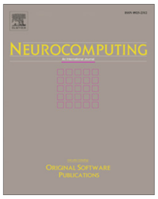
Neurocomputing 432 (2021) 91–100

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



3-D Relation Network for visual relation recognition in videos

Qianwen Cao [a,*,1], Heyan Huang [a,b], Xindi Shang [c], Boran Wang [a], Tat-Seng Chua [c]

[a] *School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China*
[b] *Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing, China*
[c] *School of Computing, National University of Singapore, Singapore*

ARTICLE INFO

Article history:
Received 30 July 2020
Revised 10 October 2020
Accepted 10 December 2020
Available online xxxx
Communciated by Zidong Wang

Keywords:
Video visual relation recognition
Visual relation detection
Computer vision
Deep neural network

ABSTRACT

Video visual relation recognition aims at mining the dynamic relation instances between objects in the form of $\langle subject, predicate, object\rangle$, such as “person1-towards-person2” and “person-ride-bicycle”. Existing solutions treat the problem as several independent sub-tasks, i.e., image object detection, video object tracking and trajectory-based relation prediction. We argue that such separation results in the lack of information flow between different sub-models, which creates redundant representation while each sub-task cannot share a common set of task-specific features. Toward this end, we connect these three sub-tasks in an end-to-end manner by proposing the 3-D relation proposal that serves as a bridge for relation feature learning. Specifically, we put forward a novel deep neural network, named 3DRN, to fuse the spatio-temporal visual characteristics, object label features, and spatial interactive features for learning the relation instances with multi-modal cues. In addition, a three-staged training strategy is also provided to facilitate large-scale parameter optimization. We conduct extensive experiments on two public datasets with different emphasis to demonstrate the effectiveness of the proposed end-to-end feature learning method for visual relation recognition in videos. Furthermore, we verify the potential of our approach by tackling the video relation detection task.

1. Introduction

Video visual relation explores the interaction knowledge between two objects in the video, which can be regarded as a bridge between vision and language in video content understanding. By providing the structured fine-grained semantics, these relations can be used as support for high-level video applications, such as video retrieval [7,29,35], video captioning [21,44,24], video question answering [8,9] and video summarization [27,50,1].

In order to make the video visual relation explicit and standard, the task of video visual relation detection is proposed [31], which aims at detecting multiple relation instances in the form of $< subject, predicate, object >$ over time. As the example illustrated in Fig. 1, it can be an action *(e.g. person1 – ride – motorcycle)*, a state change *(e.g. person1 – towards – person2, person1 – past – person2)*, or a relative position *(e.g. person1 – above – motorcycle)*. Compared with the related task in image domain [22,17], visual relation detection in videos is confronted with more challenges due to the existence of time dimension. It not only needs to spatio-temporally localize the object's position in order to extract the entity-level visual features, but also needs to detect the rich and dynamic interactions between objects, which makes the feature representation and relation prediction complicated and difficult.

Due to these intrinsic complexities of the task, existing models [31,40,25] resort to pipeline structure by dividing the problem into several sub-problems as shown on the left side in Fig. 1, with each being addressed by a developed technique. We argue that these step-by-step approaches are not able to share the features from different levels and modalities among different technical components, thus preventing the low-level features from being adaptive to the high-level supervision. In addition, it is inefficient with respect to both computation and data utilization. For example, due to the lack of temporal information in the visual features obtained by object detection on frames, the relation prediction needs to compute the hand-crafted visual features additionally based on the trajectories, which results in redundant representation of visual features. Moreover, cascade failure will also be encountered in the pipeline approaches. In fact, the failure of object detection may severely affect the later relation prediction.

In this paper, we propose a neural network model, named 3-D Relation Network (3DRN) as shown on the right side in Fig. 1, to tackle the aforementioned problems. In order to make the whole

[*] Corresponding author.
E-mail address: qwcao@bit.edu.cn (Q. Cao).
[1] ORCID: 0000-0001-7491-5201

https://doi.org/10.1016/j.neucom.2020.12.029
0925-2312/© 2020 Elsevier B.V. All rights reserved.

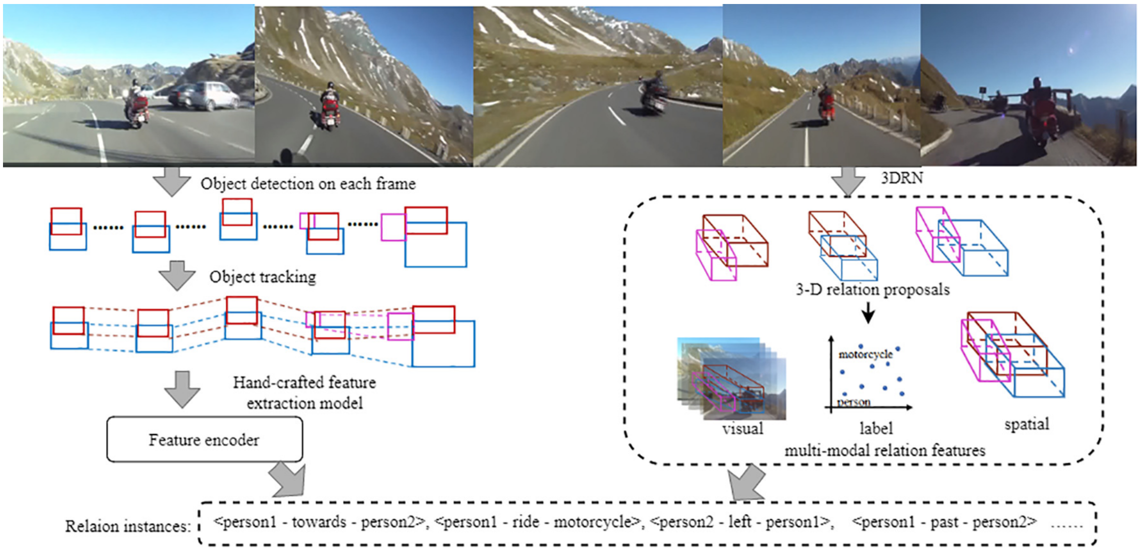

Object detection on each frame
Object tracking
Hand-crafted feature extraction model
Feature encoder
3DRN
3-D relation proposals
motorcycle
person
visual
label
spatial
multi-modal relation features
Relaion instances:
<person1 - towards - person2>, <person1 - ride - motorcycle>, <person2 - left - person1>, <person1 - past - person2> ......

Fig. 1. The left side represents the general pipeline framework, where the relations are obtained through individual components step by step. The right side represents our proposed framework that predicts the relations with multi-modal features in an end-to-end fashion.

model differentiable, 3DRN introduces the 3-D relation proposal as an approximate representation of video-level object pair to achieve the end-to-end feature learning. With the help of the relation proposals, their spatio-temporal visual features, which can encode the dynamic properties of the video, are not only used to detect the objects but also reused in relation feature representation. These visual features are fused with the object category features and relative spatial features to provide multi-modal cues for predicting multiple relation instances. We further propose an effective training strategy so that the features in different level are flowed and interacted under the supervision of relation detection. Such design empowers our proposed network to efficiently learn the discriminative and fine-grained representations of relation instances from large scale datasets and reduce the computation complexity.

