Few-Shot Referring Relationships in Videos

Yogesh Kumar and Anand Mishra
Indian Institute of Technology Jodhpur
{kumar.204, mishra}@iitj.ac.in
https://vl2g.github.io/projects/refRelations/

Figure 1. The proposed problem setup. (a) Given a query visual relationship as <subject, predicate, object> and a test video, our goal is to localize the subject and object on the test video using a support set containing a few videos sharing the same predicate. In this example, the goal is to spatiotemporally localize the plane (subject), and person (object) that are connected via fly above (predicate), using a support set containing only four videos sharing predicate fly above. It should be noted here that fly above is unseen during training. We refer to this problem as few-shot referring relationship in videos. This problem setup is inspired by the real-world scenario where obtaining large-scale annotations for every visual relationship is practically infeasible. As shown in (b), a popular visual relationship video dataset, namely ImageNet-VidVRD [27], contains many predicates with very few examples, i.e., it has long-tail distribution. Further, as shown in (c), the success of a recent visual relationship localization technique (vRGV) [35] is clearly proportional to predicate distribution in the train set. This calls for solving referring relationship tasks in a few-shot setup. We propose this task and present a novel principled solution.

Abstract

Interpreting visual relationships is a core aspect of comprehensive video understanding. Given a query visual relationship as <subject, predicate, object> and a test video, our objective is to localize the subject and object that are connected via the predicate. Given modern visio-lingual understanding capabilities, solving this problem is achievable, provided that there are large-scale annotated training examples available. However, annotating for every combination of subject, object, and predicate is cumbersome, expensive, and possibly infeasible. Therefore, there is a need for models that can learn to spatially and temporally localize subjects and objects that are connected via an unseen predicate using only a few support set videos sharing the common predicate. We address this challenging problem, referred to as few-shot referring relationships in videos for the first time. To this end, we pose the problem as a minimization of an objective function defined over a T-partite random field. Here, the vertices of the random field correspond to candidate bounding boxes for the subject and object, and T represents the number of frames in the test video. This objective function is composed of frame-level and visual relationship similarity potentials. To learn these potentials, we use a relation network that takes query-conditioned translational relationship embedding as inputs and is meta-trained using support set videos in an episodic manner. Further, the objective function is minimized using
a belief propagation-based message passing on the random field to obtain the spatiotemporal localization or subject and object trajectories. We perform extensive experiments using two public benchmarks, namely ImageNet-VidVRD and VidOR, and compare the proposed approach with competitive baselines to assess its efficacy.

1. Introduction

Consider the following problem: given a video, a visual relationship query represented as a <subject, predicate, object> tuple, and a support set of a few videos containing the same predicate but not necessarily the same subjects and objects, our objective is to spatially and temporally localize both subjects and objects that are related via the predicate within the video. We refer to this problem as Few-shot Referring Relationship and illustrate it in Figure 1. Solving this problem has the potential to benefit cross-task video understanding [41] and video retrieval [5, 7], among other applications. Identifying its utility, referring relationship task for images has been first introduced by [15]. However, referring relationships in videos pose additional video-specific challenges, such as understanding dynamic visual relationships. Some of these challenges have been addressed in recent research by Xiao et al. [35], but with a reliance on strong supervision. Referring relationships in videos within a few-shot setup is an under-explored area. We aim to fill this research gap via our work.

Visual relationships inherently have long-tail distributions in any video collection. For example, ImageNet VidVRD [27] dataset includes approximately 18.9% predicates with more than 100 instances but 20.5% predicates with less than 10 instances. This phenomenon is also shown in Figure 1, where most predicates belong to the tail side of the distribution. The methods that work best for frequent visual relationships do not necessarily generalize well to unseen visual relationships. Moreover, in a real-world scenario annotating visual relationships for each combination of subject, object, and predicate are cumbersome, expensive, and possibly infeasible. Therefore, there is a need to study visual relationship tasks in a few-shot setup. For instance: only with a few examples of the fly above predicate, such as videos containing <bird, fly above, person>, <helicopter, fly above, train> as shown in Figure 1 (a), a model should be able to generalize to the unseen visual relationship, such as <plane, fly above, person>. We propose a solution for Few-shot Referring Relationship in videos in this work.

