MULTI-FOCUS GUIDED SEMANTIC AGGREGATION FOR VIDEO OBJECT DETECTION

Haihui Ye, Guangge Wang, Yang Lu⋆*, Yan Yan, Hanzi Wang*

Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen, China
haihui_ye1@163.com, guanggew@stu.xmu.edu.cn, {luyang, yanyan, hanzi.wang}@xmu.edu.cn

ABSTRACT

For the task of video object detection, it is useful to aggregate semantic information from supporting frames. However, existing methods only focus on the current frame during the semantic aggregation, called **Single-Focus** methods. They neglect semantic information among supporting frames and deteriorate overall performance. In this work, we propose a method called Multi-Focus guided Semantic Aggregation (MFSA) for video object detection. We introduce a novel Relation Propagation Module (RPM) to capture and propagate proposal-to-proposal semantic dependencies. Moreover, we propose a simple yet effective **Multi-Focus** strategy to leverage captured dependencies to guide feature enhancement at a batch level. Aided by this strategy, our method can greatly improve aggregation efficiency of **Single-Focus** methods and enhance the accuracy of a per-frame detector significantly with negligible computing overhead. We perform extensive experiments on the ImageNet VID dataset. The results show that MFSA achieves excellent performance and a superior speed-accuracy tradeoff among the competing methods.

Index Terms— Video Object Detection, Semantic Aggregation, Relation Propagation

1. INTRODUCTION

Object detection is a vital task in computer vision and achieves remarkable successes on images. However, directly applying image detectors [1, 2] into videos leads to poor performance. This is due to low-quality frames caused by motion blur, occlusion or camera defocus. These frames make video object detection non-trivial and challenging. However, supporting frames among the video can provide valuable location and appearance information that helps detection for the target video frame. Towards how to utilize the semantic information, we classify existing methods into aggregation-based and propagation-based methods.

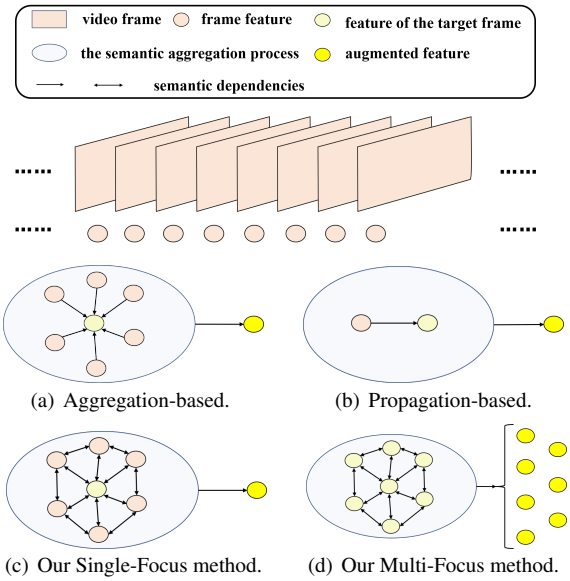

Fig. 1. Comparisons between our methods (c and d) and others when aggregating information from other frames. Best viewed in color.

As shown in Figure 1(a), aggregation-based methods [3, 4, 5, 6, 7] store complete-sized feature maps of multiple supporting frames in the memory. They aim at providing sufficient high-level semantic information for the target frame. These methods have the strength of better accuracy performance. As shown in Figure 1(b), propagation-based methods [8, 9, 10, 11, 12] split video frames into sparse keyframes and dense non-keyframes. They usually propagate high-level semantic information of keyframes to low-level features of the current frame for feature alignment. These methods have the advantage of faster inference speed.

However, during aggregation or propagation, these methods only focus on the target frame. We call this strategy as the **Single-Focus** strategy. In general, this strategy has two main drawbacks. First, existing solutions usually overlook valuable semantic information among the supporting frames. Without the information, it is difficult to enhance features by considering long-range dependencies and build an overall relation for the target frame. We refer to this problem as *focus ineffectiveness*. Second, **Single-Focus** methods only utilize the

⋆Corresponding author: Yang Lu, luyang@xmu.edu.cn.

This work was supported in part by the Open Research Projects of Zhejiang Lab (NO. 2021KB0AB03); in part by the National Natural Science Foundation of China under Grant 62002302 and 62071404; in part by the China Fundamental Research Funds for the Central Universities under Grant 20720210099.

semantic information preliminarily, in which the information is leveraged to provide guidance only for the target frame. In this case, they neglect the relation between the captured semantic dependencies and supporting frames, and thus aggregation efficiency deteriorates drastically. We refer to this problem as *focus inefficiency*.

