Visual Relation Grounding in Videos

Junbin Xiao[1], Xindi Shang[1], Xun Yang[1], Sheng Tang[2], and Tat-Seng Chua[1]

[1] Department of Computer Science, National University of Singapore, Singapore
{junbin,shangxin,chuats}@comp.nus.edu.sg, xunyang@nus.edu.sg
[2] Institute of Computing Technology, Chinese Academy of Sciences, China
ts@ict.ac.cn

Abstract. In this paper, we explore a novel task named visual Relation Grounding in Videos (vRGV). The task aims at spatio-temporally localizing the given relations in the form of *subject-predicate-object* in the videos, so as to provide supportive visual facts for other high-level video-language tasks (*e.g.*, video-language grounding and video question answering). The challenges in this task include but not limited to: (1) both the subject and object are required to be spatio-temporally localized to ground a query relation; (2) the temporal dynamic nature of visual relations in videos is difficult to capture; and (3) the grounding should be achieved without any direct supervision in space and time. To ground the relations, we tackle the challenges by collaboratively optimizing two sequences of regions over a constructed hierarchical spatio-temporal region graph through relation attending and reconstruction, in which we further propose a message passing mechanism by spatial attention shifting between visual entities. Experimental results demonstrate that our model can not only outperform baseline approaches significantly, but also produces visually meaningful facts to support visual grounding. (Code is available at https://github.com/doc-doc/vRGV).

1 Introduction

Visual grounding aims to establish precise correspondence between textual query and visual contents by localizing in the images or videos the relevant visual facts depicted by the given language. It was originally tackled in language-based visual fragment-retrieval [9,12,13], and has recently attracted widespread attention as a task onto itself. While lots of the existing efforts are made on referring expression grounding in static images [8,19,22,23,28,41,42,44], recent research attempts to study visual grounding in videos by finding the objects either in individual frames [10,32,47] or in video clips spatio-temporally [1,3,46]. Nonetheless, all these works focus on grounding in videos the objects depicted by natural language sentences. Although the models have shown success on the corresponding datasets, they lack transparency to tell which parts of the sentence help to disambiguate the object from the others, and thus hard to explain whether they truly understand the contents or just vaguely learn from data statistics. Furthermore, the models fail to effectively reason about the visual details (*e.g.*, relational semantics), and thus would generalize poorly on unseen scenarios.

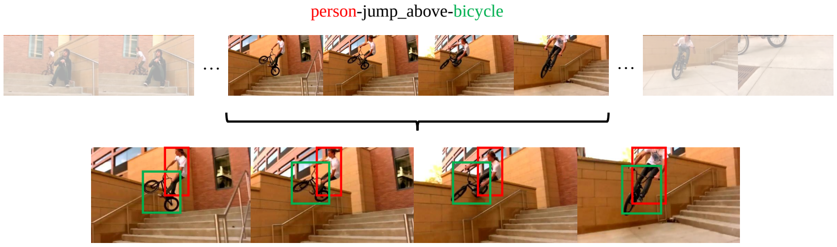


Fig. 1: Illustration of the vRGV task. For the query relation *person-jump_above-bicycle* and an untrimmed video containing the relation, the objective is to find a video segment along with two trajectories corresponding to the subject (red box) and object (green box) that match the query relation.

To achieve understandable visual analytics, many works have showed the importance of comprehending the interactions and relationships between objects [8,16,31]. To this end, we explore explicit relations in videos by proposing a novel task of visual Relation Grounding in Videos (vRGV). The task takes a relation in the form of *subject-predicate-object* as query, and requires the models to localize the related visual subject and object in a video by returning their trajectories. As shown in Fig. 1, given a query relation *person-jump_above-bicycle* and an untrimmed video containing that relation, the task is to find in the video the subject (person in white) and object (bicycle) trajectory pairs that hold the query relationship *jump_above*.[3] Considering that annotating fine-grained region-relation pairs in videos is complicated and labor-intensive [30], we define the task as weakly-supervised grounding where only video-relation correspondences are available. That is, the models only know that a relation exists but do not know where and when it is present in the video during training.

vRGV retains and invokes several challenges associated with video visual relation and visual grounding. First, unlike existing video grounding [1,3] which is to localize a specific object according to its natural language description, in visual relation grounding, the models are required to jointly localize a pair of visual entities (subject and object) conditioned on their local relationships. Second, unlike a coarsely global description, relations are more fine-grained and change over time, *i.e.*, even a same object would have different relationships with different objects at different time. For example, the enduring relations (*e.g.*, *person-drive-car*) may exist for a long time but the transient ones (*e.g.*, *person-get_off-car*) may disappear quickly. Besides, static relationships like spatial locations and states (*e.g.*, *hold* and *hug*) can be grounded at frame level, whereas the dynamic ones such as *lift_up* and *put_down* can only be grounded based on a short video clip. Such dynamic nature of relations in videos will cause great challenge for spatio-temporal modeling. Third, the requirement for weakly-supervision also challenges the models to learn to ground the relation without reliance on any spatial (bounding boxes) and temporal (time spans) supervisions.

[3] The word 'untrimmed' is regarding to relation. We refer to relation as the complete triplet *subject-predicate-object*, and relationship as the *predicate* only.

