Video Relation Detection via Multiple Hypothesis Association

Zixuan Su[1], Xindi Shang[2], Jingjing Chen[1], Yu-Gang Jiang[1*], Zhiyong Qiu[3], Tat-Seng Chua[2]
[1]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
[2] School of Computing, National University of Singapore [3] Tencent
{suzx16, chenjingjing, ygj}@fudan.edu.cn, {shangxin, chuats}@comp.nus.edu.sg, samuelqiu@tencent.com

ABSTRACT

Video visual relation detection (VidVRD) aims at obtaining not only the trajectories of objects but also the dynamic visual relations between them. It provides abundant information for video understanding and can serve as a bridge between vision and language. Compared with visual relation detection on image, VidVRD requires one more step at last called visual relation association which associates relation segments across time dimension into video relations. This step plays an important role in the task but is less studied. Nevertheless, visual relation association is a difficult task as the association process is easily affected by inaccurate tracklet detection and relation prediction in the former steps. In this paper, we propose a novel relation association method called Multiple Hypothesis Association (MHA). It maintains multiple possible relation hypothesis during the association process in order to tolerate and handle the inaccurate or missing problem in the former steps and generate more accurate video relations. Our experiments on the benchmark datasets (Imagenet-VidVRD and VidOR) show that our method outperforms the state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Information extraction**; • **Computing methodologies** → **Scene understanding**; **Activity recognition and understanding**.

KEYWORDS

video visual relation detection; data association

ACM Reference Format:
Zixuan Su, Xindi Shang, Jingjing Chen, Yu-Gang Jiang, Zhiyong Qiu, Tat-Seng Chua. 2020. Video Relation Detection via Multiple Hypothesis Association. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12-16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413764

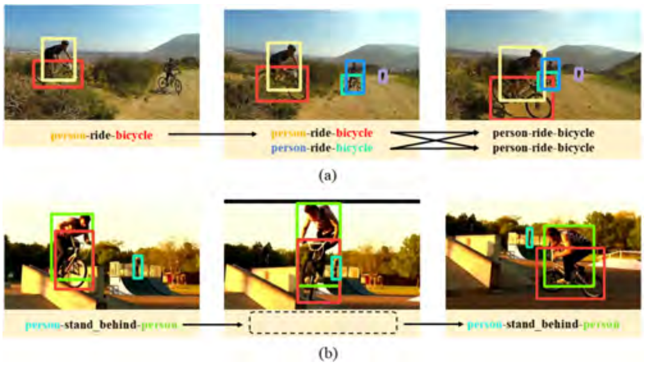

Figure 1: Examples showing the challenges of visual relation association. The first example (a) shows a scene that multiple relation segments with same triplet and similar tracklets are detected in a video segment, which makes association choices uncertain. The second example (b) shows the scene that an exist relation segment is not detected in a video segment due to missing detection of tracklet or unprecise prediction, which interrupts the association.

1 INTRODUCTION

Although there has been much progress on entity level recognition in videos, such as video object detection [6, 20, 24, 37] and action detection [2, 12, 22] etc., it is still difficult for the machine to understand video content in a fine-grained and structured level. To tackle this issue, visual relation is one of the most important and useful information that can help describe the dynamic interactions between the objects in a video. With objects as nodes and visual relations as edges connecting the nodes, video can be represented as a spatio-temporal graph, which can help building the connection between vision and language [4, 29, 32, 36]. Hence, understanding the visual relations (*i.e.* visual relation detection) will benefit many downstream video tasks and applications, such as visual question answering and video captioning.

Unlike visual relation detection in image (ImgVRD) that has been widely studied for years [5, 13, 15, 33–35], its counterpart in video domain has just attracted researchers' attention [16, 19, 23]. Video visual relation detection (VidVRD) requires to track the objects and their pairwise relations in a video. Specifically, it aims to detect all the relation triplets <subject, predicate, object> of interest and spatio-temporally localize their corresponding subject and object trajectories [19]. Compare to ImgVRD, VidVRD considers the time dimension and thus leads to many extra challenges, such as detecting relations of different duration and discriminating relation instances in terms of events. For example, the relation, *A hit B*, usually lasts for a short duration while the duration of the relation, *A speak to B*, can vary from short to long. Meanwhile, the relations of *A speak to B* happening in different time periods of a video shall also be regarded as different instances, because their connections are weak due to discontinuity and different contexts. This additionally

[*]Yu-Gang Jiang is the corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413764

requires the model to robustly detect and associate corresponding relations across the video frames.

To tackle the aforementioned problems, existing methods for VidVRD mainly focus on the visual relation prediction model on short video segments like the research on image, and use simple greedy association algorithm to greedily connect the identical relation triplets in adjacent video segments that have high trajectory overlap [19, 23]. However, such methods unavoidably produces inaccurate prediction and missing detection because of their heavy reliance on the performance of the prediction models. Though these models can be improved over short video segments by considering spatio-temporal context [16, 25], they may still suffer from the bias and noise in learning and modeling long-tail data distribution, which is quite common in visual relations [9, 18]. Alternatively, we take a different perspective by studying a more robust inference algorithm through multiple hypotheses.

In fact, the multiple detected relation instances of same semantic type can easily confuse the model to establish the correspondences. For example, Figure 1(a) shows that multiple "person-ride-bicycle" are detected in the third video segment, whose geometric positions are close. As can be seen, it is very hard to decide the correspondences to the relations in the neighbouring videos segments under this situation, and making decision immediately may lead to false association results. Moreover, such situation may happen even more frequently in practice because (1) multiple objects with the same category commonly exist in the scene; (2) redundant object tracklets are likely detected; and (3) false relationships can be easily predicted. Therefore, we propose to use the idea of multiple hypotheses to preserve all probable correspondences and wait for more information from latter segments. The idea can be illustrated by Figure 1(a) where four connection hypotheses between the relations in the second and third video segments can be made.

