Video Visual Relation Detection via Iterative Inference

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

Yicong Li
liyicong@u.nus.edu
National University of Singapore
Singapore

Junbin Xiao
junbin@comp.nus.edu.sg
National University of Singapore
Singapore

Wei Ji[*]
jiwei@nus.edu.sg
National University of Singapore
Sea-NExT Joint Lab
Singapore

Tat-Seng Chua
dcscts@nus.edu.sg
National University of Singapore
Sea-NExT Joint Lab
Singapore

ABSTRACT

The core problem of video visual relation detection (VidVRD) lies in accurately classifying the relation triplets, which comprise of the classes of subject and object entities, and the predicate classes of various relationships between them. Existing VidVRD approaches classify these three relation components in either independent or cascaded manner, thus fail to fully exploit the inter-dependency among them. In order to utilize this inter-dependency in tackling the challenges of visual relation recognition in videos, we propose a novel iterative relation inference approach for VidVRD. We derive our model from the viewpoint of joint relation classification which is light-weight yet effective, and propose a training approach to better learn the dependency knowledge from the likely correct triplet combinations. As such, the proposed inference approach is able to gradually refine each component based on its learnt dependency and the other two's predictions. Our ablation studies show that this iterative relation inference can empirically converge in a few steps and consistently boost the performance over baselines. Further, we incorporate it into a newly designed VidVRD architecture, named VidVRD-II (Iterative Inference), which generalizes well across different datasets. Experiments show that VidVRD-II achieves the start-of-the-art performance on both of ImageNet-VidVRD and VidOR benchmark datasets.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**.

KEYWORDS

video understanding, visual relation, relation inference, knowledge

ACM Reference Format:
Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2021. Video Visual Relation Detection via Iterative Inference. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475263

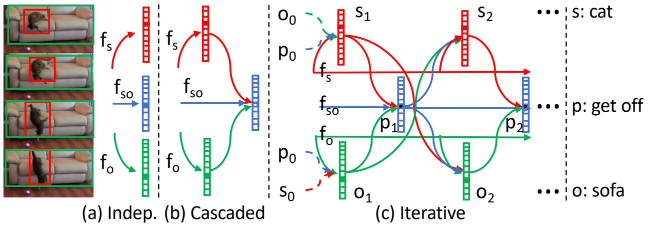

Figure 1: Illustration of different relation inference approaches for VidVRD. (a) Independent classification of subject (s), object (o) and predicate (p). (b) Cascaded classification. (c) Our proposed iterative relation inference.

1 INTRODUCTION

Comprehensive video understanding is crucial to bridging the gap between vision and language. The recent technique has been unprecedentedly advanced by the progress of video object detection [6, 14, 29, 51], segmentation [21, 23] and action detection [9, 16, 32, 35] etc. However, these approaches mainly provide entity-level visual information, which can hardly describe the various relationships and interactions between the object entities. To steadily advance video understanding towards a more comprehensive level, recent research efforts have investigated the more challenging task of detecting visual relations in video [20, 25, 31, 38, 39].

Video visual relation detection (VidVRD) focuses on recognizing triplets in the form of $\langle subject, predicate, object \rangle$ for each detected object pair in the video. This involves correctly classifying all the three components in the relation triplet, which is quite challenging due to the following reasons. (1) The latent pattern is hard to learn because of large variance in predicate representation. Classifying even a simple predicate like "walk past" need to consider many factors simultaneously, such as the visual feature, relative position changes and the ambiguity with (near-)synonyms (*e.g.* "move past"). (2) The object entities in videos are not always visually clear for classification. This is mainly because their poses may change frequently, as well as the many motion blurring, occlusion and illumination issues in videos [51]. (3) Last but not least, the problem is highly affected by the bucket effect. Just a mistake on any component will

[*]Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8651-7/21/10.
https://doi.org/10.1145/3474085.3475263

cause the task to fail and may change the final semantics drastically, or even produce an unreasonable relation triplet.

Existing approaches classify the three components of relations either in an independent (Figure 1 (a), *e.g.*[20, 25, 28]) or cascaded (Figure 1 (b), *e.g.*[2, 47]) manner. The limitation of the independent approach is because the predicate classifier has been found to be hard to learn with purely visual appearance features [22]. More preferably, the cascaded approach takes advantage of language priors by firstly classifying the subject/object and then the predicate based on the subject/object classes. While such approach has demonstrated its superiority for image data, its performance for videos is far from satisfactory. This is because the approach heavily relies on robust subject/object classification, which is not of high quality for videos as compared to images.

Given the likely errors and ambiguities in subject/object recognition, and their impact on predicate prediction, as well as the intuition that their classes are likely to be inter-dependent, we explore the joint relation classification approach in this paper. In fact, when classifying an object (subject) that is visually unclear, it is also important and necessary to also consider the related subject (object) and predicate classes. This is because given the subject and predicate (*e.g.*"person-eat"), only a constraint set of classes can be predicted as the object (*e.g.*"rice" rather than "plate"). This observation is helpful to resolve the visual ambiguity. While such intuition has been utilized in the cascaded approach to classify the predicate, the idea is not fully realized because none of the components are accurately classified or refined in advance.

In order to address this critical issue, we propose an iterative relation inference approach (Section 3) to gradually refine the class of each component based on that of the other two. As shown in Figure 1 (c), the subject/object and predicate classes will be alternatively updated at each iteration by utilizing the dependency among them. This inter-dependency is modelled by three novel preferential predictors with learnable tensors alongside the normal visual predictors (Section 3.2). We also propose a novel training approach for the preferential predictors to better learn the dependency knowledge from the likely correct triplet combinations (Section 3.3). Overall, our iterative relation inference approach can be viewed as the joint classification of the relation triplet. We empirically show that the inference process can be efficiently converged in a few iterative steps. Eventually, we propose an improved VidVRD architecture and incorporate the proposed relation inference (Section 4). We term our final approach VidVRD-II (**I**terative **I**nference).

The contributions of this paper are as follows: (1) we propose an iterative relation inference that can exploit the inter-dependency of relation components for better visual relation recognition; (2) we propose an effective training approach to better learn the dependency knowledge from the training data; and (3) we devise VidVRD-II, an improved VidVRD architecture that steadily and significantly improves the results on ImageNet-VidVRD and achieves competitive performances on the challenging VidOR dataset.

2 RELATED WORK

A majority of studies on visual relation focus only on images (ImgVRD). The initial works mainly address the problem of modelling the huge label space of relation [7, 18, 22, 48]. Some works have also explored the pruning of unlikely relation candidates for efficiency [17, 43] or attempted to improve the training objectives [3, 5, 10, 24, 49]. Compare to ImgVRD, VidVRD is more difficult because the visual relations in video are more diverse (*e.g.* relations involving actions or position changes). Most existing works [2, 8, 25, 31, 33, 38, 41] in this area followed the three-stage detection framework proposed in [28]. Some recent works explored end-to-end feature learning [2] or better temporal relation localization algorithms [8, 31].