We pose the problem of a few-shot referring relationship in the video as a minimization of an objective function defined over a T-partite random field where T is the number of frames in the test video. Furthermore, the vertices of the random field are treated as random variables and represent candidate bounding boxes for the subject and objects. The objective function consists of frame-level potentials and visual relationship similarity potentials, both of which are learned using a relation network that takes query-conditioned translational relationship embeddings as inputs. We meta-train the relation network using support set videos in an episodic manner. Further, the objective function is minimized using a belief propagation-based message passing on the random field to obtain subject and object trajectories. We perform extensive experiments on two public benchmarks, namely ImageNet-VidVRD [27] and VidOR [31], and report the accuracy of localizing subject, object, and relation, denoted by $A_{sub}$, $A_{obj}$, and $A_r$, respectively, along with other popular measures used in the literature. Our proposed approach clearly outperforms the related baselines.

The contributions of this work are three folds. (i) We propose a novel problem setup for referring relationship in videos, where the model must learn to localize the subject and object corresponding to a query visual relationship that was unseen during training using only a few support videos. (ii) We propose a new formulation to solve this task based on the minimization of an objective function on T-partite random field where T is the number of frames in the test video, and the vertices of the random field representing potential bounding boxes for subject and objects correspond to the random variables. (Section 3.1). (iii) Additionally, to enrich query-conditioned relational embeddings, we present two aggregation techniques, namely global semantic and local localization aggregations. The use of these aggregation techniques results in enhanced relationship representations, which helps to obtain better trajectories for objects and subjects related via the query visual relationship. This is evidenced by extensive experiments and ablations. (Sections 3.2 and 4.5).

2. Related Work

Visual Relationships: Interpreting visual relationships in images [19, 21, 38] as well as videos [4, 9, 17, 18, 40] have gained huge attention over the last few years. They have also been key components of large-scale popular datasets such as Visual Genome [16] and Action Genome [11]. Visual relationships were studied with respect to object segmentation to leverage spatial relations [8] and to understand human-object interactions [36] via human-centric relationships. Krishna et al. [16] proposed the concept of scene graphs by combining multiple visual relationships in a graph structure. The structured representation of the scene graph is exploited for several tasks, including image retrieval [12]. Shang et al. [27] extended scene graphs from images to videos. Videos have spatiotemporal nature introducing dynamic relations that are not present in images. Scene graphs in videos represent fine-grain information that helps in the downstream task for spatiotemporal reasoning.
Several works have been introduced that utilize visual relationships to improve downstream tasks such as image or video captioning [37], visual grounding [10], visual question answering [20]. To localize subjects and objects connected via a query visual relationship, Krishna et al. [15] proposed referring relationship for images. Their proposed method used an iterative message-passing mechanism between subject and object using language priors to ground the query relations. Further, Xiao et al. [35] extended referring relationship to videos and refer it as visual relationship grounding in videos. In a similar spirit, but, in a few-shot set-up, we propose to spatiotemporally localize subjects and objects in a video for a given relationship.

**Few-Shot Learning:** Few-shot learning in the literature can be grouped into: (i) metric-based [14, 28, 32] and (ii) model-based [6, 23, 29] methods. We limit our discussion to only closely related works. Vinyals et al. [32] proposed matching networks that learn to compare using a small support set. Further, Sung et al. [29] proposed a relation network that learns from a few labeled images of the support set. Given support set videos, and this enabled us to obtain both metric-based and model-based visual relationships, and closer to ground truth. Several works have introduced that utilize visual relationships as edges. More precisely, the set of vertices \( V \) for a video \( v \) contains ordered pair of bounding boxes as follows: \( \{u_{ik}^{jk} = (B_{i}^{k}, B_{j}^{k}) : i \in [1, T]; j, k \in [1, M]; j \neq k\} \), and the set of edges is defined as \( E = \{(u_{i}^{jk}, u_{i+1}^{jk}) : i \in [1, T-1]; j, k \in [1, M]; j \neq k\} \). Each vertex \( u_{ik}^{jk} \) of this graph is a binary random variable that takes one of two labels \{select (1), reject (0)\} for the given query \( r \). Figure 2 illustrates the construction of graph \( G \) for a three-frame test video. It should be noted here that the selected nodes (corresponding subject and object bounding box) from each frame form subject and object trajectories. To obtain optimal subject and object trajectories (\( T_{s}^{*} \) and \( T_{o}^{*} \)) for a given visual relationship \( r \), we solve following optimization problem:

\[
T_{s}^{*}, T_{o}^{*} = \arg \min_{\theta,s,j,k} \sum_{i=1}^{T} \sum_{j,k=1}^{M} \left( \Psi_{i}(u_{ik}^{jk}, r, \theta) \right) + \sum_{t \in u(i)} \Psi_{it}(u_{it}^{jk}, u_{i+1}^{jk}, r, \theta).
\]

