Video Visual Relation Detection via Multi-modal Feature Fusion

Xu Sun[1,2], Tongwei Ren[1,2,*], Yuan Zi[1], and Gangshan Wu[1]
[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] Shenzhen Research Institute of Nanjing University, Shenzhen, China
sunx@smail.nju.edu.cn,rentw@nju.edu.cn,ziyuan@smail.nju.edu.cn,gswu@nju.edu.cn

ABSTRACT

Video visual relation detection is a meaningful research problem, which aims to build a bridge between dynamic vision and language. In this paper, we propose a novel video visual relation detection method with multi-model feature fusion. First, we detect objects on each frame densely with the state-of-the-art video object detection model, flow-guided feature aggregation (FGFA), and generate object trajectories by linking the temporally independent objects with Seq-NMS and KCF tracker. Next, we break the relation candidates, *i.e.*, co-occurrent object trajectory pairs, into short-term segments and predict relations with spatial-temporal feature and language context feature. Finally, we greedily associate the short-term relation segments into complete relation instances. The experiment results show that our proposed method outperforms other methods by a large margin, which also earned us the first place in visual relation detection task of Video Relation Understanding Challenge (VRU), ACMMM 2019.

CCS CONCEPTS

• **Computing methodologies → Computer vision**.

KEYWORDS

Video visual relation detection; object trajectory detection; relation prediction

ACM Reference Format:
Xu Sun[1,2], Tongwei Ren[1,2,*], Yuan Zi[1], and Gangshan Wu[1]. 2019. Video Visual Relation Detection via Multi-modal Feature Fusion. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3343031.3356076

1 INTRODUCTION

Video visual relation describes dynamic interactions between co-occurrent objects in video with object trajectory pair and ⟨subject, predicate, object⟩ triplet, with the ability to provide comprehensive semantic understanding of video content [11]. It can be utilized by numerous high level visual-language tasks, such as video captioning [16], video summurization [18] and video retrieval [7].

Comparing with visual relation detection (VRD) on static image [8], video visual relation detection (VidVRD) is much more practical and challenging than VRD. Firstly, the interactions between objects are dynamic in video, therefore, effective temporal feature is required for relation recognition. Secondly, the variability of visual relations over time is another thorny issue, not a problem though, in static image situations.

To tackle these problems, several well designed models have been proposed. Shang *et al.* [11] introduce the first VidVRD method which adopts a bottom-up strategy. They split videos into segments with fixed duration and predict visual relations between co-occurrent short-term object tracklets for each video segment. Then they generate complete relation instances by a greedy associating procedure. Tsai *et al.* [6] proposed Gated Spatio-Temporal Energy Graph for video relation detection. The method models the spatial and temporal structure of relations in video by a fully-connected spatial-temporal graph. It also utilizes an energy function with adaptive parameterization to meet the diversity of relations, therefore achieves the state-of-the-art performance. However, the incomprehensive features utilized by existing methods still leave room for improvement of performance.

In this paper, we propose a novel video visual relation detection method explicitly combining spatial-temporal feature and language context feature, with the assistance of object trajectory detection model. Since visual relation detection is based on object detection, our proposed model consists of two modules: object trajectory detection and visual relation prediction. Figure 1 shows an overview of the proposed method.

Different to the existing works on video object detection focusing on per-frame accuracy, we build an object trajectory detection module combining the state-of-the-art video object detection model, FGFA [19], Seq-NMS [3] and KCF tracker [5] for trajectory generation. FGFA is an elaborate object detection model, which effectively exploits the temporal coherence among consecutive frames by optical flow guided feature aggregation. Seq-NMS is a post-procedure, which links object detection results predicted by FGFA on adjacent frames to generate preliminary trajectories. Then we use high-speed KCF tracker to refine the trajectories. The proposed visual relation prediction module adopts similar strategy as Shang *et al.* [11] to deal with the temporal variety of relations. We split relation candidates, i.e., co-occurrent object trajectory pairs into short-term segments and extract relative location feature and motion feature as spatial-temporal feature. Next, we embed identified object categories into high dimensional feature vectors to encode language context information using word2vec [9] pretrained with large scale textual dataset, GoogleNews. Then the spatial-temporal feature and language feature are fed into two classifiers which are trained independently. The confidences generated by the two classifiers are combined linearly to predict the predicates. Finally, we greedily associate the relation segments into complete relation instances.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3356076