Further, many recent works have leveraged contextual information for visual relation detection. To incorporate contextual cues, RNN and GCN have been utilized to model and iteratively propagate information among the detected object and relationships, such as the works done in the image domain [4, 42, 47]. For video, Qian *et al.* [25] built a spatio-temporal graph convolutional networks within adjacent video segments to refine the object and relation feature. Sun *et al.* [33] utilized language context feature along with spatial-temporal feature for predicate prediction. To capture relations involving long motions, Liu *et al.* [20] proposed a sliding-window scheme to predict short-term and long-term relationships simultaneously, and use one spatial and one temporal graph to generate the contextual embedding for tracklet proposal compatibility evaluation.

Besides, utilizing the inter-dependency among the relation components [7, 13, 36] is also a crucial and fundamental problem in visual relation recognition. Dai *et al.* [7] proposed to mimic the statistical relation inference through iterative unrolling the procedure into a deep neural network. Hwang *et al.* [13] learnt a separate triplet dependency tensor to regularize the training of visual recognition module. Tang *et al.* [36] developed a cause-effect framework to mitigate the bias from over reliance on statistical dependency rather than visual information in the existing approaches. However, these works mainly studied the problem in the image domain. Another relevant work is done by Tsai *et al.* [38] who studied the problem of utilizing the context for video relation inference. They proposed a complicated spatio-temporal energy graph to model the *global* dependency among all the entities and relationships, but the inter-dependency among the relation components is still rarely explored. Different from these works, our work is the first to explore the inter-dependency in the video domain. As it is more challenging and complicated to incorporate the inter-dependency and the visual information provided in videos, we explore the inter-dependency from a more fundamental viewpoint of joint relation classification and derive a simple yet effective model.

3 ITERATIVE RELATION INFERENCE

3.1 Problem Formulation

Given a pair of detected subject and object entities (e_s, e_o) with their entity features f_s, f_o and relation feature f_r, the problem of visual relation recognition is to model the joint probability $P(\langle s, p, o\rangle | f_s, f_r, f_o)$, where $\langle s, p, o\rangle$ belongs to a relation triplet space $\mathcal{X}\times\mathcal{Y}\times\mathcal{X}$, and $\mathcal{X}, \mathcal{Y}$ are the pre-defined sets of subject/object classes and predicate classes, respectively.

Existing works mainly factorize the above joint probability as $P(s|f_s)P(p|f_r)P(o|f_o)$ or $P(p|f_r, s, o)P(s|f_s)P(o|f_o)$, resulting in the independent or cascaded relation classifier, respectively. As can be seen, the independent relation classifier holds a strong assumption that sufficient visual information is always available for every component classifier to work well. The cascaded relation classifier relaxes such assumption for the predicate, yet still holds it for the subject and object (*i.e.* $P(s|f_s), P(o|f_o)$). However, as discussed in Section 1, such assumption is often not valid in practice. We propose to factorize the joint probability as three conditional probabilities:

$$P(s|f_s, p, o)P(p|f_r, s, o)P(o|f_o, s, p). \tag{1}$$

This is better because the classes of any two components imply certain preference over the class of the third one, and thus can help the inference when the visual information is insufficient or ambiguous. Our experiments in Section 5.3 validate that this novel formulation can lead to better visual relation recognition in videos.

3.2 Model

To model the above conditional probabilities, we use three classifiers for each of the relation components, and each of them consists of a visual predictor and a preferential predictor. The visual predictor can be any deep neural network for recognizing the visual patterns of subject/object and predicate. For simplicity, we use a single linear layer here. On the other hand, the preferential predictor is introduced for refining the prediction of one variable conditioned on the values of the other two variables, based on learnable dependency tensors.

Specifically, for subject (likewise for object), the classifier feeds f_s to the visual predictor and the prediction probability vectors $p \in \mathbb{R}^{|\mathcal{Y}|}, o \in \mathbb{R}^{|\mathcal{X}|+1}$ to the preferential predictor. Then the subject's prediction probability vector $s \in \mathbb{R}^{|\mathcal{X}|+1}$ including a background class, is computed as the sum of the predicted visual and preferential scores:

$$\begin{aligned} P(s|f_s, p, o) &= \varsigma(\mathbf{V}_e f_s + p\mathbf{W}_s o) \\ P(o|f_o, s, p) &= \varsigma(\mathbf{V}_e f_o + s\mathbf{W}_o p), \end{aligned} \tag{2}$$

where ς is the softmax function and $\mathbf{V}_e$ is the learnable weight of the visual predictor which is shared by the subject and object classifier. $\mathbf{W}_s \in \mathbb{R}^{|\mathcal{Y}|\times(|\mathcal{X}|+1)\times(|\mathcal{X}|+1)}$ is a 3-order tensor which can map two given vectors to a new vector through tensor product, and model the dependency of the subject class over the predicate and object classes (likewise for $\mathbf{W}_o \in \mathbb{R}^{(|\mathcal{X}|+1)\times(|\mathcal{X}|+1)\times|\mathcal{Y}|}$). Similarly, we formulate the predicate classifier as:

$$P(p|f_r, s, o) = \sigma(\mathbf{V}_r f_r + s\mathbf{W}_p o), \tag{3}$$

where $\mathbf{W}_p \in \mathbb{R}^{(|\mathcal{X}|+1)\times|\mathcal{Y}|\times(|\mathcal{X}|+1)}$, and σ is the sigmoid function as there might be multiple relationships between the subject and object.

3.3 Learning Dependency Knowledge

Learning effective tensors $\mathbf{W}_s, \mathbf{W}_o, \mathbf{W}_p$ is important for our model to capture the dependency among the relation components and refine them through iterative inference (Section 3.4). Intuitively, one possible approach is to consider the relation triplets seen in the training data as the training corpus, and then apply n-gram language modelling to encode the dependency knowledge into $\mathbf{W}_*$. However, since $\mathbf{W}_*$ is obtained separately, the resulting preferential predictor is not optimal for the visual predictor. On the contrary, another possible approach is to learn $\mathbf{W}_*$ to directly refine the prediction of the visual predictors, which follows the idea of [7]. This approach first computes the visual prediction $\hat{p}, \hat{o}$ and then optimizes $\mathbf{W}_s$ according to $\mathcal{L}(s^*, P(s|f_s, \hat{p}, \hat{o}))$ where $\mathcal{L}$ is the loss function and s^* is the ground truth class (likewise for the other two). Though $\mathbf{W}_*$ can be jointly learnt, they may not encode the dependency knowledge well because of the noise in visual predictions.

