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
rentw@nju.edu.cn

Jingfan Guo
State Key Laboratory for Novel Software Technology, Nanjing University, China
guojf@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 $\langle subject, predicate, object \rangle$, 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].

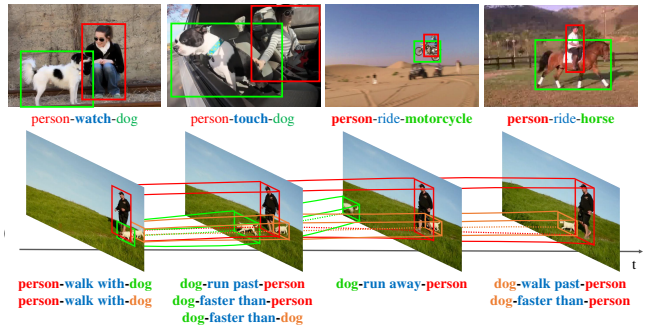

Figure 1: Examples of visual relation. The top row shows visual relations represented with relation triplets and localized objects. The bottom row shows the video visual relations.

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 $\langle subject, predicate, object \rangle$ 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,

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: http://dx.doi.org/10.1145/3123266.3123380

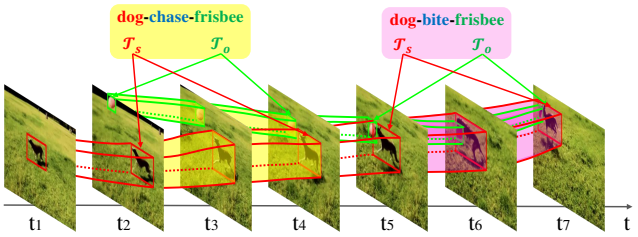

Figure 2: An example of temporally changing visual relation in videos. Two visual relation instances containing their relation triplets and object trajectories of the subjects and objects are illustrated with background color in yellow and magenta, respectively.

such as "walk" and "run". Meanwhile, some visual relations such as the dynamic relations can only be detected in videos, such as "dog-run past-person" and "dog-faster than-person". Hence, video visual relation detection (VidVRD) is a more general and feasible task as compared to ImgVRD.

Another significant difference between VidVRD and ImgVRD is that the visual relations in a video are usually changeable over time while that of images are fixed. The objects may be occluded or out of frame temporarily, which causes the occurrence and disappearance of visual relations. Even when two objects consistently appearing in the same video frames, the interactions between them may be temporally changed. Figure 2 shows an example of temporally changing visual relation between two objects within a video, in which *dog* and *frisbee* are simultaneously appearing between t_2 and t_7 while their interaction changes from *chase* to *bite*. Hence, the VidVRD task should be redefined to handle the changeability in visual relations.

Definition. To be consistent with the definition of ImgVRD, we define VidVRD task as follows: Given a set of object categories of interest C and predicate categories of interest $\mathcal{P}$, VidVRD aims to detect instances of visual relations of interest $C \times \mathcal{P} \times C$ in a video, where a *visual relation instance* is represented by a relation triplet $\langle subject, predicate, object \rangle \in C \times \mathcal{P} \times C$ with the trajectories of the subject and object, $\mathcal{T}_s$ and $\mathcal{T}_o$. Specifically, $\mathcal{T}_s$ and $\mathcal{T}_o$ are two sequences of bounding boxes, that respectively enclose the subject and object, within the maximal duration of the visual relation. In Figure 2, two visual relation instances in the given video can be represented with the relation triplets, "dog-chase-frisbee" and "dog-bite-frisbee", and the *dog* and *frisbee* are localized with the red and green trajectories between (t_2, t_4) and (t_5, t_7), respectively.