To validate the effectiveness of our proposed approach (Section 4 and 5), we also propose evaluation method named video retrieval by relation as a supplement to the existing evaluation task. It has practical application significance but is not considered in previous works.We conduct extensive experiments on three evaluation tasks and two public datasets, ImageNet-VidVRD [31] and VidOR [30]. In summary, the main contributions of this paper include:

- We propose a neural framework (3DRN) and an effective end-to-end relation feature learning method by the use of the 3-D relation proposals to tackle the complicated problems in video relation recognition.
- We propose a new evaluation task, video retrieval by relation, as a new evaluation method for video relation recognition.
- We achieve the state-of-the-art performance in terms of video relation tagging and video retrieval by relation on the above two datasets.

2. Related work

In this section, we review some state-of-the-art approaches about visual relation detection and the three-dimensional CNN.

2.1. Visual relation detection

Recent image visual relation detection research divides the task into two parts. The first is to detect the objects and the second explores the interactions between these observed objects. As a mature object detection model, Faster R-CNN [26] is commonly used to recognize and locate objects which exist in the given image. In these approaches [20,46,47], researchers pay attention to constructing models to explore various relationships between these detected objects. Another few works focus on treating the image relation detection as a whole. Li et al. [19] propose a phrase-guided message passing structure for mining the visual dependencies between subjects and objects, which sets up a horizontal information flow among different components and predicts them simultaneously. Zhang et al. [48] introduce a three-branch region proposal network to produce candidate object pairs and predict their relationships.

As a middle-aware video visual understanding task, video visual relation detection is drawing widely attention. The relation instances can not only be used for further high-level video content understanding tasks, such as video captioning [49,13] and video question answering [34,14], but also be used as semantic supports to help improve video object detection [4,33]. A large-scale dataset for video visual relation detection is first proposed by Shang et al. [31], and a pipeline framework is proposed as illustrated in the left side of Fig. 1. The follow-up works also follow the framework but their research points are different. Sun et al. [37] are concerned with how to locate accurate trajectories in order to extract discriminative features. Therefore, they assemble the effectual methods of video object detection and tracking, such as flow-guided feature aggregation (FGFA) [51], Seq-NMS [11] and KCF tracker [12], to generate accurate object trajectories. As a result, their approach won the first price in VRU'19 challenge. To capture the dependencies between relation instances over the video, Tsai et al. [40] construct a gated fully-connected graph as the feature encoder whose nodes denote objects in the video and edges denote their interactions. Similarly, Qian et al. [25] achieve even better performance by propagating the spatio-temporal information via Graph Convolution Networks (GCNs). However, the object trajectories and intermediate representations in these works still divide the proposed approaches into independent components, which results in the waste of computation and feature representation. In contrast, our proposed 3DRN is designed in a differentiable and end-to-end fashion, so that can reduce redundant representations and extract efficient video features.

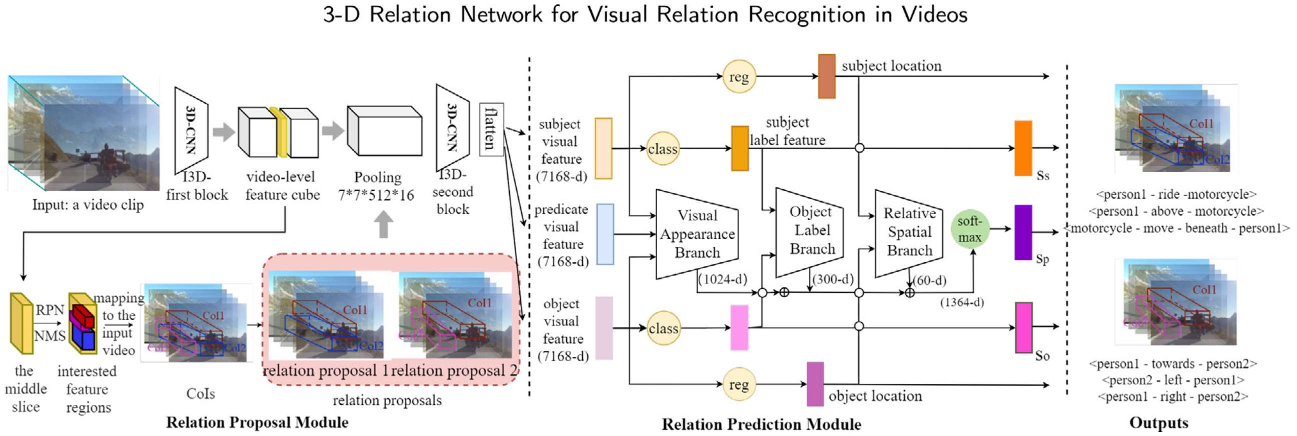


Fig. 2. The framework of the proposed 3-D Relation Network (3DRN) for video relation recognition. An input video is first passed through the relation proposal module, where 3-D CNN (specifically I3D) is utilized as the backbone. Then, the relation proposals as the intermediate results are fed into the relation prediction module to extract features with three multi-modal cues for predicting multiple relation instances.

2.2. Generalizing CNN from 2-D to 3-D

Our work aligns closely with the recent efforts in generalizing CNN from 2-D to 3-D. In particular, the approaches proposed by these works can be mainly divided into two categories. The first category uses the 2-D convolution kernels to extract features at the frame level and then connects them across the whole timeline. Examples of such connection methods include temporal pooling operation [15], RNN [6] and tracking [43], etc. On the contrary, the other category replaces the 2-D kernels with 3-D ones and directly extract the spatio-temporal features. Among the methods in this category, C3D [39] is the pioneer to propose 3 × 3 × 3 kernel in all convolution layers and leverage 3-D pooling to propagate the features with temporal information. However, due to the lack of video data for training the large numbers of parameters, the construction of C3D is relatively shallow. To tackle this dilemma, I3D [2] is proposed to reflate the kernels from 2-D to 3-D. Specifically, the filters and pooling kernels are expanded into 3-D based on deep image classification architectures. This allows I3D to take advantage of the parameters pre-trained on large-scale image datasets like ImageNet.

Most of the above 3-D networks have been applied in action recognition and video classification tasks, e.g [42,3,36]. Also, in the field of video relation recognition, current efforts still focus on 2-D networks in images. Hence, it is intuitive to generalize the 2-D networks to 3-D to extract effective visual relation features from videos by accounting for the time dimension. As far as we know, our work is the first to utilize the 3-D CNN as the backbone network for video visual relation recognition.

