Inverse Compositional Learning for Weakly-supervised Relation Grounding

Huan Li, Ping Wei,[*] Zeyu Ma, Nanning Zheng
National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Xi’an Jiaotong University, Xi’an, China
{huanli@stu., pingwei@, zeyu98@stu., nnzheng@}xjtu.edu.cn

Abstract

Video relation grounding (VRG) is a significant and challenging problem in the domains of cross-modal learning and video understanding. In this study, we introduce a novel approach called inverse compositional learning (ICL) for weakly-supervised video relation grounding. Our approach represents relations at both the holistic and partial levels, formulating VRG as a joint optimization problem that encompasses reasoning at both levels. For holistic-level reasoning, we propose an inverse attention mechanism and a compositional encoder to generate compositional relevance features. Additionally, we introduce an inverse loss to evaluate and learn the relevance between visual features and relation features. At the partial-level reasoning, we introduce a grounding by classification scheme. By leveraging the learned holistic-level features and partial-level features, we train the entire model in an end-to-end manner. We conduct evaluations on two challenging datasets and demonstrate the substantial superiority of our proposed method over state-of-the-art methods. Extensive ablation studies confirm the effectiveness of our approach.

1. Introduction

The objective of the Video Relation Grounding (VRG) task is to determine the spatial and temporal extents of a given query relation within an untrimmed video. A relation is represented as a three-tuple linguistic phrase $\langle \text{subject}, \text{predicate}, \text{object} \rangle$, where the subject and object are interconnected by the predicate, such as $\langle \text{person}, \text{ride}, \text{bicycle} \rangle$, as illustrated in Fig. 1 (a). VRG is approached as a weakly supervised problem [40], where only the relation phrase is provided during training, while the spatial bounding boxes of the subject and object, and the temporal duration of the relation in the video, are not available. VRG is a crucial task for various multimodal applications, including video captioning [34] and visual question answering [14].

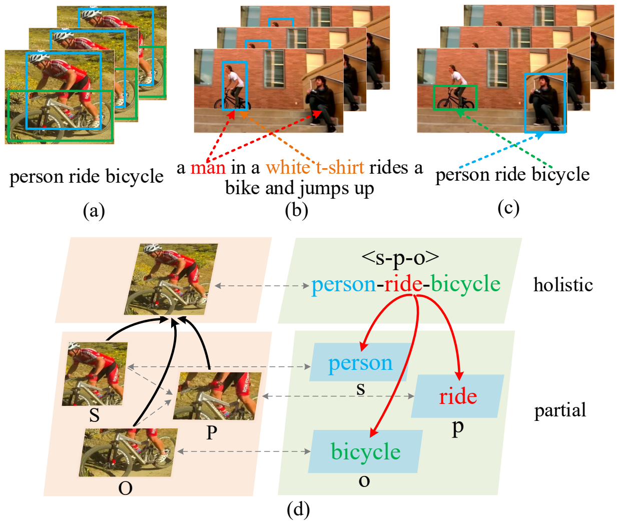


Figure 1. (a) VRG task. (b) An ambiguous result. (c) Video grounding. (d) Holistic-partial structure of a relation.

Existing visual grounding approaches adopt the paradigm of linguistic reconstruction to ground targets [40, 25, 46, 18, 12]. They learn a linguistic description representation to match visual features of the targets and then reconstruct the linguistic description with the matched visual features. This paradigm has demonstrated effectiveness in previous visual grounding tasks involving lengthy and intricate linguistic descriptions [25, 18], as depicted in Fig. 1 (b). The effectiveness of this paradigm can be attributed to the fact that sophisticated linguistic descriptions contain abundant semantic and intricate information, imposing stringent constraints on feature matching and enabling precise target grounding. However, this paradigm proves to be less productive in the VRG, primarily due to the simplicity and sparsity of the information present in a 3-tuple relation phrase. These relation phrases impose weaker constraints on feature matching. As a result, certain visual features that are weakly related or ambiguous, as illustrated in Fig. 1 (c), can still effectively reconstruct the relation phrase. This phenomenon can potentially cause the

[*] Ping Wei is the corresponding author.

model to become 'lazy' in terms of accurately localizing the precise targets, thereby impeding the performance.

Another issue of the current grounding methods is that they treat a relation as a whole and reason about it at a holistic level, without accounting for partial aspects [40, 25, 2, 18]. Indeed, a relation encompasses both holistic attributes and partial attributes, as exemplified in Fig. 1 (d). The relation $person\text{-}ride\text{-}bicycle$ can be perceived as a holistic concept that represents the overall relation. It can also be disassembled into three constituent parts: $person$, $ride$, and $bicycle$, allowing for a description at the partial level. Neglecting the partial-level relation may result in a failure to identify the crucial cues and information embedded within these constituent parts.

In this paper, we rethink video relations from a new perspective and propose a novel inverse compositional learning (ICL) approach for video relation grounding. A relation is represented both at the holistic level and the partial level, acknowledging their distinct characteristics. We formulate VRG as a joint optimization problem that incorporates both holistic-level reasoning and partial-level reasoning. For the holistic-level reasoning, we propose an inverse composition learning method, where both the attention and inverse attention are computed for the subject and object, respectively. The attention encodes the distribution of relevant visual features, while the inverse attention captures the distribution of irrelevant visual features. By paring different attention and inverse attention features of the subject and object, the compositional visual features are generated. We devise an inverse loss function to learn the compositional relevance between the relation and visual features, which encourages the model to extract and emphasize the visual features that are most pertinent to the given relation. To facilitate partial-level reasoning, we decompose the relation phrase into three parts: subject, predicate and object. A grounding by classification scheme is proposed to learn the partial-level feature. By employing this partial-level reasoning approach, we aim to enhance the accuracy of the localization process, enabling a more precise identification and tracking of the subject and object throughout the video.

With the holistic-level and partial-level features, the model is trained in an end to end way. In inference, the grounding results are computed by jointly optimizing the holistic-level reasoning and partial-level reasoning. The proposed method is tested on two challenging datasets: ImageNet-VidVRD [27] and HICO-Det [1]. It outperforms the state-of-the-art methods by a large margin.

