_k)$ computed by Eq. 1. Following MIL, a bag is represented by its most positive instance, which can be achieved by a *max* operation. Thus, the bag level score is computed as $S(V_t, Q_k) = \max_n a(v_t^n, Q_k)$, which denotes the frame-query similarity. Following [7, 25, 35], spatial MIL is formulated as a pair-wise ranking:

$$S(V_t, Q_k) > max\left(S(V_t', Q_k), S(V_t, Q_j')\right), \tag{2}$$

where (V_t, Q_j') and (V_t', Q_k) are the two cases of unaligned frame-query pairs. Q_j' is the j-th unaligned query w.r.t. region proposal V_t, while V_t' consists of region proposals in a video frame unaligned with current query Q_k. Eq. 2 encourages the correct proposal matching for a query Q_k by $S(V_t, Q_k) > S(V_t', Q_k)$ and encourages the correct query matching for a frame V_t by $S(V_t, Q_k) > S(V_t, Q_k')$.

[3]We consider a video and its descriptive sentence as an aligned video-sentence pair and define a video and the query in its descriptive sentence as an aligned video-query pair. Similarly, an unaligned video-sentence pair is that the sentence does not describe the video but describes other video in the current batch.

To achieve the pair-wise ranking in Eq. 2, the frame-query ranking loss with margin Δ_s needs to be minimized in training:

$$\mathcal{L}_{rank}^{t,(k,j)} = max\left(0, max\left(S(V_t', Q_k), S(V_t, Q_j')\right) - S(V_t, Q_k) + \Delta_s\right). \tag{3}$$

This objective encourages the similarities of aligned pairs larger than those of unaligned pair with gap Δ_s. Furthermore, by aggregating every query in the unaligned description sentence, the spatial-level ranking loss is defined as:

$$\mathcal{L}_{rank}^{t,k} = \frac{1}{K'}\sum_{j=1}^{K'} \mathcal{I}(Q_j' \neq Q_k)\mathcal{L}_{rank}^{t,(k,j)}, \tag{4}$$

where the negative query set Q' contains K' queries. Note that if Q_k meets the query Q_j' in negative query set Q', it will not contribute to the ranking loss. But if the queries in Q and Q' only share the object and have different activities, such as "mash the potato" and "peel the potato", it still contributes to the ranking loss, because of the large discrepancy in appearance.

Spatial MIL considers each frame as a bag. However, in untrimmed videos, query only appears in a part of frames. The frames without the query occurring are actually noisy positive bags. In Sec 3.3.2 and 3.3.3, we will discuss how to alleviate the false positive bags with the guidance of temporal grounding.

3.3.2 Temporal level MIL. In temporal grounding, we aim at predicting the temporal occurrence of the queries across frames. In our weakly-supervised setting, we do not have access to the temporal occurrence annotations. Thus, we still formulate it as a MIL problem. In this case, each video is considered as a bag and the frames of the video are considered as instances in the bag.

Instance score is frame-query similarity. But query object only occurs in a small region of the frame. It is not effective to align the query with the entire frame. Thus, we resort to spatial level MIL results as guidance and propose to represent each instance as the best matched region proposal such that the instance score is denoted as $S(V_t, Q_k) = \max_n a(v_t^n, Q_k)$.

In an untrimmed video, the query object may appear discontinuously across frames. Thus, it is not appropriate to represent the bag by the best-matched instance, as it ignores other positive instances that contain the query. This is different from the spatial level MIL where an object tends to appear concentrated in frame. Thus, in temporal level MIL, the bag score which is video-query pair-wise similarity should be the overall score of all positive instances in the bag $\frac{1}{T}\sum_{t\in T} S(V_t, Q_k)$, instead of the best matched instance. Since we have video-query pair as bag-level annotation, temporal level MIL is formulated as a ranking problem:

$$\frac{1}{T}\sum_{t=1}^{T} S(V_t, Q_k) > max\left(\frac{1}{T}\sum_{t=1}^{T} S(V_t', Q_k), \frac{1}{T}\sum_{t=1}^{T} S(V_t, Q_j')\right), \tag{5}$$

which is also a pair-wise ranking. V_t' and Q_j' indicate the negative video frame and the j-th query in the negative query set, respectively. Eq. 5 encourages an aligned video-query pair (V, Q_k) to be better matched than two other types of unaligned video-query pairs (V', Q_k) and (V, Q_j'). The number of instances in each bag is T. Using average instance scores to represent a bag score helps avoid a degenerate solution where we predict most of frames irrelevant to the queries, compared with *max* operation.