To address the above challenges, we devise a model of video relation grounding by reconstruction. Concretely, we incorporate a query relation into a given video which is modeled as a hierarchical spatio-temporal region graph, for the purpose of identifying the related subject and object from region proposals in multi-temporal granularity. For weakly-supervised video relation grounding during training, we optimize the localization by reconstructing the query relation with the subject and object identified by textual clues and attention shift. During inference, we dynamically link the subject and object which contribute most to the re-generated relations into trajectories as the grounding result. Our insight is that visual relations are data representations of clear structural information, and there is a strong cross-modal correspondence between the textual subject-object and the visual entities in the videos. Thus, it is reasonable to ground the relevant subject and object by re-generating the relation.

Our main contributions are: (1) we define a novel task, visual Relation Grounding in Videos (vRGV), to provide a new benchmark for the research on video visual relation and underpin high-level video-language tasks; (2) we propose an approach for weakly-supervised video relation grounding, by collaboratively optimizing two sequences of regions over a hierarchical spatio-temporal region graph through relation attending and reconstruction; and (3) we propose a novel message passing mechanism based on spatial attention shifting between visual entities, to spatially pinpoint the related subject and object.

2 Related Work

In this section, we briefly recap the history in visual relation, visual grounding and video modeling, which are either similar in spirit to the task definition or technically relevant to our approach.

Visual Relation. Early attempts on visual relation either leveraged object co-occurrence and spatial relationships for object segmentation [4], or focused on human-centric relationships for understanding human-object interactions [40]. Recently, many works started to study visual relations as a task onto itself to facilitate cognitive visual reasoning. Lu *et al.* [20] firstly formulated visual relations as three separated parts of *object_1-predicate-object_2*, and classified visual relationships as spatial, comparative, preposition and verb predicates. Krishna *et al.* [16] formalized visual relations as a scene graph for image structural representation, in which visual entities are corresponding to nodes and connected by edges depicted by object relationships. Shang *et al.* [31] introduced visual relations from images to videos (video scene graph). Apart from the relations in static images, they added relationships that are featured with dynamic information (*e.g.*, *chase* and *wave*), so as to emphasize spatio-temporal reasoning of fine-grained video contents. According to their definition, a valid relation in videos requires both the subject and object to appear together in each frame of a certain video clip.

While a handful of works have successfully exploited relations to improve visual grounding [8] and visual question answering [21], relation as an independent problem is mostly tackled in the form of detection task and the advancements are mostly made in the image domain [17,20,45]. In contrast, relation as a detection task in video domain has earned little attention, partly due to the great challenges in joint video object detection and relation prediction with insufficient video data [26,35]. In this paper, instead of blindly detecting all visual objects and relations in videos, we focuses on the inverse side of the problem by spatio-temporally grounding a given relation in a video.

Visual Grounding. Visual grounding has emerged as a subject under intense study in referring expression comprehension [8,15,22,23,28,41,42,44]. Mao *et al.* [22] first explored referring expression grounding by using the framework of Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) network [7] for image and sentence modeling. They achieved grounding by extracting region proposals and then finding the region that can generate the sentence with maximum posterior probability. Similarly, Rohrbach *et al.* [28] explored image grounding by reconstruction to enable grounding in a weakly-supervised scenario. Krishna *et al.* [15] explored referring relationships in images by iterative message passing between subjects and objects. While these works focus on image grounding, more recent efforts [1,3,10,32,39,46,47] also attempted to ground objects in videos. Zhou *et al.* [47] explored weakly-supervised grounding of descriptive nouns in separate frames in a frame-weighted retrieval fashion. Huang *et al.* [10] proposed to grounding referring expression in temporally aligned instructional videos. Chen *et al.* [3] proposed to perform spatio-temporal object grounding with video-level supervision, which aims to localize an object tube described by a natural language sentence. They pre-extracted the action tubes, and then rank and return the tube of maximal similarity with the query sentence. Instead of grounding a certain object in trimmed videos by a global object description [3], we are interested in localizing a couple of objects conditioned upon their relationships in untrimmed videos, which is more challenging and meaningful in reasoning real-world visual contents.

Video Modeling. Over the decades, modeling the spatio-temporal nature of video has been the core of research in video understanding. Established hand-crafted feature like iDT [37] and deep CNN based features like C3D [34], two-Stream [33] and I3D [2], have shown their respective strengths in different models. However, all these features mainly capture motion information in a short time interval (*e.g.*, 16 frames as the popular setting in C3D). To enable both long and short-term dependency capturing, researchers [36,43] also attempted to model the video as an ordered frame sequence using Recurrent Neural Networks (RNNs). While RNN can deal with dynamic video length in principle, it was reported that the preferable number of frames with regard to a video should be ranged from 30 to 80 [24,36]. As a result, Pan *et al.* [24] further proposed a hierarchical recurrent neural encoder to achieve temporal modeling in multiple granularity. Yet, they focused on generating a global description of the video by extracting frame-level CNN feature, which can hardly be applied to relation understanding where fine-grained regional information is indispensable. Recently, there is a tendency of modeling videos as spatio-temporal graphs [11,26,35,38], where the nodes correspond to regions and edges to spatial-temporal relationships. Nonetheless, all of them model the video as a flat and densely connected graph. Instead, we retain the temporal structure (ordered frames and clips) of videos by modeling it as a hierarchical spatio-temporal region graph with sparse directed connections.