3. Methodology

In order to illustrate the key challenge for end-to-end training in video relation recognition, we formalize the general processing framework from input video clip v to output relation instances r as:

$$P(r_{<o_1^*,p,o_2^*>}|v) = P(o|v) \cdot P(o^*|o,v) \cdot P(p|v,(o_1^*,o_2^*)), \qquad (1)$$

where o denotes one of the objects detected on video frames via image object detection model $P(o|v)$, o^* denotes one of the video-level object candidates generated from o, and p denotes one of the predicates predicted via relation prediction model $P(p|v,(o_1^*,o_2^*))$. As can be found in the existing pipeline approaches such as [31,40,37,25], it is actually not difficult to implement gradient back propagation in either $P(o|v)$ or $P(p|v,(o_1^*,o_2^*))$. Yet $P(o^*|o,v)$ can only be implemented via some tracking algorithm and ad hoc post-processing steps in these approaches, which makes the models disconnected and impossible to be optimized with a uniform supervision. Therefore, we can find that the biggest obstacle to achieve the end-to-end training lies in how to integrate $P(o^*|o,v)$ into the whole differentiable framework.

In order to resolve this challenge, we propose the use of relation proposals to replace the complicated post-processing steps for connecting the models in an end-to-end manner. The relation proposals are a series of pairs composed from the approximate expressions of video-level objects, which actually enable us to design a feasible differentiable framework 3DRN (Section 3.2) with an end-to-end training strategy (Section 3.3). In next section, we will first elaborate the details of relation proposal module.

3.1. Relation proposal

Our idea of relation proposals is to approximately represent the object pair (o_1^*,o_2^*) by Cubes of Interest (CoIs). Formally, the CoI is represented by a 6-dimensional vector $(x_1,y_1,x_2,y_2,t_1,t_2)$, indicating a spatial location (x_1,y_1,x_2,y_2) from t_1 to $t2$, where $t_1 = 1, t_2 = length(clip)$, and (x_1,y_1,x_2,y_2) denotes the up-left and bottom-right box coordinates. As shown in Fig. 2, we obtain CoI candidates by mapping the interested feature regions divided from the middle feature slice by the region proposal network [26] to the input video clip, and performing Non-Maximum Suppression (NMS) with IoU (set to 0.8) to filter redundant feature regions, which is similar to the previous work [10]. In this way, our relation proposals (o_1^*,o_2^*) in Eq. (1) are implemented.

In fact, adopting the above approximation is reasonable due to two folds. First, the dense deep features have large receptive fields. Specifically, each slice in the video-level visual feature cube, as shown in Fig. 2, has the receptive fields that can span the entire input video clip, and thus the relation proposals are temporally mapped to the whole time sequence. Second, the movement of the object is limited in short clips, so the relation proposals spatially overlap the object motion regions.

Although the method seems straightforward, there are technical difficulties in sampling at the training phase. In the end-to-end framework, $P(o^*|o,v)$ in Eq. (1) is replaced with the relation proposal sampling. In particular, relation proposals are sampled from those CoIs that have sufficient overlap with the bounding boxes in the ground truth, and whose classified labels are the same with the subject or object in the annotation triplets. If this random sampling is carried out without any control, the model is difficult to converge, and hence we propose the following sampling method. On one hand, due to the sparseness of labeled relations and the strictness of selecting relation proposals, the positive relation proposals are quite limited. Therefore, we propose to manually generate some "pseudo" proposals to increase the relation proposals by modifying the spatial size of each qualified proposals. This is done by extending the coordinates in four directions, especially top-left, top-right, bottom-left and bottom-right, by some pixels (limited to 5% of the total length and width). In addition, for those objects without any corresponding detected CoI but appear in the ground truth, we also include them in the training set. On the other hand, there are always objective relations between two objects, such as the spatial relationship ("front", "left", etc.). Thus we restrict the training batch to contain only positive examples. Furthermore, in order to avoid the results from being biased in favour of high-frequency relations, we ensure that the predicate categories of these positive examples are equally distributed. In training phase, we compare them with ground truth, and extract two CoIs that meet the IoU condition (higher than 0.5).

3.2. The 3DRN architecture

In order to elaborate how the proposed relation proposal approach is plugged with the existing separate networks, we introduce a new 3D Relation Network for video relation recognition below.

As shown in Fig. 2, I3D is utilized as the backbone in the relation proposal module. The advantage of this backbone lies in the 3-D convolution kernel. It can capture temporal dependence naturally in order to obtain the video-level feature cube whose the receptive field covers the time length of the input video clip. After obtaining the relation proposals and their visual features, we feed them into the relation prediction module, a network with three branches, which has been successfully applied in image visual relation detection [48,20,46,51]. In the following paragraphs, we briefly describe the details of the three branches.

Visual Appearance Branch. In this branch, with the help of these relation proposals, the entity-level visual features for detecting objects can be reused in the relation prediction module for learning the visual expressions of the relationships. In particular, the two 7168-dimensional visual features of the relation proposal, which are obtained from the second block in I3D and then flattened, are utilized as the appearance characteristics for the corresponding component *subject* or *object* in the relation proposal. Similar to previous works on image relation detection [48,22,47,51], we also integrate the visual features extracted from the union region of the *subject* and *object*, and use the features as the *predicate* visual appearance to provide surrounding contexts. These visual features of the three components are put into the visual appearance branch, which concatenates them together and pads with a fully connected layer, to output the visual feature representation of the relation.

Object Label Branch. Identifying the collocation patterns has been found to play an essential role in visual relation detection [45]. Thus, in 3DRN, the label features of *subject* and *object* are used as inputs of this branch for learning the joint semantic pattern. Specifically, they go through the concatenation operation and a fully connected layer to obtain the relation label features. Because we aim at building a basic general model, we initialize the label features using the one-hot vectors according to the categorical results, rather than the pre-train word embedding vectors utilizing BERT [5] or word2vec [23] algorithms on the external common knowledgebase.

Relative Spatial Branch. We suppose that (w_s, h_s, w_o, h_o) denotes the width and height of the *subject* and *object* respectively. The two-dimensional vector (w_v, h_v) is used to denote the video screen size, and $(w_\vee, h_\vee)$ represents the union location. The inputs of this branch are these location coordinates and the output is the relation spatial feature. According to the calculation methods of spatial features in [47,18,46], the interactive feature of the relation is computed based on the subject location and the object location. The feature is represented as:

$$\left[\frac{(w_s+w_o)/2}{w_v},\frac{(h_s+h_o)/2}{h_v},\frac{w_\vee}{w_v},\frac{h_\vee}{h_v},\frac{w_s}{w_o},\frac{h_s}{h_o}\right]. \tag{2}$$

It is passed through a fully connected layer for learning the relation spatial feature representation.

These three kinds features has been transformed into similar dimensional vectors by each branch. Then they are concatenated together and fed into a *softmax* layer to generate a probabilistic distribution of predicates. The final score S_r for a relation instance is the sum of object detection confidence and predicate probability score: $S_r = S_s + S_p + S_o$, and 3DRN outputs the top-K relations sorted by this score ($K = 100$ as used in [31]).

3.3. Details of end-to-end training

Training all weights with back-propagation is an important capability of 3DRN, leading to learning efficient video features. However, due to the large amount of parameters and the lack of empirical models, it is not feasible to carry out a one-stage training under the supervision of multi-task training loss from the beginning. Thus, we propose a three-stage training approach. Generally, the parameters of the two models in the framework are determined first with general ranges, and then all parameters are fine-tuned towards the multi-task loss. The training details of each stage are illustrated below.

