Object-Aware Multi-Branch Relation Networks for Spatio-Temporal Video Grounding

Zhu Zhang[1] , **Zhou Zhao**[1*] , **Zhijie Lin**[1] , **Baoxing Huai**[2] and **Jing Yuan**[2]
[1]College of Computer Science, Zhejiang University, China
[2]Huawei Cloud & AI, China
{zhangzhu, zhaozhou, linzhijie}@zju.edu.cn, {huaibaoxing, nicholas.yuan}@huawei.com

Abstract

Spatio-temporal video grounding aims to retrieve the spatio-temporal tube of a queried object according to the given sentence. Currently, most existing grounding methods are restricted to well-aligned segment-sentence pairs. In this paper, we explore spatio-temporal video grounding on unaligned data and multi-form sentences. This challenging task requires to capture critical object relations to identify the queried target. However, existing approaches cannot distinguish notable objects and remain in ineffective relation modeling between unnecessary objects. Thus, we propose a novel object-aware multi-branch relation network for object-aware relation discovery. Concretely, we first devise multiple branches to develop object-aware region modeling, where each branch focuses on a crucial object mentioned in the sentence. We then propose multi-branch relation reasoning to capture critical object relationships between the main branch and auxiliary branches. Moreover, we apply a diversity loss to make each branch only pay attention to its corresponding object and boost multi-branch learning. The extensive experiments show the effectiveness of our proposed method.

1 Introduction

Spatio-temporal video grounding is an emerging task in the cross-modal understanding field. Given a sentence depicting an object, this task aims to retrieve the spatio-temporal tube of the queried object, i.e. a sequence of bounding boxes. Most existing spatio-temporal grounding methods [Chen *et al.*, 2019b; Shi *et al.*,] are restricted to well-aligned segment-sentence pairs, where the segment has been trimmed from the raw video and is temporally aligned to the sentence. Recently, researchers [Zhang *et al.*, 2020] begin to explore spatio-temporal video grounding (STVG) on unaligned data and multi-form sentences. Concretely, as shown in Figure 1, either the declarative sentence or interrogative sentence describes a short-term action of the target object "child" and

Figure 1: An example of spatio-temporal video grounding on unaligned data and multi-form sentences.

matches with the spatio-temporal tube within a small segment. To localize the target existence in a fleeting clip, we need to distinguish the subtle status of the object according to the textual information. Specifically, the sentence often illustrates diverse relationships between the queried object with other objects, thus the key of this task is to capture these crucial relations in video contents to identify the spatio-temporal tube. Particularly, the interrogative sentences depict unknown objects and lack the explicit information of the object, e.g., the direct characteristics "a child in yellow" in Figure 1. Grounding these sentences can only depend on the object relationships such as the action relation "kicking a ball" and spatial relation "in front of the goal".

Although existing grounding methods [Yamaguchi *et al.*, 2017; Chen *et al.*, 2019b] have achieved excellent performance on aligned segment-video pairs, they are ineffectively applied to unaligned data and multi-form sentences. On the one hand, they heavily rely on the tube pre-generation to extract a series of candidate tubes and then select the most relevant one according to the sentence. But without the temporal alignment, it is difficult to pre-generate reasonable candidate tubes. On the other hand, these approaches always ignore the relation construction between objects and model each tube individually. Recently, Zhang et al. [Zhang *et al.*, 2020] explore spatio-temporal grounding on unaligned data. They incorporate the textual clues into region features and employ spatio-temporal graph reasoning to retrieve the spatio-temporal tube. Despite this method tries to capture object relations by cross-modal region interactions, it does not exclude inessential regions from massive region proposals and may lead to severe obstruction for effective relation modeling. Concretely, there are a large number of objects in videos but most of them are irrelevant to the textual descriptions. [Zhang *et al.*, 2020]

*Zhou Zhao is the corresponding author.

fails to filter out the unnecessary ones and remain in the coarse relation modeling for all regions. Hence, we need to pay more attention to the crucial objects mentioned in the sentence and build sufficient cross-modal relation reasoning between them for precise video grounding.

In this paper, we propose a novel Object-Aware Multi-Branch Relation Network (OMRN) for object-aware fine-grained relation reasoning. We first extract region features from the video and learn object representations corresponding to the nouns in the sentence. We then employ multiple branches to develop object-aware region modeling and discover the notable regions containing informative objects, where the main branch corresponds to the queried object and each auxiliary branch focuses on an object mentioned in the sentence. Concretely, we apply the object-aware modulation to strengthen object-relevant region features and weaken unnecessary ones in each branch. Next, we can conduct object-region cross-modal matching in each branch. After it, we propose the multi-branch relation reasoning to capture critical object relationships between the main branch and auxiliary branches, where the irrelevant regions are filtered out by preceding matching scores. Further, considering each branch should only focus on its corresponding object, we devise a diversity loss to make different branches pay attention to different regions, that is, have diverse distributions of matching scores. Eventually, we apply a spatio-temporal localizer to determine the temporal boundaries and retrieve the target tube. The main contributions of this paper are as follows:

- We propose a novel object-aware multi-branch relation network for spatio-temporal video grounding, which can discover object-aware fine-grained relations and retrieve the accurate tubes of the queried objects.
- We devise multiple branches with a diversity loss to develop object-aware region modeling, where each branch focuses on a crucial object mentioned in the sentence and the diversity loss makes different branches focus on their corresponding objects.
- We employ the multi-branch relation reasoning to capture critical object relationships between the main branch and auxiliary branches.
- The extensive experiments show the effectiveness of our object-aware multi-branch framework.

2 Related Work

In this section, we briefly review some related work on visual grounding and video grounding.

Visual grounding is to localize the object in an image according to the referring expression. Early approaches [Hu *et al.*, 2016; Nagaraja *et al.*, 2016] often model the language information by RNN, extract region features through CNN and then learn the object-language alignment. Recent methods [Yu *et al.*, 2018; Hu *et al.*, 2017] parse the expression into multiple parts and compute cross-modal alignment scores for each part. Furthermore, [Deng *et al.*, 2018; Zhuang *et al.*, 2018] employ the co-attention mechanism to develop cross-modal interactions for fine-grained matching. And [Yang *et al.*, 2019b; Yang *et al.*, 2019a] capture the relations between regions to boost the grounding accuracy.

Video grounding can be categorized into temporal grounding and spatio-temporal grounding. Given a sentence, temporal grounding localizes a temporal clip in the video. Early methods [Hendricks *et al.*, 2017; Gao *et al.*, 2017] apply a proposal-and-selection framework that first samples massive candidate clips and then select the most relevant one semantically matching with the sentence. Recently, [Chen *et al.*, 2019a; Zhang *et al.*, 2019c; Lin *et al.*, 2020b] develop frame-by-word interactions between visual and textual contents and discover the dynamical clues by attention mechanism. [Zhang *et al.*, 2019a] explicitly model moment-wise relations as a structured graph and employ an iterative graph adjustment. [Yuan *et al.*, 2019] propose a semantic conditioned dynamic modulation for better correlating video contents over time and [Zhang *et al.*, 2019b] adopt a 2D temporal map to cover diverse moments with different lengths. In the weakly-supervised setting, [Mithun *et al.*, 2019] leverage the attention scores to align moments with sentences, and [Lin *et al.*, 2020a] propose a semantic completion network to estimate each clip by language reconstruction. Besides natural language queries, Zhang et al. [Zhang *et al.*, 2019d] try to detect the unseen video clip according to image queries.

Spatio-temporal video grounding is a natural extension of temporal grounding, which retrieves a spatio-temporal tube from a video corresponding to the sentence. Most existing approaches are designed for well-aligned segment-sentence data. [Yamaguchi *et al.*, 2017] only ground the person tube in multiple videos and [Zhou *et al.*, 2018; Chen *et al.*, 2019b] further retrieve the spatio-temporal tubes of diverse objects from trimmed videos by weakly-supervised MIL methods. Different from single-object grounding, [Huang *et al.*, 2018; Shi *et al.*,] localize each noun or pronoun of sentences in frames. Recently, [Zhang *et al.*, 2020] explore spatio-temporal grounding on unaligned data by spatio-temporal cross-modal graph modeling. But it still fails to capture the critical objects and remains in the coarse relation reasoning. In this paper, we explore object-aware fine-grained relation modeling and further improve the grounding accuracy.

3 The Proposed Method

As shown in Figure 2, we propose the object-aware multi-branch relation network (OMRN) for this STVG task, where we develop object-aware multi-branch region modeling to discover the notable regions containing informative objects and devise multi-branch relation reasoning to capture critical object relationships.

3.1 Region and Object Extraction

Given a video $\mathbf{v}$, we extract region features $\{\{\mathbf{r}_k^n\}_{k=1}^K\}_{n=1}^N$ by a pre-trained Faster R-CNN, where the video has N frames and each frame contains K regions. The feature $\mathbf{r}_k^n$ corresponds to the k-th region in frame n. And each region is also associated with a bounding box $\mathbf{b}_k^n = (x_k^n, y_k^n, w_k^n, h_k^n)$, where (x_k^n, y_k^n) are the center coordinates and (w_k^n, h_k^n) are the box width and height. Considering video grounding requires to capture the object dynamics but region features extracted from still frames lack motion information, we then