Compared to ImgVRD, VidVRD faces more technical challenges. First, VidVRD requires to localize objects with bounding box trajectories. This is more difficult than providing a bounding box for each object in ImgVRD, because the accuracy of an object bounding box trajectory is influenced by both the performances of object localization on each frame and object tracking. Our proposed VidVRD method (Figure 3) tackles the difficulty by generating the object tracklets within each of overlapping short segments of a video, and then associating them into the object trajectories based on predicted visual relations. Second, VidVRD needs to temporally localize the visual relations within maximal duration. For this purpose, we propose a greedy association algorithm that 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

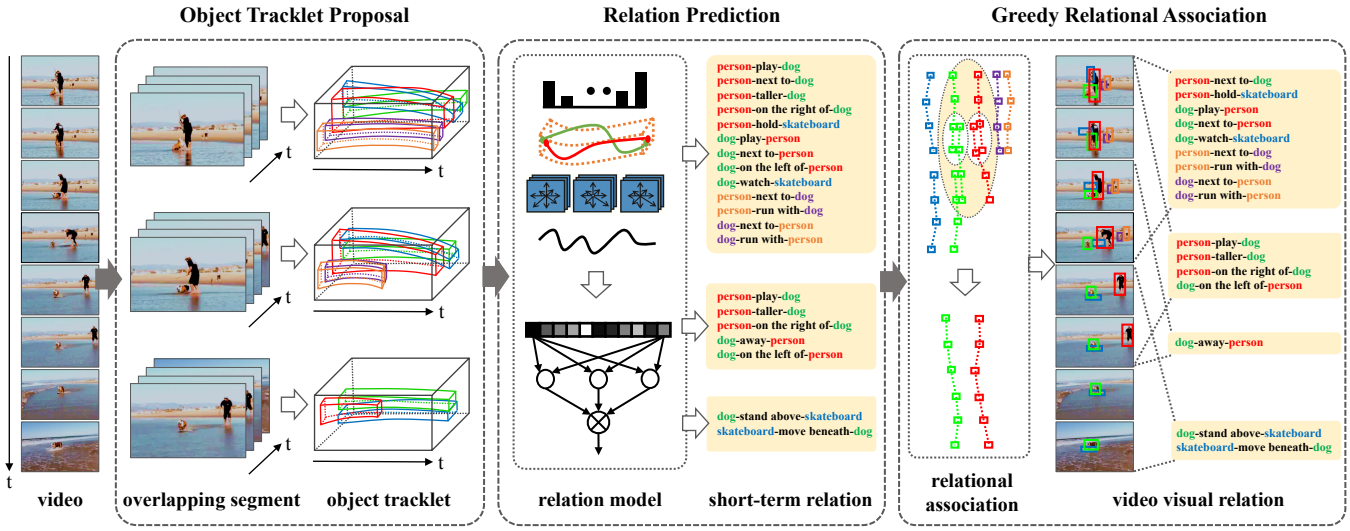

Figure 3: An overview of our VidVRD method. A given video is first decomposed into a set of overlapping segments, and the object tracklet proposals are generated on each segment. Next, short-term relations are predicted for each object pair on all the segments based on feature extraction and relation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

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 to resolve this problem. Inspired by recent progress, we utilize improved dense trajectory (iDT) [37] as a part of the features in our proposed method, because iDT still achieves outstanding performance on most action recognition datasets, especially when the training data is insufficient. Note that our proposed method aims to detect more general relations between objects than action, such as the spatial and comparative 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, train, turtle, watercraft, whale, zebra.

intransitive verbs usually describes the attributes of objects only, they are expressive 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[2], 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 [3]. 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 1: Statistics of our VidVRD dataset. The number of video-level visual relation instances is not available for the training set because it is only sparsely labeled.

	training set	test set
video	800	200
subject/object category	35	35
predicate category	132	132
relation triplet	2,961	1,011
visual relation instance (video-level)	-	4,835
segment	15,146	3,202
labeled segment	3,033	2,801
visual relation instance (segment-level)	25,917	29,714