Stage 1: We start from optimizing the parameters of the backbone and RPN in the relation proposal module with the standard object detection loss function L^{de}_{rel}. The basic training configuration is the same as Faster R-CNN.

Stage 2: Through the previous step, high-quality relation proposals and their classified labels are determined primarily. However, the visual features are optimized towards the target of object detection. Therefore, after initializing the relation proposal module with the trained parameters, we fix them except for the parameters of the second block in I3D. The relation prediction module is trained with the supervision of cross entropy loss function:

$$L^{p}_{rel} = -\sum_{j=1}^{M}\sum_{k=1}^{N} y^{j}_{k} log p^{j}_{k}, \tag{3}$$

where N is the number of predicate categories, M denotes the number of relation proposals, y_k and p_k are the kth ground truth and the predicate label. If the predicted predicate is annotated in the ground truth, then $y_k = 1$; otherwise, $y_k = 0$.

Stage 3: The two modules trained independently will modify the important parameters in the expected ranges and pave the way for end-to-end training. We eventually optimize 3DRN as a whole under the supervision of the multi-task cost function:

$$L_{rel} = L^{p}_{rel} + \lambda L^{de}_{rel} \tag{4}$$

where λ is the weight to balance the relative significance between the two task losses. In each training iteration, the forward pass generates relation proposals as pre-computed proposals when training 3DRN. For the shared layers about visual features, the back propagated signals are past through both two task losses.

4. Video relation recognition

As a semantic representation of video content, relation triplet $< subject, predicate, object >$ can support many applications such as video tagging and retrieval. In this section, we conduct experiments from the perspective of accurate relation triplet prediction, to validate the effectiveness of 3DRN in learning discriminative relation feature for video relation recognition, and demonstrate its potential in the real world applications. In the following, we will first introduce the datasets used by our experiments in this paper.

4.1. Datasets

We adopt two public datasets, ImageNet-VidVRD [31] and VidOR [30], to comprehensively evaluate our approach through performance comparison.

ImageNet-VidVRD has diverse predicates in the relation instances to describe multiple interactions between each two observed objects. The dataset contains 1,000 videos from ILSVRC2016-VID [28], densely annotated with relation instances in 132 categories of predicates and 35 categories of objects. We follow the train/test split in the dataset, i.e. 800 videos for training and 200 videos for testing. In particular, the test set contains 4,835 annotated relation instances in 1,011 triplet categories.

VidOR is a large-scale video relation dataset and has been used in ACM MM'19 Video Relation Understanding (VRU) Challenge [32]. It consists of 98.6 h of user-generated videos from YFCC-100M [38], densely annotated with relation instances in 80 categories of objects and 50 atomic categories of predicates. The dataset is offically split into 7,000 videos for training, 835 videos for validation and 2,165 videos for testing. However, since the test set is private for the grand challenge, our evaluation is conducted on the validation set, which contains 30,142 annotated relation instances in 2,410 triplet categories. Compared with ImageNet-VidVRD, VidOR has longer average length of video (35.73 s) with more complex scenes, and more annotated relation instances, which poses greater challenge in visual relation recognition.

4.2. Evaluation tasks and metrics

We specifically conduct evaluation on the following tasks, which are common in real world applications.

Video Relation Tagging. This task requires a method to output a series of relation triplets given an video. Following [31], we use the metric mP@K (mean Precision@K, K equals 1, 5 and 10) as the evaluation metric, which is expressed as:

$$mP@K = \frac{\sum_{video \in N} Precision@K}{|N|}, \quad (5)$$

where N is the set of all videos in the test set. In particular, for each given video, Precision@K is the fraction of top-K predicted relation triplets that are correct, where a relation triplet is considered to be correct if there is the same triplet annotated in the ground truth of the video.

Video Retrieval by Relation. This task uses a certain category of relation triplet as the query, and asks a method to retrieve a series of relevant videos from a database. Since there are highly imbalanced distribution of triplet categories in both of the datasets, this evaluation task with relation triplets as query can avoid the evaluation score from being dominated by the frequent labeled relations. Following the popular metrics in measuring content based visual retrieval task, we use mAP (mean Average Precision) and mR@K (mean Recall@K, K equals 10 and 20) as the evaluation metrics, which are defined as:

$$mAP = \frac{\sum_{triplet \in C} AP}{|C|}, \quad (6)$$

$$mR@K = \frac{\sum_{triplet \in C} Recall@K}{|C|} \quad (7)$$

where C is the set of relation triplet in test set. In particular, for each given relation query, AP is the average precision in retrieving correct videos from the test set, where a video is considered to be correct if it is annotated with the same relation triplet as the query in ground truth. Similarly, $Recall@K$ is the fraction of correct videos that are successfully retrieved in the top K results.

4.3. Compared methods

To generate relation triplets for the videos of variable length, it is essential to merge the relation triplets recognized in short video clips by 3DRN. Basically, there are two ways to achieve this. One is to directly merge the relation triplets of same category into one, and then assign a new confidence score with the average score over those relation triplets. This is a basic approach and we call the method implemented by this approach as **3DRN**. The other way is to go through the greedy association algorithms proposed by [31]. The approach determines whether the association needs to be performed according to not only the relation triplets but also the object locations. We call the method implemented by this approach as **3DRN+GA**.

In the experiments, our method is compared with the following methods on the two evaluation tasks. **Shang's** [31] first generates object tracklets on each clip by tracking the detects objects on frames. Then, iDT features [41] and self-designed relative spatial features are extracted to predict the relation instances on each video clip. **VRU'19#1** [37] is the top-1 solution in ACM MM'19 Video Relation Understanding challenge. The method makes huge efforts to ground object locations. Compared with VidVRD, the method first spatio-temporally detects object trajectories with ensembled methods, and then split the trajectories over time dimension to extract features and predict the relations for each clip. **GSTEG** [40] focuses on modifying the feature encoding in the pipeline approach. The method constructs a Conditional Random Field on a gated fully-connected graph and defines a gated pairwise energy function to explore the statistical dependency between the objects spatially and temporally. **VRD-GCN** [25] is also devoted to optimizing the feature encoding step. It takes advantage of GCN [16] to pass the messages from context entities for integrating the temporal information in neighbour clips.

Similarly, we implement these baselines on the evaluation tasks by using not only the basic merging approach but also the greedy association approach in [31], which are denoted as **Shang's+GA**, **VRU'19#1+GA**, **GSTEG+GA** and **VRD-GCN+GA**, respectively.

4.4. Details of implementation

We divide each video into short clips of 30 frames and pad the first and last frames until the temporal length satisfies the requirement for training I3D (*i.e.* 64 frames). On ImageNet-VidVRD, at the first stage of training phase, we set 2 frames to be the stride of consecutive segments and then filter out noisy samples where the objects appear in less than 70% of the duration. As for training of relation prediction module and fine-tuning of 3DRN, we use 15 frames as the sample stride, which is consistent with the setting of Shang's [31]. In order to ensure the diversity of relation learning, all the clips are then sent to 3DRN without any filtering. On VidOR, since the data volume is too large, there will be nearly 3 million clips if we use the same stride 2 as the above. Therefore, we increase the stride size from 2 to 30 frames for splitting the training set used in these three training stages, and the filtering of samples is the same as that for ImageNet-VidVRD.