2. Related Work

Modeling Visual Relation. Relation detection [19, 41, 43, 21, 17, 35, 6, 10, 31, 9] has attracted growing attention in recent years. By effectively modeling the relations between different objects, the models have the capability to acquire and comprehend fine-grained scene information [38, 39, 13, 20, 32]. Li *et al.* [16] proposed an integration-decomposition network for human-object interaction detection. Visual relation understanding has been extended to the domain of videos [27, 33, 30, 15, 23]. Video relation detection is a task that focuses on detecting relation instances within videos. Recently, there have been several advancements in this field. Shang *et al.* [26] introduced a novel iterative inference method for video relation detection. Gao *et al.* [7] proposed a transformer-based method for relation detection in videos. This approach formulates the relation detection task as a set prediction problem. Instead of focusing on video relation detection, Xiao *et al.* [40] introduced the video relation grounding task, which aims to localize the spatial and temporal positions of query relations within videos. They proposed an attention-based learning structure and utilized phrase reconstruction to accomplish this task. Recently, ARC [12] presents an asymmetrical reasoning pattern for grounding relation instances, which addresses the challenge of grounding relation instances by leveraging an asymmetric reasoning strategy. We propose a novel inverse compositional learning method to address the challenge of sparse relation semantics.

Weakly-Supervised Video Grounding. Temporal video grounding [4, 29, 44, 42, 45] aims to localize the temporal duration of given query sentence in an untrimmed video. Since fine-grained annotating of videos is time-consuming, several weakly-supervised video grounding methods [36, 3, 37] have been proposed. These methods aim to address the challenge of training without access to the temporal labels. For instance, the work [28] introduced a contextual similarity model and visual clustering loss to facilitate feature alignment between two frames. Lin *et al.* [18] formulated a semantic completion network for weakly-supervised video moment retrieval. Zheng *et al.* [46] proposed a novel hard negative sample mining method based on sentence reconstruction. This approach evaluates the temporal interval using a gaussian distribution, enhancing the mining of challenging negative samples. Existing methods primarily focus on learning rich semantic representations from query sentences. However, these methods may not be directly suitable for video relation grounding. Modeling the semantic integrity between the video and the given relation is the key problem for video relation grounding.

3. Method

3.1. Problem Formulation

A relation R is represented as a three-tuple $R = \langle s, p, o \rangle$, where s, p, and o denote *subject*, *predicate*, and *object*, respectively. Following the previous work [40, 12], we represent a video V consisting of N frames as a region proposal set $V = (B_1, ..., B_N)$. Here, $B_i = \{B_{i,j} \mid j = 1, ..., m\}$

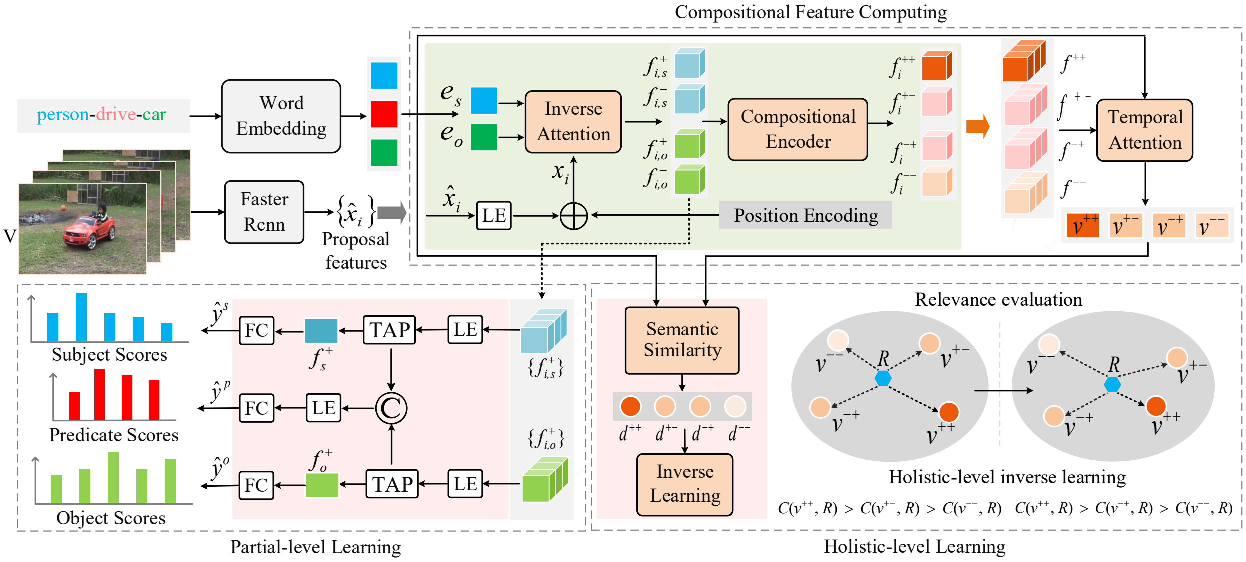


Figure 2. The architecture of the inverse compositional learning model.

denotes the set of region proposals in the i-th frame. $B_{i,j}$ refers to the j-th region proposal in the i-th frame, and m represents the total number of proposals in each frame. In our work, the region proposals are extracted with a pre-trained object detector Faster R-CNN [24].

Given the video V and the query relation R, VRG aims to spatially and temporally localize the subject (s) and object (o) entities with respect to the predicate (p) within the given video. With the proposed bounding box sets, VRG is equivalent to computing two box sequences: $S = (S_k, ..., S_l)$ and $O = (O_k, ..., O_l)$ for s and o, respectively, where $k, l \in [1, N]$ ($k < l$) are the start time and end time of the relation instance, respectively. S_i and O_i ($i \in [k, l]$) are the bounding boxes of s and o in the ith frame, respectively. We formulate the VRG task from two levels: the holistic level and the partial level. At the holistic level, the relation phrase R is taken as a whole to match the visual features. At the partial level, the relation is decomposed into subject (s), predicate (p), and object (o), and each component is individually matched with corresponding visual features. The grounding results are then obtained by jointly optimizing the holistic-level reasoning and the partial-level reasoning:

$$(S^*, O^*) = \arg\max_{S,O} h(S, O, V, R)g(S, O, V, R), \quad (1)$$

where $h(S, O, V, R)$ is the holistic-level reasoning function and $g(S, O, V, R)$ is the partial-level reasoning function.