In order to tackle the drawbacks of the above training approaches, we propose a better approach to learn the dependency knowledge for $\mathbf{W}_*$. Specifically, our approach optimizes $\mathcal{L}(s^*, P(s|f_s, p^*, o^*))$, $\mathcal{L}(o^*, P(o|f_o, s^*, p^*))$ and $\mathcal{L}(p^*, P(p|f_r, s^*, o^*))$ where s^*, p^*, o^* are the ground truth classes. This permits learning the dependency directly from the ground truth triplet combination as well as jointly optimizing with the visual predictors. From the viewpoint of masked language modelling which is widely adopted in training BERT models, this proposed training approach borrows a similar idea where the target token is explicitly masked out in the input and predicted based on the rest of tokens as well as the visual context. Therefore, the approach is able to learn the dependency knowledge more effectively from the training data, which is validated in our ablation study. To avoid overfitting, we apply dedicated weight decay α to $\mathbf{W}_*$ during training. Also, it is worth noting that some entries of $\mathbf{W}_*$ will never be updated as some pair combinations among s^*, p^*, o^* do not exist in the training data. Hence, we set these entries to the median value of those updated entries after training.

3.4 Iterative Inference

Empirically, it is hard to exactly estimate the joint probability $P(\langle s, p, o\rangle|f_s, f_r, f_o)$ in Eq.(1), as none of the conditional variables s, p, o can be accurately estimated in advance. Therefore, we propose an iterative inference for our model, which can lead to an approximate estimation of the joint probability. The inference is started by setting the initial conditional variables as follows:

$$s_{(0)} = o_{(0)} = p_{(0)} = \mathbf{0}. \tag{4}$$

In this way, the subject and object classifiers in Eq.(2) are degenerated to vanilla visual predictors, which predict simply based on the visual features f_s, f_o. Then the three conditional probabilities can be estimated through iterative steps:

$$\begin{aligned} s_{(i+1)} &= P(s|f_s, p_{(i)}, o_{(i)}) \\ o_{(i+1)} &= P(o|f_o, s_{(i)}, p_{(i)}) \\ p_{(i+1)} &= P(p|f_r, s_{(i+1)}, o_{(i+1)}). \end{aligned} \tag{5}$$

As can be seen, at each step, the prediction of the subject and object are first refined from the previous step, and then used for refining the prediction of the predicate. According to the experiments, with a few more steps, our approach can generally converge and improve over the 1-step inference. After N iterative steps, we can generate relation triplets for the given entity pair. We take the subject, object and predicate classes with probability (Eq.(2-3)) larger than a threshold $\delta = 0.2$ as the predictions.

Temporal Initialization. Eq.(4) initializes the iterative inference purely from the visual information. As for video, however, it is also intuitive to utilize the predictions in the previous time window for the initialization. First, for each given pair of subject and object entity $(e_s^{(t)}, e_o^{(t)})$, we use the entities in the last time window to

Algorithm 1 Iterative Relation Inference in a Video

```
Parameter: the maximum number of inference steps $N$
Input: a set of entity pair $\mathcal{P}^{(t)} = \{(e_s^{(t)}, e_o^{(t)})\}$
Output: the relation triplets of each entity pair
for $t = 1, \ldots, T$ do
    for $(e_s^{(t)}, e_o^{(t)})$ in $\mathcal{P}^{(t)}$ do
        Build linkage graph $G$ from $\mathcal{P}^{(t-1)}$
        Initialize conditional probabilities $s_{(0)}, p_{(0)}, o_{(0)}$
        if $G = \emptyset$ then
            Initialize by Eq.(4)
        else
            Initialize from $\mathcal{R}^{(t-1)}$ and $G$ by Eq.(6)
        end if
        for $i = 1, \ldots, N$ do
            Update conditional probabilities by Eq.(5)
        end for
        Generate relation triplets from $s_{(N)}, p_{(N)}, o_{(N)}$
        Add the results to $\mathcal{R}^{(t)}$
    end for
end for
```

build a corresponding **linkage graph**. The entities that sufficiently overlap $e_s^{(t)}$ (IoU>0.7), which is denoted as $E_s^{(t-1)}$, will be linked to $e_s^{(t)}$ in the graph. Each linking edge is assigned with a weight by the corresponding IoU value. Likewise, the entities in $E_o^{(t-1)}$ are linked to $e_o^{(t)}$. Also the relationship between each pair $e_s^{(t-1)} \in E_s^{(t-1)}$ and $e_o^{(t-1)} \in E_o^{(t-1)}$ will be regarded as a node and linked to the relationship node of $(e_s^{(t)}, e_o^{(t)})$, $r^{(t)}$. We denote those relationship nodes as $R^{(t-1)}$, and the weight of these linking edges are assigned by the corresponding subject or object IoU value, whichever is smaller. Then, we propagate the prediction probability of the nodes in $E_s^{(t-1)}, E_o^{(t-1)}, R^{(t-1)}$ to the corresponding node at t. We aggregate the probabilities through weighted average based on the edge weights. Finally, the aggregated probability is assigned to the initial conditional variable in the iterative inference for the given pair. For example,

$$s_{(0)} = \sum_{e_s^{(t-1)}} \frac{exp(\text{IoU}(e_s^{(t-1)}, e_s^{(t)}))}{\sum_{e'^{(t-1)}_s} exp(\text{IoU}(e'^{(t-1)}_s, e_s^{(t)}))} s_{(N)}(e_s^{(t-1)}), \tag{6}$$

where $s_{(N)}(e_s^{(t-1)})$ is the estimated subject probability for $e_s^{(t-1)}$. Nevertheless, if the given pair $(e_s^{(t)}, e_o^{(t)})$ does not have any linked predictions in the last time window, we still initialize it as in Eq.(4).

4 ARCHITECTURE OF VIDVRD-II

Algorithm 1 summarizes the procedure of the proposed iterative relation inference for a video. To address the VidVRD task completely, we integrate the algorithm into the following proposed architecture, namely VidVRD-II. As shown in Figure 2, our architecture follows the similar framework as proposed in VidVRD [28], but introduces several improvements for generalization.

Sliding Time Window. Given a video of any length, we apply sliding time windows with the size of 30 frames and stride of 15 frames for subsequent processing. This is similar to [28] but we use the terminology of time window as it could better illustrate the overall framework. In each given time window, we will generate the object tracklet proposals as the detected entities, and predict the relation triplets for each pair of them.

Object Tracklet Proposal. We resort to the lightweight Seq-NMS [12] instead of complex tracking techniques for tracklet generation. In particular, for frame-level object detection, we adopt Faster-RCNN with Inception-ResNet backbone [34] and the pretrained model on the Open Images dataset [15]. As Open Images dataset covers a wide range of object classes (600) for training the object detection model, the resulting model can be better generalized to more domains and datasets. Since the model is regarded as a generic object detector in our architecture, we only use the output bounding boxes and the corresponding region features for subsequent processing. Then we use Seq-NMS to generate a compact set of object tracklets. We simply use the average region features from the comprising bounding boxes as the visual feature for each tracklet, as we do not focus on the sophisticated visual feature extraction in this paper.