We first initialize the 3D-CNN with the parameters in I3D-RGB pretrained on Kinetics dataset [2]. The relation proposal module is optimized using stochastic gradient descent algorithm (SGD) with the momentum set to 0.9, and other basic configurations are the same with Faster R-CNN. The learning rate is set to 0.01. At the second training phase, the model is optimized via the same optimizer with the same learning rate. The batch size is set to 20 relation proposals which is limited by the GPU memory. Then at the last joint training phase, we train the entire 3DRN in an end-to-end manner for obtaining the task-specific features towards the same goal of visual relation recognition. λin the joint multi-task loss function in Eq. (4) is set to 1.

4.5. Quantitative results and analysis

Tables 1 and 2 report the experimental results on ImageNet-VidVRD and VidOR, respectively, where the best performances regarding each metric are highlighted in bold. We have tested the results from their available model and code. As the training source codes of GSTEG is not available, we cannot apply the method on video retrieval by relation task and evaluate the method without greedy association processing. Its published results are put at the end of the table.

Through the comparison of these models for the video relation recognition under the two evaluation tasks, we have the following observations.

i). From experimental results of video relation tagging evaluation on two datasets, our approach outperforms all the existing methods on all metrics, except VRD-GCN+GA which performs slightly better than 3DRN+GA on mP@1. Specifically, on ImageNet-VidVRD dataset, our 3DRN+GA yields 1.30% and 1.30% gains on mP@5 and mP@10 compared with the state-of-the-art baseline VRD-GCN+GA. On VidOR dataset, 3DRN+GA exceeds VRU'19#1 by 1.39% and 1.60% on mP@1 and mP@5, respectively. On the whole, these improvements demonstrate that the end-to-end feature learning method proposed by us is effective to predict accurate and diverse relation triplets in videos thanks to the information flow and interact between different layers.

ii). Comparing the results of video retrieval by relation on datasets, we find that our 3DRN is superior to other methods. Especially on the comprehensive evaluation metric mAP, our 3DRN is 1.10% and 7.00% better than VRD-GCN and VRU'19#1 which won the first price in VRU'19 grand challenge, respectively. This verifies the advantage of our approach in extracting the fine-grained feature patterns for the relation triplet in different scenes, which makes our model have the potential to be applied in video retrieval.

iii). In addition, for the two different video relation generation strategies, the strategy directly based on relation triplets makes the methods achieve better performance in retrieval test, while the strategy based on the greedy association algorithm (GA) [31] makes the methods performs better in relation tagging test. The reason for this difference is mainly because the focuses of the metrics in two evaluation tasks are different. The evaluation metric in relation tagging is precision, which focuses on relations with high scores. Therefore, relations with correct triplets but cannot be associated do not affect the performance. However, the evaluation metric in video retrieval by relation is recall, and is based on the top 10 and 20 results. Although the evaluation is more comprehensive, these relations tend to lower the test result. This is because compared with the direct merging triplets of same category, association criteria of GA also require high overlap between corresponding entities in the two adjacent clips. For the relations that do not satisfy the 2 conditions, they cannot be merged. However, their scores may not be low; and therefore they are include redundantly in the top-100 result list. They do not affect the relation tagging test, because we focus on precision of the top-k relations.

4.6. Qualitative results and analysis

From the visualization examples for video relation tagging in Fig. 3, we can see that our 3DRN correctly predicts the relations *adult-push-bicycle* and *adult-towards-adult*, which depend on the sufficient visual features to determine. It indicates that our proposed deep relation feature extraction method makes 3DRN effective to mine the detail dependency between objects from the video visual information itself. While VRU'19#1 and Shang's do not predict the predicates *push, ride* and *walk toward* between objects, and incorrectly output the relations *adult-ride-bicycle*. From the results of video retrieval by relation, one can find that the relations with clear motion feature patterns in videos, like *adult-caress-dog* and *red panda-walk toward-red panda*, are effectively identified by our 3DRN. It benefits from the end-to-end optimization of the relation proposal and their deep multi-modal features towards the same goal, which makes the feature patterns of the relation triplet embedded closely.

5. Video relation detection with spatial temporal grounding

This section particularly studies from the perspective of fine-grained visual relation recognition in videos. Specifically, we investigate not just the relation triplets, but also their temporal positions and spatial locations in the video. Thus we conduct experiments to test how well 3DRN can detect and spatio-tempoally ground each relation instance of interests, through the video visual relation detection (VidVRD) task as defined in [31]. This can demonstrate the potential of 3DRN for some high-level tasks such as video captioning and video question answering. Similarly, we use both ImageNet-VidVRD and VidOR datasets as in Section 4. As VidVRD requires the output (relation instances) in the form of both relation triplet and the corresponding bounding-box trajectories, all the methods we used are processed by the greedy association approach as mentioned in Section 4.3.

5.1. Evaluation metrics

Following [31], we use mAP, mR@K (mean Recall@K, K equals 50 and 100) as the evaluation metrics, where mAP is also served as the official metric in ACM MM'19 VRU Challenge [32]. The metrics are defined as:

$$mAP = \frac{\sum_{video \in N} AP}{|N|}, \tag{8}$$

$$mR@K = \frac{\sum_{video \in N} Recall@K}{|N|}, \tag{9}$$

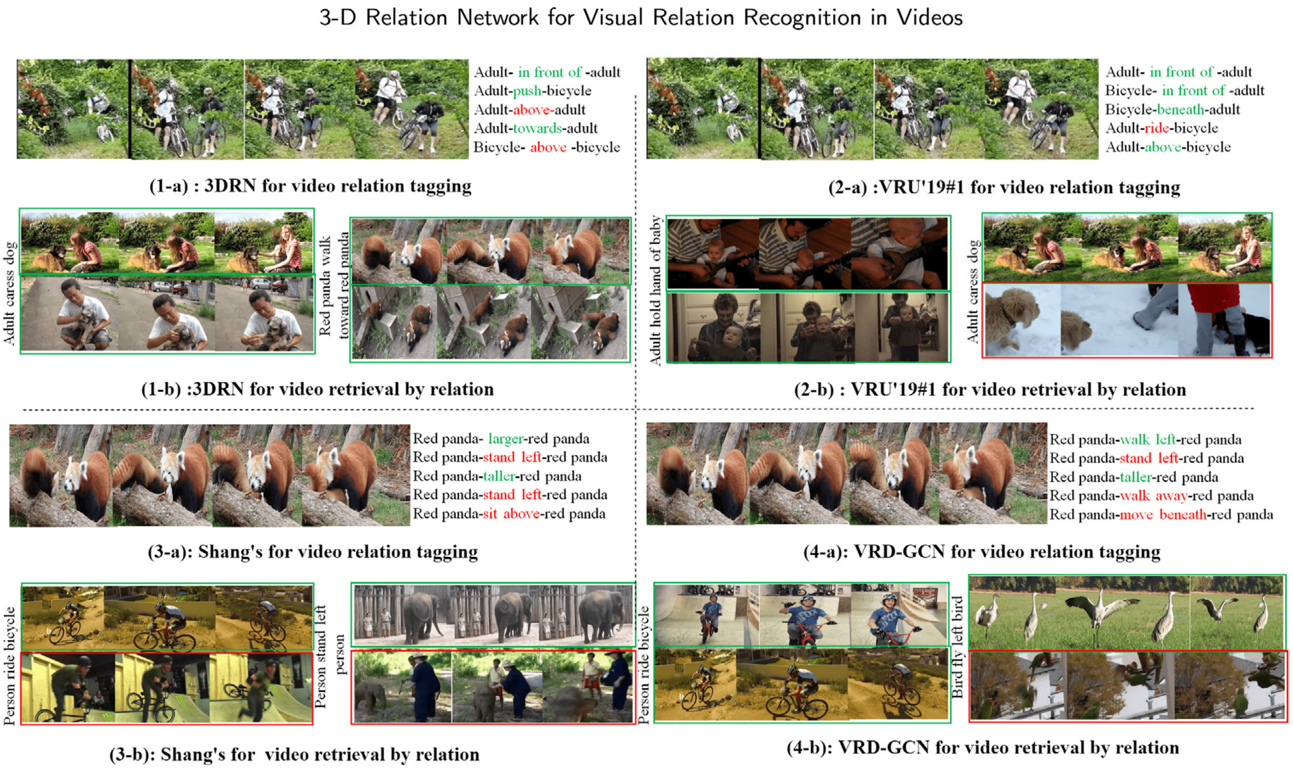
where N is the set of all videos in the test set. Note that the metrics defined here are different from those defined in Eq. (6) and (7). In particular, AP is the average precision in detecting correct relation instances from a video in N, where a relation instance is regarded as correct if its relation triplet is in the ground truth and the corresponding trajectories of *subject* and *object* have sufficient vIoU (voluminal IoU, set to be 0.5) with the ground truth. Similarly, given a video in $N, Recall@K$ is the fraction of the correct relation instances that are successfully detected in the top K results.

Table 1
Performance comparison of different methods on ImageNet-VidVRD dataset for video relation tagging and video retrieval by relation. "–" represents the result is not available on this metric.

Method	Relation Tagging			Video Retrieval by Relation		
	mP@1	mP@5	mP@10	mAP	mR@10	mR@20
Shang's	27.50	23.60	18.10	28.44	12.97	15.05
VRD-GCN	47.62	34.13	25.20	34.44	15.96	19.29
3DRN	46.70	35.60	27.70	**35.54**	**16.79**	**19.76**
Shang's+GA	43.00	28.90	20.80	18.65	8.70	10.72
VRD-GCN+GA	**59.50**	40.50	27.85	24.33	10.27	12.97
3DRN+GA	57.89	**41.80**	**29.15**	24.97	11.31	14.36
GSTEG+GA	51.50	39.50	28.23	–	–	–

Table 2
Performance comparison of different methods on VidOR dataset for video relation tagging and video retrieval by relation.

Method	Relation Tagging			Video Retrieval by Relation		
	mP@1	mP@5	mP@10	mAP	mR@10	mR@20
VRU'19#1	45.98	37.31	29.03	19.82	8.10	10.10
3DRN	47.68	37.90	29.55	**26.82**	**11.10**	**13.47**
VRU'19#1+GA	51.20	40.73	28.37	18.41	4.14	5.58
3DRN+GA	**52.59**	**42.33**	**29.89**	19.63	6.84	8.03


3-D Relation Network for Visual Relation Recognition in Videos
Adult- in front of -adult
Adult-push-bicycle
Adult-above-adult
Adult-towards-adult
Bicycle- above -bicycle
Adult- in front of -adult
Bicycle- in front of -adult
Bicycle-beneath-adult
Adult-ride-bicycle
Adult-above-bicycle
Adult caress dog
Red panda walk toward red panda
Adult hold hand of baby
Adult caress dog
Red panda- larger-red panda
Red panda-stand left-red panda
Red panda-taller-red panda
Red panda-stand left-red panda
Red panda-sit above-red panda
Red panda-walk left-red panda
Red panda-stand left-red panda
Red panda-taller-red panda
Red panda-walk away-red panda
Red panda-move beneath-red panda
Person ride bicycle
Person stand left person
Person ride bicycle
Bird fly left bird

(1-a) : 3DRN for video relation tagging

(2-a) :VRU'19#1 for video relation tagging

(1-b) :3DRN for video retrieval by relation

(2-b) : VRU'19#1 for video retrieval by relation

(3-a): Shang's for video relation tagging

(4-a): VRD-GCN for video relation tagging

(3-b): Shang's for video retrieval by relation

(4-b): VRD-GCN for video retrieval by relation

Fig. 3. Qualitative examples of video relation tagging (top-5 results) and video retrieval by relation (top-2 results). We compare our 3DRN with its competitors: Shang's, VRU'19#1 and VRD-GCN on the two datasets. Green and red marks denote the correct and incorrect results, respectively.

5.2. Results and discussion

Experimental results on the VidVRD task are presented in Table 3, where the best performances regarding each metric are highlighted in bold.

On the ImageNet-VidVRD dataset, the comparison on the metric mAP shows that the performance of 3DRN+GA is comparable with the state-of-the-art VRD-GCN+GA but significantly superior to Shang's+GA and GSTEG+GA. This demonstrates that 3DRN+GA can achieve the competitive performance in fine-grained video visual relation detection with spatio-temporal grounding. In addition, we can observe from the left case in Fig. 4 that 3DRN+GA can detect the dynamic relation evolution between the object pairs, for example, first "lion walk toward lion", then "lion walk past lion",

It is worth noting that although our model obtains comparative on the comprehensive metric mAP in Table 3, it does not perform well on mR@50 and mR@100. However, considering the general advantage of our approach in relation triplet prediction (as shown

Table 3
The results comparison of 3DRN+GA with the state-of-the-art methods on the ImageNet-VidVRD and VidOR datasets.