Specifically, we propose a novel association algorithm named Multiple Hypothesis Association (MHA) which uses hypothesis tree to model visual relations when performing association. In other words, our MHA uses a tree structure to model the multiple hypotheses about a video relation instance in video. Each node in the tree represents a segment of the relation observed (detected) in a short video segment. As MHA processes over the video segments in sequence, the observed relation segments will be selectively added to corresponding trees as leaf node to update or create hypotheses. In the end, each path from a tree's root node to a leaf node represents a hypothesis about the complete relation instance.

MHA differs from greedy association method in the following four aspects. First, during association, it maintains multiple relation hypotheses, *i.e.* paths in the tree, instead of only preserving the optimal one. This tolerates the cases of inaccurate prediction when some correct predictions get low scores, and delays the decision until more information obtained. Second, for more robust association, the information in existing relation hypothesis will be utilized, instead of only the short-term observations receiving connection. Third, geometry information and prediction confidence are both utilized for measuring the connecting affinity between relations instead of using only the geometry information. Fourth, MHA tolerates the missing relation detections of a complete relation and prevents it from interrupting the connection, as shown in Figure 1(b). The contributions of this paper includes:

- We propose an effective video visual relation detection method based on multi-hypothesis association.
- We achieve the state-of-the-art performance on the benchmark Imagenet-VidVRD dataset.
- We achieve competitive performance on the VidOR dataset compared with the top-1 solution at VRU'19 grand challenge.

2 RELATED WORK

Video Visual Relation Detection. Compare to ImgVRD [5, 13, 15, 33–35], VidVRD has not received sufficient attention until the recent due to its complexity and a lack of suitable dataset. [19] contributed ImageNet-VidVRD dataset which labels all relation triplets in video as well as the trajectories of corresponding subject and object and becomes the first dataset on video visual relation detection. They also proposed an effective three-stage detection method including object tracklet proposal, relation prediction and greedy relational association, which has become the most widely-used pipeline in VidVRD. Based on the Imagenet-VidVRD dataset, more works in this area have been done. [25] used conditional random field to generate a gated spatio-temporal energy graph in order to model the variation of relation triplets in video. The idea of multiple hypothesis is first applied to this task by [1] which generates hypothesis for each object pair when performing association. [16] built a spatio-temporal graph between adjacent video segments and used multiple layers of graph convolutional networks to pass messages between graph nodes. Besides, they proposed an online association method with a siamese network and obtained the state-of-the-art results by combining these two parts. [18] contributed a large-scale VidOR dataset for VidVRD. On this dataset, [23] utilized language context feature along with spatial-temporal feature for predicate prediction and won the first place at VRU'19 (Video Relation Understanding 2019) grand challenge.

Existing methods seldom concern about the third stage, *i.e.* relational association, which has the greatest difference between relation detection on video and image. The association method in [1] cannot satisfactorily handle various different predicates between each object pair while the siamese network in [16] only adds an appearance similarity score to the original greedy association method but suffers from extra complexity in the training process. In this paper, we differ from the framework of greedy association and propose a brand new effective association method which requires no training process.

Multiple Hypothesis. Multiple hypothesis has become an important idea in recent years and has been found to be useful for many different visual tasks. [7] constructed hypothesis trees to maintain multiple potential object detections at each frame for online object tracking. From another angle on the same task, [10] adopted multiple component trackers using different features to track multiple trajectory hypothesis for further comparison and selection. On the task of keypoint matching, [3] estimated multiple-hypothesis affine regions for each key point in an image in order to avoid the negative affection from the initial point and improve repeatability. [17] modified traditional single-prediction model to make multiple hypothesis predictions on four different tasks including human pose estimation, future prediction, image classification and segmentation in order to deal with the prediction uncertainty

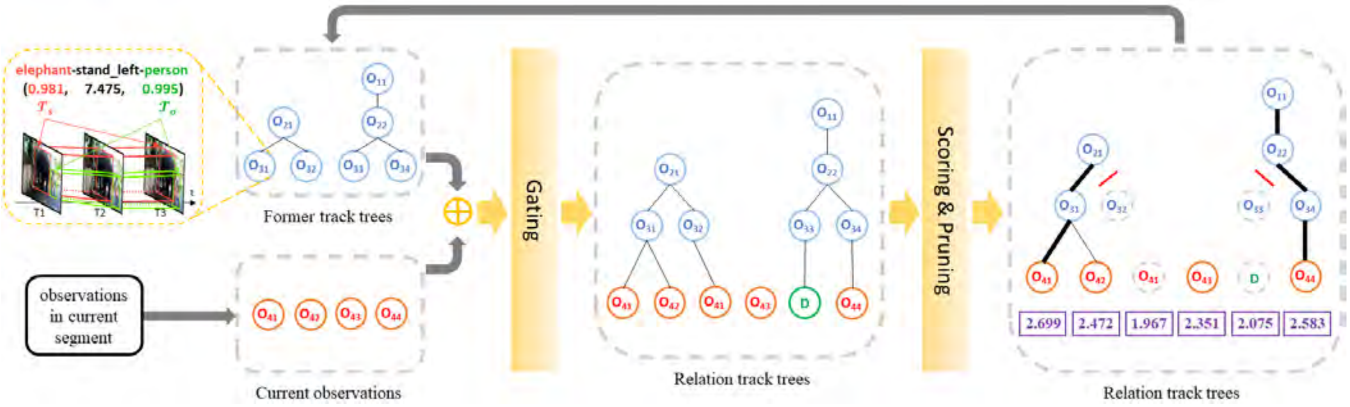

Figure 2: An overview of our MHA method. The relation hypothesis trees generated by observations in the former video segments are updated with new observations in the current video segment during each iteration. The results after each iteration will act as "former hypothesis trees" in next iteration. O_{ij} represents the j-th relation segment observed in video segment S_i. Dummy nodes are illustrated with green color. In the hypothesis trees at the right-most part, the optimal path of each tree is denoted by bold line while the paths being pruned are denoted by dotted lines.

and ambiguity. On the task of 3D human pose estimation from a single RGB image, [11] generated multiple feasible hypotheses of 3D pose from 2D joints to alleviate the problems resulting from depth ambiguity and occluded joints. Following [7], [31] built up proposal propagation tree to maintain multiple object proposal hypotheses for each object in time steps to perform data association globally in semi-supervised video object segmentation task.