3 Method

Recall that our goal is to ground relations in the given videos, which is formulated by giving a relation coupled with videos containing that relation, and returning two trajectories for each video, corresponding to the subject and object participating in the query relation[4]. We formally define the task as follows.

Task Definition: Given a set of query relations in form of $\mathcal{R} = \{< S - P - O >\}$ and a set of untrimmed videos $\mathcal{V}$ (where S, P, O denote the *subject*, *predicate* and *object* respectively), and each specific query relation $\mathcal{R}_i$ is coupled with several videos from $\mathcal{V}$ which contain that relation, the task is to spatio-temporally localize in the videos the respective subjects and objects by returning their trajectories T_s, T_o. The trajectory T is given by a sequence of bounding boxes tied to a certain visual entity across a video segment. For weakly-supervised grounding, there is no spatial (bounding box) and temporal supervisions (time spans) from the dataset during training.

3.1 Solution Overview

Given a video of N frames, we first extract M region proposals for each frame. Thus, a video can be represented by a set of regions $V = \{B_{i,j} \mid i \in [1, N],\ j \in [1, M]\}$, and a trajectory $T = \{B_i \mid i \in [k, l],\ k \in [1, N],\ l \in [k, N]\}$ can be a sequence of bounding boxes in the video. Our approach will learn to ground a given relation R by finding two trajectories T_s, T_o that indicate the subject and object of the relation. According to the task definition, we resolve the problem by maximizing the following posterior probability:

$$T_s^*, T_o^* = \underset{T_s,\ T_o}{\arg\max}\, P(R \mid T_s,\ T_o) * P(T_s,\ T_o \mid V,\ R), \tag{1}$$

where $P(T_s,\ T_o \mid V,\ R)$ aims to attend to the most relevant trajectories in V given the relation R, and $P(R \mid T_s,\ T_o)$ attempts to reconstruct the same relation R based on the relevant trajectories it attended to. During inference, our approach will output the trajectories (T_s^* and T_o^*) that contribute mostly to the re-generated relation to accomplish grounding.

The key idea of our approach is to ground the relation through reconstruction by capturing the intuition that there is a clear correspondence between subject-object in textual relation and visual instances. However, unlike image grounding [28] which can directly model and return the region proposals, effective trajectory proposals are unavailable in this task due to the complicated dynamics of

[4] If there are more than one instances that match the query relation in a video, it is a correct grounding by returning any one of them.

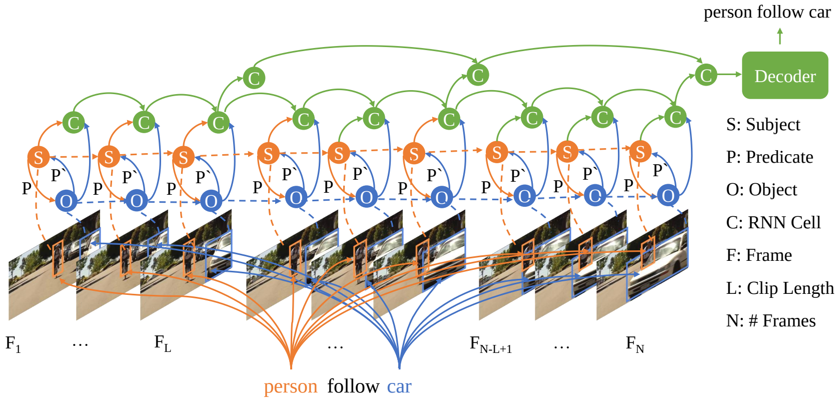


Fig. 2: Schematic diagram of video relation grounding by reconstruction. The model takes the query relation as guidance to pinpoint regions of subjects and objects over the hierarchical spatio-temporal region graph, where the regions correspond to nodes which are spatially connected by visual relationships and are temporally connected by hierarchical RNN over different frames and clips.

the relations in videos. To achieve the optimization in Equ. (1), we model the trajectory proposals implicitly over a hierarchical spatio-temporal region graph and online-optimize them through relation attending and reconstruction. As shown in Fig. 2, two sequences of regions corresponding to the query subject and object will be identified from the video (region graph) and jointly embedded into the final graph representation which will be fed to the decoder to reconstruct the relation. The reconstruction loss from the decoder will be back-propagated to the graph encoder, to penalize the incorrect object pairs with respect to the relation. During inference, we dynamically link the regions which response significantly to the reconstructed relations into explicit trajectories to accomplish the grounding, where the importance of the regions are determined by spatial and temporal attention over the hierarchical space-time region graph. In this way, our model can in principle ground visual relations in multi-temporal granularity without the bottleneck of off-line trajectory proposal extraction [3].

We next elaborate how to learn to spatio-temporally attend to the correct sequences of regions for a given relation, and then how to obtain the final trajectories based on the attention values, to accomplish the grounding.