Relation Triplet Classification. For each pair of object tracklets, we form the subject and object entity pair for relation feature (f_r) extraction and iterative relation inference with $N = 3$ steps (Section 3). To make the model aware of the relative temporal distance between the subject and object, we encode both the spatial and temporal relative positions into the relation feature. Specifically, we compute the relative positional feature between the beginning bounding boxes of the subject and object as:

$$f_r^{PB} = \left[\frac{x_s - x_o}{x_o}, \frac{y_s - y_o}{y_o}, \log \frac{w_s}{w_o}, \log \frac{h_s}{h_o}, \log \frac{w_s h_s}{w_o h_o}, \frac{t_s - t_o}{30} \right], \tag{7}$$

where $(x_s, y_s, w_s, h_s, t_s)$ is the subject's beginning bounding box coordinates and its frame ID in the time window (likewise for $(x_o, y_o, w_o, h_o, t_o)$). We also compute the relative positional feature f_r^{PE} for the ending bounding boxes of the pair in the same way. To obtain the final relation feature f_r, we fuse the subject and object's visual features f_s, f_o and f_r^{PB}, f_r^{PE} using a 2-layer feed forward network (FFN)[1]. During model training, we adopt focal loss [19] to mitigate the effect of imbalanced class distribution in the training data.

Training Entity Pair Sampling. It is worth noting that the generated object tracklets are not guaranteed to be temporally aligned with the time window (*i.e.* begin and end at same frames). However, previous works commonly omit this and simply assume that the tracklets expand to the whole time window. To consider this in our architecture, we generalize the training entity pair sampling strategy as follows. We compare each entity pair with the ground truths. If the pair is sufficiently enclosed by some ground truth pair, we regard it as the positive training sample for that ground truth and use the corresponding relation triplet as the training target. In particular, we compute the Proportion of Intersection (PoI) for each tracklet w.r.t. the corresponding ground truth:

$$\text{PoI}(\mathcal{T}, \mathcal{T}^{gt}) = \frac{\text{Intersection}(\mathcal{T}, \mathcal{T}^{gt})}{\text{Volume}(\mathcal{T})}. \tag{8}$$

[1] f_s and f_o will be concatenated and fed into the first layer of FFN. Its output will be further concatenated with f_r^{PB} and f_r^{PE} and then fed into the second layer of FFN.

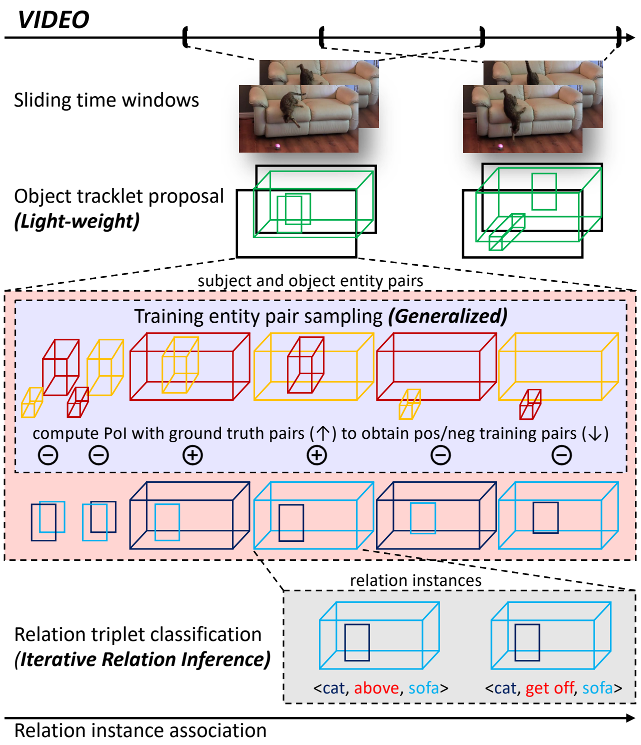
VIDEO
Sliding time windows
Object tracklet proposal
(Light-weight)
subject and object entity pairs
Training entity pair sampling *(Generalized)*
compute PoI with ground truth pairs (↑) to obtain pos/neg training pairs (↓)
⊖ ⊖ ⊕ ⊕ ⊖ ⊖
relation instances
Relation triplet classification
(Iterative Relation Inference)
<cat, above, sofa>
<cat, get off, sofa>
Relation instance association

Figure 2: Our VidVRD-II architecture with three main improvements: 1) lightweight object tracklet proposal; 2) generalized training entity pair sampling strategy; and 3) iterative relation inference.

We require both PoI($\mathcal{T}_s, \mathcal{T}_s^{gt}$) and PoI($\mathcal{T}_o, \mathcal{T}_o^{gt}$) to be larger than a threshold $\rho = 0.9$.

Relation Instance Association. Following [28], we denote a visual relation instance as $(\langle s, p, o\rangle, (t_b, t_e), (\mathcal{T}_s, \mathcal{T}_o))$, where t_b and t_e are the beginning and ending frame of the relation, and $\mathcal{T}_s$ and $\mathcal{T}_o$ are the bounding box trajectories of the subject and object during (t_b, t_e), respectively. Thus, the performance of the system can be measured by evaluating how many ground truth relation instances are successfully detected (detailed in Section 5.2). Note that the different relations between the same subject and object pair are regarded as different instances. This is mainly because different relations normally exist in different time periods of the video. In order to temporally localize the relation instances, we perform simple greedy relation association as proposed in [28] to associate the detected relation instances across the time windows.

5 EXPERIMENTS

5.1 Datasets

We conduct experiments on the two VidVRD benchmark datasets: ImageNet-VidVRD [28] and VidOR [27]. ImageNet-VidVRD is the first dataset for VidVRD, which has been widely used by the previous works. It consists of 1,000 videos collected from ILSVRC2016-VID [26] and is manually annotated with video relation instances. VidOR is a recently released large-scale benchmark. It contains 10,000 social media videos from YFCC-100M [37], whose contents are more human-centric and complicated. Since the dataset is used by Video Relation Understanding (VRU) grand challenge [30], its testing set is not publicly available. Therefore, we use the validation set for testing in our experiments. Table 1 summarizes the statistics of the two datasets. Compare to ImageNet-VidVRD, the videos in VidOR are generally longer and annotated with more relation instances.

	ImageNet-VidVRD	VidOR
# hours	3.0	98.6
# training videos	800	7,000
# validation videos	-	835
# testing videos	200	2,165
# object categories	35	80
# object instances	3,017	49,258
# predicate categories	132	50
# relation instances	4,835	378,546

Table 1: Summary of the statistics of the ImageNet-VidVRD and VidOR datasets.

