Video Visual Relation Detection

Xindi Shang
School of Computing, National University of Singapore, Singapore
shangxin@comp.nus.edu.sg

Tongwei Ren*
State Key Laboratory for Novel Software Technology, Nanjing University, China
rent@nju.edu.cn

Jingfan Guo
State Key Laboratory for Novel Software Technology, Nanjing University, China
guoj@smail.nju.edu.cn

Hanwang Zhang
Department of Computer Science, Columbia University, USA
hanwangzhang@gmail.com

Tat-Seng Chua
School of Computing, National University of Singapore, Singapore
chuats@comp.nus.edu.sg

ABSTRACT
As a bridge to connect vision and language, visual relations between objects in the form of relation triplet (subject, predicate, object), such as “person-touch-dog” and “cat-above-sofa”, provide a more comprehensive visual content understanding beyond objects. In this paper, we propose a novel vision task named Video Visual Relation Detection (VidVRD) to perform visual relation detection in videos instead of still images (ImgVRD). As compared to still images, videos provide a more natural set of features for detecting visual relations, such as the dynamic relations like “A-follow-B” and “A-towards-B”, and temporally changing relations like “A-chase-B” followed by “A-hold-B”. However, VidVRD is technically more challenging than ImgVRD due to the difficulties in accurate object tracking and diverse relation appearances in video domain. To this end, we propose a VidVRD method, which consists of object tracklet proposal, short-term relation prediction and greedy relational association. Moreover, we contribute the first dataset for VidVRD evaluation, which contains 1,000 videos with manually labeled visual relations, to validate our proposed method. On this dataset, our method achieves the best performance in comparison with the state-of-the-art baselines.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence; Computer vision;

KEYWORDS
Visual relation detection; video visual relation; relational association; visual relation tagging

1 INTRODUCTION
Bridging the gap between vision and language is essential in multimedia analysis, which has attracted a lot of research efforts ranging from visual concept annotations [3, 24], semantic description with captioning [7], and visual question_answering [1], etc. Visual relation detection (VRD), a recent effort in offering more comprehensive understanding of visual content beyond objects, aims to capture the various interactions between objects [22]. It may effectively underpin numerous visual-language tasks, such as captioning [15, 34], visual search [2, 8], and visual question_answering [1, 21].

Visual relation involves a pair of objects localized by bounding boxes together with a predicate to connect them. Figure 1(a) shows several examples of visual relations, in which two objects can be connected with various predicates and the same predicate can connect different object pairs with different appearances. In this paper, we use the term relation triplet to denote a type of visual relation represented by a unique combination of (subject, predicate, object) triplet. Due to the combinatorial complexity, the possible space for relation triplets is much larger than that of objects. Because of this, existing methods that could obtain significant performance in object detection, are not applicable to VRD. Several methods have been proposed for VRD [17, 22, 42]. However, to the best of our knowledge, they all applied to still images only. Compared to still images, videos provide a more natural set of features for detecting visual relations, such as the dynamic interactions between objects. As shown in Figure 1(b), motion features extracted from spatial-temporal content in videos help to disambiguate similar predicates,
merges the detected visual relation instances in adjacent segments if they have the identical relation triplets and their object tracklets have sufficiently high overlaps. Third, VidVRD needs to predict more types of visual relations than ImgVRD because some visual relations can only be detected in videos, such as “A-towards-B” and “A-faster-than-B”. For effective relation prediction, we propose a relation prediction model, which extracts multiple features from the subject/object tracklet pairs. The features include appearance, motion, and relative characteristics. We encode these features into a relation feature, and predict visual relations using separate subject, predicate and object predictors.

As far as we know, there is no dataset for VidVRD, although several datasets for ImgVRD exist, such as Visual Relationship dataset [22] and Visual Genome [12]. Hence, we construct a VidVRD dataset for evaluation. We design a predicate description mechanism and construct the dataset from ILSVRC2016-VID [31]. It contains 1,000 videos with manually labeled visual relations and object bounding box trajectories. On the dataset, we validate the performance of our proposed VidVRD method. The experimental results show that our method outperforms the state-of-the-art baselines.