4 VIDEO VISUAL RELATION DETECTION

A major challenge of VidVRD is to handle the changeability of visual relations over time. To this end, we propose a VidVRD method that detects visual relation instances in short-term, followed by an association algorithm to form the overall visual relation instances in a video (as illustrated in Figure 3). The assumption behind the proposed method is that the basic visual relations can 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 used 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 $(\mathcal{T}_s, \mathcal{T}_o)$ 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 $\langle subject, predicate, object \rangle$ involves recognizing the object categories of $\mathcal{T}_s$ and $\mathcal{T}_o$, and the interactions between them. In practice, it is impossible to learn a separate

[2] **Transitive verb**: bite, chase, drive, fall (off), feed, fight, follow, hold, kick, play, pull, ride, touch, watch; **Comparative**: faster, larger, taller; **Spatial**: above, away, behind, beneath, inside, in front of, next to, on the left of, on the right of, past, towards; **Intransitive verb**: creep, fly, jump, lie, move, run, sit, stand, stop, swim, walk.

[3] Available at https://lms.comp.nus.edu.sg/research/VidVRD.html

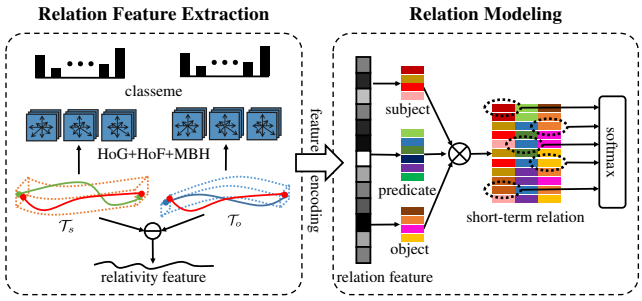

Figure 4: Illustration of relation prediction. For a pair of objects, a set of features are extracted to describe each object and their relativity, and are encoded to a relation feature. Based on the relation feature, separate predictors for subject, predicate and object are trained under softmax loss.

predictor for each single relation triplet due to the huge number of combination and insufficiency of training data. Our model (Figure 4) learns separate subjet, predicate and object predictors to reduce the modeling complexity and exploit the common components in various relations. The model also leverages a rich relation feature that combines the appearance and motion characteristics of the subject and object, as well as the relative characteristics between them.

Relation Feature Extraction. We extract the object features for $\mathcal{T}_s$ and $\mathcal{T}_o$ to describe their appearance and motion characteristics. In particular, we first extract the improved dense trajectory (iDT) features [37] with HoG, HoF and MBH in segments, which capture both the motion and low-level visual characteristics. To encode the features, we train a codebook for each of the four descriptor types in iDT using 100,000 randomly sampled features. The size of each codebook is set to 1,000. Then the object feature for $\mathcal{T}$ is computed as a bag of iDT features enclosed in $\mathcal{T}$, where half of an iDT locates within the area of $\mathcal{T}$ is considered as being enclosed. Additionally, we append the object feature with a classeme feature [36], which is a N-d vector of classification probabilities (*i.e.*, N classes) predicted by deep neural networks, to encode the semantic attributes in the visual appearance.

To extract the relative characteristics between $\mathcal{T}_s$ and $\mathcal{T}_o$, we propose a relativity feature which describes the relative position, size and motion between the two objects. Denoting $C_s^t = (x_s^t, y_s^t)$ and $S_s^t = (w_s^t, h_s^t)$ as the centeral point and size of $\mathcal{T}_s$ at time t respectively (resp. $\mathcal{T}_o$), we compute the relative position ΔC, relative size ΔS and relative motion ΔM descriptors as

$$\begin{aligned} \Delta C &= (C_s^1 - C_o^1, \ldots, C_s^L - C_o^L), \\ \Delta S &= (S_s^1 - S_o^1, \ldots, S_s^L - S_o^L), \\ \Delta M &= (\Delta C^2 - \Delta C^1, \ldots, \Delta C^L - \Delta C^{L-1}). \end{aligned} \tag{1}$$