Moreover, consecutive frames in a video are correlated but their visual context does not necessarily to be continuous due to the frequent camera shot cut. Therefore, the simple arithmetic mean in Eq. 5 is not adaptive for temporal grounding in untrimmed videos. In this paper, we propose an attention module to learn the the weight of each frame. Specifically, we extract the each frame's features from the last layer of VGG-16 backbone and then encode the frames' features by a self-attention layer as $f_t, t \in [1, T]$. Based on that, we compute the weight of each frame as w_t by a linear layer and a sigmoid activation w.r.t f_t. Note that the weight of each frame is agnostic to different queries while the video content continuity can be addressed by the temporal consistency of frame weights.

To achieve the goal of Eq. 5 by considering the temporal context, we propose to minimize the following temporal ranking loss:

$$\mathcal{L}_{\text{tem}}^{k,j} = max\left(0, max\left(\Gamma(V', Q_k), \Gamma(V, Q_j')\right) - \Gamma(V, Q_k) + \Delta_t\right), \tag{6}$$

where $\Gamma(V, Q_k) = \frac{1}{T}\sum_{t=1}^{T} w_t S(V_t, Q_k)$ indicates the query specific bag score computed by weighted sum of query specific frame scores. $\mathcal{L}_{\text{tem}}^{k,j}$ encourages video V and its paired query Q_k to be better aligned than a query Q_j' in negative query set Q'. Δ_t serves as similarity margin for temporal-level grounding. Finally, the temporal video-query ranking loss is defined as the average over the entire unaligned query set:

$$\mathcal{L}_{\text{tem}}^{k} = \frac{1}{K'}\sum_{j=1}^{K'} \mathcal{I}(Q_k \neq Q_j')\mathcal{L}_{\text{tem}}^{k,j}. \tag{7}$$

The temporal MIL allows the queries to find the frames where they occur in a given video.

3.3.3 Overall objective function. In an untrimmed raw video, query does not necessarily occur in every frame of a video. Previous work [25] propose a contextual similarity to weight the importance of frames corresponding to a query. In our work, we have the temporal MIL to learn the query specific attention over the temporal domain. Thus, only the query-related frame should contribute to the spatial grounding. We propose to utilize our temporal level grounding results as guidance to mask out query-irrelevant frames' contributions in spatio-level MIL ranking loss:

$$\mathcal{L}_{\text{spatio}}^{k} = \sum_{t=1}^{T} \mathcal{I}\left(S(V_t, Q_k) > 0\right)\mathcal{L}_{rank}^{t,k}, \tag{8}$$

where $\mathcal{L}_{rank}^{t,k}$ is computed by Eq. 4. The temporal grounding result $\mathcal{I}\left(S(V_t, Q_k) > 0\right)$ is incorporated into the spatial ranking loss so that spatial and temporal MILs are mutually guided.

We add a penalty term to avoid the trivial solution $S(V_t, Q_k) = 0$. The final objective function of our model is formulated as:

$$\mathcal{L}^{k} = \mathcal{L}_{\text{spatio}}^{k} + \mathcal{L}_{\text{tem}}^{k} + \frac{\lambda}{T}\sum_{t=1}^{T} -w_t S(V_t, Q_k), \tag{9}$$

where λ is the weight for the sparsity constraint. And the final objective is the average of ranking loss on each query and is summarized as $\mathcal{L} = \frac{1}{K}\sum_{k\in K}\mathcal{L}^k$.

4 EXPERIMENTS

Following [25], we train and evaluate our model on YouCookII dataset [36] in a weakly-supervised setting. Besides, we validate the generalization ability of our model on RoboWatch dataset [24].

4.1 Dataset

YouCookII [36] contains 2,000 cooking videos from 89 recipes. Each video recipe consists of 3 to 15 steps. Each step is described by a sentence including multiple queries. We follow [25, 35] to extract 15K video-description pairs from the steps. Training, validation and testing splits contain 5161, 3483 and 1560 pairs, respectively. The average duration of each step is 19.6s. Bounding box annotations [35] for the most 67 frequently appearing objects in the description sentence for the validation and testing split are used. The presence and bounding boxes of objects are labeled every second in a video, which can be used to evaluate spatio-temporal grounding models.