3.2 Message Passing by Attention Shifting

Spatial Attention. This unit takes as input all the region proposals $B = \{B_j \mid j \in [1, M]\}$ in a frame and the query relation $R =< S - P - O >$. It learns two spatial attentions $(\alpha_s^{M\times1},\ \alpha_o^{M\times1})$ corresponding to the subject and object. Concretely, the spatial attention unit (SAU) is formulated as:

$$\alpha_s = SAU(\ f(B),\ g(S)\ ),\ \alpha_o = SAU(\ f(B),\ g(O)\ ), \tag{2}$$

in which $g(\cdot)$ returns the textual word feature for subject (S) or object (O), and is achieved by embedding the respective GloVe [25] vector: $g(S) = Emb(GloVe(S))$.

Besides, $f(\cdot)$ means feature extraction for the region proposals. In our implementation, we utilize several kinds of feature related to the object appearances, relative locations and sizes, which are not only important in visual relation understanding, but also crucial in identifying the same object in different frames.

The object appearance is captured by the ROI-aligned feature from object detection models, *i.e.*, $f_{app} = CNN(B_j)$. The object relative location and size are useful to identify the spatial and comparative relationships, and are given by $f_B = [\frac{x_{min}}{W}, \frac{y_{min}}{H}, \frac{x_{max}}{W}, \frac{y_{max}}{H}, \frac{area}{W*H}]$, in which W, H are respectively the width, height of the frame and *area* is the area of the bounding box represented by the top-left (x_{min}, y_{min}) and bottom-right (x_{max}, y_{max}) coordinates. The final feature of a region proposal $f(B_j)$ is thus obtained by element-wise addition of the transformed visual appearance feature f_{app} and bounding box feature f_B. The transform operation is achieved by linear mapping with ReLU activation.

We take the subject S as an example to introduce how to obtain the representation f_s and attention distribution α_s for it (similar way to obtain α_o and f_o for object O). Given the textual subject representation $g(S)$ and each region proposal $f(B_j)$, the attention score s_j is obtained by

$$s_j = W_2 tanh(W_1[f(B_j),\ g(S)] + b_1), \tag{3}$$

in which W_1, b_1, W_2 are model parameters. Then, the attention distribution over different region B_j, and the final representation for subject S are give by

$$\alpha_{s_j} = softmax(s_j) = \frac{exp(s_j)}{\sum_{z=1}^{M} exp(s_z)},\ f_s = \sum_{j=0}^{M} \alpha_{s_j} f(B_j). \tag{4}$$

Attention Shifting. Although the aforementioned attention unit is capable of identifying the subjects and objects semantically related to the query relations, they are not necessarily the exact visual entities that hold the relationships. Take the instance in Fig. 2 as an example, there is one *person* but several *cars* on the street, and only one *car* match the relationship *follow* with the *person.* It is the relationship *follow* that helps to disambiguate the *car* of interest from the other *cars.* Another intuition is that, given the subject *person* and relationship *follow*, the searching space of the object *car* can be narrowed to the areas in front of the *person*, and vice verse.

As shown in Fig 2, we capture these insights by modeling the relationships as attention shifting (message passing) between the visual entities, so as to accurately pinpoint the subject and object participating in the query relation. Specifically, we learn two independent transfer matrices W_{so} and W_{os} tied to the forward relationship (P, message from subject to object) and backward relationship (P', message from object to subject) respectively.

$$f_{so} = ReLU(W_{so}\alpha_s),\ f_{os} = ReLU(W_{os}\alpha_o). \tag{5}$$

The transferred location feature from subject (object) to object (subject) will be added to the attention based object (subject) feature:

$$f_s = f_s + f_{os},\ f_o = f_o + f_{so}. \tag{6}$$

Finally, the subject and object representations will be concatenated and transformed to obtain the node input at time step i, *i.e.*, $f_i = W_3([f_s, f_o]) + b_3$, where W_3, b_3 are learnable parameters.

3.3 Hierarchical Temporal Attention

To cope with the temporal dynamics of video relations, we devise two relation-aware hierarchical temporal attention units TAU_1 and TAU_2 (which work in a way similar to SAU) over the frames and clips respectively. As shown in Fig. 2, a video is firstly divided into $H = \frac{N}{L}$ short clips of length L. The frame-wise temporal attention β^{l1} (of dimension N) is obtained by

$$\beta^{l1} = TAU_1(f^{l1},\ f^{l2}_H), \tag{7}$$

in which f^{l1} denotes the sequence of frame-wise feature, and is achieved by sequence modeling the subject and object concatenated feature f

$$f^{l1}_i = LSTM_{l1}(f_{1,\cdots,i}),\ i \in [1, N]. \tag{8}$$

Besides, f^{l2}_H denotes the output at the last time step of the clip-level neural encoder, which is obtained by

$$f^{l2}_H = LSTM_{l2}((\beta^{l2}_i f^c_i)_{i=1,\cdots,H}), \tag{9}$$

where f^c denotes the sequence of clip-level inputs which are obtained by selecting the output of the first layer of LSTM at every L steps, *i.e.*, $f^c = \{f^{l1}_i \mid i \in \{1, L, \cdots, N\}\}$. β^{l2} (of dimension H) is the clip-level temporal attention distribution, and is obtained by

$$\beta^{l2} = TAU_2(f^c,\ f_R), \tag{10}$$

where f_R denotes the query relation which is obtained by concatenating the GloVe feature of each part in the relation (average for phrase) and further transformed to the same dimension space as feature vectors in f^c.