In order to characterize the abundant spatial relations, such as "behind", "larger" and "past", as well as their various combinations, such as "past behind", we use dictionary learning to train a codebook for each type of descriptor. Specifically, for each codebook, we set the size to 1,000 and randomly sample 100,000 descriptors for training. The elements in the obtained codebooks can be interpreted as the atomic relative features, so that complicated relative features can be represented by their linear combination. For a pair of $\mathcal{T}_s$ and $\mathcal{T}_o$, 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 $\mathcal{T}_s$ and $\mathcal{T}_o$ 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-occurence 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:

$$L = \sum_{\langle s_i, p_j, o_k \rangle} -\log \text{softmax}_{\mathcal{R}} \left(P^s(f, s_i) \cdot P^p(f, p_j) \cdot P^o(f, o_k) \right), \tag{2}$$

where f is the relation feature of a specific relation triplet $\langle s_i, p_j, o_k \rangle$, 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 $(\mathcal{T}_s, \mathcal{T}_o)$, 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 $\{(c^t, \langle s, p, o \rangle, (\mathcal{T}_s^t, \mathcal{T}_o^t))\}_t$ $(t = m, \ldots, n)$ detected from the m-th segment to the n-th segment, which have identical relation triplet $\langle s, p, o \rangle$ and with sufficient overlapping between successive ones, our goal is to merge them into a single visual relation instance $(\hat{c}, \langle s, p, o \rangle, (\hat{\mathcal{T}}_s, \hat{\mathcal{T}}_o))$ with confidence score:

$$\hat{c} = \frac{1}{n - m + 1} \sum_{t=m}^{n} c^t, \tag{3}$$

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 $\mathcal{S} = \{(c, \langle s, p, o\rangle, (\mathcal{T}_s, \mathcal{T}_o))\}$
Output: the set of merged instances $\mathcal{L} = \{(\hat{c}, \langle s, p, o\rangle, (\hat{\mathcal{T}}_s, \hat{\mathcal{T}}_o))\}$
Initialize: $\mathcal{L} = \emptyset$, $\gamma = 0.5$
for $t = 1, \ldots, T$ **do**
  $\mathcal{A}$ = instances in $\mathcal{L}$ that end at the $(t-1)$-th segment
  $\mathcal{B}$ = instances in $\mathcal{S}$ that detected at the t-th segment
  Descending sort $\mathcal{A}$ accordint to $\hat{c}$
  Descending sort $\mathcal{B}$ accordint to c
  for $(c, \langle s, p, o\rangle, (\mathcal{T}_s, \mathcal{T}_o))$ `in` $\mathcal{B}$ **do**
    for $(\hat{c}', \langle s', p', o'\rangle, (\hat{\mathcal{T}}_s', \hat{\mathcal{T}}_o'))$ `in` $\mathcal{A}$ **do**
      if $\langle s, p, o\rangle = \langle s', p', o'\rangle$ AND vIoU$(\mathcal{T}_s, \hat{\mathcal{T}}_s') > \gamma$ AND vIoU$(\mathcal{T}_o, \hat{\mathcal{T}}_o') > \gamma$ **then**
        Recompute $\hat{c}'$ using Eq. (3)
        Append $(\mathcal{T}_s, \mathcal{T}_o)$ to $(\hat{\mathcal{T}}_s', \hat{\mathcal{T}}_o')$
        Remove $(\hat{c}', \langle s', p', o'\rangle, (\hat{\mathcal{T}}_s', \hat{\mathcal{T}}_o'))$ from $\mathcal{A}$
        Break
      end if
    end for
    if NOT merged **then**
      Add $(c, \langle s, p, o\rangle, (\mathcal{T}_s, \mathcal{T}_o))$ to $\mathcal{L}$
    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 $(\hat{\mathcal{T}}_s, \hat{\mathcal{T}}_o)$. 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 most confident ones 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.

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.