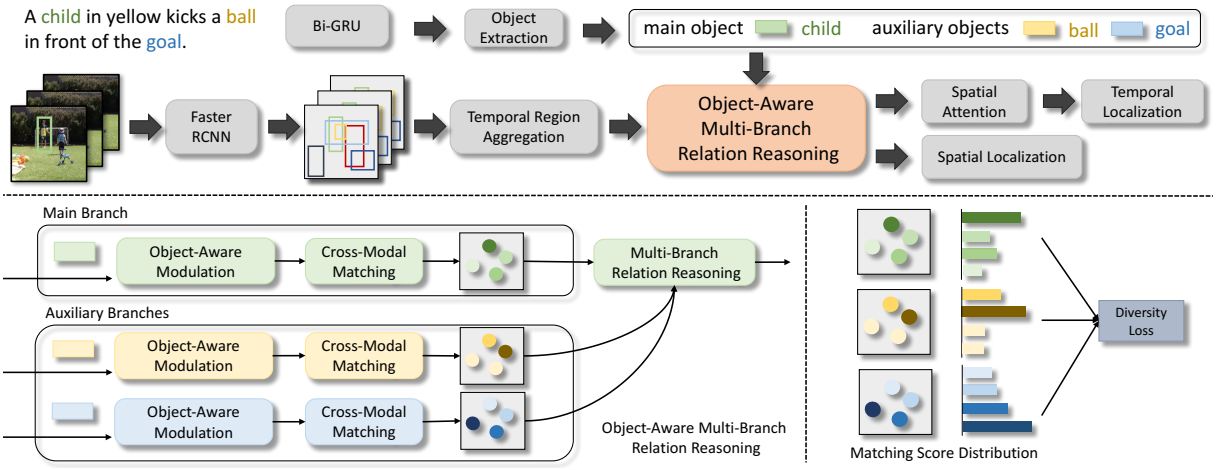

Figure 2: The Overall Architecture of Object-Aware Multi-Branch Relation Network.

adopt a temporal region aggregation method to incorporate dynamic clues from forward L frames and backward L frames into each region. Concretely, if two regions in adjacent frames have similar semantic features and spatial locations, we expect them to contain the same object. Thus, the linking score between region $\mathbf{r}_i^{n_1}$ and $\mathbf{r}_j^{n_2}$ by

$$s(\mathbf{r}_i^{n_1}, \mathbf{r}_j^{n_2}) = \cos(\mathbf{r}_i^{n_1}, \mathbf{r}_j^{n_2}) + \frac{\alpha}{|n_2 - n_1|} \cdot \mathrm{IoU}(\mathbf{b}_i^{n_1}, \mathbf{b}_j^{n_2}),$$

where the $\cos(\cdot)$ means the cosine similarity of two features and $\mathrm{IoU}(\cdot)$ is the IoU score of their bounding boxes. The $|n_2 - n_1|$ is the temporal distance of two regions and applied to limit the IoU score. And α is a balanced coefficient. For each region, we select the region with the maximal linking score from each adjacent frame and obtain $2L$ relevant features. Next, we apply a mean pooling on the $2L + 1$ features to absorb region dynamics, followed by a linear layer for feature transformation. For simplicity, we still denote the pooled region features with temporal dynamics by $\{\{\mathbf{r}_k^n\}_{k=1}^K\}_{n=1}^N$.

For language modeling, we input word embeddings of the sentence into a Bi-GRU to learn the word-level semantic features $\{\mathbf{s}_m\}_{m=1}^M$ with M words. After it, we learn object features with context information for each object mentioned in the sentence. Specifically, we first identify all nouns in the sentence using the library of NLTK. Assuming there are T nouns in the sentence, each noun points to an object in the video and the t-th noun corresponds to the word-level feature $\mathbf{s}_t$. In Figure 2, the sentence contains three objects "child", "ball" and "goal", and we aim to retrieve the spatio-temporal tube of the main object "child" (i.e. the queried object). For interrogative sentences, we regard the interrogative words (e.g. "who" and "what") as the main objects. Next, we apply a context attention to aggregate the language context for each object by

$$\beta_{t,m} = \mathbf{w}^\top \tanh(\mathbf{W}_1^s \mathbf{s}_t + \mathbf{W}_2^s \mathbf{s}_m + \mathbf{b}^s),$$
$$\widetilde{\mathbf{o}}_t = \sum_{m=1}^{M} \mathrm{softmax}(\beta_{t,m}) \cdot \mathbf{s}_m, \ \mathbf{o}_t = [\mathbf{s}_t; \widetilde{\mathbf{o}}_t],$$

where $\mathbf{W}_1^s$, $\mathbf{W}_2^s$ are projection matrices, $\mathbf{b}^s$ is the bias and $\mathbf{w}^\top$ is the row vector. The $\beta_{t,m}$ is the attention weight of object t for the m-th word. Finally, we obtain the object features $\{\mathbf{o}_t\}_{t=1}^T$, where $\mathbf{o}_1$ is the feature of the main object and $\{\mathbf{o}_t\}_{t=2}^T$ are auxiliary object features.