3.4 Train and Inference

During training, we drive the final graph embedding by an attention-guided pooling of the node representations across the video, *i.e.*, $feat_v = \sum \beta^{l1} f^{l1}$. The graph embedding will be fed to the decoder to re-generate the query relation. The decoder part of our model is similar to [28] by treating the relation as a textual phrase and reconstruct it by a single LSTM layer. The model was trained with the cross-entropy loss

$$L_{rec} = -\frac{1}{n_{vr}} \sum_{n=1}^{n_{vr}} \sum_{t=1}^{n_w} logP(R_t | R_{0:t-1},\ feat_v), \tag{11}$$

where R_t denotes the t^{th} word in the relation. n_{vr} and n_w denote the number of video-relation samples and number of words in the relation respectively.

During inference, we base on the learned spatio-temporal attention values to achieve the relation-aware trajectories. First, we temporally threshold to obtain a set of candidate sub-segments for each relation-video instance in the test set. The segments are obtained by grouping the successive frames in the remaining frame set after thresholding with value σ, *i.e.*, $B = \{B_{i,1:M} | \beta_i >= \sigma\}$, in which B_i denotes regions in frame i. β is temporal attention value obtained by $\beta = \beta^{l1} + \beta^{l2}$. Note that the clip-level attention value will be propagated to all frames belonging to that clip. (Refer to appendix for more details.)

Table 1: Statistics of ImageNet-VidVRD.

Dataset		#Videos	#Objects	#Predicates	#Relations	#Instances
ImageNet-VidVRD [31]	Train	800	35	132	2961	25,917
	Val	200	35	132	1011	4835

Then, for each sub-segment, we define a linking score $s(B_{i,p}, B_{i+1,q})$ between regions of successive frames (after sampling)

$$s(B_{i,p}, B_{i+1,q}) = \alpha_{i,p} + \alpha_{i+1,q} + \lambda \cdot IoU(B_{i,p}, B_{i+1,q}), \tag{12}$$

where α is the spatial attention value, it can be α_s (subject) or α_o (object) depending on the linking visual instances. IoU denotes the overlap between two bounding boxes, and $\lambda = \frac{1}{D}$ is a balancing term related to the distance ($D \in [1, 10]$) of the two successive frames. By defining λ, we trust more on the attention score when the distance between the two frames are larger. Our idea is to link the regions which response strongly to the subject or object, and their spatial extent overlaps significantly. The final trajectory can thus be achieved by finding the optimal path over the segment

$$T^* = \arg\max_{T} \frac{1}{K-1} \sum_{i=1}^{K-1} s(B_{i,p}, B_{i+1,q}), \tag{13}$$

where T is a certain linked region sequence of length K for the subject or object. Similar to [5], we solve the optimization problem using Viterbi algorithm. Finally, the linking scores associated with the subject and object are averaged to obtain the score for the corresponding sub-segment. The grounding is achieved by returning the segment (subject-object trajectory pair) of maximal score.

4 Experiments

4.1 Dataset and Evaluation

We conduct experiments on the challenging video relation dataset ImageNet-VidVRD [31]. It contains 1000 videos selected from ILSVRC-VID [29], and is annotated with over 30,000 relation instances covering 35 object classes and 132 predefined predicates. Our preliminary investigation shows that over 99% of relations do not appear throughout the video, with 92% (67%) appearing in less than 1/2 (1/5) length of the video, and the shortest relation only exists in 1 second, while the longest relation lasts for 40 seconds. Besides, each video contains 2 to 22 objects (3 on average), excluding those un-related objects which also matter due to the weakly-supervised setting. The dataset statistics are listed in Table 1, others details are given in the appendix. Note that the object trajectories are provided but are not used during training.

We report accuracy (in percentage) as the grounding performance. Specifically, for each query relation, there might be one or more videos, with each having one or more visual instances corresponding to that relation. For a video-relation pair, a true-positive grounding is confirmed if the returned subject-object trajectory has an overlap ratio of larger than 0.5 with one of the ground-truth visual relation instances. The overlap is measured by the temporal Intersection over Union (tIoU), which is based on the average number of three different spatial IoU thresholds (*i.e.*, sIoU = 0.3, 0.5 and 0.7). Aside from the joint accuracy for the whole relation (Acc_R), we also separately report the accuracy for the subject (Acc_S) and object (Acc_O) for better analysis of algorithms.

4.2 Implementation Details

For each video-relation instance, we uniformly sample N=120 frames from the video and further divide them into H=10 clips of length L=12 frames, and extract M=40 region proposals for each frame. We apply Faster R-CNN [27] with ResNet-101 [6] as backbone (pretrained on MS-COCO [18]) to extract the region proposals, along with the 2048-D regional appearance feature. The final dimension of each region representation is transformed to 256. For each word in the textual relation, we obtain the 300-D GloVe feature and then embed it to 256 dimension space. The hidden size for the encoder and decoder LSTM is 512. Besides, the models are trained using Adam [14] optimization algorithm based on an initial learning rate of 1e-4 and batch size of 32. We train the model with maximal 20 epochs, and use dropout rate 0.2 and early stopping to alleviate over-fitting. During inference, we first obtain the spatial and temporal grounding results on the basis of all the sampled frames, and then propagate the adjacent bounding boxes to the missing frames based on linear interpolation (see appendix for details). The temporal attention threshold is set to 0.04 , and is greedily searched on a validation split of the training data.