Here, \( n(i) \) represents the neighboring nodes to frame \( i \), and \( \theta \) is a learnable parameter of a neural network that needs to be trained using support set videos. Specifically, we use support set videos to meta-train relation network \( (R_{\theta}) \) in a few-shot way. We describe the relation network used in our framework and its training strategy in detail in Section 3.3. Further, in the aforementioned objective function, the terms \( \Psi_{l} \) and \( \Psi_{il} \) denote frame-level and visual relationship similarity potentials that are defined next.

### 3.1.1 Frame-level Potentials (\( \Psi_{l} \))

We compute frame-level potentials \( \Psi_{l}(u_{ik}^{jk}, r, \theta) \) such that the cost of selecting a node to form a trajectory is low if the selected node is semantically similar to query predicate \( r \), otherwise high. Mathematically,

\[
\Psi_{l}(u_{ik}^{jk}, r, \theta) = \frac{-\sum_{m=1}^{K} R_{\theta}(f_{r}(u_{ik}^{jk}), f_{r}(u_{im}^{jk}))}{K},
\]

where, \( u_{im}^{jk} \) is the relationship pair in the \( m^{th} \) support set video connected via query predicate \( r \) and \( f_{r}(u_{ik}^{jk}) \) is a visual relationship embedding in video \( v \) for \( j^{th} \) and \( k^{th} \) bounding boxes of \( i^{th} \) frame. Thus, \( \Psi_{l}(u_{ik}^{jk}, r, \theta) \) gives the negative of average similarity of node \( u_{ik}^{jk} \) with respect to the relationship of \( S_{sup} \) having predicate as \( r \). The \( R_{\theta} \) returns a value closer to 1 when it represents visually and semantically similar visual relationships, and closer to 0, otherwise.
3.2 Query-conditioned Relationship Embedding

We describe the representation of a frame-level visual relationship in this section. To this end, we first obtain $M$ extracted objects for each frame of the video using Faster-RCNN [24]. Then, given a query relationship $r = <s, p, o>$, for each frame, an object or subject representation corresponding to $j^{th}$ and $k^{th}$ bounding box of $i^{th}$ frame respectively are obtained by the following equations:

$$ f_s(B_i^j) = W_{s_2}\left(ReLU\left(W_{s_1}[f_{app}^s(B_j^j); f_{spa}^s(B_j^j); G(s)]\right)\right), $$

$$ f_o(B_i^j) = W_{o_2}\left(ReLU\left(W_{o_1}[f_{app}^o(B_j^j); f_{spa}^o(B_j^j); G(o)]\right)\right). $$

Here, $W_{s_1}$, $W_{s_2}$, $W_{o_1}$ and $W_{o_2}$ are learnable parameters. In addition, the variables $f_{app} \in \mathbb{R}^{2048}$ and $f_{spa} \in \mathbb{R}^4$ represent the ROI appearance and spatial features of the corresponding bounding boxes, while $G(s) \in \mathbb{R}^{300}$ and $G(o) \in \mathbb{R}^{300}$ represent the GloVe word embeddings for the subject and object, respectively. Moreover, $[\cdot;\cdot;\cdot]$ denotes the concatenation operation.

From here onwards, for the sake of simplicity of notations, we represent $f_s(B_i^j)$ and $f_o(B_i^j)$ as $f_s$ and $f_o$ respectively. It should be noted that $f_s$ and $f_o$ can be used directly for obtaining relationship embeddings that can be used to compute frame-level and visual relationship similarity potentials. However, to further enrich these representations, we present the following two attention-based aggregation techniques:

(i) Global Semantic Aggregation (GSA): The subject and object representations, i.e. $f_s$ and $f_o$, learned using eq. (4) and (5) are independent of other frames in the video. However, the global semantic context information from other frames may help in enriching subject and object representation. Therefore, we fused the $I3D$-features [1] of every frame by weighting them with a global attention vector $\alpha_s^g$ which is computed as follows:

$$ \alpha_s^g = G_{Att}(f_{I3D}, G(s)). $$

Where $G_{Att}$ is a learnable attention unit, and it is defined as follows:

$$ s_j^g = W_{gs_1}ReLU(W_{gs_2}[f_{I3D}^j; G(s)] + b_{gs}), $$

$$ \alpha_s^g = \text{softmax}(s_j^g), $$

where, $W_{gs_1}$, $W_{gs_2}$ and $b_{gs}$ are learnable parameters and $f_{I3D}^j$ represent the $I3D$-feature for $j^{th}$ frame. Finally, after aggregating the global semantic information the subject representation is obtained as:

$$ f_s^g = \sum_{j=1}^{T}(\alpha_s^g \odot f_{I3D}^j) \cdot f_s. $$

Similarly, object representation after global semantic aggregation is obtained as:

$$ f_o^g = \sum_{j=1}^{T}(\alpha_s^g \odot f_{I3D}^j) \cdot f_o. $$

3.1.2 Visual Relationship Similarity Potentials ($\Psi_{il}$)

To ensure that coherent (both semantically and visually) visual relationships are being selected across frames, we define visual relationship similarity potential such that the cost of selecting a similar visual relationship in the optimal tracklet is low. To this end, the visual relationship similarity potential $\Psi_{il}(u_i^{jk}, u_i^{j'k'}, \vartheta)$ is computed as follows:

$$ \Psi_{il}(u_i^{jk}, u_i^{j'k'}, \vartheta) = -R\theta \left( f_r(u_i^{jk}), f_r(u_i^{j'k'}) \right), $$

where $f_r(u_i^{jk})$ is a visual relationship embedding in the video $v$ for $j^{th}$ and $k^{th}$ bounding boxes of $i^{th}$ frame. The computation of these relationship embeddings and details of the relation network follows next.
Here, $f_s$ and $f_o$ are initial subject and object features obtained using eq. (4) and (5), respectively.

(ii) Local Localization Aggregation (LLA): For adding the context from the adjacent frames as a local localization context, we considered a window size of five, i.e., for $i^{th}$ frame, we considered $i - 2$ to $i + 2$ frames. The local context helps with partially visible or occluded objects. We fuse the local spatial and ROI features from the adjacent frames weighted by the local attention vector for the subject $s$ ($\alpha_s^t$) which is computed as follows:

$$\alpha_s^t = L_{Att}(f_s^{t-2} \leq s \leq f_s^{t+2}, f_s).$$ (11)

Here, $f_s$ is the initial representation of subjects obtained from eq. (4) and $f_s^{t-2} \leq s \leq f_s^{t+2}$ is obtained as:

$$f_s^{t-2} \leq s \leq f_s^{t+2} = W_{ls2} ReLU(W_{ls1}[(f_s^{t-2} - f_s^{t-1}) \cdot (f_s^{t+1} - f_s^{t+2})]),$$ (12)

where $f_s^{t-2}$ to $f_s^{t+2}$ are stacked ROI and spatial features of all objects from their respective frame numbers. $W_{ls1}$, $W_{ls2}$, $W_{ls3}$ and $W_{ls4}$ are learnable parameters. Then, from eqs. (7) to (10), local localization information-aggregated subject and object features, $f_s^l$ and $f_o^l$, are obtained.

Obtaining Relationship Embedding: To obtain relationship embedding, we first enrich subject and object representations as follows:

$$f_s = W_{sr1} ReLU(W_{sr2}[f_s^g \cdot f_s^l]),$$ (13)

$$f_o = W_{or1} ReLU(W_{or2}[f_o^g \cdot f_o^l]).$$ (14)

Where, $W_{sr1}$, $W_{sr2}$, $W_{or1}$, and $W_{or2}$ are learnable parameters. Finally, for any pair of subject and object, $(j,k)$ of $i^{th}$ frame of a video $v$, the relationship embedding with respect to query visual relationship $r$ is computed as a translation vector in lower dimensional relation space similar to VTrasE [39] as follows:

$$f_r(w_{ij}^{jk}) = W_{r2} ReLU(W_{r1} ReLU([W_{rs}f_j - W_{ro}f_k]; G(p)),$$ (15)

where, $W_{r2}$, $W_{r1}$, $W_{rs}$, and $W_{ro}$ are learnable parameters and $f_j$, $f_k$ are subject and object representation obtained from eq. (13) and (14) respectively.

3.3. Learning Relation Network with Few Examples

The proposed problem formulation has to learn the similarity between object pairs as the visual relationships using a few videos from $S_{sup}$. We have selected the Relation Network ($R_\theta$) [29], a metric-based meta-learning approach, as our method of choice. After learning the similarity measure for the unseen predicate, $R_\theta$ is used to compute both frame-level as well as visual relationship similarity potentials.