Though the above methods are various and used in different tasks, they share the similar idea to maintain multiple hypothesis of an uncertain result when suffering from different types of ambiguity. Instead of making an early decision, this way of process delays it until more information arrives that allows the model to make better comparison and choices at a global scale. In this paper, we utilize the same idea on the VidVRD task.

Multi-Object Tracking. Multi-object tracking (MOT) is a famous task with the purpose of tracking multiple different objects in the video simultaneously. Recent methods are mostly based on the tracking-by-detection paradigm, which firstly detect the objects on each frame before associating them together across the video. In this pipeline, detection association is one of the most important parts and has attached much attention from researchers. [7] applied the classical multiple hypothesis tracking algorithm for association, based on which, [8] added the bilinear LSTM to improve the learning of long-term appearance information. [28] generated special non-uniform hypergraphs to model and associate detections on each frame. [30] proposed the spatial-temporal relation network in order to encode various cues across different domains simultaneously and generate better similarity scores.

MOT and VidVRD both require a similar association process. Relational association in VidVRD can be seen as "multi-relation tracking" in a way. That is why we can learn from MOT during our research. In this paper, we apply the multiple hypothesis tracking algorithm as a framework for relation association which shows satisfying performance.

3 MULTIPLE HYPOTHESIS ASSOCIATION

For relatively precise relation segment detection, videos are usually split into segments of 30 frames and with 15 frames of overlap between two adjacent segments, denoted as $V = \{S_1, S_2, \ldots, S_k\}$ where S_i are sorted by time. For each video segment, the detected relation segments are denoted as $S_i = \{O_{i1}, O_{i2}, \ldots, O_{iN_i}\}$. O_{ij} $(1 \leq i \leq k, 1 \leq j \leq N_i)$ represents a relation segment observed in video segment S_i and N_i is the total number of observations in S_i.

Figure 2 shows the framework of the proposed Multiple Hypothesis Association (MHA) method which builds a dynamically-growing hypothesis tree for each probable video relation. As shown in the figure, each node in tree represents an observed relation segment O_{ij} and the nodes in the same level come from the same video segment S_i. The information in each node O_{ij} includes their relation triplets, tracklets (subject and object), and prediction confidence scores (subject, predicate and object). The path of a tree from its root to a leaf node means a possible constitution of the corresponding relation, *i.e.* a hypothesis. During MHA process, video segments are operated by turns. In each iteration, the hypothesis trees built from the former video segments are updated with all new observations in the current video segment by gating, scoring and pruning process.

3.1 Gating

After new relation segments are detected (observed), they should first be connected to some existing hypothesis trees which are constituted by the observations in the former video segments. The process of judging the connection condition and selecting the eligible observations for each leaf on each hypothesis tree is called gating. Here we need to measure the connecting affinity between the new observations and each leaf node of the hypothesis trees in order to perform gating.

$$s_{con,s} = \alpha * vIoU_s + \beta * s_s \tag{1}$$

$$s_{con,o} = \alpha * vIoU_o + \beta * s_o \tag{2}$$

Eq.1 and Eq.2 are used to compute the connecting scores of subject and object separately, which are served as the connecting affinity measurements. vIoU calculates the total IoU in concurrent frames of two trajectories. In the equations, $vIoU_s$ and $vIoU_o$ denote the vIoU between the observation and leaf node of the subject and object tracklet respectively, where s_s and s_o represents the confidence score of the subject and object of observation, and α and

β are hyper-parameters. vIoU provides the geometry information and confidence score provides the prediction confidence. By combining them together, the connecting score becomes more robust by considering the inaccuracy of trajectory detection and object prediction. It is noteworthy that the connecting score does not contain the confidence score of predicate because of the relatively low accuracy of predicate prediction. By setting less strict connecting condition for predicate here, the observations with low predicate prediction score can also be successfully connected.

During actual processing, if a leaf node's relation triplet is the same as an observation's and the two connecting scores between them are both higher than the threshold, a new node will be built according to this observation and connected to this leaf node. After the whole gating process, the observations remained isolated will be the root of a new tree. On the other hand, when a leaf node receives no connection in an iteration, which means a missing detection may have happened, a dummy node containing the same information as this leaf node will be connected to it as its child and become the new leaf node, as the green node shown in Figure 2.

3.2 Scoring

After gating, there may be multiple paths in each relation hypothesis tree which represent multiple hypothesis about the corresponding video relation up to the current video segment. In this section, we design a path score in order to measure the reliability of each hypothesis and for conveniently performing the subsequent steps.

$$s_{rel} = \frac{s_{con,s} + s_{con,o} + \gamma * s_p}{2 + \gamma} \tag{3}$$

$$s_{path} = AVG(s_{rel}) \tag{4}$$

According to Eq.3, each new node's node score is the weighted average of two connecting scores and predicate prediction score, which contains the information of both connection affinity and predicate similarity with former nodes. In the equation, s_p denotes the confidence score of predicate and γ is a hyper-parameter. Specially, we set the node score of the root to be its triplet confidence score which equals to $s_s * s_p * s_o \ / \ 10^f$ for it has no former node for computing connecting score. f is a scale factor which is used to make triplet confidence score have the same order of magnitude as the node score computed by Eq.3.

Eq.4 defines the path score for each hypothesis using average. In this way, it models the reliability of forming the corresponding hypothetical video relation, which considers both long-term connecting fitness and detection confidence.

3.3 N-Scan Pruning

After the former steps in each iteration, we have a forest in which each relation hypothesis tree has multiple paths representing multiple relation hypothesis. In order to ensure that our trees always maintain multiple uncertain hypothesis of controllable size, we need to prune some branches. Specifically, each iteration of the connection can be regarded as a process of aggregating information from a new video segment to all the hypothesis. When there is sufficient information for judgement, a number of hypothesis with lower reliability can be confirmed wrong and should be pruned. Since the branches formed in recent video segments lack information from the latter part of video, we are not able to affirm their reliability or make choices among them. Thus, we can only prune the branches that appeared N video segments ago. Such process is called N-scan pruning.