4.3 Compared Baselines

As there is no existing method for the vRGV task, we adapt several related works as our baselines. (1) **Object Co-occurrence**, it was applied in [15] for referring relationship in images. We can equivalently achieve it in the video scenario by removing all the predicates in our model and only grounding the two categories. This baseline is to study how much the object co-occurrence will contribute to the relation grounding performance. (2) **Trajectory Ranking**. We adapt the method proposed in video language grounding (namely WSSTG [3]) to the relation scenario. This can be achieved by regarding the relation as a natural language sentence and transforming grounding to sentence matching with the pre-extracted object tubes. Specifically, we consider two implementation variants: (a) $\mathbf{V_1}$, which optimizes the similarity between each trajectory proposal and the query relation during training, and outputs the top-2 ranked trajectories as the grounded subject and object during inference; and (b) $\mathbf{V_2}$, which concatenates the trajectories pair-wisely to compare their similarity with the query sentence during training, and returns the top-1 trajectory-pair as grounded results during inference. (More details can be found in the appendix.)

Table 2: Results of visual relation grounding in videos. We add bold and underline to highlight the best and second-best results under each metric respectively.

Methods	sIoU=0.3			sIoU=0.5			sIoU=0.7			Average		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
T-Rank V_1 [3]	33.55	27.52	17.25	22.61	12.79	4.49	6.31	3.30	0.76	20.27	10.68	3.99
T-Rank V_2 [3]	34.35	21.71	15.06	23.00	9.18	3.82	7.06	2.09	0.50	20.83	7.35	3.16
Co-occur* [15]	27.84	25.62	18.44	23.50	20.40	13.81	17.02	14.93	7.29	22.99	19.33	12.80
Co-occur [15]	31.31	30.65	21.79	28.02	27.69	18.86	21.99	21.64	13.16	25.90	25.23	16.48
vRGV* (ours)	37.61	37.75	27.54	32.17	32.32	21.43	21.34	21.02	10.62	31.64	30.92	20.54
vRGV (ours)	**42.31**	**41.31**	**29.95**	**37.11**	**37.52**	**24.77**	**29.71**	**29.72**	**17.09**	**36.77**	**36.30**	**24.58**

4.4 Result Analysis

Table 2 shows the performance comparisons between our approach and the baselines, where vRGV* (similar for Co-occur*) denotes our model variant that greedily links the regions of maximal attention score in each frame by setting λ in Equ. (12) to 0. We conduct this experiment to validate that our model is capable of learning the object identity across different frames, because we implicitly model and optimize the trajectories on the spatio-temporal graph. When the model is complemented with explicit object locations during post-linking, it can achieve better performances as shown in the bottom row.

From the results, we can see that our methods significantly outperform the baselines, and both methods adapted from WSSTG [3] (*i.e.*, T-Rank V_1 and T-Rank V_2) perform poorly on this task. We speculate the reasons are two folds: (1) the method in WSSTG is designed for single object grounding, they fail to jointly ground two visual entities and further to disambiguate between subject and object (see Fig. 3). In our approach, we collaboratively optimize two sequences of objects on the spatio-temporal graph with relation attending and message passing mechanisms, and thus cope well with the joint grounding problem. This is supported by the observation that the two baselines [3] obtain relatively closer results to ours on separate grounding accuracy (Acc_S), but much lower results than ours regarding the joint accuracy Acc_R; (2) The method in WSSTG aims for object grounding in trimmed video clips, and it pre-extracts relation agnostic object tube proposals and keeps them unchanged during training. In contrast, our approach enables online optimization of object trajectories regarding relation and post-generates relation-aware trajectory pairs. Thus, we can generate better trajectories tailored for relations. This is supported by the observation that the two baselines can get closer results to ours at a relatively lower overlap threshold, but their results degenerate significantly at higher thresholds. (Please also refer to our results on different temporal overlap thresholds shown in the appendix.)

Another observation is that T-Rank V_2 performs better than T-Rank V_1 in grounding the subject (Acc_S), but gets much worse results in terms of object (Acc_O) and hence acts poorly on the joint grounding results for relations (Acc_R). This indicates that the two objects in the top-ranked trajectory pair usually do not correspond to the subject and object mentioned in the sentence, and

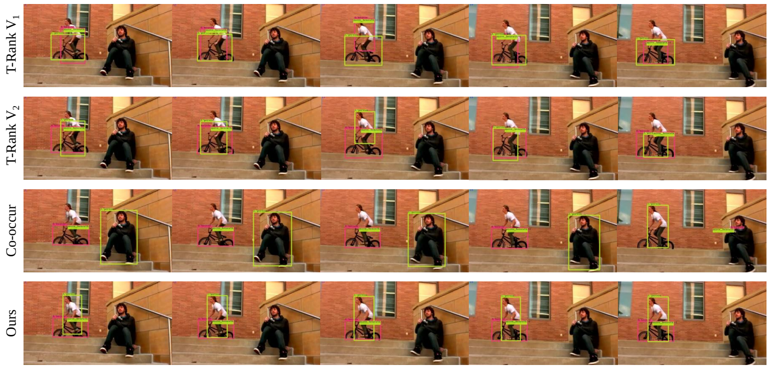


Fig. 3: Qualitative results on the query relation *bicycle*-*move_beneath*-*person*.