Dataset	Method	Video Relation Detection		
		mAP	mR@50	mR@100
ImageNet -VidVRD	Shang's+GA	8.58	5.54	6.37
	GSTEG+GA	9.52	7.05	8.67
	VRD-GCN+GA	14.23	**7.43**	**8.75**
	3DRN+GA	**14.68**	5.53	6.39
VidOR	VRU'19#1+GA	**6.56**	**6.89**	**8.83**
	3DRN+GA	2.47	2.58	2.75

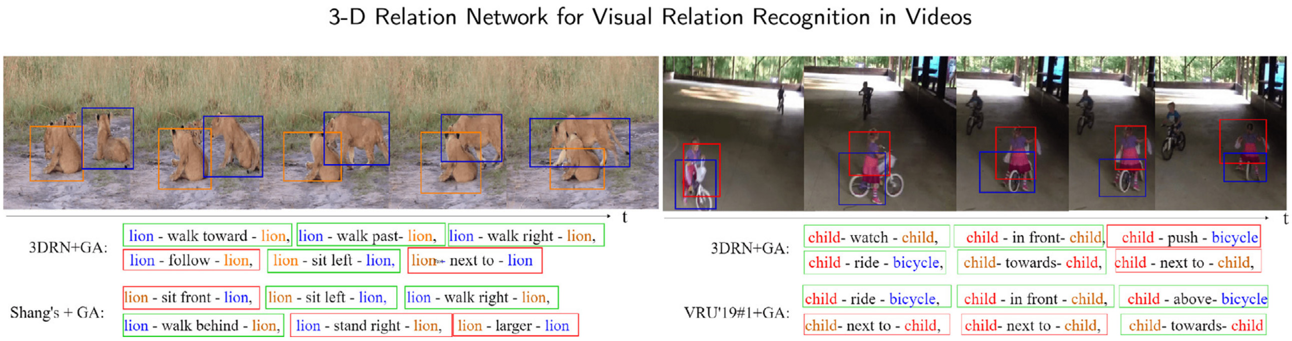


Fig. 4. Qualitative examples of VidVRD using different methods on the ImageNet-VidVRD (shown on the left side) and VidOR (shown on the right side) datasets. The top-6 results are provided, and the trajectories recorded in the examples are generated by 3DRN+GA. Green and red borders denote the correct and incorrect results.

Table 4
Ablation study on the influence of GA on the VidVRD task. The results are reported on short video clips.

Dataset	Method	mAP	mR@50	mR@100
ImageNet -VidVRD	Shang's-asso	14.16	15.53	18.42
	3DRN-asso	**17.05**	**16.80**	**19.01**
VidOR	VRU'19#1-asso	**17.28**	**27.83**	**34.76**
	3DRN-asso	16.23	20.30	23.39

in Section 4.5), we think that it can be the greedy association algorithm (GA) that causes such loss. This is because 3DRN adopts an approximate representation of the object's trajectory during generating the relation proposal, which is not harmonious with the used GA to achieve an optimal performance. In other words, the process of GA aggravates the inaccuracy of the object localization, and thus weaken the overall performance.

Similar phenomenon is also observed on VidOR (Table 3). As the right case shown in Fig. 4, after GA, since the generated trajectory of the child in orange boarder has not sufficient vIoU with the ground truth, therefore, the relation instances related to this child are regarded as incorrect results. Actually, the video length in VidOR is much longer than that of ImageNet-VidVRD (more than 3 times), which presents even greater challenge to associate the relation instances detected in short video clips. This means that our model needs to go through more association processes to localize the relation instances spatially and temporally, thus causing a greater loss in performance. However, VRU'19#1+GA adopts a tricky approach to overcome the challenge by ensembling many object detection and tracking techniques for better object localization before the relation recognition.

5.3. Ablation study

Although 3DRN+GA does not achieve competitive results as VRU'19#1+GA in Table 3, we argue that it does not contradict the effectiveness of our approach, which focuses on facilitating the end-to-end learnt features for better visual relation recognition in videos. In order to minimize the influence of association in our evaluation, we conduct an additional ablation study on short video clips directly. Specifically, we replace the test sets with a set of short video clips, that are obtained by dividing each video into length 30 with 15 overlapping frames. Moreover, the judging condition about the location in VidVRD task is weakened from object trajectories to frame-level bounding boxes. Thus a relation instance is counted as correct if the relation triplet is correct and both bounding boxes have sufficient IoU (set to be 0.5) with the ground truth. We similarly use mAP, mR@50 and mR@100 as the metrics.

The models which do not pass through the association process are recorded as Shang's-asso, VRU'10#1-asso and 3DRN-asso in Table 4 respectively. From the results displayed in Table 4, we can see that when testing directly with video clips as input, the methods generally perform better than which with associating. Particularly, we can find that under this test setting, 3DRN is superior to Shang's on ImageNet-VidVRD dataset, and achieves comparable performance with VRU'19#1 on VidOR dataset. This observation can confirm our previous argument about the challenge in associating relations in long videos. The results also demonstrate that our proposed relation proposal module can effectively localize relevant regions for feature extraction.

6. Conclusion

In this paper, we explore the essential role of the end-to-end feature learning method in visual relation recognition. For achieving the method, the relation proposals are introduced as an intermediate bridge to integrate two sub-modules that cannot be connected together in the pipeline framework. We particularly design a deep structured model, named 3DRN, to predict multiple relations from the multi-modal cues between two observed objects. And a suitable training strategy is provided for achieving effective optimization. Extensive experiments on ImageNet-VidVRD and VidOR datasets demonstrate the effectiveness of the proposed feature learning method in the 3DRN framework for predicting the diverse and accurate relations. In addition, we apply the model in VidVRD task to spatio-temporally ground the relations. Experimental results also verify the potential of 3DRN to tackle the task. In the future, we are going to explore applying these relations to video content understanding tasks, such as video question answering and video captioning.

CRediT authorship contribution statement

Qianwen Cao: Conceptualization, Methodology, Software, Writing - original draft. **Heyan Huang:** Supervision. **Xindi Shang:** Data curation, Writing - review & editing. **Boran Wang:** Validation, Software. **Tat-Seng Chua:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work is supported by National Key R&D Plan (No. 2018YFB1005100), NSFC (No. U19B2020, 61772076, 617 51201 and 61602197), NSFB(No. Z181100008918002), the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative, Google Cloud and China Scholarship Council CSC, No. 201806030102. 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] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, W. Shi, Real-time video super-resolution with spatio-temporal networks and motion compensation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4778–4787.

[2] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[3] Y.W. Chao, S. Vijayanarasimhan, B. Seybold, D.A. Ross, J. Deng, R. Sukthankar, Rethinking the faster r-cnn architecture for temporal action localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1130–1139.

[4] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, T. Mei, Relation distillation networks for video object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7023–7032.

[5] J. Devlin., M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. arXiv preprint arXiv:1810.04805.

[6] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[7] M. Douze, J. Revaud, J. Verbeek, H. Jégou, C. Schmid, Circulant temporal encoding for video retrieval and temporal alignment, International Journal of Computer Vision 119 (2016) 291–306.

[8] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, H. Huang, Heterogeneous memory enhanced multimodal attention model for video question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1999–2007.

[9] J. Gao, R. Ge, K. Chen, R. Nevatia, Motion-appearance co-memory networks for video question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6576–6585.

[10] R. Girdhar, J. Carreira, C. Doersch, A. Zisserman, A better baseline for ava, 2018. arXiv preprint arXiv:1807.10066.

[11] W. Han, P. Khorrami, T.L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, T.S. Huang, Seq-nms for video object detection, 2016. arXiv preprint arXiv:1602.08465.

[12] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2014) 583–596.

[13] J. Hou, X. Wu, X. Zhang, Y. Qi, Y. Jia, J. Luo, Joint commonsense and relation reasoning for image and video captioning, in: AAAI,2020, pp. 10973–10980.