3.2 Object-Aware Multi-Branch Relation Reasoning

We next devise object-aware multi-branch relation reasoning with a diversity loss to capture object-aware fine-grained relations. Concretely, we first take multiple branches to learn object-aware region features and then apply multi-branch relation reasoning to capture critical object relationships from multiple branches, where the main branch corresponds to the main object and auxiliary branches correspond to auxiliary objects. And the diversity loss makes different branches focus on their corresponding objects.

Object-Aware Multi-Branch Region Modeling

For branch t with object $\mathbf{o}_t$, we first apply the object-aware modulation to strengthen object-relevant region features and weaken unnecessary ones. Concretely, we produce the object-aware modulation vectors by

$$\boldsymbol{\gamma}_t = \tanh(\mathbf{W}^\gamma \mathbf{o}_t + \mathbf{b}^\gamma), \ \boldsymbol{\delta}_t = \tanh(\mathbf{W}^\delta \mathbf{o}_t + \mathbf{b}^\delta),$$

where $\boldsymbol{\gamma}_t$ and $\boldsymbol{\delta}_t$ are the modulation gate and bias based on object t. We then modulate all region features by

$$\mathbf{r}_{tk}^n = \boldsymbol{\gamma}_t \odot \mathbf{r}_k^n + \boldsymbol{\delta}_t,$$

where $\odot$ is the element-wise multiplication and $\mathbf{r}_{tk}^n$ means the object-aware region feature in branch t. The modulation vectors are expected to emphasize region features containing the object t and weaken unrelated ones.

In branch t, we then conduct cross-modal matching between region features $\mathbf{r}_{tk}^n$ and the object feature $\mathbf{o}_t$ by

$$d_{tk}^n = \mathbf{w}^\top \tanh(\mathbf{W}^c[\mathbf{r}_{tk}^n; \mathbf{o}_t; \mathbf{r}_{tk}^n \odot \mathbf{o}_t; \mathbf{r}_{tk}^n - \mathbf{o}_t] + \mathbf{b}^c),$$

where d_{tk}^n is the matching score of region k in frame n on branch t. Next, we apply the softmax function on d_{tk}^n to obtain the matching score distribution over regions, given by $\hat{d}_{tk}^n = \frac{\exp(d_{tk}^n)}{\sum_{k=1}^K \exp(d_{tk}^n)}$. It is used to multi-branch relation reasoning for critical object relation discovery and is also applied to construct the diversity loss between multiple branches.

Multi-Branch Relation Reasoning

Next, we develop the multi-branch relation reasoning in each frame to capture critical object relationships by integrating auxiliary branches into the main branch. Concretely, the branch t focuses on its corresponding object t and has learnt the object-aware region feature $\mathbf{r}^n_{tk}$ with the matching score $\hat{d}^n_{tk}$. To build the relation between the main object (i.e. object 1) and auxiliary object t in each frame, we absorb crucial clues of notable regions from the branch t into the main branch. We estimate the attention weight between the k-th region $\mathbf{r}^n_{1k}$ in the main branch (i.e. branch 1) and the l-th region $\mathbf{r}^n_{tl}$ in the auxiliary branch t by

$$\epsilon^n_{1k,tl} = \mathbf{w}^\top \tanh(\mathbf{W}^m_1 \mathbf{r}^n_{1k} + \mathbf{W}^m_2 \mathbf{r}^n_{tl} + \mathbf{W}^m_3 \mathbf{b}^n_{1k,tl} + \mathbf{b}^m),$$

where $\mathbf{b}^n_{1k,tl}$= $[x^n_{1k,tl}; y^n_{1k,tl}; w^n_{1k,tl}; h^n_{1k,tl}]$ is the relative geometry vector between region $\mathbf{r}^n_{1k}$ and $\mathbf{r}^n_{tl}$, given by

$$x^n_{1k,tl} = (x^n_{1k} - x^n_{tl})/w^n_{tl},\ y^n_{1k,tl} = (y^n_{1k} - y^n_{tl})/h^n_{tl},$$
$$w^n_{1k,tl} = \log(w^n_{1k}/w^n_{tl}),\ h^n_{1k,tl} = \log(h^n_{1k}/h^n_{tl}).$$

Thus, the attention weight $\epsilon^n_{1k,tl}$ is built on object-aware features $\mathbf{r}^n_{1k}$ and $\mathbf{r}^n_{tl}$ with the spatial location information. After it, we aggregate the regions relevant to object t from the auxiliary branch t by

$$\hat{\epsilon}^n_{1k,tl} = \frac{\exp(\epsilon^n_{1k,tl})}{\sum_{l=1}^K \exp(\epsilon^n_{1k,tl})},\ \mathbf{r}^n_{1k,t} = \sum_{l=1}^K \hat{d}^n_{1k} \cdot \hat{d}^n_{tl} \cdot \hat{\epsilon}^n_{1k,tl} \cdot \mathbf{r}^n_{tl},$$

where $\mathbf{r}^n_{1k,t}$ is the aggregation feature from branch t for the region k in the main branch. We first apply the softmax on the attention weights $\epsilon^n_{1k,tl}$ and then aggregate region features with the matching scores $\hat{d}^n_{1k}$ and $\hat{d}^n_{tl}$ as weighting terms. Thus, the relation reasoning between the main object and auxiliary object t will focus on notable regions with higher matching scores $\hat{d}^n_{tl}$ and filter out inessential ones. Simultaneously, by the prior confidence $\hat{d}^n_{1k}$ of the main object, we can enhance the relation modeling for these regions with higher $\hat{d}^n_{1k}$ in the main branch.