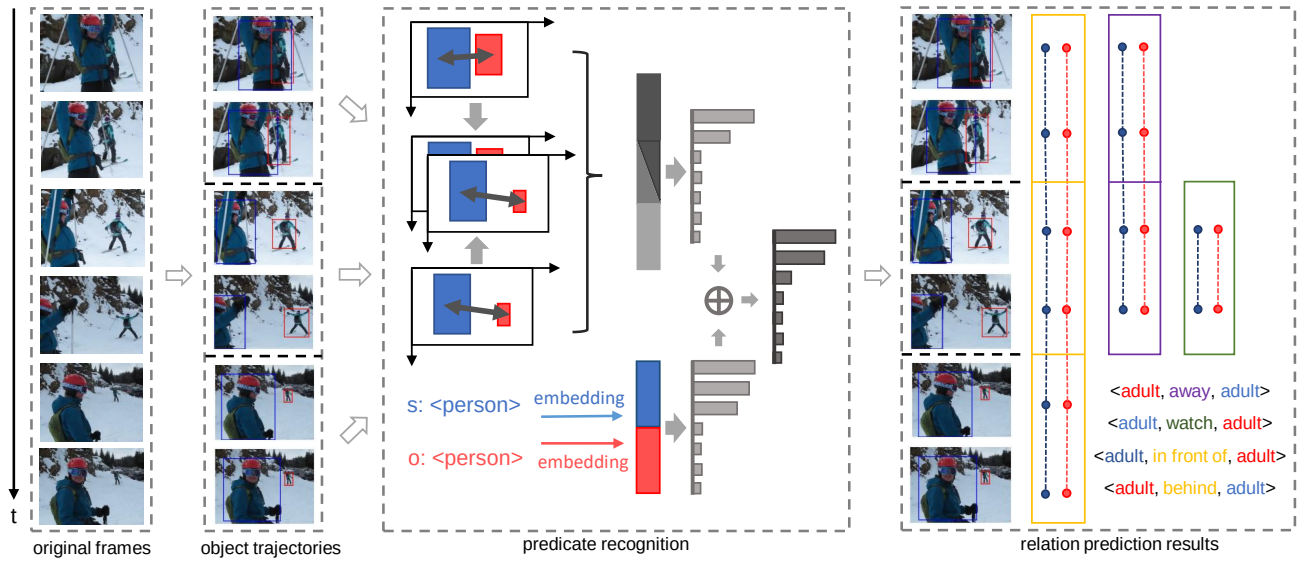

Figure 1: An overview of the proposed method. The boxes with different colors represent different predicates in the relation prediction results.

2 PRELIMINARY

2.1 Flow-guided feature aggregation

Flow-guided feature aggregation (FGFA) [19] is an elaborate object detection method, derived from a classic image object detector R-FCN [1] for video. It improves the per-frame feature by fusing nearby features under guidance of dense optical flow [14]. In this way, the accuracy of detection is significantly improved, since the intractable problems existed in video detection, such as motion blur, rare poses, are effectively relieved. FGFA consists of a feature extraction network [4] and optical flow network [14]. Specifically, the feature extraction network is applied on individual frames for per-frame feature maps. The optical flow network estimates motion between nearby frames and reference frame. The feature maps from consecutive frames are warped and fused with the reference frame with motion flow.

2.2 Seq-NMS

Seq-NMS [3] is a post-process procedure for video object detection. The main purpose of Seq-NMS is to boost the hard detection instances with rare poses or blurry appearance by exploiting the temporal coherence in adjacent video frames. Specifically, when a video sequence and per-frame object detections are given, it associates the bounding boxes in the same category on consecutive frames according to their overlaps, to generate trajectories and to rescore the individual detections. Due to the plain nature of trajectory generation criterion, we modify and utilize it to generate short-term preliminary trajectories at a low computing cost. Then we extend and associate the preliminary trajectories with visual tracking algorithm.

2.3 KCF

High-speed tracking with kernelized correlation filter (KCF) [5], the third prize winner in VOT Challenge in 2014, is one of the most powerful visual tracking algorithm. KCF tracker adopts discriminative tracking strategy who utilizes circulant matrix to sample patch and discrete fourier transform to accelerate. The greatest strength of KCF is that it achieves real time without compromising its accuracy, particularly the case in short duration. In this paper, we use KCF tracker to extend the preliminary trajectories generated by Seq-NMS in a concurrent way and generate complete object trajectories by further associate the short-term ones.

3 METHOD

3.1 Object trajectory detection

Object trajectory detection is an essential part of video visual relation detection model, which determines the ceiling of the performance. A small number of high-quality trajectory detection instances are required to avoid combinatorial explosion in relation candidate generation. To this end, we comprehensively combine several advanced techniques including video object detection and visual tracking to solve this problem.