The **holistic-level** reasoning function is defined as:

$$h(S, O, V, R) = P(S, O|V, R)P(R|S, O, V). \quad (2)$$

$P(S, O|V, R)$ describes the joint probability of the subject box sequence and object box sequence. $P(R|S, O, V)$ characterizes the similarity between the relation phrase R and the visual feature given S, O and V. Eq. (2) follows the previous studies [40, 12]. The **partial-level** reasoning function is defined as:

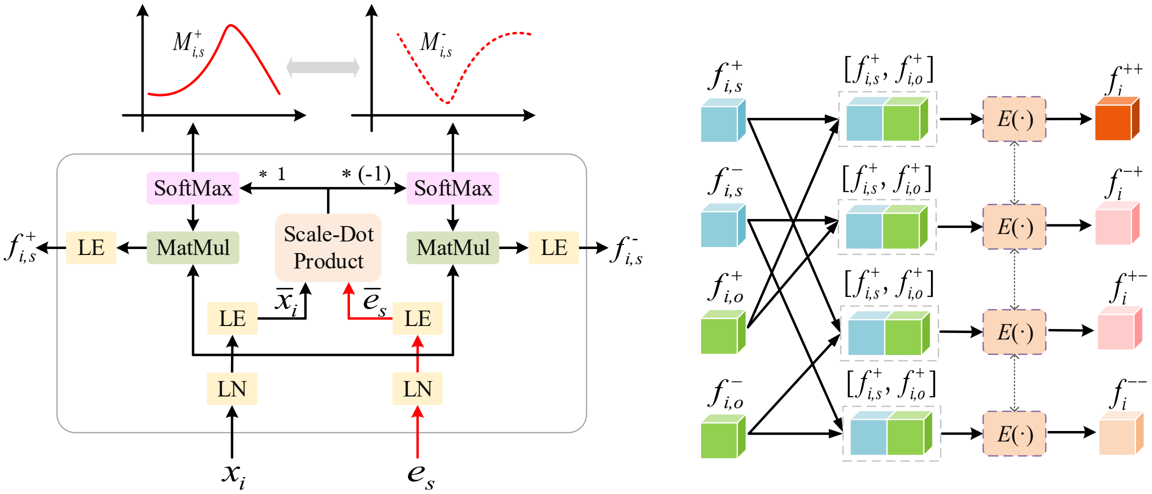
$$g(S, O, V, R) = P(s|S, V)P(o|O, V)P(p|S, O, V). \quad (3)$$

$P(s|S, V)$ represents recognition of subject s given S, V. Similarly, $P(o|O, V)$ represents recognition of object o given O, V. $P(p|S, O, V)$ describes recognition of predicate p. Our framework reasons about a relation by integrating the holistic-level and partial-level information, which not only utilizes the global patterns of the relation but also the inner structures of it.

3.2. Architecture Overview

Fig. 2 shows the overall architecture of the proposed model. Our model contains four major correlative functional components: **feature extraction, compositional feature computing, holistic-level learning**, and **partial-level learning**. The feature extraction component extracts the raw visual features and relation word embeddings from the input video and the given relation, respectively. Given a video V of N frames and a query relation $R = \langle s, p, o \rangle$, a pre-trained Faster R-CNN [24] is used to propose m regions and the region features in each video frame. Suppose $\hat{x}_i$ is the set of the m region features in frame i, and the entire video proposal feature set is represented as $\{\hat{x}_i\}$ ($i \in [1, N]$). The Glove 300 [22] is used to extract the relation word embeddings. The word embeddings of s and o are denoted as e_s and e_o, respectively.

The compositional feature computing component takes the raw visual features and relation word embeddings as inputs and outputs the compositional features by the proposed



(a) Inverse Attention

(b) Compositional Encoder

Figure 3. Illustration of the inverse attention and compositional encoder.

inverse attention and compositional encoder. The holistic-level learning component and the partial-level learning component take the compositional features as inputs and reason about video relations at the holistic and partial level, respectively. The four functional components form a closely-interacting loop. In inference, the grounding results are computed by jointly optimizing the holistic-level reasoning and partial-level reasoning.

3.3. Compositional Computing

As Fig. 2 shows, the compositional computing component takes as inputs the raw video proposal features $\{\hat{x}_i\}$ and the relation word embeddings e_s, e_o. $\hat{x}_i$ is embedded into a $m \times d$ feature which adds the position encoding [40] to generate the feature $x_i \in \mathbb{R}^{m \times d}$ for the frame i. Given the feature x_i and the word embeddings e_s, e_o, an Inverse Attention module is proposed to compute the attention feature $f^+_{i,s}$ and the inverse attention feature $f^-_{i,s}$ for subject s, and $f^+_{i,o}$, $f^-_{i,o}$ for object o. $f^+_{i,s}$, $f^-_{i,s}$, $f^+_{i,o}$, and $f^-_{i,o}$ are grouped in pair and encoded into four fusion features f^{++}_i, f^{+-}_i, f^{-+}_i and f^{--}_i by a Compositional Encoder. Then, the fusion features of all frames are utilized to generate the compositional features v^{++}, v^{+-}, v^{-+}, and v^{--} by a Temporal Attention module for the video.