However, there may be different nodes in each tree or among different trees that come from the same observation, which may lead to conflict when selecting optimal path. Hence, it should be done globally instead of inside a tree, which is called global hypothesis formation. Specifically, it should be done among all the trees before the actual pruning to select the globally-optimal path (hypothesis) for each tree at present, with which we know the chosen preserved branch. Global hypothesis formation can be formulated as the following optimization problem:

$$\begin{aligned} &\max_{z} \sum_{j_1=0}^{N_1} \sum_{j_2=0}^{N_2} \cdots \sum_{j_k=0}^{N_k} s_{O_{1j_1}O_{2j_2}\cdots O_{kj_k}} z_{O_{1j_1}O_{2j_2}\cdots O_{kj_k}} \\ &\text{s.t.} \sum_{j_1=0}^{N_1} \cdots \sum_{j_{u-1}=0}^{N_{u-1}} \sum_{j_{u+1}=0}^{N_{u+1}} \cdots \sum_{j_k=0}^{N_k} z_{O_{1j_1}O_{2j_2}\cdots O_{kj_k}} = 1 \\ &\text{for} \quad j_u = 0, 1, 2, \ldots, N_u \quad \text{and} \quad u = 1, 2, \ldots, k \end{aligned} \tag{5}$$

where s denotes the score of the corresponding path, and z is the binary variable. When $j_i = 0$, it means that this path does not contain observation in video segment i.

Algorithm 1 Greedy Global Hypothesis Formation

```
Input: the set of tree: T; the set of all observations: O
Output: The leaf of the chosen optimal path for each tree: L =
    {(tree: leaf node)}
  InitiateList(Path_record)
  Append all paths in all trees to Path_record
  Descending sort Path_record according to path score
  for each path in Path_record do
      t ⟵ the corresponding tree
      obs ⟵ the observations in path
      if t in T and obs in O then
          T ⟵ T - t
          O ⟵ O - obs
          Append (t: the leaf node in path) to L
      end if
  end for
  for each tree t in T do
      Prune the conflict part of all paths (from conflict node to leaf
      node)
      Select the best path from remaining tree
      Append (t: the leaf of the best path) to L
  end for
```

Unlike [7] which modeled this as a Maximum Weighted Independent Set (MWIS) problem, we perform a greedy selection as in Algorithm 1 to simplify global hypothesis formation in a more reasonable way for our task. In VidVRD, multiple relation proposals are predicted from each pair of object tracklets, which forms abundant observations. After using them to form multiple hypothesis trees, the number of trees is usually larger than actual relation numbers in the video, which means only a part of trees can produce the correct video relation that we need. In order to select the path in these trees, we should greedily select path with high score which reflects high reliability. Imagining two trees both have a path containing observation O, when considering all existing trees for a global optimization like in Eq.5, O may be allocated to a less reliable path which results in the removal of a more reliable path. But, greedy algorithm will select the more reliable path first.

Having the results of global hypothesis formation which are shown as the bold lines in Figure 2, the pruning step can be done easily. Here we set N to be 2. Non-optimal branches formed 2 video segments ago which are shown as the dotted lines in Figure 2 will be pruned in each tree. After this 2-scan pruning process, only the last level in each tree have branches and each former levels will have only one node.

3.4 Relation Generating

After MHA processes all the video segments, the final hypothesis trees will be used to generate the final relation results. Using the optimal hypothesis selected by global hypothesis formation, we only have to convert them into video relations by connecting the nodes (relation segments) in the selected path together into an individual video relation.

Dummy nodes in the paths will be skipped when generating the results. For two adjacent nodes in a path that has skipped dummy nodes, if their trajectories have overlap, we can easily connect them together by averaging their bounding boxes in overlapping duration. Otherwise, if their trajectories have no overlap, which means there are dummy nodes and missing detections between them, we manually generate the missing trajectories using linear interpolation. We use the path score in Eq.4 for evaluation since it comprehensively measures the forming reliability of the hypothesis about a video relation.

4 EXPERIMENTS

4.1 Datasets

We evaluate our method on two datasets: the benchmark ImageNet-VidVRD dataset [19] and the newly released VidOR dataset [18]. ImageNet-VidVRD is the first dataset for VidVRD, which consists of 1,000 videos collected from ILSVRC2016-VID and is split into 800 training videos and 200 test videos. Covering 35 categories of subject/objects and 132 categories of predicates in total, the videos are densely annotated with relation triplets in the form of <subject, predicate, object> as well as the subjects and objects' corresponding trajectories. Particularly, the test set contains 258 relation triplets out of 1,011 that are unseen in the training set, which form a zero-shot set. VidOR, used by VRU'19 grand challenge [21], contains 10,000 videos and is split into 7,000 videos for training, 835 videos for validation, and 2,165 videos for testing. Since the test set is not release, we use validation set for testing in this paper. Comparing to ImageNet-VidVRD, VidOR collects much longer videos. The average video length in train/val set is 35.73 seconds and the longest video lasts 180.01 seconds. The annotations are organized in a similar way but covering 80 categories of subject/objects and 50 categories of predicates. Among the videos, 56.34% of subject/object labels appeared are human (adult, child, baby) which shows that VidOR is more human-centric and complex than ImageNet-VidVRD.