The main contributions of this paper include: 1) we propose a novel VidVRD task that aims to explore various relationships between objects in videos, which provides a more feasible VRD task as compared to ImgVRD; 2) we propose a VidVRD method which detects the visual relations in videos through object tracklet proposal, relation prediction and greedy relational association; and 3) we contribute the first VidVRD evaluation dataset, consisting of 1,000 videos with manually labeled visual relations.

The rest of the paper is organized as follows. In Section 2, we survey the related works, including visual relation detection, video object detection, and action recognition. In Section 3, we introduce the first dataset for the VidVRD task. Then, we present the details of the proposed methods in Section 4, and show the constructed evaluation benchmark and some preliminary experimental results in Section 5. Finally, we conclude the paper in Section 6.

2 RELATED WORK

Video object detection. Video object detection aims to detect objects belonging to the pre-defined categories and localize them with bounding box trajectories in a given video [11]. The state-of-the-art methods address this problem by integrating the latest techniques in both image object detection [28] and multi-object tracking [23, 41]. Recent sophisticated deep neural networks have achieved mature performances in image object detection [9, 18, 30, 39]. However, object detection in videos still suffers from low accuracy, because of the existence of blur, camera motion and occlusion in videos, which hamper accurate object localization with bounding box trajectories. On the other hand, multi-object tracking with tracking-by-detection strategy tends to generate short trajectories due to the high miss detection rate of object detectors, and thus additional merging algorithms are developed to obtain more temporally consistent object trajectories [4, 14, 25]. Inspired by [27, 33], our proposed method utilizes video object detectors to generate object tracklet proposals in short-term duration, which dodges their common weaknesses. Note that our approach can be
applied on top of any image object detection and multiple object tracking methods.

**Visual relation detection.** Recent research works have focused efforts on VRD in images. It has been commonly observed that a fundamental challenge in VRD lies on how to model and predict the huge number of relations by learning from few training examples. To tackle the problem, most existing methods separately predict the subject, predicate and object in the visual relation triplet \([5,16,17,22,42,43]\), reducing the complexity from \(O(N^2K)\) to \(O(N + K)\), where \(N\) and \(K\) are the numbers of objects and predicates respectively. Some of these methods further improve the performance by leveraging language prior \([17,22]\) and regularizing relation embedding space \([42,43]\). Extracting relation related features is another crux of VRD. \([5,42]\) particularly used coordinate or binary mask based features to enhance the performance of detecting spatial relation. \([5,16]\) also studied the visual feature level connection among the components of relation triplet to exploit additional statistical dependency, but required \(O(NK)\) parameters for the modeling. Hence, in order to address these problems of VidVRD, we will propose a video specific relation feature and a new training criterion for learning separate prediction models. It should be noted that the existing ImgVRD methods are unable to tackle the specific challenges of VidVRD, such as the dynamic relations and the changeability of video relations. To the best of our knowledge, our work is the first attempt to perform VRD on video. Note that although some previous works \([29,44]\) are related to video visual relations, they pursue completely different goals to VidVRD.

**Action recognition.** As action is one primary type of predicate in visual relation \([22]\), VidVRD can draw on the advances in action recognition. In action recognition, feature representation plays a crucial role in handling large intra-class variation, background clutter, and camera motion \([20,26,40]\). Both hand-crafted features \([10,37]\) and deep neural networks \([35,38]\) are developed

Figure 3: An overview of our VidVRD method. A given video is first decomposed into a set of overlapping segments, and the object tracklet proposals are generated on each segment. Next, short-term relations are predicted for each object pair on all the segments based on feature extraction and relation modeling. Finally, video visual relations are generated through greedily associating the short-term relations.

3 DATASET

We construct the first evaluation dataset for VidVRD based on the training set and validation set of ILSVRC2016-VID \([31]\), which contains videos with the manually labeled bounding boxes for 30 categories of objects. After carefully viewing and analyzing the contents of videos, we selected 1,000 videos which contain clear and plentiful visual relations, while the videos with single object and ambiguous visual relations were ignored. We randomly split the video set into the training set and test set, which contain 800 videos and 200 videos, respectively.

Based on the 1,000 videos, we supplement the 30 object categories with additional five object categories that frequently appearing in visual relations, namely person, ball, sofa, skateboard and frisbee. All the resulting 35 object categories \(^1\) describe independent object, that is, we do not include the part-of relationship between objects, such as “bicycle-with-wheel”, in the constructed dataset.