First, we train the FGFA model [19] with ResNet-101 network [4] as backbone, which is pretrained with ImageNet [2]. Then we apply the FGFA model to densely detect 300 individual objects on each video frame. The detections whose confidences are lower than a threshold are filtered out, which is 0.01 in our experiments. With filtered per-frame object detections, we are able to generate short-term preliminary trajectories in association by applying Seq-NMS [3] at a low computing cost. Since the criterion for association is simply based on overlap in Seq-NMS, the quality of generated trajectories is hard to control. Therefore, we make the bounding box association criterion stricter. Two boxes B_i and B_{i+1} categorized as C_i and C_{i+1} on frame i and $i+1$ can be associated, if and only if:

(1) they are in the same category, $C_i = C_{i+1}$,

(2) the overlap between the two boxes is larger than a threshold, $IoU(B_i, B_{i+1}) > \alpha$,

(3) the different between the scales of the two boxes is less than a threshold, $|\frac{B_i^h}{B_i^w} - \frac{B_{i+1}^h}{B_{i+1}^w}| < \beta$,

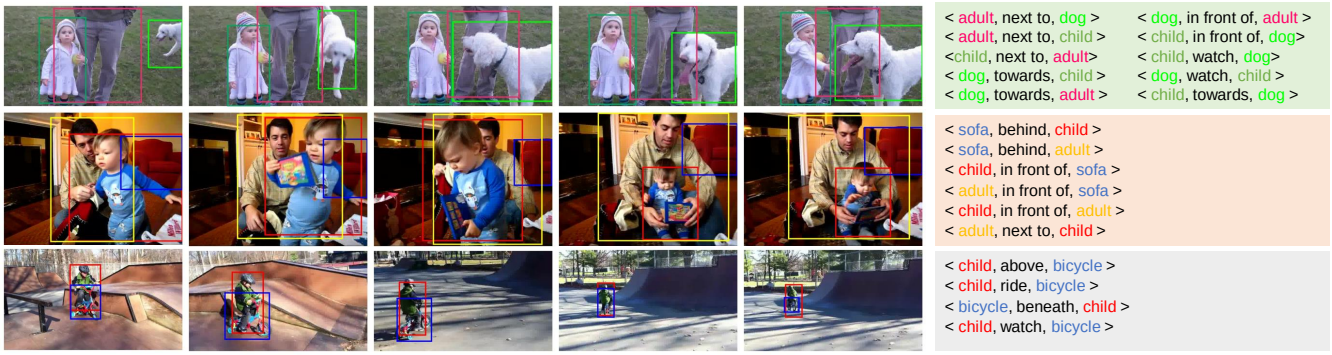

Figure 2: Qualitative results on VidOR validation set.

where B^h and B^w are height and width of box B, α is 0.8 and β is 0.3, $|\cdot|$ is absolute value.

After Seq-NMS, we obtain a set of short-term preliminary trajectories, which may be incomplete. We track both the head and tail of each preliminary trajectories and try to further associate them into complete ones. To reduce the processing time, we implement this procedure in a concurrent way, so that all CPUs can be thoroughly utilized. Finally, we further filter out the trajectory instances whose confidences are lower than a threshold (0.05 in our experiments). Only 20 trajectory instances at most with the highest scores are reserved for relation prediction.

3.2 Relation instance generation

The proposed visual relation prediction module adopts the similar bottom-up strategy as Shang *et al.* [11], which consists of three steps: relation candidate breaking, predicate recognition on segments and relation segment association.

For each object trajectory pair, we break the co-occurrent part into segments with fixed duration, which is 15 frames, as relation candidate segments. We explicitly combine spatial-temporal information and language context information extracted from each segment for predicate prediction. The spatial-temporal feature consists of relative location feature and motion feature. The relative location feature, $f_{Loc} = [s_x, s_y, s_w, s_h, s_a]$, utilized by the proposed method is a widely used spatial feature [17]. It is calculated as follows: $s_x = \frac{x-x'}{w}$, $s_y = \frac{y-y'}{h}$, $s_w = \log\frac{w}{w'}$, $s_h = \log\frac{h}{h'}$, $s_a = \log\frac{h\cdot w}{h'\cdot w'}$, where (x, y, w, h) and (x', y', w', h') are the box coordinates of subject and object respectively. The proposed motion feature captures relative location variety over time between subject and object:

$$f_{Mot} = f^e_{Loc} - f^s_{Loc}, \tag{1}$$

where f^s_{Loc} and f^e_{Loc} is the relative location feature extracted on the start frame and the end frame of relation candidate segment respectively. The spatial-temporal feature f_{ST} is generated by concatenating the features mentioned above as $[f^e_{Loc}, f^s_{Loc}, f_{Moc}]$. The language context feature f_{Lan} is a 1200-D vector, which is the concatenation of subject and object category embedding vectors. We use a word2vec [9] model pretrained on large scale textual data, GoogleNews, for category embedding. It successfully captures both statistical prior and language context information, the effectiveness of which has been proved by numerous visual relation methods [8, 20].