4.2 Tasks and Evaluation Metrics

On both datasets, we evaluate our method on two standard tasks as used in [19]: relation detection and relation tagging. Relation detection requires the output of all the detected relation triplets as well as the corresponding subject/object trajectories within the video. During evaluation, a detected relation is considered correct if it has the same relation triplet with a relation in the ground truth and their trajectory vIoUs of the subject and object are both higher than a threshold. Following [19], we set the threshold to be 0.5 and use mean Average Precision (mAP) as well as Recall@K (K=50, 100) for quantitative evaluation. Relation tagging removes the requirement for object localization and only focus on the precision of relation triplet. The detected relation can be considered correct if its relation triplet appears in the ground truth. For this task, we use Precision@K (K=1, 5, 10) as evaluation metric, which also follows [19]. As mentioned in 4.1, a subset of ImageNet-VidVRD can be used for zero-shot learning. We additionally evaluate the results of this part of videos under the same tasks and metrics in order to test the performance of our method in zero-shot setting.

4.3 Compared Methods

ImageNet-VidVRD. We compare the performance of our method with five existing methods: Shang's [19], GSTEG [25], MHRA [1], VRD-GCN [16], VRD-GCN+Siamese [16]. Among them, Shang's, GSTEG and VRD-GCN utilize the greedy algorithm to associate relation segments into final results. Their difference lies in the design of relation prediction model. By using the GCN module, VRD-GCN achieves the best segment-level prediction results. The other two methods both make progress towards relation association. MHRA uses a different multi-hypothesis relational association method based on the prediction results of Shang's. It generates the hypothesis for each object pair instead of each probable relation and thus it needs an additional step to split the hypothesis into multiple relations. VRD-GCN+Siamese obtained the state-of-the-art result by adopting the siamese network for association based on the prediction results of VRD-GCN. In order to make fair comparison among different association methods, we use the same detected object trajectories and prediction models as the existing methods (*e.g.* Shang's, VRD-GCN) in detecting relation segments in video segments. We then perform relation association using our MHA and obtain the results of Shang's+MHA and VRD-GCN+MHA.

VidOR. Since it is a newly-released dataset, there are not many existing methods that can used for comparison. The state-of-the-art method on the VidOR dataset is [23] which obtained the top-1 result at VRU'19 grand challenge [21]. The pipeline in [23] is not exactly the same as the usual methods. Specially, it directly performs multi-object tracking on the whole video in the first step to detect the complete object trajectories and predicts relation segments in the overlapping video segments between each object trajectory pairs. It designs two models to handle spatial-temporal feature and language feature respectively and combines their output together for the segment-level relation detection. In the association step, it performs association between each pair of objects instead of

Table 1: Summarized implementation details of our experiments.

Dataset	Method	Object tractklet proposal	Relation prediction
ImageNet-VidVRD	Shang's+MHA VRD-GCN+MHA	Faster-RCNN + Tracker (Dlib) + NMS	iDT & classeme & relativity feature GCN feature
VidOR	VRU'19 Top-1+MHA	FGPA + Seq-NMS + KCF	ST & Language feature

performing association on the whole video. Similarly, in order to do fair comparison, we adopt the detected object trajectories and train a prediction model following [23] for relation prediction and uses our MHA for relation association.

Table 2: Evaluation for different methods on the ImageNet-VidVRD dataset

Method	relation detection			relation tagging		
	mAP	R@50	R@100	P@1	P@5	P@10
Shang's [19]	8.58	5.54	6.37	43.00	28.90	20.80
GSTEG [25]	9.52	7.05	7.67	51.50	39.50	28.23
MHRA [1]	13.27	6.82	7.39	41.00	28.70	20.95
VRD-GCN [16]	14.23	7.43	8.75	**59.50**	40.50	27.85
VRD-GCN+Siamese [16]	16.26	8.07	9.33	57.50	41.00	28.50
Shang's+MHA (Ours)	15.71	7.40	8.58	40.00	26.70	18.25
VRD-GCN+MHA (Ours)	**19.03**	**9.53**	**10.38**	57.50	**41.40**	**29.45**

Table 3: Evaluation for different methods on the ImageNet-VidVRD dataset in zero-shot setting

Method	relation detection			relation tagging		
	mAP	R@50	R@100	P@1	P@5	P@10
Shang's [19]	0.40	1.62	2.08	**4.11**	1.92	1.92
GSTEG [25]	0.15	1.16	2.08	2.74	1.92	1.92
MHRA [1]	0.51	1.85	2.47	**4.11**	1.93	1.92
VRD-GCN [16][1]	0.67	3.94	7.18	4.11	1.11	1.37
VRD-GCN+Siamese [16]	0.74	4.63	7.64	4.11	1.11	1.23
Shang's+MHA (Ours)	0.77	**2.08**	**3.47**	1.37	1.92	1.64
VRD-GCN+MHA (Ours)	**1.18**	**2.08**	2.31	2.74	**2.19**	**2.33**

4.4 Implementation Details

We set $\alpha = 0.6$, $\beta = 0.4$, $\gamma = 0.6$ and $N = 2$ in MHA. The connecting score threshold in gating step is set to be 0.5. Besides the main steps as described in Section 3, we perform simple pruning immediately after each iteration of gating which allows only 5 leaves on each tree in order to reduce the computing complexity and memory cost. We also add a tree checking process after each iteration of N-scan pruning in order to avoid wrongly connecting two disconnected relations together. In this step, we abandon the tree that has not been updated by a non-dummy node for a long time and generate a video relation with the best path in this tree. The maximum allowable consecutive dummy node number in a path is set to be 3. For fairly comparing association method, we adopt the object trajectories, features and prediction models used by the previous methods on each dataset.

ImageNet-VidVRD. We adopts the object trajectory results from [19] which are also used by [16]. The relation prediction module of Shang's [19] and VRD-GCN [16] are both utilized by us for separate comparison. These two methods adopts the same relation feature design which concatenates two object features and a relativity feature for each tracklet pairs. Object feature is a concatenation of the improved dense trajectory (iDT) [27] feature and the classeme feature [26] for each object, while the relativity feature computes the relative position, size and motion between two two objects. Shang's designs three simple predictors for subject, predicate and object respectively, and trains them jointly using paired tracklets that have overlap with ground truth by more than 0.5 in vIoU as training samples. Top 20 prediction results for each pair and top 200 results in each video segment are reserved for testing. The scale factor f in scoring step is set to be 1 because the confidence score output of VidVRD is one order of magnitude larger. As for VRD-GCN method, following [16], top 5 tracklets in each video segment are kept and the object features of them in three adjacent video segments are used as input for the 3 layers of ST-GCN module. The resulting updated features are then fed into a linear transformation layer to predict the object categories. The combination of two objects' updated feature and their relativity feature is used to predict their predicate categories through another linear transformation layer. The whole process is trained jointly with the batch size of 5 and initial learning rate of 0.001. The scale factor is set to 0.