Next, we build the set of predicate categories as follows: we directly use transitive verbs as predicate, such as “ride”; we transfer the adjectives to predicates in the format of comparative, such as “faster”; and we manually define common spatial predicates from camera viewpoints to ensure consistency, such as “above”. While

\(^1\)Object: airplane, antelope, ball, bear, bicycle, bird, bus, car, cat, cattle, dog, elephant, fox, frisbee, giant panda, hamster, horse, lion, lizard, monkey, motorcycle, person, rabbit, red panda, sheep, skateboard, snake, sofa, squirrel, tiger, tram, turtle, watercraft, whale, zebra.
intransitive verbs usually describes the attributes of objects only, they are expressible in relation representation. For example, “walk behind” provides more information than “behind” in visual relation. Thus, we also include the combination of intransitive verbs and spatial predicates, as well as the combination of an intransitive verb and “with”, which represents two objects acting in the same manner.

We exclude prepositions in the predicate definition, because the prepositions of spatial kind can be covered by the defined spatial predicates, while the remaining types of prepositions are mainly related to part-of relationship, which has already been excluded according to the object definition. According to the above predicate definition mechanism and video content, we selected 14 transitive verbs, 3 comparatives, 11 spatial descriptors, and 11 intransitive verbs, which is able to derive 160 categories of predicates. In the constructed dataset, 132 predicate categories appear in the videos. The number is more than that in previous works [16, 17, 22, 42].

Eight volunteers contributed to video labeling, and another two volunteers took charge of labeling checking. In object labeling phase, the objects belonging to the additional five categories in all the videos were manually labeled with their categories and bounding box trajectories. In predicate labeling phase, in order to consider the fact that visual relations are temporally changeable, all videos were decomposed into segments of 30 frames with 15 overlapping frames in advance. Then, all the predicates appearing in each segment were required to be labeled to obtain segment-level visual relation instances. To save labeling labor, we only labeled typical segments in the training set and all the segments in the test set. For the test set, the visual relation instances in adjacent segments with the same object pairs and predicate were automatically linked to generate the video-level visual relation instances.

Table 1 shows the statistics of the constructed VidVRD dataset. Overall, our dataset contains a total of 3,219 relation triplets (i.e. the number of visual relation types), and the test set has 258 relation triplets that never appear in the training set. At the instance level, the test set contains 4,835 visual relation instances, among which 432 instances are unseen in the training set. Note that although the videos in the test set are fully labeled, there is still a small portion of content without any visual relation because some parts of these videos contain less than two objects. From the segment-level statistics available in the lower part of Table 1, the numbers of visual relation instances per segment in our dataset is 9.5, which is higher than 7.6 instances per image in the Visual Relationship dataset [22], suggesting that our dataset is more completely labeled.

<table>
<thead>
<tr>
<th>Video</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video</td>
<td>800</td>
<td>200</td>
</tr>
<tr>
<td>Subject/Object Category</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Predicate Category</td>
<td>132</td>
<td>132</td>
</tr>
<tr>
<td>Relation Triplet</td>
<td>2,961</td>
<td>1,011</td>
</tr>
<tr>
<td>Visual Relation Instance (Video-level)</td>
<td>-</td>
<td>4,835</td>
</tr>
<tr>
<td>Segment</td>
<td>15,146</td>
<td>3,202</td>
</tr>
<tr>
<td>Labeled Segment</td>
<td>3,033</td>
<td>2,801</td>
</tr>
<tr>
<td>Visual Relation Instance (Segment-level)</td>
<td>25,917</td>
<td>29,714</td>
</tr>
</tbody>
</table>

always be recognized in a short duration, while more complicated relations can be inferred from the sequence of basic visual relations. Detecting visual relations in short-term can also help to detect the emergence and disappearance of the visual relations in a video, and alleviate the computational burden of directly analyzing a long-term duration. The following sub-sections introduce the details of our method.

4.1 Object Tracklet Proposal

Given a video, we decompose it into segments of length \( L \) with \( L/2 \) overlapping frames (e.g. \( L = 30 \)), and generate object tracklet proposals in each segment. Comparing to generating the proposals for the object trajectories in a whole video and then doing the segmentation, our object tracklet proposal in short-term can reduce the drifting problem commonly observed in object tracking algorithms, where the drifting is caused by variations in illumination and occlusion, etc. Also, individual object tracklet proposal in each segment can generate a more diverse set of candidates. The diversity is important for the subsequent relation modeling, because it provides various appearance and motion aspects of objects for robust modeling.