After multi-branch relation reasoning from all auxiliary branches, we learn the multi-branch aggregation features $\{\mathbf{r}^n_{1k,t}\}^T_{t=2}$ for each region k of frame n in the main branch. We then obtain the final object-aware multi-branch features $\{\{\widetilde{\mathbf{r}}^n_k\}^K_{k=1}\}^N_{n=1}$ by

$$\widetilde{\mathbf{r}}^n_k = \mathrm{ReLU}(\mathbf{r}^n_{1k} + \sum_{t=2}^T \mathbf{r}^n_{1k,t}).$$

Diversity Loss Between Branches

Considering each branch should only focus on its corresponding object, we devise a diversity loss to make different branches have diverse score distributions over regions. Specifically, we denote the score distribution in the frame n on branch t as $\hat{\mathbf{d}}^n_t = [\hat{d}^n_{t1}, \ldots, \hat{d}^n_{tK}]^\top$ and calculate the diversity loss by the distribution similarity between multiple branches as follows:

$$\mathcal{L}_d = \tfrac{1}{Z} \textstyle\sum_{n \in \mathcal{S}_{gt}} \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} (\hat{\mathbf{d}}^n_i)^\top (\hat{\mathbf{d}}^n_j),$$

where $\mathcal{S}_{gt}$ is the set of frames in the ground truth segment and $Z = \frac{1}{2}|\mathcal{S}_{gt}|T(T-1)$ is the normalization factor. This diversity loss encourages each branch to pay more attention to the region matching with the corresponding object.

3.3 Spatio-Temporal Localization

In this section, we apply a spatio-temporal localizer to predict the spatio-temporal tube of the queried object based on final region features $\{\{\widetilde{\mathbf{r}}^n_k\}^K_{k=1}\}^N_{n=1}$, including the temporal moment localization and spatial region localization.

For spatial localization, we estimate the region confidence scores according to the main object feature $\mathbf{o}_1$ by

$$p^n_k = \sigma((\mathbf{W}^r \widetilde{\mathbf{r}}^n_k)^\top (\mathbf{W}^o \mathbf{o}_1)),$$

where σ is the sigmoid function and p^n_k is the confidence score of region k in frame n. Next, we apply the spatial loss to guide the spatial localization where we only consider frames in the set $\mathcal{S}_{gt}$, i.e. in the ground truth segment. We first compute the IoU score IoU^n_k of each region with the corresponding ground truth region and then calculate the spatial loss by

$$\mathcal{SL}^n_k = (1 - \mathrm{IoU}^n_k) \cdot \log(1 - p^n_k) + \mathrm{IoU}^n_k \cdot \log(p^n_k),$$
$$\mathcal{L}_s = -\frac{1}{|\mathcal{S}_{gt}|K} \sum_{n \in \mathcal{S}_{gt}} \sum_{k=1}^K \mathcal{SL}^n_k.$$

For temporal localization, we sample a set of candidate segments and estimate their confidence scores with the boundary adjustment. Concretely, we first adopt a spatial attention to aggregate the final region features by

$$\zeta^n_k = \mathbf{w}^\top \tanh(\mathbf{W}^f_1 \widetilde{\mathbf{r}}^n_k + \mathbf{W}^f_2 \mathbf{o}_1 + \mathbf{b}^f),$$
$$\mathbf{f}^n = \textstyle\sum_{k=1}^K \mathrm{softmax}(\zeta^n_k) \cdot \widetilde{\mathbf{r}}^n_k,$$

where $\mathbf{f}^n$ is the object-aware feature of frame n. We then input $\{\mathbf{f}^n\}^N_{n=1}$ into a Bi-GRU to learn context features $\{\widetilde{\mathbf{f}}^n\}^N_{n=1}$. Next, we regard each frame as a sample center and define H candidate segments with widths $\{w^h\}^H_{h=1}$ at each center. After it, we generate the confidence scores and their boundary offsets by

$$\mathbf{c}^n = \sigma(\mathbf{W}^c \widetilde{\mathbf{f}}^n + \mathbf{b}^c),\ \mathbf{l}^n = \mathbf{W}^l \widetilde{\mathbf{f}}^n + \mathbf{b}^l,$$