After that, f_{Lan} and f_{ST} are fed into two independent two-layered fully-connected neural networks with leaky ReLU [15] and softmax for predicate prediction, which are trained separately. The loss function used in training stage is defined as follows:

$$h = W_h\theta(f) + b_h, \tag{2}$$

$$\mathcal{L} = -\log softmax(W_o\theta(h) + b_o), \tag{3}$$

where $\theta(\cdot)$ denotes the activation function, f denotes spatial-temporal feature or language context feature, $W_h \in \mathbb{R}^{1024\times len(f)}$, $W_o \in \mathbb{R}^{50\times 1024}$, $b_h \in \mathbb{R}^{1024}$, $b_h \in \mathbb{R}^{50}$. In training stage, we use SGD optimizer with learning rate defaulted at 0.01. We train the two classifiers respectively for 40 epochs and divide the learning rate by 10 for each 10 epochs. The connections between neuron are randomly dropped out by 50 percent in order to improve the performance. In testing stage, we linearly combine the confidences, *i.e.*, $P(c_p|f_{Lan})$ and $P(c_p|f_{ST})$, generated by the two classifiers to predict the predicates:

$$P(c_p|f_{Lan}, f_{ST}) = \lambda P(c_p|f_{Lan}) + (1-\lambda)P(c_p|f_{ST}), \tag{4}$$

where c_p denotes predicate category, λ is set to 0.7.

Finally, we generate complete relation instances by greedily associating the relation segments with same triplet predictions.

4 EXPERIMENTS

4.1 Dataset and experiment settings

In this paper, we use a large-scale video object relation dataset, VidOR [10] for experiments. The dataset consists of 10,000 user-generated videos from social media on 80 object categories and 50 predicate categories, densely annotated with object trajectories and visual relation triplets. The dataset is divided into three parts: 7,000 for training, 835 for evaluation, 2,165 for final testing. The average length of the videos in VidOR is 35.73 seconds, which is much longer than that of ILSVRC-VID dataset. The large scale of the dataset poses great challenge on both prediction accuracy and computing effectiveness.

We use Recall@50, Recall@100, tagging precision@1, tagging precision@5 and mAP for evaluation, in which mAP is the official metric used in VRU Challenge. To match a predicted relation instance$(\langle s, p, o\rangle^p, T^p_s, T^p_o)$ and groundtruth instance$(\langle s, p, o\rangle^g, T^g_s, T^g_o)$, the following requirements should be satisfied:

(1) the triplets are exactly same, $\langle s, p, o\rangle^p = \langle s, p, o\rangle^g$,

Table 1: Component analysis results on VidOR validation set.

method	tagging precision@1	tagging precision@5	Recall@50	Recall@100	mAP
Ours\KCF	50.21	40.34	5.49	6.72	5.27
Ours\Lan	43.75	35.32	6.00	7.60	4.95
Ours\ST	50.84	40.41	6.66	8.64	6.33
Ours	**51.20**	**40.73**	**6.89**	**8.83**	**6.56**

Table 2: Comparison with the state-of-the-arts on VidOR validation set.

method	tagging precision@1	tagging precision@5	Recall@50	Recall@100	mAP
OTD+CAI	48.31	38.49	6.19	8.16	5.65
OTD+GSTEG	51.20	37.26	6.40	8.43	5.58
Ours	**51.20**	**40.73**	**6.89**	**8.83**	**6.56**

(2) $vIoU(T_s^p, T_s^g) >= 0.5$ and $vIoU(T_o^p, T_o^g) >= 0.5$, where $vIoU$ refers to the volume intersection over union [12, 13],

(3) $ov_{pg} \leq ov_{pg'}$, where g' indicates any mismatch groundtruth instances, $ov_{pg} = \min(vIoU(T_s^p, T_s^g), vIoU(T_o^p, T_o^g))$.

4.2 Component analysis

The proposed method consists of two essential parts including object trajectory detection and relation prediction. We analyze the two modules respectively on VidOR validation set.