Our object tracklet proposal is implemented based on a video object detection method similar to [11] on each segment. First, we employ an object detector for 35 categories in our dataset to detect objects in the segment frames. The object detector is trained using a Faster-RCNN [30] with ResNet101 [9] on an image set consisting of the train/validation images for the 35 categories from MS-COCO [19] and ILSVRC2016-DET [31] datasets. Second, we track the frame-level detection results across the segment using the efficient implementation of [6] in Dlib. To reduce the number of overlapping proposals, we perform non-maximum suppression (NMS) with vIoU > 0.5 on the generated tracklets, where vIoU denotes the voluminal intersection over union of two tracklets. As a result, we generate 19.7 object tracklet proposals per segment on average.

4.2 Relation Prediction

Suppose \((T_s, T_o)\) are a pair of object tracklet proposals in a segment, each of which is in the form of a sequence of bounding boxes. Predicting the relation triplet \((subject, predicate, object)\) involves recognizing the object categories of \(T_s\) and \(T_o\), and the interactions between them. In practice, it is impossible to learn a separate
can be represented by their linear combination. For a pair of \( T_s \) and \( T_o \), the proposed relativity feature is the concatenation of the three sparse representations with respect to the corresponding codebooks.

The overall relation feature vector for a pair of object tracklet proposals is the concatenation of the object features of \( T_s \) and \( T_o \) and their relativity feature.

**Relation Modeling.** Given a relation feature, our relation model predicts the likely relation triplets by integrating the scores of subject, predicate and object predictors. One approach to our relation modeling is to train the predictors under separate training criteria as in [42]. However, the predictors trained in this way will produce different types of scores under independent scales, which makes the integrated score less discriminative to the co-occurrence of subjects, predicates and objects. For example, the scores of impossible relation triplets, such as “cat-drive-car”, may not be guaranteed to be lower than those of other possible relation triplets.

In order to produce good ranked scores for relation triplets, we jointly train the predictors under a unified training loss. In particular, we integrate the scores by multiplication, and formulate the training objective to classify among the observed relation triplets \( \mathcal{R} \) in the training data:

\[
L = \sum_{(s, p, o) \in \mathcal{R}} -\log \text{softmax}_s \left( P^s(f, s) \cdot P^p(f, p) \cdot P^o(f, o) \right),
\]

(2)

where \( f \) is the relation feature of a specific relation triplet \( (s, p, o) \), and \( P^s, P^p, P^o \) are respectively the predictors for subject, predicate and object. Since we are only interested in the top relation prediction scores, we use softmax loss which has recently been proved to be effective in this case, both theoretically and empirically [13, 42].

In this paper, we keep the top 20 prediction results for each pair \( (T_s, T_o) \), and the top 200 ones for each segment.

To obtain the training samples, we sample pairs of object tracklet proposals that overlap with a ground truth pair, where each tracklet of a pair overlaps with the ground truth by more than 0.5 in vIoU, and extract the relation feature for each pair.

### 4.3 Greedy Relational Association

After obtaining the relation prediction results for all the pairs of object tracklet proposals, we adopt a relational association algorithm to merge the relations detected in short-term. Supposing there is a sequence of short-term visual relation instances \(((c^t, (s, p, o), (T^t_s, T^t_o)))_t \quad (t = m, \ldots, n)\) detected from the \( m \)-th segment to the \( n \)-th segment, which have identical relation triplet \((s, p, o)\) and with sufficient overlapping between successive ones, our goal is to merge them into a single visual relation instance \((\hat{c}, (s, p, o), (T_s, T_o))\) with confidence score:

\[
\hat{c} = \frac{1}{n - m + 1} \sum_{t=m}^{n} c^t,
\]

(3)

where \( c^t \) is the short-term score predicted by our relation model.