Method	relation detection			relation tagging		
	R@50	R@100	mAP	P@1	P@5	P@10
VidVRD-C	4.36	5.36	7.17	30.00	22.60	16.33
VidVRD-CT	4.78	5.79	6.73	31.00	23.50	18.55
VidVRD-CR	5.07	5.98	8.35	41.00	27.50	19.00
VidVRD-M	0.97	1.68	1.99	17.00	11.30	9.05
VidVRD	5.54	6.37	8.58	43.00	28.90	20.80
VidVRD-T$_{gt}$	**12.51**	**16.55**	**15.53**	**43.50**	**29.70**	**23.20**

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.

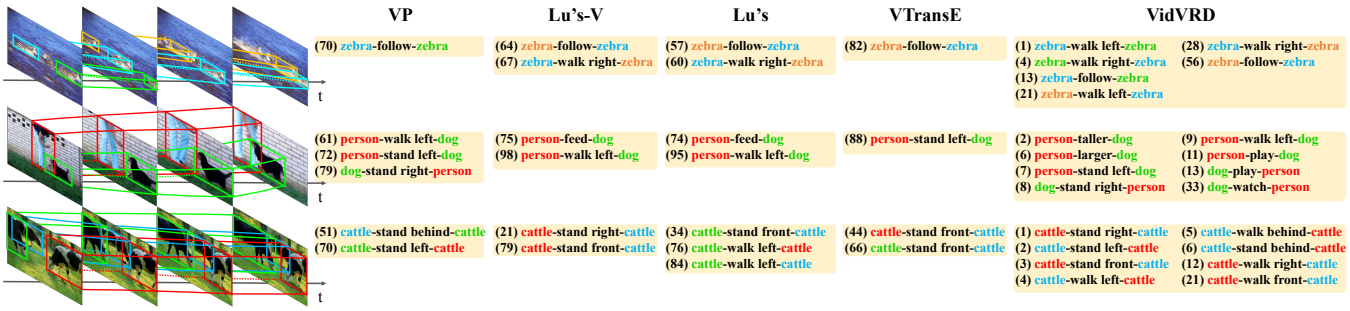

Figure 5: Qualitative examples of visual relation detection using different methods. The correct visual relation instances in the top 100 results are shown, and their ranks are marked in front of them with parentheses. Note that the object localization results are not shown due to space limitation, but they are all required to have sufficiently high vIoUs to the ground truth object trajectories, which are shown on the video frames with different colors.

Relation modeling. Our proposed method explores the interdependency of subject, predicate and object prediction by joint modeling. It combines the predictions of the three components to optimize the rank of relation triplets instead of their ranks independently. To validate its effectiveness, we generate a baseline by modeling subject, predicate and object independently (VidVRD-M).

The fourth row in Table 2 shows the performance of VidVRD-M. 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.

Method	relation detection			relation tagging		
	R@50	R@100	mAP	P@1	P@5	P@10
VP [32]	0.89	1.41	1.01	36.50	25.55	19.20
Lu's-V [22]	0.99	1.80	2.37	20.00	12.60	9.55
Lu's [22]	1.10	2.23	2.40	20.50	16.30	14.05
VTransE [42]	0.72	1.45	1.23	15.00	10.00	7.65
VidVRD	**5.54**	**6.37**	**8.58**	**43.00**	**28.90**	**20.80**

Table 4: Evaluation of different methods on zero-shot visual relation detection and visual relation tagging. Note that VP does not applicable to zero-shot learning because it can only detect seen relation triplets.

Method	relation detection			relation tagging		
	R@50	R@100	mAP	P@1	P@5	P@10
Lu's-V [22]	0.93	0.93	0.40	2.74	0.82	0.82
Lu's [22]	0.69	1.16	**0.47**	1.37	1.37	1.23
VTransE [42]	0.69	0.69	0.03	1.37	1.37	0.96
VidVRD	**1.62**	**2.08**	0.40	**4.11**	**1.92**	**1.92**

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 relativity 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.