To alleviate these two problems, in this paper, we propose a Multi-Focus guided Semantic Aggregation (MFSA) method. In this work, we regard frames participating in semantic aggregation as a batch. MFSA explores the overall relation and further leverages the built relation to provide semantic guidance for a batch level of frames. To be specific, we introduce a Relation Propagation Module (RPM) to refine valuable semantic information by fully considering the proposal-to-proposal semantic correspondence among a batch of frames, including the supporting frames. Based on RPM, we propose a base method with the **Single-Focus** strategy (see Figure 1(c)). An improvement towards solving the problem of *focus ineffectiveness* is obtained. Then, we leverage the information to provide semantic guidance for feature alignment towards the batch of frames simultaneously. We call this strategy as the **Multi-Focus** strategy. Aided by the strategy, we propose a formal method called Multi-Focus guided Semantic Aggregation, referred to as MFSA (see Figure 1(d)). MFSA innovatively extends the focus from only one target frame (see Figure 1(c)) to a batch of frames (see Figure 1(d)) during one semantic aggregation. Then, MFSA performs the semantic aggregation at a batch level, which can avoid the repeated and time-consuming aggregation computation. Thus, the problem of *focus inefficiency* can be mitigated significantly.

To sum up, our contributions are listed as follows: (1) we introduce a novel Relation Propagation Module (RPM) to build the semantic dependencies by fully considering the overall correlation among a batch of frames. (2) we introduce a simple yet effective **Multi-Focus** strategy to leverage the distilled semantic information to provide semantic guidance for feature augmentation at a batch level. (3) based on RPM and the **Multi-Focus** strategy, we propose a MFSA method for video object detection. MFSA can expand the detection range from an isolated frame to a batch of frames.

From a per-frame perspective, MFSA can be considered as a lightweight but powerful variant to a per-frame detector that can perceive global relation among the video. It is worth noting that MFSA improves the accuracy significantly while introducing negligible computing overhead and maintaining similar inference speed as the per-frame detector. We perform extensive experiments on the ImageNet VID dataset [13]. The results demonstrate the superiority of our method. To the best of our knowledge, MFSA achieves state-of-the-art accuracy performance of 83.6% mAP among methods using ResNet-101 as the backbone, without using complex data augmentation and offline post-processing techniques. Moreover, our method achieves an outstanding tradeoff among accuracy, inference speed and model size.

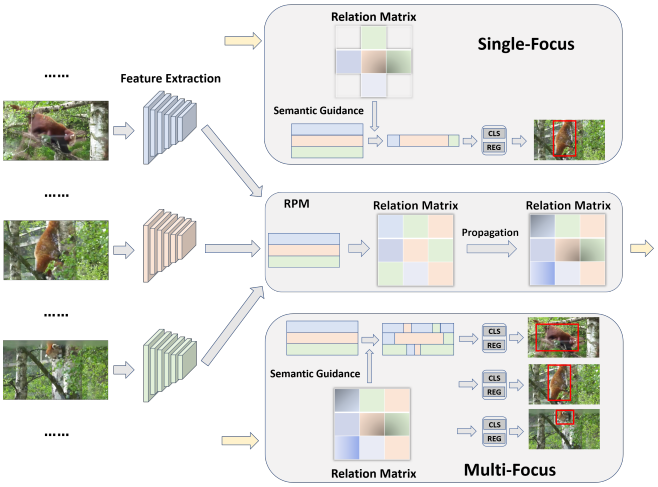

Fig. 2. The framework of our proposed Single-Focus and Multi-Focus methods. Yellowish arrows denote information flow of the semantic relation matrix. Best viewed in color.

2. PROPOSED METHOD

2.1. Framework Overview

The overall framework is shown in Figure 2. Faster-RCNN is adopted as the basic detector. Assume that the aggregation size (the number of frames in Figure 1) is 3 and the frames sampled globally from a video are denoted as F_{t-l}, F_t and F_{t+l}. Here, l is the sampling stride. We refer to these frames during one semantic aggregation as a batch and the batch is denoted as $F = \{F_{t-l}, F_t, F_{t+l}\}$. The feature extraction $\mathcal{N}_{feat}$ receives F as input and produces the intermediate features f ($f = \{f_{t-l}, f_t, f_{t+l}\}$). Then, we perform semantic aggregation on f to obtain the relation matrix at the proposal level. We introduce a Relation Propagation Module (RPM) to mine and propagate the overall semantic coherence among frames of the batch. Then, we obtain a more accurate relation matrix and leverage the relation to provide semantic guidance for feature enhancement for f. During the semantic aggregation, we extend the focus from only one target frame (Single-Focus) into a batch of frames (Multi-Focus). We refer to the two methods as Single-Focus and Multi-Focus guided semantic aggregation, namely SFSA and MFSA respectively. Finally, we generate classification and regression results for the enhanced feature simultaneously.