Table 3: Model ablation results on ImageNet-VidVRD.

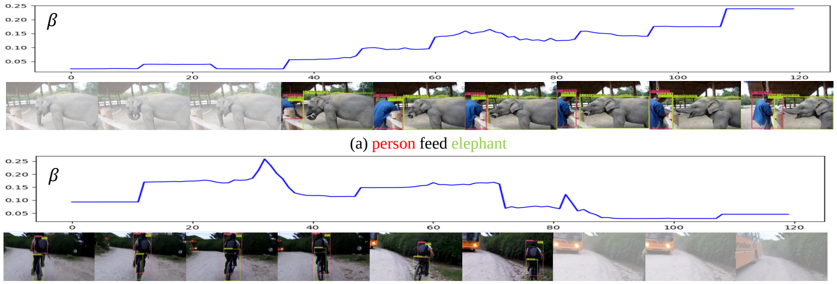
Models	sIoU=0.3			sIoU=0.5			sIoU=0.7			Average		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
vRGV	**42.31**	**41.31**	**29.95**	**37.11**	**37.52**	**24.77**	**29.71**	**29.72**	**17.09**	**36.77**	**36.30**	**24.58**
w/o Msg	34.72	33.23	23.96	31.60	29.15	19.43	22.56	21.36	11.78	29.41	27.46	17.63
w/o Clip	41.08	39.64	27.15	36.31	35.05	21.77	28.19	27.11	13.72	35.05	34.03	20.58
w/o TAU	32.99	32.76	20.34	22.36	19.99	7.61	15.29	13.27	4.83	21.75	19.26	7.06

they are more likely the redundant proposals of the main objects. As shown in Fig. 3, the model T-Rank V_2 can successfully find the subject *bicycle*, but fails to localize the object *person*. We think the reason is that WSSTG [3] is oriented for grounding the main object in a natural language sentence (*i.e.*, the subject), and when concatenating the representations of two trajectories, it can enhance the representation for the main object, but not sure for other supportive objects. According to the results, it even confuses the model and thus jeopardizes the grounding performance for the object (Acc_O).

Relatively, the co-occurrence baseline performs better than [3] on this task. Yet, its performances are still worse than ours (Co-occur* v.s. vRGV* and Co-occur v.s. vRGV). This demonstrates that the "predicate" in the relation is crucial in precisely disambiguating the subjects and objects. As shown in Fig. 3, the occurrence baseline wrongly grounds the person sitting as the object, whereas our method successfully grounds the object to the person riding the bicycle. We also note that the co-occurrence baseline beats our weak model variant under the metric with threshold 0.7 (Co-occur v.s. vRGV*), which in fact, shows the superiority of our overall framework for joint grounding of two objects. Also, it indicates the importance of object locations in generating better trajectories.

4.5 Model Ablations

We ablate our model in Table 3 to study the contribution of each component. Results in the 2nd row are obtained by removing the message passing module. We



(a) person feed elephant

(b) person ride bicycle

Fig. 4: Qualitative results based on temporal threshold 0.04.

can see that the performance Acc_R drops from 24.58% to 17.63%. This is mainly because the model without explicit message communication between subject and object cannot cope well with the scenario where there are multiple visual entities of same category present in the video. Another reason could be that the ablated model is weak in detecting objects under complex conditions (*e.g.*, occlusion and blur) without the contextual information from their partners.

Results in the 3rd row are obtained by removing the clip-level attention. We do this ablation to prove the importance of the hierarchical structure of our model. Note that the temporal threshold σ for this experiment is set to 0.0001, because we only have the frame-level attention β^{l1}. Comparing with results in the first row, we can see the results drop under all criteria without the hierarchical structure. When we further remove the frame-level attention (shown in the 4th row), and thus delete the whole temporal attention unit (TAU), the results degrade significantly from 24.58% to 7.06%. This is because our model will blindly link the objects throughout the video regardless of relation without the temporal grounding module. These findings demonstrate the importance of relation-aware temporal grounding in untrimmed videos.

We also analyze the temporal threshold σ by changing it from 0.01 to 0.05, the corresponding results Acc_R are 24.51%, 24.76%, 24.43%, 24.58% and 22.94%. From the results, we can see that there is no significant difference when the threshold changes from 0.01 to 0.05, and the best result is achieved at the threshold of 0.02. We show some qualitative results in Fig. 4 based on temporal threshold of 0.04 (More results can be found in the appendix), from which we can see that our model can ground the subjects and objects in the videos when the query relations exist.

4.6 Zero-shot Evaluation

In this section, we analyze the models' capability of grounding the new (unseen during training) relation triplets. Specifically, in zero-shot relation grounding, we consider the case that the complete relation triplet is never seen, but their separate components (*e.g.*, *subject*, *predicate* or *object*) are known during training. For example, the model may have seen the relation triplets *person-ride-bicycle* and *person-run_behind-car* during training, but it never knows *bicycle-run_behind-*

Table 4: Results of zero-shot visual relation grounding.

Methods	Acc_S	Acc_O	Acc_R
T-Rank V_1 [3]	4.05	4.08	1.37
T-Rank V_2 [3]	7.09	4.13	1.37
Co-occur [15]	11.60	10.99	7.38
vRGV (ours)	**18.94**	**17.23**	**10.27**

car. As a result, we can find 432 relation instances of 258 unseen relation triplets in 73 videos from the test set.