[14] W. Jin, Z. Zhao, M. Gu, J. Yu, J. Xiao, Y. Zhuang, Multi-interaction network with object relation for video question answering, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1193–1201.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[16] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016. arXiv preprint arXiv:1609.02907.

[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.J. Li, D.A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International Journal of Computer Vision 123 (2017) 32–73.

[18] B. Li, Y. Wang, Visual relationship detection using joint visual-semantic embedding, in: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, IEEE. pp. 3291–3296.

[19] Y. Li, W. Ouyang, X. Wang, X. Tang, Vip-cnn: Visual phrase guided convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1347–1356.

[20] W. Liao, B. Rosenhahn, L. Shuai, M. Ying Yang, Natural language guided visual relationship detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[21] S. Liu, Z. Ren, J. Yuan, Sibnet: Sibling convolutional encoder for video captioning, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 1425–1434.

[22] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, Visual relationship detection with language priors, European Conference on Computer Vision, Springer (2016) 852–869.

[23] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems (2013) 3111–3119.

[24] J. Mun, L. Yang, Z. Ren, N. Xu, B. Han, Streamlined dense video captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6588–6597.

[25] X. Qian, Y. Zhuang, Y. Li, S. Xiao, S. Pu, J. Xiao, Video relation detection with spatio-temporal graph, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 84–93.

[26] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems (2015) 91–99.

[27] M. Rochan, Y. Wang, Video summarization by learning from unpaired data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7902–7911.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (2015) 211–252.

[29] K. Schoeffmann, M.J. Primus, B. Muenzer, S. Petscharnig, C. Karisch, Q. Xu, W. Huerst, Collaborative feature maps for interactive video search, International Conference on Multimedia Modeling, Springer (2017) 457–462.

[30] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, T.S. Chua, Annotating objects and relations in user-generated videos, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, ACM, 2019, pp. 279–287.

[31] X. Shang, T. Ren, J. Guo, H. Zhang, T.S. Chua, Video visual relation detection, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM, 2017, pp. 1300–1308.

[32] X. Shang, J. Xiao, D. Di, T.S. Chua, Relation understanding in videos: A grand challenge overview, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 2652–2656.

[33] M. Shvets, W. Liu, A.C. Berg, Leveraging long-range temporal relationships between proposals for video object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9756–9764.

[34] G. Singh, Spatio-temporal relational reasoning for video question answering. Ph.D. thesis. University of British Columbia, 2019.

[35] J. Song, L. Gao, L. Liu, X. Zhu, N. Sebe, Quantization-based hashing: a general framework for scalable image and video retrieval, Pattern Recognition 75 (2018) 175–187.

[36] C. Sun, A. Shrivastava, C. Vondrick, K. Murphy, R. Sukthankar, C. Schmid, Actor-centric relation network, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 318–334.

[37] X. Sun, T. Ren, Y. Zi, G. Wu, Video visual relation detection via multi-modal feature fusion, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2657–2661.

[38] B. Thomee, D.A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.J. Li, Yfcc100m: The new data in multimedia research, 2015. arXiv preprint arXiv:1503.01817.

[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[40] Y.H.H. Tsai, S. Divvala, L.P. Morency, R. Salakhutdinov, A. Farhadi, Video relationship reasoning using gated spatio-temporal energy graph, 2019. arXiv preprint arXiv:1903.10547.
[41] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.
[42] L. Wang, W. Li, W. Li, L. Van Gool, Appearance-and-relation networks for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1430–1439.
[43] P. Weinzaepfel, Z. Harchaoui, C. Schmid, Learning to track for spatio-temporal action localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3164–3172.
[44] J. Xu, T. Yao, Y. Zhang, T. Mei, Learning multimodal attention lstm networks for video captioning, in: Proceedings of the 25th ACM International Conference on Multimedia, ACM, 2017, pp. 537–545.
[45] R. Zellers, M. Yatskar, S. Thomson, Y. Choi, Neural motifs: Scene graph parsing with global context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5831–5840.
[46] Y. Zhan, J. Yu, T. Yu, D. Tao, On exploring undetermined relationships for visual relationship detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5128–5137.
[47] H. Zhang, Z. Kyaw, S.F. Chang, T.S. Chua, Visual translation embedding network for visual relation detection, in: CVPR, 2017, p. 5.
[48] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, A. Elgammal, Relationship proposal networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5678–5686.
[49] J. Zhang, Y. Peng, Hierarchical vision-language alignment for video captioning, International Conference on Multimedia Modeling, Springer (2019) 42–54.
[50] B. Zhao, X. Li, X. Lu, Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7405–7414.
[51] X. Zhu, Y. Wang, J. Dai, L. Yuan, Y. Wei, Flow-guided feature aggregation for video object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 408–417.

Qianwen Cao received the B.S. degree from Beijing Institute of Technology in 2017, and now she is a Ph.D. candidate in computer science and technology from Beijing Institute of Technology. Her research interests include machine learning, computer vision, visual content understanding. She was funded by China Scholarship Council to participate in research as a visiting student in National University of Singapore multimedia laboratory.

Heyan Huang is currently professor and dean of School of Computer Science and Technology in Beijing Institute of Technology of China. She received her Ph.D. degree in Computer Science and Technology in 1989 from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her current research interests include machine learning, information retrieval, and natural language processing. She has published over 100 research papers in reputed journals and conferences, such as TKDE, IJCAI, AAAI, ACL, WWW, COLING. She serves on the editorial boards of International Journal of Advanced Intelligence and Journal of Computer Research and Development. She has undertaken 20 more research projects including National 863 Project of China, National 973 Project of China, National Natural Science Foundation of China, etc.

Xindi Shang received the B.S. degree from Zhejinag University (Zhejiang, China) in 2016, and now he is a Ph. D. candidate in School of Computing from National University of Singapore. His research interests include computer vision, crowdsourcing and explainable AI.

Boran Wang received the B.S. degree from Jilin University (Jilin, China) in 2017. In 2020, he received the M.S.degree in computer science and technology from Beijing Institute of Technology, Beijing, China. His research interests include machine learning, dialogue system, natural language processing.

Tat-Seng Chua is the KITHCT Chair Professor at the School of Computing, National University of Singapore. He was the Founding Dean of the School from 1998-2000. His main research interest is in multimedia information retrieval and social media analytics. He is also the Director of a joint research Center between NUS and Tsinghua (NExT) to research into big unstructured multi-source multimodal data analytics. He is active in the international research community. He was the recipient of ACM SIGMM Technical Achievement Award 2015. He has also organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia and text processing. He was the conference co-chair of ACM Multimedia 2005, ACM CIVR 2005, ACM SIGIR 2008 and ACM Web Science 2015. He serves in the editorial boards of: ACM Transactions of Information Systems (ACM), The Visual Computer (Springer Verlag), and Multimedia Tools and Applications (Kluwer). He is the chair of steering committee of ICMR (International Conference on Multimedia Retrieval) and Multimedia Modeling conference series.