For a pair of object pairs $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$, their representation as a visual relationship, $f_{p_1}^r, f_{p_2}^r$ are obtained using eq. (15). Then pairwise similarity score is computed as:

$$R_\theta(f_{p_1}^r, f_{p_2}^r) = W_r^T \Phi(f_{p_1}^r, f_{p_2}^r) + b,$$ (16)

where $W_r$, $b$ are learnable parameters matrix and bias vector respectively. Further, $\Phi$ is computed using the following equation:

$$\Phi(f_{p_1}^r, f_{p_2}^r) = \tanh(W_1([f_{p_1}^r; f_{p_2}^r]) + b_1) \sigma(W_2[f_{p_1}^r; f_{p_2}^r]) + b_2) + ((f_{p_1}^r + f_{p_2}^r)/2),$$ (17)

where $W_1$, $W_2$, and $b_1$, $b_2$ are learnable parameters matrices, and bias vectors, and $\sigma$ and $\tanh$ are sigmoid and hyperbolic tangent activation functions, respectively.

For each support set video, we extract ground truth positive and negative object pairs. Positive object pairs are connected via the query relation predicate $p$, while negative object pairs are picked randomly from a set containing pairs of objects that are not connected via predicate $p$. We used ground-truth spatial and its corresponding ROI features to get the relation embedding using eq. (15). $R_\theta$ learns to return higher similarity between semantically similar pairs of relationships while lower similarity for other pairs. Let us consider a set of positive object pairs $p_i^r = \{(x_{ij}^p, y_{ij}^p)\}_{j=1}^k$ and negative object pairs $p_i^v = \{(x_{ij}^v, y_{ij}^v)\}_{j=1}^k$ extracted from the $S_{sup}$, where $l$ is the number of positive or negative object pairs extracted from each support set video. Finally, the $R_\theta$ is meta-trained on $S_{sup}$ using the following episodic loss.

$$L = \sum_{a=1}^{K} \sum_{b=1}^{l} \sum_{c=1}^{l} \log \left(1 + e^{-R_\theta(f_{p_i}^{+bc}, f_{p_i}^{-bc})}\right),$$ (18)

where $f_{p_i}^{+bc}, f_{p_i}^{-bc}$ are the positive and negative object pairs extracted from each support set video.

3.4. Trajectory Generation

To generate the final trajectories of the subject and object, we used three different optimization approaches. In the first technique, we used only frame-level potential $\Psi$, to select a node with the optimal potential value, which was determined as the minimum value among all the nodes at frame $i$. We skipped a frame if the optimal potential value
is higher than the fixed threshold (= −0.5). As an alternative to this, we also used a greedy solver by selecting a node from frame $t$ using the first technique, and then, we find the next node in the next frame $j$ with optimal $\Psi_{ij}$. We skipped a frame $j$ if the optimal $\Psi_{ij}$ potential value is higher than the fixed threshold (= −0.5). In our full model, we used both frame-level potential and visual relationship similarity potential and used belief propagation using message passing [22] to find the optimal trajectories.

4. Experiments and Results

4.1. Datasets

We have used two video benchmark datasets, namely, VidOR [26, 31] and ImageNet-VidVRD [27] in our experiments. ImageNet-VidVRD contains 1000 videos obtained from the ILVSRC2016-VID dataset [25]. It has 132 predicates and 35 object categories. VidOR is a large-scale dataset containing 10,000 videos obtained from YFCC100M collection. VidOR contains 50 predicates and 80 object categories. For our problem setting, we split both datasets into disjoint sets based on predicates and videos by randomly assigning 35 and 15 predicates to the train and test sets of VidOR, while assigning 88 and 22 predicates to the train and test sets of ImageNet-VidVRD, respectively.