RoboWatch [24] contains 255 YouTube instructional videos, each of which also contains multiple steps. Huang et al. [14] extends the bounding box annotation for a part of those videos, and the query can be either a word or a phrase. We follow [25] to evaluate the generalization ability of our model trained on YouCookII [35] dataset. Following [25], we evaluate our model on the aligned pairs of video and query in RoboWatch. Since each query appears in all of the annotated frames of its video, we only evaluate spatial grounding on RoboWatch dataset.

4.2 Evaluation Metric

We follow [5, 25, 35] to evaluate spatial grounding performance using *box accuracy* and *query accuracy*. The *box accuracy* is defined as the ratio of correctly grounded boxes to all of the grounded boxes by setting a threshold, i.e., 50%, for Intersection-over-Union (IoU) between the grounded box and its corresponding ground-truth. *Query accuracy* is defined as the ratio of correctly grounded queries to all queries. Following [25], the average of each class accuracy and the global accuracy without considering the class are evaluated, which are denoted as *macro-accuracy* and *micro-accuracy*, respectively. In addition, we follow an existing temporal grounding method [19] and compute the temporal IoU (tIOU) between the grounded and ground-truth temporal occurrence as the temporal grounding metric.

However, in previous work [25, 35], the *box accuracy* and *query accuracy* consider only the frames with query occurring. These two evaluation metrics ignore the frames that no query object appears, and thus are not suitable for evaluating the performance of spatio-temporal grounding on untrimmed videos. We propose the following metric to evaluate spatio-temporal grounding models for untrimmed videos:

$$\text{stACC} = \frac{1}{\left|\mathcal{S}^{(U)}\right|} \sum_{t \in \mathcal{S}^{(I)}} \mathcal{I}\left(\text{IoU}(\hat{r}^t, r^t) > R\right), \tag{10}$$

where $\mathcal{S}^{(U)}$ is the union set of frames in which either ground-truth or the grounded bounding boxes are located for a query in an entire video. $\mathcal{S}^{(I)}$ is the intersection set of frames in which both the ground-truth and the grounded bounding boxes occur simultaneously for a query. To compute the intersection, for each grounded box, we count the intersected box by computing the IoU between the grounded box $\hat{r}^t$ and its corresponding ground truth r^t with threshold R. Similar to existing grounding metrics, our proposed stACC can be used to compute *box accuracy* and *query accuracy* by considering the class of each query. Function $\mathcal{I}(\cdot)$ is an indicator function. The proposed metric in Eq. 10 will be used for evaluating the spatio-temporal grounding performance.

4.3 Implementation Details

Following [25], the description sentence is parsed by Stanford CoreNLP parser [18] into nouns and pronouns. We also parse the predicates of nouns/pronouns in the description sentence by SpaCy [13]. A pre-trained BERT [8] model is applied to encode the sentence. For visual modality, a Faster R-CNN framework [22] with VGG-Net [27] as backbone pre-trained on Visual Genome [16] is applied to extract top-20 confident region proposals for each frame, which is the same setting as [25, 35]. We uniformly sample 16 frames from each video. Hyperparameter v_a in spatial prior is set to 40. Visual and textual features are embedded to a joint feature space with 512-dimension. *tanh* is used as the activation function for both visual and text embedding.

We use TITAN Xp and implement the network using PyTorch. Adam with learning rate 0.001 is used for optimization. The ranking margin Δ_t and Δ_s is set to 10 and 5. Constraint weight λ is set to 0.9. We use a batch-size of 8 in all experiments. Thus, each of positive sample is coupled with 7 negative samples. Following [5, 25, 35], we report the grounding results in both validation and test split.

4.4 Comparison

4.4.1 Spatial Grounding on YouCookII Dataset. We compare our method with the state-of-the-art weakly-supervised video grounding methods [7, 25, 35] and two extensions from image grounding methods *DVSA* [15] and *GroundR* [23]. Following [25, 35], we utilize RPN to extract region proposals. Existing work [25, 35] only evaluate the spatial grounding accuracy on the frames where the query occurs. The frames without the query are disregarded in evaluation. We follow this evaluation setting and report the results under the four metrics used in [25]. As shown in Table. 1, the proposed method on spatial grounding consistently outperforms the comparison methods. This is because we bridge the granularity gap between text and visual domains by considering the activity-effect. In addition, our spatio-temporal MIL framework ensures a spatial grounding model learned from the frames that query appears, even without temporal annotations.