We propose a greedy algorithm for relational association, which repeatedly merges two most confident visual relation instances that overlap in two successive segments. The greedy strategy can help to generate longer visual relation instances, so that the subject
Algorithm 1 Greedy Relational Association Algorithm

Input: the set of all detected short-term relation instances \( S = \{ (c, (s, p, o), (T_s, T_o)) \} \)

Output: the set of merged instances \( L = \{ (c, (s, p, o), (T_s, T_o)) \} \)

Initialize: \( L = \emptyset \), \( \gamma = 0.5 \)

for \( t = 1, \ldots, T \) do

\( \mathcal{A} \) = instances in \( L \) that end at the \((t-1)\)-th segment

\( \mathcal{B} \) = instances in \( S \) that detected at the \( t \)-th segment

Descending sort \( \mathcal{A} \) accordint to \( c \)

Descending sort \( \mathcal{B} \) accordint to \( c \)

for \( (c, (s, p, o), (T_s, T_o)) \) in \( \mathcal{B} \) do

for \((c', (s', p', o'), (T'_s, T'_o))\) in \( \mathcal{A} \) do

if \((s, p, o) = (s', p', o')\) AND \( \text{vIoU}(T_s, T'_s) > \gamma \) AND \( \text{vIoU}(T_o, T'_o) > \gamma \) then

Recompute \( c' \) using Eq. (3)

Append \((T_s, T_o)\) to \((T'_s, T'_o)\)

Remove \((c', (s', p', o'), (T'_s, T'_o))\) from \( \mathcal{A} \)

Break

end if

end for

if NOT merged then

Add \((c, (s, p, o), (T_s, T_o))\) to \( L \)

end if

end for

end for

and object of each visual relation are temporally localized more accurately. We also average the bounding boxes in the overlapping region of two associated tracklets to get a robust estimation of the \((T_s, T_o)\). The pseudocodes for the relational association are given in Algorithm 1. After merging all possible visual relation instances, we rank them according to their confidence scores \( \hat{c} \) and output the top three instances as the visual relation detection results for the video.

5 EXPERIMENTS

5.1 Tasks and Evaluation Metrics

Tasks. As defined in Section 1, the input of VidVRD is a given video, and its output is a set of visual relations with localized objects. Similar to ImgVRD [22], a detected visual relation instance is treated as correct in VidVRD, if it contains the same relation triplet as in the ground truth and both the bounding box trajectories of its subject and object have sufficiently high vIoU as compared to those in the ground truth. In our experiments, the overlapping threshold of vIoU is set to 0.5.

Considering that object localization in videos is still an open problem, we also evaluate our method under a different task, named visual relation tagging. Its input is also a given video, but its output is a set of visual relation triplets annotated to the whole video without the requirement of object localization. Obviously, visual relation tagging reduces the influence of object location in performance evaluation, and it can effectively support various visual relation based applications, such as video retrieval and visual question answering.

Note that we do not conduct experiments on the tasks of predicate detection and phrase detection introduced in [22]. For predicate detection, it requires the localized objects with their categories as the input in order to predict a set of possible predicates, which is easier than visual relation tagging in practice and less feasible in real applications. For phrase detection, it aims to predict a set of relation triplets and localize each entire visual relation instance with one bounding box trajectory. Similar to visual relation detection, its performance is also influenced by the accuracy of object localization in videos; moreover, it is less challenging than visual relation detection as it only requires to provide the union bounding box trajectory.

Evaluation metrics. Mean average precision (mAP) is used as an evaluation metric for visual relation detection, which is widely used for detection tasks. However, this metric is discarded in the previous VRD evaluation because of incomplete relation labeling of dataset, which does not exist in the construction of our dataset. Following [5, 16, 17, 22, 42], we also use Recall@K (K equals 50 and 100) as the evaluation metrics for visual relation detection; it denotes the fraction of correct visual relation instances detected in the top K detection results.

In visual relation tagging, we use Precision@K as the evaluation metric to emphasize the ability of tagging accurate visual relations. Since the average number of relation triplets per video is 10.34 in our dataset, we set K to 1, 5 and 10 in the experiments.

5.2 Component Analysis

Relation prediction is the key module in our proposed method, which consists of two main components: relation feature extraction and relation modeling. We validate their influences to the performance of our method.

Relation feature. Our proposed method extracts two types of features for VidVRD: object feature and relativity feature. The former includes object classeme and iDTs extracted from each object tracklet, and the latter includes the relative position, size and motion between a pair of object tracklets. As object classeme is crucial to subject and object prediction, we keep it in the component analysis of feature extraction, and generate three baselines: only using object classeme (VidVRD-C), using object classeme and iDT (VidVRD-CT) and using object classeme and relativity feature (VidVRD-CR).