As shown in Table 4, our approach still outperforms the baselines in the zero-shot setting. We attribute such strength of our approach under zero-shot scenario to two reasons. First, we decompose the relation and separately embed the words corresponding to the subject and object into a semantic space during relation embedding. This is different from modeling the relation holistically using LSTM as in [3], which lacks flexibility and is hard to learn with limited training data. Second, we treat the relation as a natural language in the reconstruction stage, which enhances the model's ability in visual reasoning through forcing it to infer the remaining words conditioned on the related visual content and the previously generated words in the relation.

5 Conclusion

In this paper, we defined a novel task of visual relation grounding in videos which is of significance in underpinning other high-level video-language tasks. To solve the challenges in the task, we proposed a weakly-supervised video relation grounding method by modeling the video as hierarchical spatio-temporal region graph, and collaboratively optimizing two region sequences over it by incorporating relation as textual clues and passing messages by spatial attention shift. Our experiments demonstrated the effectiveness of the proposed approach. Future efforts can either be made on how to jointly ground the subject and object in videos conditioned on their interactions, or how to better capture the temporal dynamics of relations in videos. In addition, it is also important to explore how to better optimize the video graph model based on video-level supervisions only. Another promising direction could be utilizing relation to boost video language grounding and video question answering.

Acknowledgement

This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

1. Balajee Vasudevan, A., Dai, D., Van Gool, L.: Object referring in videos with language and human gaze. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4129–4138 (2018)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
3. Chen, Z., Ma, L., Luo, W., Wong, K.Y.K.: Weakly-supervised spatio-temporally grounding natural sentence in video. ACL (2019)
4. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
5. Gkioxari, G., Malik, J.: Finding action tubes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 759–768 (2015)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
8. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1115–1124 (2017)
9. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4555–4564 (2016)
10. Huang, D.A., Buch, S., Dery, L., Garg, A., Fei-Fei, L., Niebles, J.C.: Finding" it": Weakly-supervised reference-aware visual grounding in instructional videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5948–5957. IEEE (2018)
11. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5308–5317 (2016)
12. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
13. Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in neural information processing systems. pp. 1889–1897 (2014)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Krishna, R., Chami, I., Bernstein, M., Fei-Fei, L.: Referring relationships. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6867–6876 (2018)
16. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)

17. Liang, K., Guo, Y., Chang, H., Chen, X.: Visual relationship detection with deep structural ranking. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
19. Liu, X., Li, L., Wang, S., Zha, Z.J., Meng, D., Huang, Q.: Adaptive reconstruction network for weakly supervised referring expression grounding. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2611–2620 (2019)
20. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision. pp. 852–869. Springer (2016)
21. Lu, P., Ji, L., Zhang, W., Duan, N., Zhou, M., Wang, J.: R-vqa: learning visual relation facts with semantic attention for visual question answering. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1880–1889. ACM (2018)
22. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
23. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: European Conference on Computer Vision. pp. 792–807. Springer (2016)
24. Pan, P., Xu, Z., Yang, Y., Wu, F., Zhuang, Y.: Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1029–1038 (2016)
25. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)
26. Qian, X., Zhuang, Y., Li, Y., Xiao, S., Pu, S., Xiao, J.: Video relation detection with spatio-temporal graph. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 84–93. ACM (2019)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
28. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: European Conference on Computer Vision. pp. 817–834. Springer (2016)
29. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)
30. Shang, X., Di, D., Xiao, J., Cao, Y., Yang, X., Chua, T.S.: Annotating objects and relations in user-generated videos. In: ACM International Conference on Multimedia Retrieval. Ottawa, ON, Canada (June 2019)
31. Shang, X., Ren, T., Guo, J., Zhang, H., Chua, T.S.: Video visual relation detection. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1300–1308. ACM (2017)
32. Shi, J., Xu, J., Gong, B., Xu, C.: Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10444–10452 (2019)

33. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
34. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
35. Tsai, Y.H.H., Divvala, S., Morency, L.P., Salakhutdinov, R., Farhadi, A.: Video relationship reasoning using gated spatio-temporal energy graph. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10424–10433 (2019)
36. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE international conference on computer vision. pp. 4534–4542 (2015)
37. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision. pp. 3551–3558 (2013)
38. Wang, X., Gupta, A.: Videos as space-time region graphs. In: Proceedings of the European conference on computer vision (ECCV). pp. 399–417 (2018)
39. Yamaguchi, M., Saito, K., Ushiku, Y., Harada, T.: Spatio-temporal person retrieval via natural language queries. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1453–1462 (2017)
40. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 17–24. IEEE (2010)
41. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1307–1315 (2018)
42. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision. pp. 69–85. Springer (2016)
43. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4694–4702 (2015)
44. Zhang, H., Niu, Y., Chang, S.F.: Grounding referring expressions in images by variational context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4158–4166 (2018)
45. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A.M., Elhoseiny, M.: Large-scale visual relationship understanding. In: AAAI (2019)
46. Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
47. Zhou, L., Louis, N., Corso, J.J.: Weakly-supervised video object grounding from text by loss weighting and object interaction. arXiv preprint arXiv:1805.02834 (2018)