Inverse Attention. We propose an inverse attention module to compute the attention and inverse attention features, as shown in Fig. 3 (a). The inverse attention module operates on the visual feature x_i and word embedding e_s (or e_o). Let's focus on the subject s to explain the computation in more detail. The feature x_i and e_s first pass through a layer normalization (LN) layer followed by a linear embedding (LE) to generate $\bar{x}_i$ and $\bar{e}_s$, respectively. Then the attention score of the word s with respect to a region is computed as:

$$\alpha^s_{i,j} = \frac{\exp(\beta^j_i)}{\sum_{q=1}^{m} \exp(\beta^q_i)}, \quad \beta_i = \frac{\bar{e}_s(\bar{x}_i)^{\mathrm{T}}}{\sqrt{d_x}}, \tag{4}$$

where $\frac{1}{\sqrt{d_x}}$ is a scaling factor. β^q_i corresponds to the qth region in β_i, i.e., $\beta_i = \{\beta^q_i | q = 1, ..., m\}$. $\alpha^s_{i,j}$ is the normalized relevant score of the word s to the jth region in the ith frame. The word attention distribution $M^+_{i,s} = \{\alpha^s_{i,j} | j = 1, ..., m\}$ is defined to represent the relevance scores of the word s to all m region proposals in the ith frame. Conversely, the inverse attention score of the word s that is irrelevant to a region is computed as:

$$\bar{\alpha}^s_{i,j} = \frac{\exp((-1) * \beta^j_i)}{\sum_{q=1}^{m} \exp((-1) * \beta^q_i)}. \tag{5}$$

The inverse attention distribution is denoted as $M^-_{i,s} = \{\bar{\alpha}^s_{i,j} | j = 1, ..., m\}$. Then, the attention distribution $M^+_{i,s}$ is multiplied by the input x_i and then passed through a linear embedding layer to get the attention feature $f^+_{i,s}$. Similarly, the inverse attention feature $f^-_{i,s}$ is obtained using the same operation. With the same computing, we obtain the attention distribution $M^+_{i,o}$, the inverse attention distribution $M^-_{i,o}$, the attention features $f^+_{i,o}$, and the inverse attention features $f^-_{i,o}$ for object o.

Compositional Encoder. The attention features and inverse attention features, namely $f^+_{i,s}$, $f^+_{i,o}$, $f^-_{i,s}$, and $f^-_{i,o}$, corresponding to the subject and object, are paired together. These paired features are then fed into four encoders, resulting in fusion features f^{++}_i, f^{+-}_i, f^{-+}_i, and f^{--}_i, as shown in Fig. 3 (b). This computing process is summarized as:

$$\begin{aligned} f^{++}_i &= E([f^+_{i,s}, f^+_{i,o}]), \quad f^{+-}_i = E([f^+_{i,s}, f^-_{i,o}]), \\ f^{-+}_i &= E([f^-_{i,s}, f^+_{i,o}]), \quad f^{--}_i = E([f^-_{i,s}, f^-_{i,o}]), \end{aligned} \tag{6}$$

where $[\cdot]$ denotes a concatenation operation. $E(\cdot)$ is the encoder which consists of a fully-connected layer with ReLU activation followed by a dropout layer.

Subsequently, a temporal attention module is applied to the fusion features of all frames. This process generates compositional features v^{++}, v^{+-}, v^{-+}, and v^{--} for the video. The temporal attention is defined as:

$$\tau_i = \frac{\exp(\sigma_i)}{\sum_{n=1}^{N}\exp(\sigma_n)}, \sigma = \frac{e_c(f^{++})^{\mathrm{T}}}{\sqrt{d_f}}, \tag{7}$$

where e_c is the merged word feature. It is obtained by concatenating the word embeddings e_s and e_o, and then applying a linear layer to fuse them. $f^{++} = \left\{f_i^{++}|i = 1, ...N\right\}$ and d_f is the dimension of f_i^{++}. $\{\tau_i|i = 1, ..., N\}$ is the temporal attention distribution. The four types of the compositional features for the video are computed as:

$$\begin{aligned} v^{++} &= \sum_{i=1}^{N}\tau_i f_i^{++}, \quad v^{+-} = \sum_{i=1}^{N}\tau_i f_i^{+-}, \\ v^{-+} &= \sum_{i=1}^{N}\tau_i f_i^{-+}, \quad v^{--} = \sum_{i=1}^{N}\tau_i f_i^{--}. \end{aligned} \tag{8}$$

3.4. Holistic-level Learning

3.4.1 Inverse Compositional Mechanism

As mentioned earlier, existing reconstruction-based learning methods [40, 25] struggle to accurately localize relation instances in videos due to the sparse and simplistic nature of relation phrases. To overcome this limitation, we introduce an inverse compositional mechanism to tackle this problem. As introduced in Sec. 3.3, we compute the attention and inverse attention features for subject and object, respectively. The attention mechanism quantifies the relevance of the subject or object to the regions, while the inverse attention mechanism quantifies their irrelevance. By crosswise pairing the attention features and inverse attention features of the subject and object, we aim to obtain both the 'relevant' and 'irrelevant' visual features related to the relation. Through the competition between these 'relevant' and 'irrelevant' features, the model is driven to extract the most pertinent visual features for the relation. We refer to this mechanism as the inverse compositional mechanism.

As the various combinations of subject features (relevant or irrelevant) and object features (relevant or irrelevant), the four types of compositional features v^{++}, v^{+-}, v^{-+}, and v^{--} (as defined in Eq. (8)) inherently possess different levels of relevance to the query relation R. We evaluate the relevance between each type of compositional feature and the relation R. v^{++}, composed of the relevant attention feature of subject and the relevant attention feature of object, exhibits high relevance with R. v^{+-} derives from the relevant attention feature of subject and the irrelevant attention feature of object, and thus it partly correlates with R. Similarly, v^{-+} also partly correlates with R. Since v^{--} is composed of both the irrelevant attention feature of subject and irrelevant attention feature of object, it has low relevance with R. The relevances between these features and the query relation R can be summarized:

$$\begin{aligned} C(v^{++}, R) > C(v^{+-}, R) > C(v^{--}, R), \\ C(v^{++}, R) > C(v^{-+}, R) > C(v^{--}, R), \end{aligned} \tag{9}$$

where $C(\cdot)$ represents the relevance between the compositional feature and the relation. Based on the relevance evaluation, our objective is to learn a more relevant feature v^{++}.