VidOR. We adopt the object trajectories from the state-of-the-art method [23]. We use 20 trajectories at most for relation prediction after further filtering with a confidence threshold of 0.05. Different from [23] which split the concurrent part of each pair of trajectories into segments with no overlap separately, we split the whole video into segments with 30 frames in length and with 15 frames of overlap between adjacent segments (the same strategy as on ImageNet-VidVRD). We then split the concurrent part of trajectory pairs aligning to video segments in order to perform our multiple hypothesis association on the whole video after prediction. We follow [23] to generate spatial-temporal feature and language feature for each pair of object tracklets. The spatial-temporal feature is produced by computing the relative location of two objects while the language feature for each object is the output of a word2vec [14] model pretrained on GoogleNews dataset using the object trajectory label as input. These two parts of features are then separately fed into two independent two-layered fully-connected neural networks for separate training with initial learning rate of 0.01. The predictions of the two neural networks are combined together as the final results in testing, and the results are grouped by segments for subsequent association.

4.5 Results on ImageNet-VidVRD

The quantitative results of relation detection and relation tagging are reported in Table 2. Due to the different relation prediction

[1] VRD-GCN and VRD-GCN+Siamese both utilize ground-truth tracklets for zero-shot experiments so they do not participate in the comparison.

models being used, the experimental results of different methods can be separated into two groups to facilitate fair comparison.

For the first group, VRD-GCN, VRD-GCN+Siamese and VRD-GCN+MHA all adopt the same object trajectories and relation prediction model, which means they generate exactly the same relation segments. In the association step, they use greedy association, siamese network association and MHA respectively so their performance can directly reflect the effect of these three association methods. From the relation detection results in Table 2, we can see that VRD-GCN+MHA improves the performance on mAP by 4.80% and 2.77% as compared to VRD-GCN and VRD-GCN+Siamese, respectively. Besides, VRD-GCN+MHA also improves the results of Recall@50 and Recall@10 remarkably. These results show the effectiveness of our MHA on relation detection. As for relation tagging, VRD-GCN+MHA ourperforms the other methods under the metrics of Precision@5 and Precision@10 but is inferior to VRD-GCN under the metric of Precision@1.

For the second group, Shang's, MHRA, and Shang's+MHA adopts the same object trajectories and relation prediction model. During association, they use greedy association, MHRA and MHA respectively. From Table 2, we can see that the results of MHA on relation detection outperforms the other two remarkably under all three metrics, which also prove the superiority of MHA on this task. However, Shang's+MHA performs a bit worse than the other two methods on relation tagging, which shows that our association method may have some bad effect on relation tagging. It is noteworthy that MHRA also utilizes the idea of multiple hypothesis and the inferior results of MHRA suggest that it cannot handle the large variety of predicates between each object pair well.

GSTEG is not in both groups because its relation prediction model is different from both groups. From Table 2, we can see its priority to Shang's under all metrics of both tasks, which indicates that the prediction model and feature used by GSTEG are more effective as compared to Shang's. On this premise, our Shang's+MHA outperforms GSTEG by a large margin on relation detection with less effective prediction model and feature. This comparison indicates the effectiveness of MHA as well as the significance of relation association step in this task. However, the inferior results of Shang's+MHA on relation tagging shows the ability gap between the prediction models and the limitation of MHA.

From the analysis above, for relation detection, we can see that our method can be combined with different prediction model and different features to help improve the result remarkably under different metrics. Specifically, the superior results of MHRA, Shang's+MHA and VRD-GCN+MHA on mAP as compared to the methods using the same object trajectories and relation prediction model indicates the effectiveness and significance of the idea of multiple hypothesis. As for relation tagging, the results of methods with MHA are similar with their counterpart methods with other association method. This is mainly because relation association mainly influence relation detection at instance level instead of relation triplet (*i.e.* tag) level. Nonetheless, our relation association method still have an effect on relation tagging from two aspects. On one hand, relation association will adjust the confidence score of final video relation, which may lead to improvement of tagging results such as the Precision@5 and Precision@10 of VRD-GCN+MHA. On the other hand, some correct relation triplet label may be pruned in the pruning process of MHA, which leads to the worsening of relation tagging results like Precision@1 of VRD-GCN+MHA.

Zero-shot Learning. The results under zero-shot setting are shown in Table 3 from which we can observe that it is really an intractable problem. Since more than 25% of relation triplets in the test set of ImageNet-VidVRD are unseen to relation detectors, zero-shot learning also becomes a bottleneck in standard evaluation. Though existing methods all have difficulty dealing with zero-shot set, our MHA can still get a relatively better result based on the same trajectories and relation predictions especially on relation detection. That is because 1) MHA can maintain the relation segments with lower confidence score instead of discarding them; and 2) MHA can still make complete connection instead of split it when missing detection happens.

Table 4: Evaluation for different methods on the VidOR validation dataset

Method	relation detection			relation tagging		
	mAP	R@50	R@100	P@1	P@5	P@10
VRU'19 Top-1 [23]	6.56	6.89	8.83	51.20	40.73	-
VRU'19 Top-1+MHA	6.59	6.35	8.05	50.72	41.56	32.53

4.6 Results on VidOR

Table 4 reports the results of relation detection and relation tagging on the VidOR dataset. VRU'19 Top-1 [23] utilized greedy association method. Our result is obtained based on our implementation of the relation prediction model described in [23] but using our MHA association algorithm. Comparing to VRU'19 Top-1 on relation detection, our method performs better under the evaluation metric of mAP, which shows the general effectiveness of MHA. However, MHA performs worse on the results of Recall. On relation tagging, the results of both methods are close. Our method achieves better Precision@5 result while VRU'19 Top-1 achieves better Precision@1 result. The Precision@10 result of VRU'19 Top-1 is not reported.