5.2 Evaluation Protocol

Same as prior works, our experiments adopt the evaluation protocol proposed in [28]. The protocol basically evaluates how many ground truth relation instances are detected by the approach in each testing video. In particular, there are two groups of metrics: relation detection and relation tagging metrics. The relation detection metrics measure the precision of both relation triplet and the corresponding subject/object trajectories for each detected relation instance. A detected relation instance is considered to be correct if it has the same relation triplet with a ground truth relation instance and their trajectory vIoU (voluminal Intersection-over-Union) of the subject and object are both larger than 0.5. Mean Average Precision (mAP), Recall@50 (R@50), Recall@100 (R@100) are specifically used to evaluate the detection performance. The relation tagging metrics instead focus only on the precision of relation triplet without considering the precision of its spatio-temporal location in the video. It is evaluated by the Precision@1 (P@1), Precision@5 (P@5) and Precision@10 (P@10). To reduce the randomness introduced in the experiments, we run all of our experiments 5 times with different random seeds and report the mean/std. scores.

5.3 Comparison with Non-Iterative Inference

To demonstrate the effectiveness of our proposed iterative relation inference, we compare with the non-iterative baselines, the independent and cascaded inference approaches, as formulated in Section 3.1. We implement them in our VidVRD-II architecture for fair comparison. From Table 2, under all the evaluation metrics and benchmark datasets, we can clearly find that iterative relation inference consistently outperforms those non-iterative baselines. The improvements under the relation tagging metrics are more significant. For ImageNet-VidVRD, our approach improves over the independent baseline by 3.30% and 3.70% under P@1 and P@5, respectively. For the challenging VidOR, our approach also improves over the independent baseline by 1.44% under P@1. These comparative results indicate the effectiveness of iterative relation inference in recognizing visual relations in videos.

	Relation Detection			Relation Tagging		
	mAP	R@50	R@100	P@1	P@5	P@10
ImageNet-VidVRD Dataset						
Independent Baseline	27.49 ± 0.47	18.18 ± 0.41	21.28 ± 0.56	67.10 ± 1.02	50.18 ± 1.28	38.02 ± 0.90
Cascaded Baseline	27.27 ± 0.58	18.08 ± 0.41	21.03 ± 0.44	67.00 ± 2.00	50.74 ± 0.89	38.15 ± 0.68
VidVRD-II (Ours)	**29.37 ± 0.40**	**19.63 ± 0.19**	**22.92 ± 0.48**	**70.40 ± 1.53**	**53.88 ± 0.31**	**40.16 ± 0.70**
VidOR Dataset						
Independent Baseline	8.52 ± 0.09	8.47 ± 0.09	10.53 ± 0.11	55.96 ± 0.51	43.81 ± 0.37	32.89 ± 0.37
Cascaded Baseline	8.50 ± 0.09	8.52 ± 0.06	10.58 ± 0.06	56.25 ± 0.68	43.84 ± 0.62	32.97 ± 0.26
VidVRD-II (Ours)	**8.65 ± 0.11**	**8.59 ± 0.11**	**10.69 ± 0.08**	**57.40 ± 0.57**	**44.54 ± 0.68**	**33.30 ± 0.31**

Table 2: Comparison with the non-iterative inference baselines on ImageNet-VidVRD and VidOR datasets.

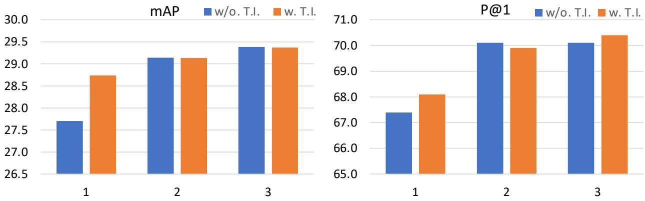

Figure 3: Ablation on the number of iterative inference steps (1-3) and the use of temporal initialization (T.I.). Results of mAP and P@1 are reported respectively.

	mAP	R@50	R@100	P@1	P@5	P@10
refine	25.89	16.89	19.89	66.60	46.74	34.54
n-gram	28.31	19.04	22.20	69.00	52.06	38.98
ours	**29.37**	**19.63**	**22.92**	**70.40**	**53.88**	**40.16**

Table 3: Ablation on different training approaches for learning the dependency knowledge.

	ImageNet-VidVRD			VidOR		
	mAP	R@50	P@1	mAP	R@50	P@1
$\alpha = 0.01$	**29.37**	**19.63**	**70.40**	8.47	8.53	57.00
$\alpha = 0.1$	28.58	19.38	70.40	8.43	8.57	57.52
$\alpha = 1$	28.04	18.50	69.30	**8.65**	**8.59**	57.40

Table 4: Ablation on different values of weight decay α in our training approach for different datasets.

5.4 Ablation Studies

We then conduct ablation studies to better understand the effect of several important parameter settings and components.

Number of Iterative Steps. We test VidVRD-II with different iterative steps ($N = 1, 2, 3$). Figure 3 shows the corresponding performances of mAP and P@1 on ImageNet-VidVRD either using the temporal initialization (orange bars) or not (blue bars). In general, we can find that the 2-step inference significantly improves over the 1-step inference and the process can converge within 3 steps. This indicates the importance of performing multiple inference steps to refine each component based on the other two of relation triplet. We choose 3-step inference as the default setting for our approach.

Temporal Initialization. From Figure 3, we can also see that the 1-step inference with temporal initialization significantly improves over that without the temporal initialization. However, such improvement cannot be observed for the inference with more iterative steps. This suggests that the main role of this component is in kick-starting the iterative inference with better information but does not provide long-term benefit.

Effectiveness of the Proposed Learning Approach. Section 3.3 addresses the weaknesses of two standard training approaches and proposes a novel training approach to better learn the dependency knowledge for $\mathbf{W}_*$. We compare them in Table 3. In particular, we denote the approach to separately learning $\mathbf{W}_*$ using n-gram language modelling as *n-gram*, and the approach to jointly learning to refine the visual prediction as *refine*. To implement *n-gram*, we first construct $\mathbf{W}_s$ (likewise for $\mathbf{W}_o, \mathbf{W}_p$) by filling it with the probability $P(s|p,o) = (c(s,p,o)+1)/(c(p,o)+d)$ where $c(s,p,o)$ is the frequency of a specific relation triplet appeared in the training data, likewise for $c(p,o)$. d indicates the uniform probability $1/d$ of Laplace smoothing when there lacks observation of (p,o), whose value is set to the corresponding number of classes. Then, $\mathbf{W}_s$ is multiplied with a learnable scalar in Eq.(2) to learn the weight of how much this estimated dependency knowledge should be used by the model. For all of the compared training approaches, we adopt Adam optimizer with learning rate of 0.001 and train the model with 50 epochs. From Table 3, we can see that our proposed training approach can generally achieve better performance.