3.4.2 Inverse Learning Loss

We introduce a novel inverse learning loss to actualize the inverse compositional mechanism and formulate the relevance evaluation defined in Eq. (9). To this end, we begin by evaluating the semantic similarity. Specifically, we first train a transformer-based encoder [5, 12] to learn the query relation feature f^R. In this process, we incorporate an additional learnable class token into the word embeddings of the relation phrase. The relation word embeddings, along with the class token, are passed through multiple encoder layers. The resulting updated class token serves as the query relation feature f^R. Then we compute the semantic similarity between the compositional features and the query relation feature. Specifically, the compositional features v^{++}, v^{+-}, v^{-+}, and v^{--} are respectively used to compute the Euclidean distance with the query relation feature f^R. Correspondingly, the four different distances are denoted as d^{++}, d^{+-}, d^{-+}, and d^{--}, respectively. The distance d^{++} measures the similarity between the query relation and the video feature v^{++}. A smaller value of d^{++} indicates a higher similarity. As shown in Eq. (9), the relevance between v^{++} and R is stronger than the relevance between v^{+-} and R, and the relevance between v^{+-} and R is greater than the relevance between v^{--} and R. Therefore, we define the inverse learning loss with respect to $\mathcal{L}^{+-}$ as follows:

$$\mathcal{L}^{+-} = \delta(d^{++} - d^{+-}) + \delta(d^{+-} - d^{--}), \tag{10}$$

where $\delta(a) = \ln(1+e^a)$ is a monotonically increasing function. The implication behind the inverse learning loss is that, by minimizing $\mathcal{L}^{+-}$, the semantic similarity of the relation with v^{++} is higher than that with v^{+-} and meanwhile the semantic similarity with v^{+-} is better than that with v^{--}. Intuitively, we can directly minimize $\delta(d^{++} - d^{--})$ and ignore the term d^{-+}. In this case, since d^{--} measures the similarity evaluated with the completely irrelevant feature v^{--}, the term of $\delta(d^{++} - d^{--})$ can be easily minimized. Consequently, the learning result would be suboptimal

mal. By introducing d^{-+} and incorporating it into the minimization of $\mathcal{L}^{+-}$, the objective is to improve the learning of semantic similarity with the relevant feature v^{++} compared to the partly relevant feature v^{+-}, and to enhance the learning of semantic similarity with v^{+-} compared to v^{--}. With this adversarial competition, the learned feature v^{++} will be more relevant and robust. Similarly, the inverse learning loss with respect to $\mathcal{L}^{-+}$ is defined as follows:

$$\mathcal{L}^{-+} = \delta(d^{++} - d^{-+}) + \delta(d^{-+} - d^{--}). \quad (11)$$

The total holistic-level inverse learning loss is:

$$\mathcal{L}^{\mathrm{w}} = \mathcal{L}^{+-} + \mathcal{L}^{-+}. \quad (12)$$

3.5. Partial-level Learning

In addition to the holistic-level learning, we propose reasoning about video relations at the partial level, which offers two main benefits. First, by reasoning about the subject and object separately, we can localize them as precisely as possible in each video frame, which reduces the difficulty of the holistic-level learning. Second, reasoning about the predicate solely from the video can learn the visual commonalities of certain actions. For example, actions like 'ride horse' and 'ride bicycle' share some visual similarities in terms of the action of 'ride'. By isolating the predicate and analyzing its visual appearance, we can capture these commonalities and extract the underlying visual cues that contribute to the understanding of similar actions.

For partial-level learning, we employ a grounding by classification scheme, which means that more relevant features result in higher classification scores. As shown in Fig. 2, we assemble the attention feature $f^{+}_{i,s}$ and $f^{+}_{i,o}$ as spatial feature sets $\{f^{+}_{i,s}\}$ and $\{f^{+}_{i,o}\}$ for the subject and object, respectively. These two feature sets are mapped by a linear embedding layer and converted as two classification features f^{+}_{s}, f^{+}_{o} by taking a temporal average pooling. Meanwhile, the two classification features are concatenated and transformed into a unified classification feature for the predicate p. The three classification features, f^{+}_{s}, f^{+}_{o}, and f_p, are fed into three fully connected layers to compute the respective classification scores $\hat{y}^s$, $\hat{y}^o$, and $\hat{y}^p$ for the current given relation. The cross-entropy loss $\mathcal{L}^{\mathrm{ce}}$ is then used to establish the partial-level learning loss, given by:

$$\mathcal{L}^{\mathrm{a}} = \frac{1}{N}\sum_{n=1}^{N}(\mathcal{L}^{\mathrm{ce}}(y^s_n, \hat{y}^s_n) + (\mathcal{L}^{\mathrm{ce}}(y^o_n, \hat{y}^o_n) + (\mathcal{L}^{\mathrm{ce}}(y^p_n, \hat{y}^p_n)). \quad (13)$$

N represents the total number of video frames. y^s_n, y^o_n, and y^p_n are the ground truth labels for the nth frame, where we assume that each frame contains the query relation and is assigned the same label.

The entire model is jointly optimized by integrating both the holistic-level learning and partial-level learning. To achieve this, we introduce a hyper-parameter λ to balance the contributions of these two components. The total loss of our model can be formalized as follows:

$$\mathcal{L} = \mathcal{L}^{\mathrm{a}} + \lambda\mathcal{L}^{\mathrm{w}}. \quad (14)$$

3.6. Inference

During the inference stage, we employ a threshold η to generate temporal candidate segments for each video, leveraging the learned temporal attention distribution $\{\tau_i | i = 1, ..., N\}$. These candidate segments represent potential regions of interest within the video timeline. For each candidate segment, we select the box pair with the maximum score based on the attention distribution $M^{+}_{i,s}$ for the subject and $M^{+}_{i,o}$ for the object, considering each frame of the segment. We average the maximum scores of all frames in each segment as the segment evaluation score. Then the segment with the highest score is selected, and the box sequences S^{*} and O^{*} chosen within this segment are considered as the grounding result.

4. Experiment

4.1. Datasets and Metrics

Following the previous methods [40, 12], we test our model on the ImageNet-VidVRD dataset [27] which consists of more than 30,000 relation instances, 35 object classes and 132 predicate classes. We also conduct experiments on the image based relation dataset HICO-DET [1] which contains 38,118 images for training and 9,658 images for testing. It contains 80 object classes and 117 predicate classes. Our model is evaluated with accuracy (Acc). We define true positives as the temporal intersection over union (tIoU) between the predicted subject box sequence and object box sequence and one of the ground-truth instances, with a tIoU greater than 0.5. Three different spatial intersection over union (sIoU) thresholds are evaluated, respectively. Following vRGV [40], we use the whole relation accuracy (Acc_R) to evaluate the performance of our model. And the separate subject accuracy (Acc_S) and object accuracy (Acc_O) are also reported to analyse models.