where $\mathbf{c}^n \in \mathbb{R}^H$ represent confidence scores of H candidates at step n and $\mathbf{l}^n \in \mathbb{R}^{2H}$ are their offsets. Similar to the spatial loss, we apply the temporal alignment loss for segment selection, which is based on the temporal IoU score tIoU^n_h of each candidate segment with the ground truth segment, given by

$$\mathcal{TL}^n_h = (1 - \mathrm{tIoU}^n_h) \cdot \log(1 - c^n_h) + \mathrm{tIoU}^n_h \cdot \log(c^n_h),$$
$$\mathcal{L}_t = -\frac{1}{NH} \sum_{n=1}^N \sum_{h=1}^H \mathcal{TL}^n_h.$$

Next, we select the segment with the highest c^n_h and adjust its temporal boundaries by the offset (l_s, l_e) from $\mathbf{l}^n_h$. Here we develop another regression loss to train the offsets. With the original boundaries (s, e) of the selected segment and ground truth $(\hat{s}, \hat{e})$, we first compute the ground truth offsets $\hat{l}_s = s - \hat{s}$ and $\hat{l}_e = e - \hat{e}$ and compute the regression loss by

$$\mathcal{L}_r = \mathrm{R}(l_s - \hat{l}_s) + \mathrm{R}(l_e - \hat{l}_e),$$

where R is the smooth L1 function.

Method	Declarative Sentence Grounding				Interrogative Sentence Grounding			
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
GroundeR + TALL		9.78%	11.04%	4.09%		9.32%	11.39%	3.24%
STPR + TALL	34.63%	10.40%	12.38%	4.27%	33.73%	9.98%	11.74%	4.36%
WSSTG + TALL		11.36%	14.63%	5.91%		10.65%	13.90%	5.32%
GroundeR + L-Net		11.89%	15.32%	5.45%		11.05%	14.28%	5.11%
STPR + L-Net	40.86%	12.93%	16.27%	5.68%	39.79%	11.94%	14.73%	5.27%
WSSTG + L-Net		14.45%	18.00%	7.89%		13.36%	17.39%	7.06%
STGRN	48.47%	19.75%	25.77%	14.60%	46.98%	18.32%	21.10%	12.83%
OMRN (Ours)	**50.73%**	**23.11%**	**32.61%**	**16.42%**	**49.19%**	**20.63%**	**28.35%**	**14.11%**

Table 1: Performance Evaluation Results on the VidSTG Dataset.

Eventually, we apply the multi-task loss to train our OMRN method in an end-to-end manner, given by

$$\mathcal{L}_{OMRN} = \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_t + \lambda_3 \mathcal{L}_r + \lambda_4 \mathcal{L}_d,$$

where λ_1, λ_2, λ_3 and λ_4 control the balance of four losses.

During inference, we first detect the segment with the highest temporal confidence score, fine-tune its boundaries by its offsets and then select the regions with the highest spatial scores within the selected segment to generate the target tube.

4 Experiments

4.1 Experimental Settings

Dataset

We conduct experiments on a large-scale spatio-temporal video grounding dataset VidSTG [Zhang *et al.*, 2020], which is constructed from the video object relation dataset VidOR [Shang *et al.*, 2019] by annotating the natural language descriptions. As we know, VidSTG is the only grounding dataset on unaligned video-sentence data and multi-form sentences. Specifically, VidSTG contains 5,563, 618 and 743 videos in the training, validation and testing sets, totaling 6,924 videos. There are 99,943 sentence annotations for 80 types of queried objects, including 44,808 and 55,135 for declarative and interrogative sentences, respectively. Moreover, the duration of the videos is 28.01s and temporal tube length is 9.68s on average. And declarative and interrogative sentences have about 11.12 and 8.98 words, respectively.

Implementation Details

During data preprocessing, we extract 1,024-d region features by a pre-trained Faster R-CNN [Ren *et al.*, 2015]. We sample 5 frames per second and extract $K = 20$ regions for each frame. For language, we apply a pre-trained Glove embedding to obtain 300-d word features and use the NLTK to recognize the nouns in sentences. As for modeling setting, we set α to 0.6, L to 5 and set λ_1, λ_2, λ_3 and λ_4 to 1.0, 1.0, 0.001 and 1.0, respectively. We define $H = 9$ candidate segments at each step with temporal widths $[3, 9, 17, 33, 65, 97, 129, 165, 197]$. We set the dimensions of all projection matrices and biases to 256 and set the hidden state of each direction in BiGRU to 128. We employ an Adam optimizer with the initial learning rate 0.0005.