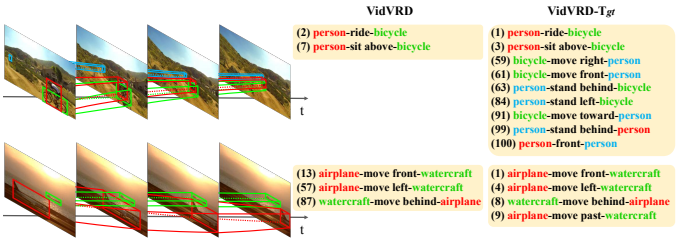

Figure 6: Our failure examples influenced by inaccurate object localization. The correct visual relation instances in the top 100 results of our method with/without the ground truth object trajectories are shown with their ranks.

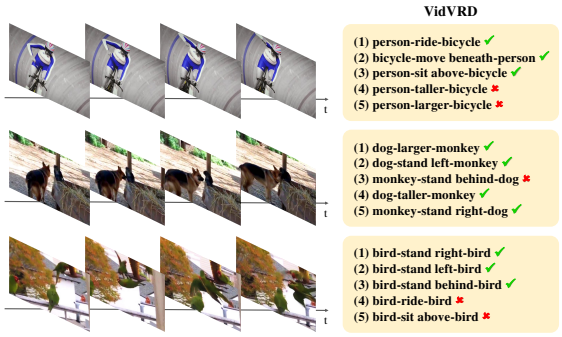

Figure 7: Examples of our results (top-5) on visual relation tagging. The correct and incorrect results are marked with ticks and crosses, respectively.

Overall performance. From the quantitative results in Table 3 and the qualitative results in Figure 5 to 7, we have:

1) 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 the performance by 6.18% and 4.14% on mAP and Recall@100 respectively as compared to the top baseline (Lu's); and on visual relation tagging, 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.

2) 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)).

3) 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 two example 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-T$_{gt}$ 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 (61321491, 61202320), Collaborative Innovation Center of Novel Software Technology and Industrialization, and China Scholarship Council.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision*. IEEE, 2425–2433.

[2] Evlampios Apostolidis, Vasileios Mezaris, Mathilde Sahuguet, Benoit Huet, Barbora Červenková, Daniel Stein, Stefan Eickeler, José Luis Redondo Garcia, Raphaël Troncy, and Lukás Pikora. 2014. Automatic fine-grained hyperlinking of videos within a closed collection using scene segmentation. In *ACM International Conference on Multimedia*. ACM, 1033–1036.

[3] Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. 2014. Object-based visual sentiment concept analysis and application. In *ACM International Conference on Multimedia*. ACM, 367–376.

[4] Wongun Choi. 2015. Near-online multi-target tracking with aggregated local flow descriptor. In *IEEE International Conference on Computer Vision*. IEEE, 3029–3037.

[5] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting Visual Relationships With Deep Relational Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

[6] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. 2014. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press.

[7] Jianfeng Dong, Xirong Li, Weiyu Lan, Yujia Huo, and Cees GM Snoek. 2016. Early Embedding and Late Reranking for Video Captioning. In *ACM International Conference on Multimedia*. ACM, 1082–1086.

[8] Haiyun Guo, Jinqiao Wang, Min Xu, Zheng-Jun Zha, and Hanqing Lu. 2015. Learning multi-view deep features for small object retrieval in surveillance scenarios. In *ACM International Conference on Multimedia*. ACM, 859–862.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[10] Yu-Gang Jiang, Qi Dai, Xiangyang Xue, Wei Liu, and Chong-Wah Ngo. 2012. Trajectory-based modeling of human actions with motion reference points. In *European Conference on Computer Vision*. Springer, 425–438.

[11] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2016. Object detection from video tubelets with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 817–825.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[13] Maksim Lapin, Matthias Hein, and Bernt Schiele. 2016. Loss functions for top-k error: Analysis and insights. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1468–1477.