4.2. Implementation Details

We sample 120 frames from each video and extract 40 proposals for each frame. The region proposal features are extracted with the pretrained Faster R-CNN [24] with backbone ResNet101 [8]. The region features and word embeddings are transformed into the same dimension of $d = 512$. The batch size is 32. We conduct the experiment using the

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
baseline	42.09	40.53	30.80	37.73	36.28	26.20	30.07	29.00	16.98	35.98	35.08	24.10
w/o L_p^a	40.15	39.37	29.39	36.41	36.17	26.20	26.82	28.48	15.43	33.52	33.30	22.68
w/o L_o^a	42.76	40.87	31.91	39.41	37.41	27.70	29.50	29.36	18.30	36.59	34.94	25.07
w/o L_s^a	41.14	40.91	31.64	37.47	37.83	26.89	29.73	30.46	17.96	35.93	36.61	25.06
w/o $\mathcal{L}^{+-}$	42.63	41.48	31.88	38.04	37.01	27.00	29.73	30.17	17.62	35.45	34.94	24.04
w/o $\mathcal{L}^{-+}$	42.30	41.62	32.20	38.21	37.70	27.94	29.51	30.03	17.76	35.53	35.04	24.95
$\delta(d^{++} - d^{--})$	42.95	41.57	31.53	37.63	36.36	26.42	29.13	29.27	16.78	35.71	34.97	24.49
Our ICL	42.38	40.95	31.85	38.66	37.88	27.51	30.16	30.30	18.46	36.54	36.51	26.05

Table 1. Ablation study experiment results dataset with different spatial overlaps on ImageNet-VidVRD dataset (Acc %).

Models	tIOU=0.3			tIOU=0.5			tIOU=0.7		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
w/o L_p^a	49.21	49.11	35.70	33.52	33.30	22.68	21.51	23.64	12.84
w/o L_o^a	50.93	50.17	36.67	36.59	34.94	25.07	23.96	23.57	14.50
w/o L_s^a	50.42	49.60	37.54	35.93	36.61	25.06	25.78	24.86	15.69
w/o $\mathcal{L}^{+-}$	50.98	50.43	36.92	35.45	34.94	24.04	25.67	24.55	15.78
w/o $\mathcal{L}^{-+}$	50.94	50.24	36.76	35.53	35.04	24.95	26.12	24.67	15.78
$\delta(d^{++} - d^{--})$	50.27	49.56	35.97	35.71	34.97	24.49	25.34	24.15	15.34
Our ICL	51.18	50.25	37.33	36.54	36.51	26.05	25.39	24.66	15.82

Table 2. Ablation study experiment results with different temporal overlaps on ImageNet-VidVRD dataset (Acc %).

Models	sIOU=0.3			sIOU=0.5			sIOU=0.7		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
baseline	42.13	73.26	30.81	30.58	64.18	18.08	22.59	49.24	9.62
w/o L_p^a	72.50	68.44	47.21	67.93	58.11	36.91	57.51	44.06	23.23
w/o L_o^a	59.93	66.75	37.74	52.07	54.50	24.21	42.02	39.60	13.97
w/o L_s^a	67.17	73.27	46.65	61.31	64.05	36.36	51.16	49.39	23.30
w/o $\mathcal{L}^{-+}$	57.88	72.65	39.92	47.53	61.31	25.49	37.76	46.45	14.29
w/o $\mathcal{L}^{+-}$	57.43	74.55	40.82	47.29	63.23	26.40	37.64	48.19	15.16
$\delta(d^{++} - d^{--})$	53.81	70.21	37.45	42.51	57.85	22.54	32.86	43.38	12.16
Our ICL	80.90	73.12	56.75	75.95	63.65	46.09	64.61	48.08	29.14

Table 3. Ablation study experiment results on HICO-DET dataset (Acc %).

Pytorch toolbox with FP16 training and Adam optimizer. λ is set to 2 for ImageNet-VidVRD dataset and 1 for HICO-DET dataset. In inference, η is set to 0.00001. For validating the effectiveness on the image-level relation dataset, we slightly modulate the model to adapt to the image-level grounding. The temporal attention module is removed from the model. We directly use the f_i^{++}, f_i^{+-}, f_i^{-+} and f_i^{--} to conduct the partial-level and holistic-level inverse learning.

4.3. Ablation Analysis

Table 1 shows the ablation study results on ImageNet-VidVRD dataset. We first compare our ICL model with our baseline model that solely focuses on relation reconstructions [40, 12]. Obviously, our ICL outperforms the baseline by a large margin under all experimental settings. This result illustrates the proposed partial-level learning and holistic-level inverse learning are effective.

We investigate the influence of each term in the total loss. We remove each component from the loss function respectively to train the model. 'w/o L_o^a', 'w/o L_s^a', and 'w/o L_p^a' represent removing the object, subject, predicate term, respectively from the partial-level loss. Table 1 shows missing any partial component degrades the performance. Specially, removing L_p^a results in a significant performance degradation. Omitting L_p^a results in the model erroneously localizing subject and object entities that share the same class as s and o but lack the predicate p. This undermines the model's ability to discern and accurately localize relation instances associated with the specific predicate. Although excluding the $\mathcal{L}^{-+}$ term improves results for sIOU thresholds of 0.5 and 0.3, our overall model achieves superior performance in the more challenging sIOU=0.7 and average settings.