4.4.2 Temporal Grounding on YouCookII Dataset. We compare our method with an existing weakly-supervised video temporal grounding method TGA [23] and an extension of [25] to temporal grounding. The extension of [25] to temporal grounding is achieved by extending its frame-query contextual similarity module, which conducts 0-1 normalization of frame-query importance across frames during training. We use it in the test phase to mask out the frame-query pair whose contextual similarity score is less than 0.5.

As shown in Table 2, our approach significantly outperforms the extension of [25] and TGA [19] by 12% and 10%, respectively. This is because these video temporal grounding method grounds a query to the relevant frames based on the similarity between the query and the entire frame. In our method, spatial grounding

Table 1: Weakly-supervised spatial grounding results on YouCookII. "pre-trained" indicates the dataset that RPN is pre-trained on. Zhou et al. [35], extended GroundR [23] and Chen et al. [5] only report macro box accuracy in their papers.

Methods	pre-trained	box accuracy%				query accuracy%			
		macro		micro		macro		micro	
		val	test	val	test	val	test	val	test
Extended GroundR [23]	MSCOCO	19.63	19.94	-	-	-	-	-	-
Zhou et al. [35]	MSCOCO	30.31	31.73	-	-	-	-	-	-
Chen et al. [5]	MSCOCO	33.24	34.90	-	-	-	-	-	-
Extended DVSA [15]	VisualGenome	36.90	37.55	44.26	44.16	38.48	39.31	46.27	46.41
Shi et al. [25]	VisualGenome	39.54	40.71	46.41	46.33	41.29	42.45	48.52	48.41
Ours (BERT)	VisualGenome	37.40	38.88	48.12	45.20	39.00	40.55	46.10	47.23
Ours (Glove+Activity)	VisualGenome	38.50	39.28	46.57	45.85	40.11	41.07	48.59	47.91
Our full model	VisualGenome	**40.66**	**41.67**	**49.11**	**48.22**	**41.43**	**42.55**	**49.71**	**48.91**

Table 2: Weakly-supervised temporal grounding results on YouCookII. tIOU is used as the evaluation metric.

Methods	tIOU
TGA [19]	29.43
Extension of [25]	27.12
Ours	**39.51**

provides guidance for temporal grounding to be focused on the query specific region. This allows us to represent the visual data more accurately.

4.4.3 Spatio-temporal Grounding on YouCookII Dataset. Since there is no existing weakly-supervised spatio-temporal grounding method for untrimmed videos, we extend [25] to ground temporal occurrences (Extension of [25]) using the method described above. We also compare with the weakly-supervised spatio-temporal grounding method [7], which is originally developed for trimmed video. The performance of spatio-temporal grounding methods is evaluated using the metric stACC described in Eq. 10.

As shown in Table. 3, our approach significantly outperforms the Extension of [25] by $5 \sim 8\%$. This shows that a direct extension to spatio-temporal grounding is far from solving this challenging problem. Our method is more effective since we solve this problem using a mutually guided MIL. When generalized to untrimmed videos, Chen et al. [7] shows inferior results to ours, because they highly rely on a visual tracker that easily fails due to the frequent camera shot cut in untrimmed videos.

Table 3: Weakly-supervised spatio-temporal grounding results on YouCookII. stACC is used as the evaluation metric.

Methods	macro		micro	
	val	test	val	test
Extension of [25]	15.89	19.10	17.35	18.54
Chen et al. [7]	7.31	7.70	8.02	8.79
Ours-max	5.17	5.25	5.93	6.11
Ours w/o attention	18.94	20.98	22.31	21.41
Our full model	**21.73**	**24.25**	**25.50**	**25.65**

4.4.4 Generalize Grounding Model to RoboWatch Dataset. Following [25], we conduct the generalization ability experiment of the grounding model trained on YouCookII dataset. We train our grounding model using the nouns and pronouns parsed in the sentences of YouCookII and directly test the grounding model on RoboWatch dataset. We compare our method with two existing methods Shi et al. [25] and extended DVSA [15] and a variant of our method that does not contain activity-driven object states encoding module The comparison is conducted on two types of data split, including the entire test set of RoboWatch and its unseen split which only consists of the objects that never occur in YouCookII such as "oreo", "flesh", "alcohol", "hanger", "tie" *etc.*

Table 4: Generalization results on RoboWatch using query micro-accuracy (%). "*" indicates the results are achieved by running the authors' code on our side. All the other comparison results are from their original papers.