4.2. Performance Measure

In this work, we adopt widely-used evaluation metrics in the referring relationship and video understanding literature, as described in [3, 15, 35]. Specifically, we calculate the subject and object accuracy, which represents the percentage of correct trajectories returned by the model for the entire test set. A relation accuracy is defined as the percentage of correct subject and object pair trajectories. Here, a pair is considered correct if both subject and object trajectories are correct. For spatiotemporal accuracy, $A_{s-t}^{sub}$, $A_{s-t}^{obj}$, of the subject or object, a trajectory is considered correct if at least 50% of the bounding boxes across frames have $\geq 0.5$ intersection over union (IoU) with respect to ground truth bounding boxes. Similarly, for spatiotemporal accuracy of relation, $A_{s-t}^{obj}$, a pair of subject and object is considered true only if both trajectories are spatiotemporally accurate (i.e., at least 50% of the bounding boxes across frames have $\geq 0.5$ IoU). For spatial accuracy, $A_{s-t}^{obj}$, a trajectory is considered true if the average of IoUs of the bounding box with respect to ground truth is at least 0.5. The mean IoU for a trajectory is an average of the IoU score for all bounding boxes with respect to ground truth.

<table>
<thead>
<tr>
<th>Method</th>
<th>$A_{s-t}^{subj}$</th>
<th>$A_{s-t}^{obj}$</th>
<th>$mIoU_{s-t}^{subj}$</th>
<th>$A_{s-t}^{obj}$</th>
<th>$mIoU_{s-t}^{obj}$</th>
<th>$A_{s-t}^{rel}$</th>
<th>$mIoU_{s-t}^{rel}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VRC [30]</td>
<td>8.4</td>
<td>10.7</td>
<td>11.9</td>
<td>8.6</td>
<td>10.3</td>
<td>11.2</td>
<td>5.8</td>
</tr>
<tr>
<td>vRGV-fs [35]</td>
<td>8.2</td>
<td>11.1</td>
<td>12.4</td>
<td>8.3</td>
<td>11.6</td>
<td>12.9</td>
<td>6.1</td>
</tr>
<tr>
<td>Tracklet-based</td>
<td>12.6</td>
<td>15.1</td>
<td>15.9</td>
<td>12.8</td>
<td>14.8</td>
<td>15.3</td>
<td>9.6</td>
</tr>
<tr>
<td>Ours w/o vis.sim. potential, w/o GSA, w/o LLA</td>
<td>19.6</td>
<td>21.4</td>
<td>20.1</td>
<td>18.2</td>
<td>18.7</td>
<td>20.8</td>
<td>16.3</td>
</tr>
<tr>
<td>w/o GSA, w/o LLA</td>
<td>25.3</td>
<td>26.9</td>
<td>25.7</td>
<td>25.1</td>
<td>26.5</td>
<td>22.8</td>
<td>24.5</td>
</tr>
<tr>
<td>w/o vis.sim. potential</td>
<td>22.6</td>
<td>22.8</td>
<td>21.9</td>
<td>22.3</td>
<td>21.4</td>
<td>22.9</td>
<td>17.2</td>
</tr>
<tr>
<td>Greedy solver</td>
<td>25.4</td>
<td>26.5</td>
<td>24.7</td>
<td>23.4</td>
<td>23.8</td>
<td>26.4</td>
<td>22.8</td>
</tr>
<tr>
<td>Full model</td>
<td>26.8</td>
<td>28.4</td>
<td>27.1</td>
<td>26.0</td>
<td>25.1</td>
<td>27.9</td>
<td>25.1</td>
</tr>
</tbody>
</table>

Table 1. Performance of few-shot referring relationship in videos. Each method is trained on the same split of train and test datasets with a support size of four videos. Ours (full model) represents the proposed method where global semantic aggregation (GSA) and local localization aggregation (LLA) are performed to enrich the query-conditioned relationship representation, a translational relation embedding is learned, and an objective function containing frame-level and visual similarity potential is optimized using belief propagation on the $T$-partite random field. For more details, refer Section 4.

4.3. Baselines

To the best of our knowledge, we are the first to approach referring relationships in videos in a few-shot setup, and there are no existing baselines in the literature that can be directly compared to our approach. Therefore, we adopt closely related methods as our baselines to compare the effectiveness of our method as follows: (i) Few-Shot Visual Relation Co-Localization (VRC): Few-shot VRC has originally been proposed for localizing common subjects and objects in a bag of images in a few-shot setup [30]. We adopt it for videos by treating object trajectory pairs in the test and support set videos as a bag and performing visual relationship co-localization at the frame level. (ii) Visual vRGV [35] tackles the same task as ours but differs significantly in supervision. We adapted vRGV for our few-shot problem setup for a fair comparison. Recall that in our problem setup, the model is trained in episodes, and each episode contains a support set of a few videos and a test video for evaluation. We utilize all available videos from the train set to train the vRGV model. During testing for each episode, we fine-tune the model on the support set and then perform localization for the given query visual relationship on the test video. We refer to these baselines as vRGV in a few-shot setup or vRGV-fs. (iii) Tracklet-based
Figure 3. Qualitative result on a selection of test videos. Each subject and object of the query relationship is spatiotemporally localized on the test videos. Red and blue color bounding boxes represent the subject and object, respectively. [Best viewed in color].