The following reasons may explain the unsatisfying results on relation detection. First, the performance of the relation prediction model we trained is a bit worse than the original paper when testing with greedy association, which means that our relation prediction result before association is worse than VRU'19 Top-1. Second, as we mentioned before, [23] performs multi-object tracking in the first step and performs association between each object pair in the last step, which is not the same as usual pipeline. When the results of multi-object tracking are not reliable enough, this kind of process may lead to more severe inaccuracy in object trajectory detection. However, the object trajectory results by [23] are satisfactory. In this condition, performing association between object pairs is better than performing association on whole video, because the trajectories of different objects will less likely be confused during association. The confusion of object trajectories results in the inaccurate trajectories of the final video relations and makes the Recall of relation detection low.

4.7 Qualitative Results

The qualitative results on the ImageNet-VidVRD dataset are shown in Figure 3 where the correct relation detection results among the

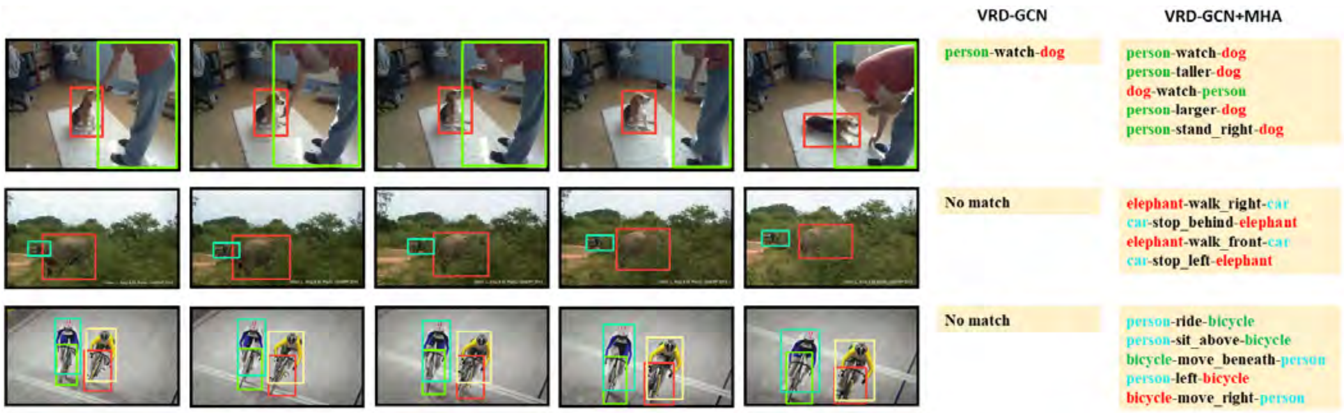

Figure 3: Visualization of relation detection results on the ImageNet-VidVRD dataset using VRD-GCN baseline and our VRD-GCN+MHA method. The correctly detected video relations among the top-20 detected relation results are shown.

top-20 detected results are presented. The methods for comparison are VRD-GCN and VRD-GCN+MHA. They generate exactly the same tracklets and relation segments on each video segment. From the results, we can obviously see that MHA produces more correct final relations which means more video relations are correctly generated via association and achieve a high confidence score after MHA. This shows that MHA can improve the final results by a lot as compared to the greedy association method. The second examples shows the situation when missing tracklet detection happens. As a result, detection results of VRD-GCN are mostly split into fragments, which lead to zero match with the ground truth relations. With MHA, our method successfully resolve this problem and avoid video relations from being interrupted. In the third example, multiple objects exist in the video and have close geometric positions, which may lead to confusion during association. The results of VRD-GCN include a few correct relation triplets but all of them are with incorrect trajectories, which leads to zero match with the ground truth relations. By applying the idea of multiple hypothesis, our MHA maintains all the confusing probable connections and delays the connection decisions until more information arrives. The results show the effectiveness of our method.

Figure 4 shows an representative failure example produced by VRD-GCN+MHA on the ImageNet-VidVRD dataset, where VRD-GCN+MHA detects no correct relation in the top-20 results. Five ground-truth relations and five relations with high confidence score detected by VRD-GCN+MHA are presented. This example exposes the following two problems that MHA cannot resolve: (1) false prediction of subject/object or predicate; (2) irrelevant tracklet detection. As shown in the figure, a giant panda is wrongly predicted to be "person" and all predicted predicate are false. MHA cannot modify the relation triplets during association, so it cannot deal with this situation. The predicted predicates shown in the figure also indicate that the model is more likely to predict easy predicates, like "watch", "larger", etc, while complex predicates like "jump_toward" are hard to predict. Besides, the object "ball" predicted by VRD-GCN+MHA comes from an irrelevant tracklet which cannot be distinguished by MHA. For further improving the performance on VidVRD, these problems need to be solved well.

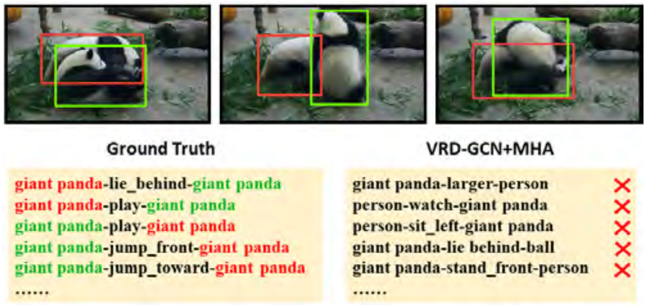
Ground Truth
VRD-GCN+MHA
giant panda-lie_behind-giant panda
giant panda-play-giant panda
giant panda-play-giant panda
giant panda-jump_front-giant panda
giant panda-jump_toward-giant panda
......
giant panda-larger-person ✗
person-watch-giant panda ✗
person-sit_left-giant panda ✗
giant panda-lie behind-ball ✗
giant panda-stand_front-person ✗
......