Methods	all	unseen split
Extended DVSA [15]	28.25	25.12*
Shi et al. [25]	31.68	26.79*
Ours w/o activity	30.11	26.56
Our full model	**34.21**	**35.97**

As shown in Table. 4, the proposed activity-driven model outperforms the variant Ours w/o activity and two other existing methods [25] and the extended DVSA [15] by a large margin. Also, our full model's performance in the unseen split is even better than the performance in the entire test set denoted as "all". This is because we model the activity effect on objects' physical states. Thus, even though our model has never seen the query during training, it can utilize the seen activity information to ground the unseen query on which the activity is performed.

4.5 Ablation Studies

4.5.1 Activity-driven Object-States Encoding. We conduct ablation study on the activity-driven object states encoding module with following two variants: 1) "**BERT**" which first encodes the entire description sentence by a pre-trained BERT model [8] and then extracts the query embedding of the objects without considering activity; 2) "**Glove+Activity**". It first extracts the predicates and nouns/pronouns from the description sentence by [18] and then encodes the predicate-object pair into 200-dimensional GloVe [20]. Note that GloVe is used as word embedding in [25, 35].

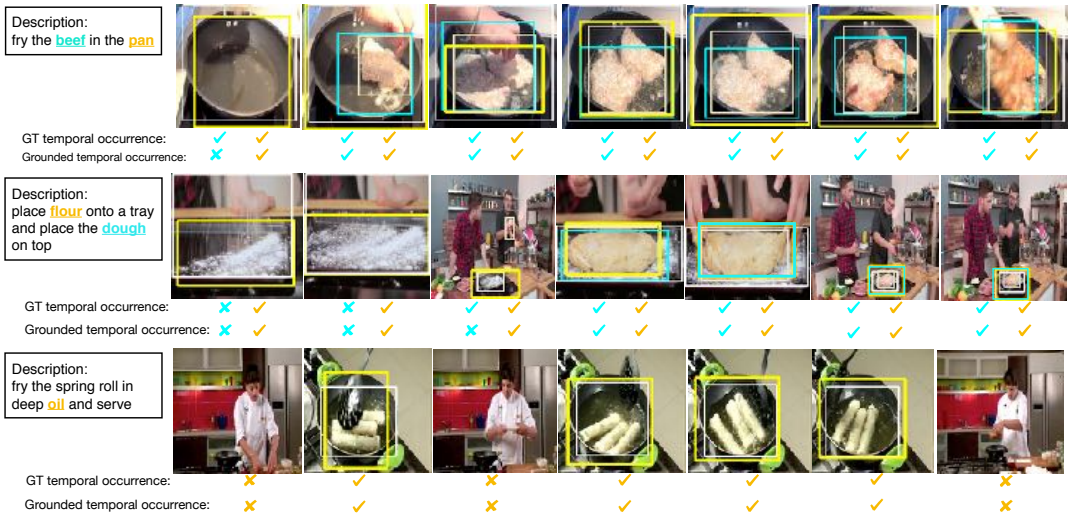

Figure 4: Qualitative results of our spatio-temporal video grounding model. The yellow and cyan boxes are the grounded results of the corresponding queries in description sentences. The white boxes are their ground-truth. Best viewed in color.

As shown in Table 1, the superiority of our full model over the variant "BERT" shows that encoding activity effect on object states benefits grounding. As expected, the activity-driven object state encoding module proves to bridge the granularity gap between the coarse text modality and the rich visual modality, by incorporating the underlying activity-effect on object states into text representation. Moreover, the performance gain from the activity cue is significantly larger than the gain from better query embeddings, i.e., Glove and BERT features. This further demonstrates the effectiveness of the proposed activity-driven encoding.

4.5.2 Temporal MIL Loss. We conduct ablation study on the temporal grounding with the following variants: 1) "**Ours-max**", which replaces the average instance operation in Eq. 2 by *max* operation, selecting the top-ranked query specific frame to represent the bag; 2) "**Ours w/o attention**", which uses Eq. 5 as the temporal ranking loss but removes the guidance of frame consistency attention block.