Weight Decay in Learning $\mathbf{W}_*$. Table 4 shows the results of using different α in learning $\mathbf{W}_*$. For ImageNet-VidVRD, it can be seen that the performance is generally better when using a smaller α. However, for VidOR, the small α can lead to performance degradation. This is because of the severe imbalance of classes in VidOR which can make the model overfit the dominant classes. Thus, we find that weight decay is particularly useful when learning $\mathbf{W}_*$ in practice where class imbalance is unavoidable. We empirically set $\alpha = 0.01$ for ImageNet-VidVRD and $\alpha = 1$ for VidOR.

PoI Threshold in Training Entity Pair Sampling. Table 6 shows the performances of using different PoI threshold ρ in our training entity pair sampling strategy (Section 4) on ImageNet-VidVRD. As can be seen, there is a significant performance drop if we use a smaller threshold, typically 0.5 adopted by most previous works. We conjecture that the tracklets sampled by these smaller thresholds (*i.e.* the positive training pairs) may not sufficiently represent the motion trajectory of the ground truth entities in the video. This is possibly because a small threshold will include many inaccurate tracklets that overlap the ground truth at most frames but deviate at some important frames; and this may cause the model to wrongly

	Relation Detection			Relation Tagging		
	mAP	R@50	R@100	P@1	P@5	P@10
ImageNet-VidVRD Dataset						
VidVRD [28]	8.58	5.54	6.37	43.00	28.90	20.80
GSTEG [38]	9.52	7.05	7.67	51.50	39.50	28.23
VRD-GCN [25]	14.23	7.43	8.75	59.50	40.50	27.85
3DRN [2]	14.68	5.53	6.39	57.89	41.80	29.15
VidVRD+MHA [31]	15.71	7.40	8.58	40.00	26.70	18.25
VRD-GCN+Siamese [25]	16.26	8.07	9.33	57.50	41.00	28.50
VRD-GCN+MHA [31]	19.03	9.53	10.38	57.50	41.40	29.45
VRD-STGC [20]	18.23	11.21	13.69	60.00	43.10	32.24
VidVRD-II (Ours)	**29.37 ± 0.40**	**19.63 ± 0.19**	**22.92 ± 0.48**	**70.40 ± 1.53**	**53.88 ± 0.31**	**40.16 ± 0.70**
VidOR Dataset						
RELAbuilder [50]	1.47	1.58	1.85	33.05	35.27	-
3DRN [2]	2.47	2.58	2.75	52.59	42.33	29.89
MAGUS.Gamma [33]	6.56	6.89	8.83	51.20	40.73	-
MAGUS.Gamma+MHA [31]	6.59	6.35	8.05	50.72	41.56	32.53
VRD-STGC [20]	6.85	8.21	9.90	48.92	36.78	-
VidVRD-II (Ours)	**8.65 ± 0.11**	**8.59 ± 0.11**	**10.69 ± 0.08**	**57.40 ± 0.57**	**44.54 ± 0.68**	**33.30 ± 0.31**

Table 5: Comparison with the state-of-the-art approaches on ImageNet-VidVRD and VidOR datasets. For each dataset and metric, the best and second best scores are highlighted by bold and underline, respectively.

	mAP	R@50	R@100	P@1	P@5	P@10
$\rho = 0.5$	18.12	13.53	16.29	63.30	43.62	33.83
$\rho = 0.7$	24.38	17.26	20.82	69.90	49.78	38.40
$\rho = 0.9$	**29.37**	**19.63**	**22.92**	**70.40**	**53.88**	**40.16**

Table 6: Ablation on using different PoI threshold ρ to sample the positive training entity pairs.

infer the change in the size or direction of the motion trajectory, which may have very different semantic meaning. Therefore, to sample more representative training entity pairs, we set $\rho = 0.9$ in our approach[2].

5.5 Comparison with the State-of-the-Arts

This section compares our VidVRD-II with the state-of-the-art approaches on the benchmark datasets. For the compared approaches, we show the performances reported in their previous papers.

Compared Approaches. For ImageNet-VidVRD, the compared approaches include VidVRD [28], GSTEG [38], VRD-GCN [25], VRD-STGC [20] and 3DRN [2]. VidVRD is the first approach for VidVRD. GSTEG constructs a fully-connected spatio-temporal graph for relation inference. VRD-GCN and VRD-STGC extract spatio-temporal features for relation recognition. 3DRN develops a 3-D CNN to learn the visual features for relation recognition in an end-to-end manner. In addition, some prior works developed sophisticated relation association algorithms to better temporally localize the relation instances in the last stage of VidVRD. We also include them in our comparison, denoted as VRD-GCN+Siamese [25], VidVRD+MHA and VRD-GCN+MHA [31].

For VidOR, we mainly compare with 3DRN [2] and VRD-STGC [20], which are the only works that reported the corresponding results as far as we know. Following [20], we also compare with the results of the top 2 performing systems in the VRU challenge at ACM MM'19: MAGUS.Gamma [33] (the top) and RELAbuilder [50] (the runner-up). In addition, we include the result of combining MAGUS.Gamma and Multi-Hypothesis Association (MAGUS.Gamma+MHA [31]).

Results. Table 5 presents the comparative results. As can be seen, our VidVRD-II simultaneously achieves competitive performances on both datasets. On ImageNet-VidVRD, our approach outperforms all the compared approaches by a significantly large margin. Such large improvement comes from both the proposed iterative relation inference and our improvements in the VidVRD-II architecture. On VidOR, VRD-STGC and 3DRN only achieve the state-of-the-art performance under one group of the evaluation metrics, *i.e.* relation detection or relation tagging, respectively. However, our VidVRD-II demonstrates its superiority over them by achieving the best performance for all the metrics.

5.6 Qualitative Results

We visualize the detected visual relations by our approach in Figures 4 and 5. Most previous works only present the correctly detected relations in the qualitative examples, which can hardly show the overall quality of the output. In this paper, we directly visualize the top 10 detected relations with their localized subject and object, to analyse the effectiveness of VidVRD-II.

As can be seen in Figure 4, VidVRD-II detected more correct visual relations than that of the VidVRD baseline, and made fewer mistakes in the top results. For instance, VidVRD-II successfully detected relations like "person-feed-dog", "person-drive-car" and "person-inside-car", with high confidence scores; while VidVRD mostly detected common relations like "person-stand left-dog" and "person-taller-dog" but failed to detect those interesting relations. It is also interesting to find that, though both examples are about a

[2] $\rho = 0.9$ also leads to the stronger baselines in Section 5.3. In fact, our iterative inference can improve much more over those baselines under $\rho = 0.5$.