The top three rows in Table 2 show the performance of these three baselines. We can see that both iDT and relativity feature can complement object classeme; and our method VidVRD obtains the best performance when fusing all the features. It shows that all the components of our relation features are effective in VidVRD.

<table>
<thead>
<tr>
<th>Method</th>
<th>relation detection</th>
<th>relation tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>VidVRD-C</td>
<td>R@50 = 4.36</td>
<td>P@1 = 5.42</td>
</tr>
<tr>
<td>VidVRD-CT</td>
<td>R@100 = 5.79</td>
<td>P@5 = 6.73</td>
</tr>
<tr>
<td>VidVRD-CR</td>
<td>R@50 = 5.07</td>
<td>P@1 = 5.98</td>
</tr>
<tr>
<td>VidVRD-M</td>
<td>R@50 = 0.97</td>
<td>P@1 = 1.68</td>
</tr>
<tr>
<td>VidVRD</td>
<td>R@50 = 5.54</td>
<td>P@1 = 6.37</td>
</tr>
<tr>
<td>VidVRD-T_50</td>
<td>R@50 = 12.51</td>
<td>P@1 = 16.55</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of our method with different components on visual relation detection and visual relation tagging. R@K and P@K are abbreviations of Recall@K and Precision@K, respectively.
We can see that the performance of VidVRD-M has significantly degraded in both the visual relation detection and visual relation tagging as compared to all other variants of VidVRD. This validates the necessity of exploring the interdependency of subject, predicate and object prediction.

**Object localization.** As mentioned in Section 1, a technical challenge in VidVRD is that VidVRD requires to localize objects with bounding box trajectories. Yet it is still an open problem in video analysis. To validate the influence of object localization on our performance, we generate a baseline by using the ground truth object trajectories (VidVRD-$T_{gt}$). These object trajectories are divided into object tracklets in video decomposition, and only the ones across the segments are retained for feature extraction. Note that only object trajectory is provided in this baseline, and the object category of each trajectory is not given.

The bottom row in Table 2 shows the performance of the baseline. We see that the ground truth object trajectories can obviously improve the performance in visual relation detection; however, it only leads to slight improvement in performance of visual relation tagging because its output does not require object localization. It shows that object localization is still a major constraint in VidVRD.

### 5.3 Comparison with State-of-the-Arts

**Comparison methods.** We compare the performance of our proposed method with four state-of-the-art methods: Visual Phrase (VP) [32], Lu’s only V (Lu’s-V) [22], Lu’s [22], VTransE [42]. Since these methods all aimed at ImgVRD, they only focus on feature extraction for still images but ignore dynamic features in videos. Moreover, most methods only retain the top one confident relation prediction for each object pair in order to obtain high recall on the sparsely labeled evaluation dataset, such as Visual Relationship dataset [22] and Visual Genome [12].

**Table 3: Evaluation of different methods on visual relation detection and visual relation tagging.**

<table>
<thead>
<tr>
<th>Method</th>
<th>relation detection</th>
<th>relation tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@50</td>
<td>R@100</td>
</tr>
<tr>
<td>VP [32]</td>
<td>0.89</td>
<td>1.41</td>
</tr>
<tr>
<td>Lu’s-V [22]</td>
<td>0.99</td>
<td>1.80</td>
</tr>
<tr>
<td>Lu’s [22]</td>
<td>1.10</td>
<td>2.23</td>
</tr>
<tr>
<td>VTransE [42]</td>
<td>1.07</td>
<td>1.45</td>
</tr>
<tr>
<td>VidVRD</td>
<td>5.54</td>
<td>6.37</td>
</tr>
</tbody>
</table>

**Table 4: Evaluation of different methods on zero-shot visual relation detection and visual relation tagging.**

<table>
<thead>
<tr>
<th>Method</th>
<th>relation detection</th>
<th>relation tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@50</td>
<td>R@100</td>
</tr>
<tr>
<td>Lu’s-V [22]</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>Lu’s [22]</td>
<td>0.69</td>
<td>1.16</td>
</tr>
<tr>
<td>VTransE [42]</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>VidVRD</td>
<td>1.62</td>
<td>2.08</td>
</tr>
</tbody>
</table>