Evaluation Criteria

We apply the criterion m_tIoU to evaluate the temporal grounding performance and use m_vIoU and vIoU@R to estimate the spatio-temporal accuracy as [Zhang *et al.*, 2020]. Concretely, the m_tIoU is the average temporal IoU of selected segments with the ground truth. We define vIoU as the spatio-temporal IoU between the predicted and ground truth tubes, given by $\text{vIoU} = \frac{1}{|\mathcal{S}_p \cup \mathcal{S}_{gt}|} \sum_{n \in \mathcal{S}_p \cap \mathcal{S}_{gt}} \text{IoU}(r^n, \hat{r}^n)$. The $\mathcal{S}_p$ is the frame set in the predicted segment and $\mathcal{S}_{gt}$ is the frame set in the ground truth. The r^n, $\hat{r}^n$ are the predicted and ground truth regions in frame n. The m_vIoU is the mean vIoU of all test samples and vIoU@R is the rate of testing samples with vIoU $>$ R.

4.2 Performance Comparison

As an emerging task, only the STGRN method [Zhang *et al.*, 2020] is designed for STVG on unaligned data. Besides it, we combine the temporal grounding methods TALL [Gao *et al.*, 2017] and L-Net [Chen *et al.*, 2019a] with spatio-temporal grounding approaches on aligned data as the baselines. Specifically, TALL employs a proposal-and-selection framework for the temporal localization and L-Net develops frame-by-word interactions for holistic segment selection. Given the predicted segment, GroundeR [Rohrbach *et al.*, 2016] is a visual grounding method to retrieve the target region in each frame. And the STPR [Yamaguchi *et al.*, 2017] and WSSTG [Chen *et al.*, 2019b] approaches apply tube pre-generation in the segment and then rank these tubes by cross-modal estimation. Concretely, the original STPR is applied to multi-video person grounding, we extend it to single-video grounding for diverse objects. The WSSTG originally use a weakly-supervised ranking loss but we replace it with a supervised triplet loss [Yang *et al.*, 2019a]. Thus, there are 6 combinations such as WSSTG+TALL and STPR+L-Net.

The overall experiment results are shown in Table 1 and we can find some interesting points:

- On the whole, the grounding performance of all models for interrogative sentences is lower than for declarative sentences, validating the unknown objects without explicit characteristics are more difficult to ground.
- For temporal grounding, region-level methods STGRN and OMRN outperform the frame-level methods TALL and L-Net, which demonstrates the fine-grained region modeling is beneficial to determine the accurate temporal boundaries of target tubes.

Methods	m_vIoU	vIoU@0.3	vIoU@0.5
GroundeR + Tem. GT	27.31%	40.53%	20.86%
STPR + Tem. GT	28.20%	42.08%	21.75%
WSSTG + Tem. GT	31.51%	46.99%	27.65%
STGRN + Tem. GT	36.75%	50.78%	32.93%
OMRN + Tem. GT	**39.57%**	**58.19%**	**37.91%**

Table 2: Evaluation Results with the Temporal Ground Truth.

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
w/o. OM	47.78%	19.85%	28.55%	12.99%
w/o. DL	48.32%	20.08%	28.80%	13.42%
w/o. CM	47.23%	19.06%	27.08%	12.25%
w/o. TA	48.92%	20.50%	29.76%	13.61%
w/o. CA	48.85%	20.73%	29.87%	14.02%
full	**49.88%**	**21.73%**	**30.26%**	**15.14%**

Table 3: Ablation Results on the VidSTG Dataset.

- For spatio-temporal grounding, the GroundR+$\{\cdot\}$ approaches ignore the temporal dynamics of objects and achieve the worst performance, suggesting it is crucial to capture the object dynamics among frames for high-quality spatio-temporal video grounding.
- In all criteria, our OMRN achieves the remarkable performance improvements compared with baselines. This fact shows our method can effectively focus on the notable regions by object-aware multi-branch region modeling with the diversity loss and capture critical object relations by multi-branch reasoning.

Furthermore, given the temporal ground truth segment during inference, we compare the spatial grounding ability of our OMRN method with baselines. The results are shown in Table 2 and we do not separate declarative and interrogative sentences here. We can see that our OMRN still achieves the apparent performance improvement on all criteria, especially for vIoU@0.3. This demonstrates our OMRN approach is still effective when applied to aligned segment-sentence data.

4.3 Ablation Study

We next verify the contribution of each part of our method by ablation study. We remove one key component at a time to generate an ablation model. The object-aware multi-branch modeling is vital in our method, so we first remove the object-aware modulation from each branch as **w/o. OM**. We then discard the diversity loss from the multi-task loss, denoted by **w/o. DL**. Further, we remove the cross-modal matching from all branches and discard the weighting terms $\hat{d}^n_{1k}$ and $\hat{d}^n_{tl}$ in multi-branch relation reasoning, denoted by **w/o. CM**. Note that in this ablation model, the diversity loss is also ineffective due to the lack the matching score distributions. Next, we develop the ablation study on the basic region and object modeling. We discard the temporal region aggregation from region modeling as **w/o. TA** and remove the context attention during object extraction as **w/o. CA**.