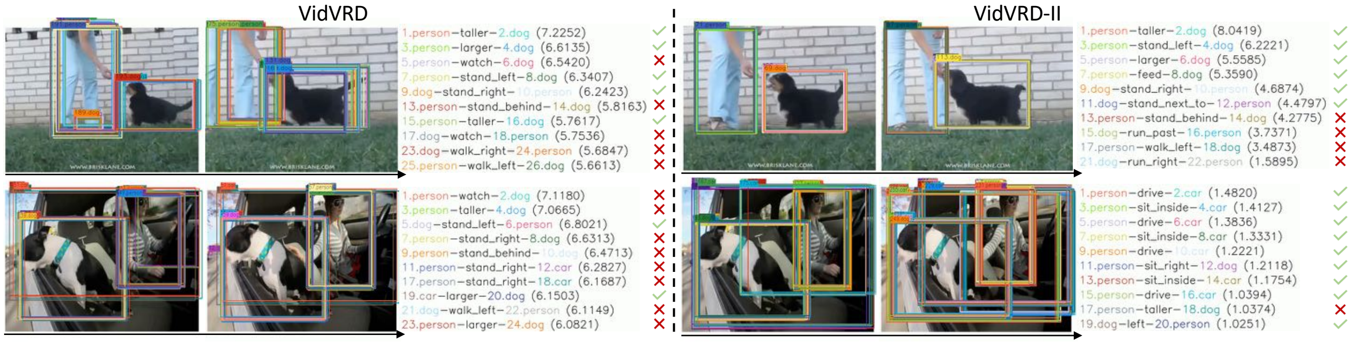
VidVRD
VidVRD-II
1.person−taller−2.dog (7.2252) ✓
3.person−larger−4.dog (6.6135) ✓
5.person−watch−6.dog (6.5420) ✗
7.person−stand_left−8.dog (6.3407) ✓
9.dog−stand_right−10.person (6.2423) ✓
13.person−stand_behind−14.dog (5.8163) ✗
15.person−taller−16.dog (5.7617) ✓
17.dog−watch−18.person (5.7536) ✗
23.dog−walk_right−24.person (5.6847) ✗
25.person−walk_left−26.dog (5.6613) ✗
1.person−watch−2.dog (7.1180) ✗
3.person−taller−4.dog (7.0665) ✗
5.dog−stand_left−6.person (6.8021) ✓
7.person−stand_right−8.dog (6.6313) ✗
9.person−stand_behind−10.dog (6.4713) ✗
11.person−stand_right−12.car (6.2827) ✗
17.person−stand_right−18.car (6.1687) ✗
19.car−larger−20.dog (6.1503) ✓
21.dog−walk_left−22.person (6.1149) ✗
23.person−larger−24.dog (6.0821) ✗
1.person−taller−2.dog (8.0419) ✓
3.person−stand_left−4.dog (6.2221) ✓
5.person−larger−6.dog (5.5585) ✓
7.person−feed−8.dog (5.3590) ✓
9.dog−stand_right−10.person (4.6874) ✓
11.dog−stand_next_to−12.person (4.4797) ✓
13.person−stand_behind−14.dog (4.2775) ✗
15.dog−run_past−16.person (3.7371) ✗
17.person−walk_left−18.dog (3.4873) ✗
21.dog−run_right−22.person (1.5895) ✗
1.person−drive−2.car (1.4820) ✓
3.person−sit_inside−4.car (1.4127) ✓
5.person−drive−6.car (1.3836) ✓
7.person−sit_inside−8.car (1.3331) ✓
9.person−drive−10.car (1.2221) ✓
11.person−sit_right−12.dog (1.2118) ✓
13.person−sit_inside−14.car (1.1754) ✓
15.person−drive−16.car (1.0394) ✓
17.person−taller−18.dog (1.0374) ✗
19.dog−left−20.person (1.0251) ✓

Figure 4: Qualitative comparison between the VidVRD baseline (left) and our VidVRD-II (right). Zoom in for better viewing. Top-10 detection results are displayed. Scores in the parentheses indicate the detection confidence.

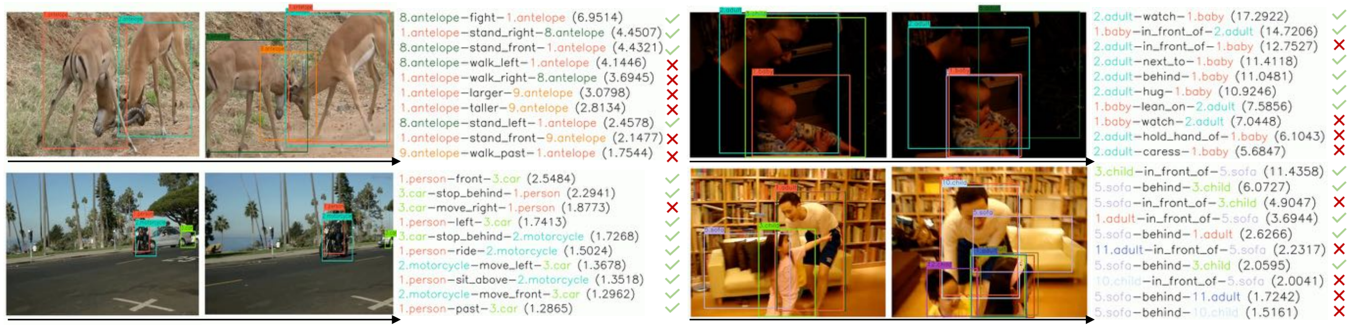
8.antelope−fight−1.antelope (6.9514) ✓
1.antelope−stand_right−8.antelope (4.4507) ✓
8.antelope−stand_front−1.antelope (4.4321) ✓
8.antelope−walk_left−1.antelope (4.1446) ✗
1.antelope−walk_right−8.antelope (3.6945) ✗
1.antelope−larger−9.antelope (3.0798) ✗
1.antelope−taller−9.antelope (2.8134) ✗
8.antelope−stand_left−1.antelope (2.4578) ✓
1.antelope−stand_front−9.antelope (2.1477) ✗
9.antelope−walk_past−1.antelope (1.7544) ✗
1.person−front−3.car (2.5484) ✓
3.car−stop_behind−1.person (2.2941) ✓
3.car−move_right−1.person (1.8773) ✗
1.person−left−3.car (1.7413) ✓
3.car−stop_behind−2.motorcycle (1.7268) ✓
1.person−ride−2.motorcycle (1.5024) ✓
2.motorcycle−move_left−3.car (1.3678) ✓
1.person−sit_above−2.motorcycle (1.3518) ✓
2.motorcycle−move_front−3.car (1.2962) ✓
1.person−past−3.car (1.2865) ✓
2.adult−watch−1.baby (17.2922) ✓
1.baby−in_front_of−2.adult (14.7206) ✓
2.adult−in_front_of−1.baby (12.7527) ✗
2.adult−next_to−1.baby (11.4118) ✓
2.adult−behind−1.baby (11.0481) ✓
2.adult−hug−1.baby (10.9246) ✓
1.baby−lean_on−2.adult (7.5856) ✓
1.baby−watch−2.adult (7.0448) ✗
2.adult−hold_hand_of−1.baby (6.1043) ✗
2.adult−caress−1.baby (5.6847) ✗
3.child−in_front_of−5.sofa (11.4358) ✓
5.sofa−behind−3.child (6.0727) ✓
5.sofa−in_front_of−3.child (4.9047) ✗
1.adult−in_front_of−5.sofa (3.6944) ✓
5.sofa−behind−1.adult (2.6266) ✓
11.adult−in_front_of−5.sofa (2.2317) ✗
5.sofa−behind−3.child (2.0595) ✓
10.child−in_front_of−5.sofa (2.0041) ✗
5.sofa−behind−11.adult (1.7242) ✗
5.sofa−behind−10.child (1.5161) ✗