We extend these methods to satisfy the requirements of VidVRD on our constructed dataset for fair comparison. First, we replace the original features in these methods with the relation features extracted on the object tracklets in our method. Specifically, the relation features are not used in VP, because it focuses on localizing each visual relation instance with an entire bounding box rather than providing two separate bounding boxes for subject and object, and hence the relativity between subject and object is not applicable to VP. Second, we retain multiple relation predictions with top confidence for each object pair in order to avoid low recall on our fully labeled dataset. In our experiments, we set the number of retained relation predictions for each object pair to 20, which is the same as the setting in our method. Third, we associate the segment-level relation predictions of these methods with our greedy relational association strategy to generate the final visual relation instances.
Our method is superior to the state-of-the-art baselines on both visual relation detection and visual relation tagging, specifically for visual relation detection, our method improves by 6.5% in Precision@1 compared to the top baseline (VP). Figure 5 shows several comparison examples to illustrate our advantages on visual relation detection, and Figure 7 presents some examples of our results on visual relation tagging.

Our relation features can help the proposed method and all the baselines to effectively detect the specific visual relations in videos. For example, our method together with four baselines detect the visual relation “zebra-follow-zebra” in the top row of Figure 5. It requires the use of dynamic video features to distinguish the predicate from “follow” to “stand (on the) left (of)”. Another example of the effectiveness of our relation features is illustrated in the middle row of Figure 5, which shows the successful detection of the changes of the person’s state from “stand” (rank (7)) to “walk” (rank (9)) and the dog’s action from “watch” (rank (33)) to “play” (rank (13)).

Object tracklet proposal used in our method can provide approximate object positions, which helps to detect the coarse spatial relations. The bottom row of Figure 5 shows the effectiveness of our method, in which 8 spatial relations combined with the subjects’ actions are correctly detected. However, inaccurate object localization prevents the detection of visual relation instances that require fine-grained position description, such as “towards” and “past”. Figure 6 shows 2 examples that our method fails to detect the visual relation “bicycle-move towards-person” and “airplane-move past-watercraft”. If we were to use the ground truth object trajectories as the input (i.e., VidVRD-Tgt in Table 2), these visual relations would be correctly detected. Moreover, we can see from Figure 6 that accurate object localization can help to detect more visual relation instances and improve the ranks as well.

Zero-Shot Learning. Since it is impractical to collect and label all possible relation triplets, a promising VidVRD method should be able to predict unseen visual relations. With this in mind, we compare our proposed method with the baseline in zero-shot learning setting. As noted earlier, our test set contains 258 relation triplets out of 1,011 that never appear in our training set, such as “dog-sit behind-person”. It means that 25.5% of relation triplets are unseen to the visual relation detectors.

We report the zero-shot results in Table 4. VP is not included in the comparison because it can only detect seen relation triplets and is not applicable to zero-shot learning. We can see that our method significantly surpasses the baselines that only use visual features: Lu’s V and VTransE, and is slightly worse than Lu’s in mAP of relation detection as it exploits language priors. Moreover, as compared to Table 3, the performances of all the methods degrade drastically, though the random guess performs even worse (e.g. Recall@100 is less than 0.062%). For example, our method has 4.29% drop in Recall@100 for visual relation detection and 38.89% drop in Precision@1 for visual relation tagging. It shows that zero-shot learning is challenging when the unseen relation ratio is high.

6 CONCLUSIONS

We proposed a new vision task named VidVRD, which aims to detect all visual relation instances in form of the relation triplets and object trajectories in videos. To handle the technical challenges in VidVRD, we presented a method consists of object tracklet proposal, relation prediction and greedy relational association. Moreover, we constructed a VidVRD dataset containing 1,000 videos with manually labeled visual relations. The experimental results on the dataset demonstrated that our method outperforms the state-of-the-art baselines on both visual relation detection and visual relation tagging. In future, we will focus on tackling the challenge of weakly supervised learning framework for VidVRD. We will also explore the role of language or linguistic resources and human knowledge for relation learning, especially in zero-shot setting.

ACKNOWLEDGMENTS

This research is part of the NExT++ project, supported by the National Research Foundation, Prime Minister’s Office, Singapore under its IRC@SG Funding Initiative. It is also supported by National Science Foundation of China (6132149, 61202320), Collaborative Innovation Center of Novel Software Technology and Industrialization, and China Scholarship Council.
REFERENCES