The ablation results are shown in Table 3. We can find all ablation models have the performance degradation compared with the full model, showing each above component

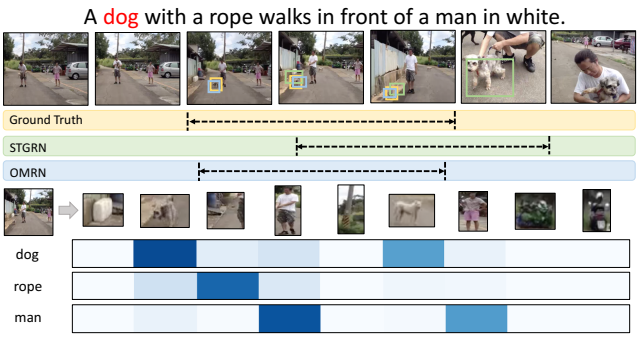

Figure 3: A typical example of the grounding result.

is helpful to improve the grounding accuracy. And the ablation models **w/o. OM**, **w/o. DL** and **w/o. CM** have the lower accuracy than **w/o. TA** and **w/o. CA**, which suggests the object-aware multi-branch relation reasoning plays a crucial role in high-quality spatio-temporal grounding. Moreover, the model **w/o. CM** achieves the worst performance, validating the cross-modal matching with the diversity regularization is very important to incorporate language-relevant region features from auxiliary branches into the main branch.

4.4 Qualitative Analysis

To qualitatively validate the effectiveness of our OMRN method, we display a typical example in Figure 3. The sentence describes a short-term state of the "dog" and requires to capture object-aware fine-grained relations. By intuitive comparison, our OMRN can retrieve the more accurate temporal boundaries and spatio-temporal tube of the "dog" than the best baseline STGRN. Furthermore, we display the object-region matching score distribution in the example, where we visualize the matching scores between three objects (i.e. "dog", "rope" and "man") and the regions of the 4-th frame. Although there are a woman and another dog in the frame, our method can still eliminate the interference and focus on the notable region containing the corresponding object, e.g., the object "dog" assigns a higher score to the 2-th region rather than the 6-th region.

5 Conclusion

In this paper, we propose a novel object-aware multi-branch relation network for STVG. The method can effectively focus on the vital regions by object-aware multi-branch region modeling and capture sufficient object relations by multi-branch reasoning for high-quality spatio-temporal grounding.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No. 2018AAA0100603, Zhejiang Natural Science Foundation LR19F020006 and the National Natural Science Foundation of China under Grant No.61836002, No.U1611461 and No.61751209. This research is partially supported by the Language and Speech Innovation Lab of HUAWEI Cloud.

References

[Chen *et al.*, 2019a] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *AAAI*, 2019.

[Chen *et al.*, 2019b] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. 2019.

[Deng *et al.*, 2018] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *CVPR*, pages 7746–7755, 2018.

[Gao *et al.*, 2017] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *ICCV*, pages 5277–5285, 2017.

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

[Hu *et al.*, 2016] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016.

[Hu *et al.*, 2017] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, pages 1115–1124, 2017.

[Huang *et al.*, 2018] De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In *CVPR*, June 2018.

[Lin *et al.*, 2020a] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, 2020.

[Lin *et al.*, 2020b] Zhijie Lin, Zhou Zhao, Zhu Zhang, Zijian Zhang, and Deng Cai. Moment retrieval via cross-modal interaction networks with query reconstruction. *TIP*, 29:3750–3762, 2020.

[Mithun *et al.*, 2019] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, pages 11592–11601, 2019.

[Nagaraja *et al.*, 2016] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807. Springer, 2016.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[Rohrbach *et al.*, 2016] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, pages 817–834. Springer, 2016.

[Shang *et al.*, 2019] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR*, pages 279–287. ACM, 2019.

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

[Yamaguchi *et al.*, 2017] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *ICCV*, pages 1453–1462, 2017.

[Yang *et al.*, 2019a] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *CVPR*, pages 4145–4154, 2019.

[Yang *et al.*, 2019b] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *ICCV*, pages 4644–4653, 2019.

[Yu *et al.*, 2018] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018.

[Yuan *et al.*, 2019] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NIPS*, pages 534–544, 2019.

[Zhang *et al.*, 2019a] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, pages 1247–1257, 2019.

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

[Zhang *et al.*, 2019c] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*, 2019.

[Zhang *et al.*, 2019d] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jingkuan Song, and Deng Cai. Localizing unseen activities in video via image query. In *IJCAI*, 2019.

[Zhang *et al.*, 2020] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. 2020.

[Zhou *et al.*, 2018] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *arXiv preprint arXiv:1805.02834*, 2018.

[Zhuang *et al.*, 2018] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *CVPR*, pages 4252–4261, 2018.