[14] Byungjae Lee, Enkhbayar Erdenee, Songguo Jin, Mi Young Nam, Young Giu Jung, and Phill Kyu Rhee. 2016. Multi-class Multi-object Tracking Using Changing Point Detection. In *European Conference on Computer Vision*. Springer, 68–83.

[15] Xiangyang Li, Xinhang Song, Luis Herranz, Yaohui Zhu, and Jiang Shuqiang. 2016. Image Captioning with both Object and Scene Information. In *ACM International Conference on Multimedia*. ACM, 1107–1110.

[16] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. 2017. ViP-CNN: Visual Phrase Guided Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

[17] Xiaodan Liang, Lisa Lee, and Eric P. Xing. 2017. Deep Variation-Structured Reinforcement Learning for Visual Relationship and Attribute Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

[18] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv:1312.4400* (2013).

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.

[20] An-An Liu, Yu-Ting Su, Wei-Zhi Nie, and Mohan Kankanhalli. 2017. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 1 (2017), 102–114.

[21] Xueliang Liu and Benoit Huet. 2016. Event-based cross media question answering. *Multimedia Tools and Applications* 75, 3 (2016), 1495–1508.

[22] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*. Springer, 852–869.

[23] Wenhan Luo, Junliang Xing, Xiaoqin Zhang, Xiaowei Zhao, and Tae-Kyun Kim. 2014. Multiple object tracking: A literature review. *arXiv:1409.7618* (2014).

[24] Foteini Markatopoulou, Vasileios Mezaris, and Ioannis Patras. 2016. Deep Multi-task Learning with Label Correlation Constraint for Video Concept Detection. In *ACM International Conference on Multimedia*. ACM, 501–505.

[25] Anton Milan, Stefan Roth, and Konrad Schindler. 2014. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 1 (2014), 58–72.

[26] Li Niu, Xinxing Xu, Lin Chen, Lixin Duan, and Dong Xu. 2016. Action and event recognition in videos by learning from heterogeneous web sources. *IEEE Transactions on Neural Networks and Learning Systems* (2016).

[27] Dan Oneata, Jérôme Revaud, Jakob Verbeek, and Cordelia Schmid. 2014. Spatio-temporal object detection proposals. In *European conference on computer vision*. Springer, 737–752.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 779–788.

[29] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics* 1 (2013), 25–36.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. IEEE, 91–99.

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[32] Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1745–1752.

[33] Xindi Shang, Tongwei Ren, Hanwang Zhang, Gangshan Wu, and Tat-Seng Chua. 2017. Object trajectory proposal. In *IEEE International Conference on Multimedia and Expo*. IEEE.

[34] Rakshith Shetty and Jorma Laaksonen. 2016. Frame-and segment-level features and candidate pool evaluation for video caption generation. In *ACM International Conference on Multimedia*. ACM, 1073–1076.

[35] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*. 568–576.

[36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2016).

[37] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*. IEEE, 3551–3558.

[38] Limin Wang, Yu Qiao, and Xiaoou Tang. 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4305–4314.

[39] Meng Wang, Changzhi Luo, Richang Hong, Jinhui Tang, and Jiashi Feng. 2016. Beyond object proposals: Random crop pooling for multi-label image recognition. *IEEE Transactions on Image Processing* 25, 12 (2016), 5678–5688.

[40] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu, and Nicu Sebe. 2014. Multitask linear discriminant analysis for view invariant action recognition. *IEEE Transactions on Image Processing* 23, 12 (2014), 5599–5611.

[41] Long Ying, Tianzhu Zhang, and Changsheng Xu. 2015. Multi-object tracking via MHT with multiple information fusion in surveillance video. *Multimedia Systems* 21, 3 (2015), 313–326.

[42] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

[43] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. 2017. PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN. In *IEEE International Conference on Computer Vision*. IEEE.

[44] C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. 2013. Learning the visual interpretation of sentences. In *IEEE International Conference on Computer Vision*. IEEE, 1681–1688.