Table 2. Ablation study to demonstrate the importance of global semantic aggregation (GSA) and local localization aggregation (LLA) on the ImageNet-VidVRD dataset. Our full model (the last row) performs better as compared to the settings where one or both of these are removed.

<table>
<thead>
<tr>
<th>GSA</th>
<th>LLA</th>
<th>$A_{A-t}^{subj}$</th>
<th>$A_{A-t}^{obj}$</th>
<th>$A_{A-t}^r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>21.3</td>
<td>22.1</td>
<td>20.8</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>24.7</td>
<td>23.5</td>
<td>22.8</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>22.3</td>
<td>21.7</td>
<td>21.1</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>26.8</td>
<td>26.0</td>
<td>25.1</td>
</tr>
</tbody>
</table>

4.4. Implementation Details

For frame-level detection, we utilized FasterRCNN [24], which was pre-trained on MS-COCO with ResNet-101 as the backbone. We extracted the 30 most confident object bounding boxes for each frame. Our implementation was done using PyTorch. We optimized the model parameters using Adam [13] with an initial learning rate of 1e-5, while also employing a dropout rate of 0.3 to reduce overfitting. To prevent overfitting, we used early stopping. We trained the model on an Nvidia-RTX A6000 GPU.

4.5. Results and Discussions

We compare our approach with baselines discussed in Section 4.3 on VidOR and ImageNet-VidVRD datasets in Table 1. We observe that the baselines modified using prior works, namely VRC and vRGV, exhibit weaker performance across all the evaluated metrics. The tracklet-based approach performs marginally better than these baselines. However, its performance is heavily bottlenecked with the quality of tracklets generated. Our approach, without the proposed global semantic and local localization aggregation itself, surpasses these baselines. We also observed a positive impact of frame-level and visual relationship similarity potential; for example, in Table 1, the accuracy of localizing relationship ($A_{A-t}^r$) has increased by 2.3% (absolute scale) in the full model with respect to the greedy solver on ImageNet-VidVRD dataset. The superior performance of our method can be attributed to the principled optimization framework, the proposed aggregation techniques, query-conditioned learning of relationship embeddings, and the effective utilization of support set videos to meta-train relation networks. To show the efficacy of our method and to justify the choice of different modules, we conduct the following ablations and analyses:

Global semantic and local localization aggregation: We remove global semantic (GSA) and local localization ag-...
In order to analyze the performance of our approach for different types of relations, we present the predicate-wise results in Table 3. The results reveal that our proposed approach generally achieves better performance for static relationships. Nonetheless, despite the additional challenges posed by dynamic relationships, such as the presence of moving objects and subjects, occlusion, motion blur, and deformation, our approach demonstrates reasonable results for dynamic predicates highlighting its robustness.

**Qualitative Results:** Figure 3 shows a selection of qualitative results of our approach on ImageNet-VidVRD dataset where query visual relationships, \(<\text{bird}, \text{fly with}, \text{bird}>\), and \(<\text{person}, \text{fall off}, \text{bicycle}>\) are referred in the respective video by localizing the subject and object connected to the query predicate. Our method is successful in localizing subjects and objects connected via dynamic relationships indicating the effectiveness of our approach. We provide more qualitative results on our project website\(^1\).

### 5. Conclusions

In this work, we approached the problem from frame-level object detection to video-level trajectory generation by optimizing an objective function on a random field in a few-shot way. We used global semantic and local localization aggregation to enhance query-conditioned translation visual relationship embedding. The objective function is minimized using belief propagation on the random field. We performed experimental comparisons along with ablation studies to show the efficacy of our approach. We firmly believe that our work will open up several future research towards the larger exciting goal of comprehensive cross-task video understanding.

**Acknowledgment:** This work is partly supported by a gift grant from Accenture Labs (project number: S/ACT/AM/20220078). Y. Kumar is supported by a UGC fellowship.

---

\(^1\)https://vl2g.github.io/projects/refRelations/
References


[26] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In ICMR, 2019. 6


[34] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In Proc. ICIP, 2017. 7