We also report experiment results with different temporal overlaps, as shown in Table 2. Our overall model consistently achieves optimal performance under both sIOU=0.5 and sIOU=0.7, as evidenced by the results. These findings

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
T-Rank V1 [2]	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 V2 [2]	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 [11]	27.84	25.62	18.44	23.50	20.40	13.81	17.02	14.93	7.29	22.99	19.33	12.80
vRGV [40]	37.61	37.75	27.54	32.17	32.32	21.43	21.34	21.02	10.62	31.64	30.92	20.54
ARC [12]	41.60	40.61	30.23	37.13	36.78	26.09	28.65	29.41	17.56	34.96	34.72	23.75
Our ICL	**42.38**	**40.95**	**31.85**	**38.66**	**37.88**	**27.51**	**30.16**	**30.30**	**18.46**	**36.54**	**36.51**	**26.05**
Co-occur$^{\#}$ [11]	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$^{\#}$ [40]	42.31	41.31	29.95	37.11	37.52	24.77	29.71	29.72	17.09	36.77	36.30	24.58
ARC$^{\#}$ [12]	**45.66**	**44.01**	32.53	40.99	40.41	27.83	**33.24**	**33.39**	**20.44**	**39.66**	**39.20**	26.42
Our ICL$^{\#}$	44.28	42.82	**33.29**	**41.90**	**40.92**	**30.31**	32.91	33.14	19.87	39.57	38.88	**27.87**

Table 4. Comparison with SOTA methods on ImageNet-VidVRD with different spatial overlap thresholds (Acc %. $^{\#}$ means Viterbi algorithm adopted in inference).

illustrate the effectiveness of our model. As discussed in section 3.4, bypassing d^{-+}, d^{-+} and only optimizing the term of $\delta(d^{++} - d^{--})$ are unfeasible to learn robust visual semantic features. The observed performance degradation only with $\delta(d^{++} - d^{--})$ provides substantial evidence to support this assumption.

Table 3 presents the ablation results on the HICO-DET dataset. Our proposed ICL method outperforms all other approaches across various spatial overlap threshold settings, confirming the effectiveness of its different components. Notably, we observe that Acc_S (subject accuracy) surpasses Acc_O (object accuracy), indicating the relatively easier localization of the subject in the HICO-DET dataset. Furthermore, removing the term of L_o^a results in a significant decrease in performance. This can be attributed to the dataset's human-centered nature, where the subjects of the relations are always 'person'. Consequently, localizing the subject tends to be relatively more straightforward compared to localizing the object involved in the relation. Based on the comprehensive ablation results discussed above, we can conclude that our ICL model demonstrates excellent performance on both video relation dataset and image relation dataset. The model's effectiveness is validated by its superior results across various metrics and settings.

4.4. Comparison with State-of-the-Art

Our model is compared with some SOTA methods, including T-Rank V1[2], Co-occur[11], vRGV[40] and ARC [12]. The results of T-Rank [2] and Co-occur[11] were reported by [40]. Table 4 shows the results on ImageNet-VidVRD dataset with different spatial overlap thresholds. The relation accuracy of our ICL obviously outperforms the SOTA method ARC [12] under all spatial threshold settings. Since some methods employ the Viterbi algorithms during the inference phrase, we initially compare our methods with variations of these approaches that greedily connect regions based on their maximum attention scores in each frame. Specifically, under the average setting, our ICL model demonstrates a significant performance improvement compared to the ARC method. The ARC method achieves an accuracy of 23.75% (Acc_R), whereas our model achieves a higher accuracy of 26.05% (Acc_R).

Additionally, we evaluate the results obtained by employing the Viterbi algorithm during inference (indicated by $^{\#}$). Across different settings, our model consistently outperforms the ARC model by a substantial margin. It achieves 27.87%, 33.29%, and 30.31% (Acc_R) under different evaluation settings, surpassing the ARC model in most cases. An important point to emphasize is that even without employing the Viterbi algorithm, our model consistently outperforms the majority of methods that do utilize the Viterbi algorithm. Our model's performances, even in the absence of Viterbi algorithm, is only slightly worse than the results achieved by the ARC model when it utilizes the Viterbi algorithm. This phenomenon highlights the superiority and the ability of our model to effectively capture temporal continuity. Under the average setting, we have observed a slight advantage of the ARC model over our model in terms of Acc_S and Acc_O. However, our model achieves a higher accuracy for relation (Acc_R). This observation suggests that the ARC method may occasionally localize incorrect subject-object pairs, leading to higher accuracy for individual entities but lower accuracy for the overall relation.

In the comparison of different temporal overlap thresholds (Table 5), our model consistently outperforms other methods. Utilizing Viterbi algorithm in inference also leads to performance improvement. Our model achieves the best results under tIOU=0.3 and tIOU=0.5, outperforming the compared methods in multiple evaluation metrics. These results highlight the effectiveness of our model, particularly in capturing and leveraging temporal dependencies.