Figure 5: Visualization of the VidVRD-II results on the examples from ImageNet-VidVRD (left) and VidOR (right). The results are post-processed by the improved graph-based association (introduced in Section 5.6), which helps to remove redundant bounding boxes and maintain the object identity. Zoom in for better viewing.

person and a dog, VidVRD-II can effectively recognize the different relationships between them, rather than guess the same set of likely relationships without much consideration of the visual input.

Visualization via Graph-Based Association. As noted in Figure 4, even though there are only two objects, the visualization contains lots of annotated object entities, *i.e.* many overlapping bounding boxes and large entity IDs. This inevitably makes it hard to visualize in the videos with more complex content for analysis. It also makes us unable to understand whether certain entity involves in multiple relations in a video. To better visualize and understand the VidVRD output, we modify the existing greedy relation association algorithm with entity association. The extension will associate the relation instances from the graph view point: we first respectively associate the subject and object of a given relation instance to the existing entities according to the spatio-temporal overlap (vIoU), and then the predicate according to the temporal overlap.

Figure 5 presents more qualitative results of VidVRD-II post-processed by the improved association algorithm. As can be seen, there are much fewer redundant bounding boxes; and the entities involving multiple relations are also presented better. Further, we can observe more interesting visual relations detected by VidVRD-II. For example, "antelope-fight-antelope" in the 1st example, "person-past-car" in the 2nd example, "adult-hug-baby" and "baby-lean on-adult" in the 3rd example. As for most of the failure cases, especially in the 1st and 4th examples, we can find that they are mainly resulted from duplicated detection of an entity or wrong entity classification between the similar classes (*e.g.*"adult" vs "child").

6 CONCLUSION

We propose to detect visual relations in video through iterative relation inference (VidVRD-II). The model in our approach explores the dependency among the three components in relation triplet and refines their classification scores in an iterative manner during the inference. A training approach is also introduced to better learn the dependency knowledge from the training data. Our experiments demonstrate the effectiveness of the proposed iterative relation inference and training approach, and show the state-of-the-art performance of our VidVRD-II on the benchmark datasets. It is worth mentioning that the improvement on ImageNet-VidVRD is remarkably large. Our qualitative analysis also shows that VidVRD-II can detect many interesting visual relations in video, which is rarely observed in the previous works. Future works include exploring the use of pretrained models, knowledge graphs or their combination to learn more comprehensive dependency knowledge. This would advance the current technique to more generalized domains and learning settings. It is also interesting to explore the use of video relations in the high-level video-language tasks, such as video captioning [1], VQA [11, 40], and multimedia retrieval [44–46].

ACKNOWLEDGMENTS

This research is funded by Sea-NExT Joint Lab, Singapore.

REFERENCES

[1] Yi Bin, Xindi Shang, Bo Peng, Yujuan Ding, and Tat-Seng Chua. 2021. Multi-Perspective Video Captioning. In *Proceedings of the 29th ACM International Conference on Multimedia.*

[2] Qianwen Cao, Heyan Huang, Xindi Shang, Boran Wang, and Tat-Seng Chua. 2021. 3-D Relation Network for visual relation recognition in videos. *Neurocomputing* 432 (2021), 91–100.

[3] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. 2019. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 4613–4623.

[4] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. 2019. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6163–6171.

[5] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. 2019. Scene graph prediction with limited labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2580–2590.

[6] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. 2020. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10337–10346.

[7] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE conference on computer vision and Pattern recognition.* 3076–3086.

[8] 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.

[9] 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.

[10] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. 2019. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1969–1978.

[11] Mao Gu, Zhou Zhao, Weike Jin, Richang Hong, and Fei Wu. 2021. Graph-Based Multi-Interaction Network for Video Question Answering. *IEEE Transactions on Image Processing* 30 (2021), 2758–2770.

[12] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. 2016. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465* (2016).

[13] 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.

[14] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 727–735.

[15] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4. *International Journal of Computer Vision* (2020), 1–26.

[16] 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.

[17] Yikang Li, Wanli Ouyang, Zhou Bolei, Shi Jianping, Zhang Chao, and Xiaogang Wang. 2018. Factorizable Net: An Efficient Subgraph-based Framework for Scene Graph Generation. In *Proceedings of the European Conference on Computer Vision (ECCV).*

[18] Xiaodan Liang, Lisa Lee, and Eric P Xing. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 848–857.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision.* 2980–2988.

[20] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. 2020. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10840–10849.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3431–3440.

[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] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 724–732.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Zixuan Su, Xindi Shang, Jingjing Chen, Yu-Gang Jiang, Zhiyong Qiu, and Tat-Seng Chua. 2020. Video Relation Detection via Multiple Hypothesis Association. In *Proceedings of the 28th ACM International Conference on Multimedia.* 3127–3135.

[32] 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.

[33] 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.

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[35] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. 2020. Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision.* Springer, 71–87.

[36] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 3716–3725.

[37] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The New Data in Multimedia Research. *Commun. ACM* 59, 2 (2016), 64–73.

[38] 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.

[39] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. 2020. Visual relation grounding in videos. In *European Conference on Computer Vision.* Springer, 447–464.

[40] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 9777–9786.

[41] Wentao Xie, Guanghui Ren, and Si Liu. 2020. Video Relation Detection with Trajectory-aware Multi-modal Features. In *Proceedings of the 28th ACM International Conference on Multimedia.* 4590–4594.

[42] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 5410–5419.

[43] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV).* 670–685.

[44] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1339–1348.

[45] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[46] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. 2020. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM International Conference on Multimedia.* 1939–1947.

[47] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5831–5840.

[48] 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.

[49] 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.

[50] Sipeng Zheng, Xiangyu Chen, Shizhe Chen, and Qin Jin. 2019. Relation understanding in videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2662–2666.

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