Figure 4: A representative failure case produced by VRD-GCN+MHA on the ImageNet-VidVRD dataset.

5 CONCLUSION

In this paper, we proposed a novel relation association method MHA for VidVRD. MHA generates dynamic relation hypothesis trees to track and maintain multiple hypothesis of relations in order to deal with the inaccurate or missing detection problem when detecting tracklets and predicting relations. The competitive experimental results on both ImageNet-VidVRD and VidOR dataset indicate the effectiveness of our method and prove that the idea of multiple hypothesis can indeed play an important role in VidVRD. However, our method still have much room for improvement on both relation detection and tagging tasks. In future, we will explore how to utilize language or other information to correct the false relationship predicted on segments in the relation association step, to further tolerate the unsatisfactory results in former steps and improving the final results.

ACKNOWLEDGMENTS

This work was supported in part by National Key Research and Development Program of China under Grant 2018YFB1004300 and the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

[1] Donglin Di, Xindi Shang, Weinan Zhang, Xun Yang, and Tat-Seng Chua. 2019. Multiple Hypothesis Video Relation Detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 287–291.

[2] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6047–6056.

[3] Takahiro Hasegawa, Mitsuru Ambai, Kohta Ishikawa, Gou Koutaki, Yuji Yamauchi, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2015. Multiple-Hypothesis Affine Region Estimation with Anisotropic LoG Filters. *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 585–593.

[4] Yaosi Hu, Zhenzhong Chen, Zheng-Jun Zha, and Feng Wu. 2019. Hierarchical Global-Local Temporal Modeling for Video Captioning. In *Proceedings of the 27th ACM International Conference on Multimedia*. 774–783.

[5] Seong Jae Hwang, Sathya Ravi, Zirui Tao, Hyunwoo J. Kim, Maxwell D. Collins, and Vikas Singh. 2018. Tensorize, Factorize and Regularize: Robust Visual Relationship Learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 1014–1023.

[6] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. 2017. Object detection in videos with tubelet proposal networks. In *Proc. CVPR*, Vol. 2. 7.

[7] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. 2015. Multiple Hypothesis Tracking Revisited. *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 4696–4704.

[8] Chanho Kim, Fuxin Li, and James M. Rehg. 2018. Multi-object Tracking with Neural Gating Using Bilinear LSTM. In *ECCV*.

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

[10] Dae-Youn Lee, Jae-Young Sim, and Chang-Su Kim. 2015. Multihypothesis trajectory analysis for robust visual tracking. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 5088–5096.

[11] Chen Li and Gim Hee Lee. 2019. Generating Multiple Hypotheses for 3D Human Pose Estimation With Mixture Density Network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 9879–9887.

[12] Dong Li, Ting Yao, Zhaofan Qiu, Houqiang Li, and Tao Mei. 2019. Long Short-Term Relation Networks for Video Action Detection. In *Proceedings of the 27th ACM International Conference on Multimedia*. 629–637.

[13] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual Relationship Detection with Language Priors. *ArXiv* abs/1608.00187 (2016).

[14] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

[15] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2019. Detecting Unseen Visual Relations Using Analogies. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 1981–1990.

[16] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. 2019. Video relation detection with spatio-temporal graph. In *Proceedings of the 27th ACM International Conference on Multimedia*. 84–93.

[17] Christian Rupprecht, Iro Laina, Robert S. DiPietro, and Maximilian Baust. 2016. Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses. *2017 IEEE International Conference on Computer Vision (ICCV)* (2016), 3611–3620.

[18] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 279–287.

[19] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *Proceedings of the 25th ACM international conference on Multimedia*. 1300–1308.

[20] Xindi Shang, Tongwei Ren, Hanwang Zhang, Gangshan Wu, and Tat-Seng Chua. 2017. Object trajectory proposal. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 331–336.

[21] Xindi Shang, Junbin Xiao, Donglin Di, and Tat-Seng Chua. 2019. Relation Understanding in Videos: A Grand Challenge Overview. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2652–2656.

[22] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. 2018. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 318–334.

[23] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu. 2019. Video Visual Relation Detection via Multi-modal Feature Fusion. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2657–2661.

[24] Xu Sun, Yuantian Wang, Tongwei Ren, Zhi Liu, Zheng-Jun Zha, and Gangshan Wu. 2018. Object Trajectory Proposal via Hierarchical Volume Grouping. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM, 344–352.

[25] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. 2019. Video relationship reasoning using gated spatio-temporal energy graph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10424–10433.

[26] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2017), 652–663.

[27] Heng Wang and Cordelia Schmid. 2013. Action Recognition with Improved Trajectories. *2013 IEEE International Conference on Computer Vision* (2013), 3551–3558.

[28] Longyin Wen, Dawei Du, Shengkun Li, Xiao Bian, and Siwei Lyu. 2019. Learning Non-Uniform Hypergraph for Multi-Object Tracking. In *AAAI*.

[29] Xian Wu, Guanbin Li, Qingxing Cao, Qingge Ji, and Liang Lin. 2018. Interpretable video captioning via trajectory structured localization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6829–6837.

[30] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. 2019. Spatial-Temporal Relation Networks for Multi-Object Tracking. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 3987–3997.

[31] Shuangjie Xu, Daizong Liu, Linchao Bao, Wei Liu, and Pan Zhou. 2019. MHP-VOS: Multiple Hypotheses Propagation for Video Object Segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 314–323.

[32] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. 2019. STAT: spatial-temporal attention mechanism for video captioning. *IEEE transactions on multimedia* (2019).

[33] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural Motifs: Scene Graph Parsing with Global Context. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 5831–5840.

[34] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 3107–3115.

[35] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. 2019. Graphical Contrastive Losses for Scene Graph Parsing. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 11527–11535.

[36] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Video Question Answering via Hierarchical Spatio-Temporal Attention Networks.. In *IJCAI*. 3518–3524.

[37] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 408–417.