Table. 3 shows that our full model achieves the best performance. Its superiority over "Ours w/o attention" demonstrates that the temporal context information w.r.t frame similarity plays an important role in video grounding with large visual inconsistency. The variant "Ours-max" is inferior to others, indicating selecting the top-ranked frame as the video representation is not appropriate for temporal grounding. This is because in an untrimmed video, the query may appear discontinuously across frames, leading to multiple frames for the query in the video.

4.5.3 Region Proposal Refinement. We conduct ablation study for the activity-driven region proposal refinement module. On YouCookII dataset, only 22% frames are captured distant view. Thus, we evaluate this module on the distant view split that only contains the frames with human and the entire test set, correspondingly. In the variant "Ours w/o region refine", we simply keep the top-N proposals without refinement. As shown in Table 5, our method outperforms the variant without region refinement, which elaborates the effectiveness of our region refinement module. Our full model refines the proposals to include more query related proposals. This makes the query more likely to be grounded.

Table 5: Ablation study on YouCookII for our activity-driven region proposal refinement module. The results are box/query micro-accuracy. Results of the distant view frame split in the test set and the entire test set are reported.

Methods	distant view split		all	
	box	query	box	query
Ours w/o region refine	20.54	21.35	48.07	48.72
Our full model	**21.03**	**21.85**	**48.22**	**48.91**

4.6 Qualitative Results

The qualitative results of YouCookII dataset are shown in Fig. 4. Each row depicts 6 frames sampled from a video. Camera shot cut frequently appears in these videos. But even though the large visual inconsistency appears, our method is able to ground each query in terms of its temporal occurrence and spatial locations.

5 CONCLUSION

In this paper, we investigate the spatio-temporal grounding in untrimmed videos with frequent visual inconsistency in a weakly-supervised manner. We develop two novel MIL ranking losses for the spatial and temporal domains. Furthermore, to bridge the granularity gap between the coarse text information and the detailed visual information, we introduce an activity-driven object state encoding module to enhance textual representation. Experiments on two popular datasets demonstrate the superiority of our method and its generalization ability to other dataset with unseen queries.

Acknowledgement: We thank Nvidia for the GPU donation.

REFERENCES

[1] Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. 2017. Natural language acquisition and grounding for embodied robotic systems. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[2] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1989–1998.

[3] Z Cao, G Martinez Hidalgo, T Simon, SE Wei, and YA Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[4] Kan Chen, Jiyang Gao, and Ram Nevatia. 2018. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4042–4050.

[5] Lei Chen, Mengyao Zhai, Jiawei He, and Greg Mori. 2019. Object Grounding via Iterative Context Reasoning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

[6] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. 2020. Look Closer to Ground Better: Weakly-Supervised Temporal Grounding of Sentence in Video. *arXiv preprint arXiv:2001.09308* (2020).

[7] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. 2019. Weakly-supervised spatio-temporally grounding natural sentence in video. *ACL* (2019).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*. 4171–4186.

[9] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems*. 3059–3069.

[10] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5374–5383.

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

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[13] Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1373–1378. https://aclweb.org/anthology/D/D15/D15-1162

[14] De-An Huang*, Shyamal Buch*, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding "It": Weakly-Supervised, Reference-Aware Visual Grounding in Instructional Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[17] Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, and Cordelia Schmid. 2015. Unsupervised object discovery and tracking in video collections. In *Proceedings of the IEEE international conference on computer vision*. 3173–3181.

[18] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

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

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

[21] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. 2012. Learning object class detectors from weakly annotated video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3282–3289.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[23] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*. Springer, 817–834.

[24] Ozan Sener, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2015. Unsupervised semantic parsing of video collections. In *Proceedings of the IEEE International Conference on Computer Vision*. 4480–4488.

[25] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. 2019. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *CVPR*. 10444–10452.

[26] Mohit Shridhar and David Hsu. 2018. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831* (2018).

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

[28] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. 2019. Unsupervised deep tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1308–1317.

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

[30] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4145–4154.

[31] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*. 4683–4693.

[32] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, and Jiebo Luo. 2019. Grounding-Tracking-Integration. *arXiv preprint arXiv:1912.06316* (2019).

[33] Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 53–63.

[34] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. 2019. Grounded Video Description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Luowei Zhou, Nathan Louis, and Jason J Corso. 2018. Weakly-supervised video object grounding from text by loss weighting and object interaction. *BMVC* (2018).

[36] Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.