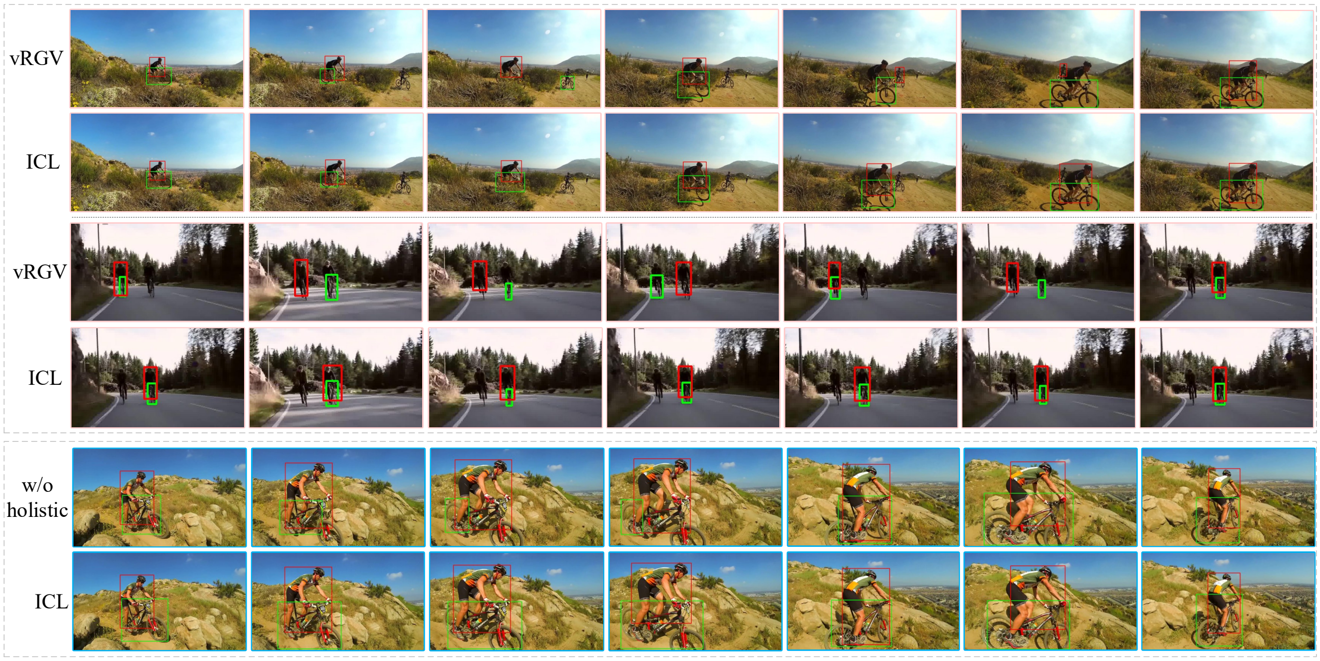


Figure 4. Visualization results comparison between vRGV [40] and our ICL (Top), and comparison between the model without using the holistic-level learning and our ICL (Bottom). Query relation: $person\text{-}ride\text{-}bicycle$.

Models	tIOU=0.3			tIOU=0.5			tIOU=0.7		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
T-Rank V1 [2]	36.51	28.67	15.05	20.27	10.68	3.99	6.15	2.67	0.55
T-Rank V2 [2]	36.99	20.70	12.81	20.83	7.35	3.16	6.19	1.30	0.21
ARC [12]	49.61	49.43	35.68	34.96	34.72	23.75	24.14	25.25	14.46
Our ICL	**51.18**	**50.25**	**37.33**	**36.54**	**36.51**	**26.05**	**25.39**	**24.66**	**15.82**
Co-occur# [11]	35.30	35.50	23.23	25.90	25.23	16.48	16.81	15.04	8.94
vRGV# [40]	49.97	48.98	33.16	36.77	36.30	24.58	24.27	22.11	13.69
ARC# [12]	52.74	52.41	35.61	**39.66**	**39.20**	26.42	**28.68**	**28.68**	**17.67**
Our ICL#	**53.38**	**52.66**	**37.62**	39.57	38.88	**27.87**	28.57	27.37	17.61

Table 5. Comparison with SOTA methods on ImageNet-VidVRD with different temporal overlap thresholds (Acc %).

4.5. Visualization

Fig. 4 shows the grounding results from the vRGV model [40] and our ICL model (Top). Our ICL model demonstrates superior performance compared to the vRGV model. Our model excels in accurately localizing relation instances, even in challenging multi-instance scenarios. vRGV model often generates incorrect matching pairs. Additionally, we compare our ICL model with a variant that does not employ the holistic-level learning (Bottom). The comparison showcases that our ICL model produces more precise bicycle bounding boxes, indicating the effectiveness of holistic-level learning.

5. Conclusion

This paper addresses the challenging task of weakly-supervised relation grounding in videos and images. We propose a novel inverse compositional learning (ICL) model that combines the holistic-level learning and partial-level learning. The partial-level learning adopts a grounding by classification strategy, while the holistic-level learning utilizes an inverse attention module and a compositional encoder, guided by the proposed inverse loss. Extensive experiments validate the effectiveness of our method. Future work will investigate the application of inverse compositional learning in other grounding tasks.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (No. 2018AAA0102501), the Youth Innovation Team of Shaanxi Universities, and the Fundamental Research Funds for the Central Universities.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *ieee winter conference on applications of computer vision (wacv)*, 2018.

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

[3] Cheng Da, Yanhao Zhang, Yun Zheng, Pan Pan, Yinghui Xu, and Chunhong Pan. Asynce: Disentangling false-positives for weakly-supervised video grounding. In *29th ACM International Conference on Multimedia*, 2021.

[4] Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. Support-set based cross-supervision for video grounding. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[6] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, 2020.

[7] Kaifeng Gao, Long Chen, Yifeng Huang, and Jun Xiao. Video relation detection via tracklet based visual transformer. In *29th ACM International Conference on Multimedia*, 2021.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*. Springer, 2020.

[10] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[11] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[12] Huan Li, Ping Wei, Jiapeng Li, Zeyu Ma, Jiahui Shang, and Nanning Zheng. Asymmetric relation consistency reasoning for video relation grounding. In *ECCV 2022*, 2022.

[13] Jiapeng Li, Ping Wei, Yongchi Zhang, and Nanning Zheng. A slow-i-fast-p architecture for compressed video action recognition. In *The 28th ACM International Conference on Multimedia*, 2020.

[14] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *European Conference on Computer Vision (ECCV)*, 2018.

[15] Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. Interventional video relation detection. In *29th ACM International Conference on Multimedia*, 2021.

[16] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 2020.

[17] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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

[19] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016.

[20] Zeyu Ma, Ping Wei, Huan Li, and Nanning Zheng. Hoig: End-to-end human-object interactions grounding with transformers. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022.

[21] Li Mi and Zhenzhong Chen. Hierarchical graph attention network for visual relationship detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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

[23] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *27th ACM International Conference on Multimedia*, 2019.

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

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

[26] Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Video visual relation detection via iterative inference. In *29th ACM International Conference on Multimedia*, 2021.

[27] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM International Conference on Multimedia*, 2017.

[28] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[30] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu. Video visual relation detection via multi-modal feature fusion. In *27th ACM International Conference on Multimedia*, 2019.

[31] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[32] Haowen Tang, Ping Wei, Huan Li, Jiapeng Li, and Nanning Zheng. Relation reasoning for video pedestrian trajectory prediction. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2022.

[33] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[34] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence – video to text. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[35] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[36] Wei Wang, Junyu Gao, and Changsheng Xu. Weakly-supervised video object grounding via stable context learning. In *29th ACM International Conference on Multimedia*, 2021.

[37] Yuechen Wang, Wengang Zhou, and Houqiang Li. Fine-grained semantic alignment network for weakly supervised temporal language grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[38] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for event and object recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[39] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1165–1179, 2017.

[40] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. Visual relation grounding in videos. In *European Conference on Computer Vision*, 2020.

[41] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *IEEE international conference on computer vision*, 2017.

[42] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019.

[43] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. On exploring undetermined relationships for visual relationship detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[44] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. Cascaded prediction network via segment tree for temporal video grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[45] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. 2022.

[46] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.