Object trajectory detection. We combine FGFA [19] and Seq-NMS [3] to generate preliminary object trajectories and refine the results by applying KCF tracker [5]. To prove the effectiveness of the refinement procedure, we construct a degraded implementation without tracking as baseline (Ours\KCF). Table 1 indicates the performance of the baseline and the proposed method. In comparison with Ours and Ours\KCF, Recall@50 and Recall@100 decrease by 1.17% and 1.92% respectively. The experiment results prove that the tracking and the association process introduced in Sec 3.1 effectively improve the integrity of object trajectory detection results and further boost relation detection performance.

Relation prediction. Relation prediction relies on spatial-temporal feature and language context feature in our proposed model, encompassing complementary information. To illustrate the effectiveness of the two types of features, we construct two variations from the proposed model. (1) Ours\ST eliminates spatial-temporal stream and predicts predicate with language context stream only. (2) Ours\Lan eliminates language context stream and utilizes spatial-temporal stream only. The evaluation results are shown in Table 1. Both Ours\ST and Ours\Lan are inferior in performance than Ours. The results illustrate that both spatial-temporal information and language context information improve the recognition accuracy. The superiority of Ours\ST's performance over that of Ours\Lan leads to the conclusion that language context information plays a vital role in the proposed model.

4.3 Comparison with state-of-the-arts

To evaluate the effectiveness of the proposed method, we construct two baselines with the state-of-the-art visual relation detection methods on image and video respectively. CAI [20] is one of the most well-known visual relation detection models on static image. By proposing a context-aware relation recognition model boosted with attention mechanism, it achieves outstanding performance. We replace the relation prediction module of our proposed framework with CAI model to construct a video relation detection baseline (OTD+CAI). GSTEG [6] is a novel visual relation recognition method on video. We extend GSTEG with our object trajectory detection module to meet the requirements of video relation detection task and construct another state-of-the-art baseline (OTD+GSTEG). We compare the performances using Vidor validation set, the experiment results is indicated in Table 2. It shows that the proposed method outperforms all the state-of-the-art baselines on most of the evaluation metrics. Despite its satisfying performance in VRD task, CAI [20] is still unable to overcome the difficulties in VidVRD task without the assistance of temporal feature. GSTEG [6] exploits spatial-temporal structure of relations in video, but the model only utilizes visual feature for predicate recognition, leaving room for improvement. Compared with the baselines, the multi-model feature used by the proposed model is more feasible for dynamic relation recognition. Figure 2 provides some qualitative results generated by the proposed method.

Table 3: Component analysis results on VidOR test set.

method	tagging precision@5	mAP
RELAbuilder	23.60	0.546
Ours	**42.10**	**6.310**

Table 3 indicates the evaluation results on the test set of VidOR dataset. Our method is superior than the method comes second by a large margin.

5 CONCLUSION

We introduced a novel visual relation detection method on video, which consists of an object trajectory detection module and a visual relation prediction module. Specifically, the object trajectory detection module comprehensively combines FGFA, Seq-NMS and KCF tracker. The visual relation prediction module, on the other hand, adopts bottom-up strategy and recognizes relation with multi-model feature. The experiment results indicate that our superiority over the state-of-the-art baselines and other solutions competed in visual relation detection task of Video Relation Understanding Challenge.

6 ACKNOWLEDGEMENTS

This work is supported by National Science Foundation of China (61202320), Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20180307151516166), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*. 379–387.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[3] 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).

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

[5] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2014. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2014), 583–596.

[6] 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 *IEEE Conference on Computer Vision and Pattern Recognition*. 10424–10433.

[7] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. 2007. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM International Conference on Image and Video Retrieval*. 494–501.

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

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

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

[11] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *ACM International Conference on Multimedia*. 1300–1308.

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

[13] Xu Sun, Yuantian Wang, Tongwei Ren, Zhi Liu, Zheng-Jun Zha, and Gangshan Wu. 2018. Object trajectory proposal via hierarchical volume grouping. In *ACM International Conference on Multimedia Retrieval*. 344–352.

[14] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. 2013. DeepFlow: Large displacement optical flow with deep matching. In *IEEE International Conference on Computer Vision*. 1385–1392.

[15] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* (2015).

[16] Haonan Yu, Wang Jiang, Zhiheng Huang, Yang Yi, and Xu Wei. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *IEEE Conference on Computer Vision Pattern Recognition*. 4584–4593.

[17] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual translation embedding network for visual relation detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5532–5540.

[18] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *European Conference on Computer Vision*. 766–782.

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

[20] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. 2017. Towards context-aware interaction recognition for visual relationship detection. In *IEEE International Conference on Computer Vision*. 589–598.