2.2. Multi-Focus Guided Semantic Aggregation

Relation Propagation Module for Focus Effectiveness. Previous methods usually adopt self-attention to build the dependencies. They ignore the overall semantic information among supporting frames and suffer from *focus ineffectiveness*. In this section, we elaborate how to promote *focus effectiveness* by introducing the Relation Propagation Module

(RPM) to build semantic dependencies among frames, which is inspired by the relation mechanism [14, 15, 16, 17, 18].

Formally, suppose a batch of video frames is $F = \{F_1, F_2, ..., F_K\}$. Here K denotes the number of aggregated frames (the size of the batch) and $F_i \in F$ indicates the i-th input frame. At fisrt, we input F to $\mathcal{N}_{feat}$ ($e.g.$ ResNet-101) and obtain complet-sized features of F, denoted as $f = \{f_1, f_2, ..., f_K\}$. $R^t = \{R_1^t, R_2^t, ..., R_M^t\}$ denote region features generated by the RPN network of f_t. Here M denotes the number of selected regions according to the confidence score. f_t means the t-th frame feature of f and $R_i^t \in \mathbb{R}^{1\times D}$ indicates the i-th region of f_t. D is the dimention of region feature and we set it as 1024 in this work. Take $K = 2$ for example. We measure the semantic affinity $a_{i,j}$ between either specific pair of candidate boxes (R_i, R_j) with the modified cosine similarity:

$$a_{i,j} = \sigma(R_i)^T \cdot \varphi(R_j), \tag{1}$$

where $\sigma(.)$ and $\varphi(.)$ represent the transformation with the nonlinear and normalization function. Here $a_{i,j}$ lies in the range of $(0, 1)$. Then, the obtained semantic affinities can be used to represent the proposal-to-proposal affinities. The measurement of regions is not limited to different frames because we argue that regions within the same frame can also provide valuable semantic information. $A^{m,n}$ is the affinity matrix between f_m and f_n, where $A^{m,n} \in \mathbb{R}^{M\times M}$.

Assume R and A are selected regions and overall semantic affinities of a batch of frames, where $R \in \mathbb{R}^{(KM)\times D}$ and $A \in \mathbb{R}^{(KM)\times(KM)}$. To capture long-range dependencies to build a more accurate and stable relation, we perform semantic aggregation on R for a batch of frames. Then we can update A by:

$$\hat{A} = Norm(A \times A + A), \tag{2}$$

where $\times$ means the matrix multiplication and $Norm$ means the normalization function. Here $\hat{A} \in \mathbb{R}^{(KM)\times(KM)}$. Eq. (3) is the core of relation propagation. For j-th column of A, $a_{.,j}$ denotes the semantic affinity between all regions of the batch and R_j. For i-th row of A, $a_{i,.}$ is the semantic affinity between R_i and all regions. After the matrix multiplication between $a_{i,.} \times a_{.,j}^T$, the global semantic information has been propagated and the newly obtained $a_{i,j}$ contains the global proposal-to-proposal semantic relation. We perform the semantic relation propagation based on the valuable semantic parts of the selected regions and propagate potentially valuable semantic information. To avoid the propagation loss of semantic information, we adopt the residual connection of ResNet [19]. We plus A after propagation, which is helpful to keep the original semantic relation between regions and makes the updating process more accurate and stable. Then, we normalize A to obtained $\hat{A}$.

Finally, we update the region features by:

$$\hat{R} = \hat{A} \times R, \tag{3}$$

where $\hat{R} \in \mathbb{R}^{(KM)\times D}$ contains valuable semantic information not only among the target frame and supporting frames but also among the supporting frames. Then, we append the refined regions into the following pipeline for classification and regression. Aided by RPM, we can model long-range semantic consistency among frames at the batch level. It is helpful to alleviate the problem of *focus ineffectiveness*.

Multi-Focus for Focus Efficiency. With RPM, the overall semantic relation has been built. However, we aim to empower semantic information more value for a batch of frames during semantic aggregation. Then, we propose a **Multi-Focus** strategy, in which we leverage the captured information to provide semantic guidance at the batch level.

SFSA only focuses on the current frame. But our formal proposed MFSA focuses on multiple frames simultaneously and performs a unified semantic aggregation for them to benefit detection. During the aggregation, semantic dependencies among the batch ($F = \{F_i\}_{i=0}^{K-1}$) are modeled as follows:

$$Relation = Map(F \to F). \tag{4}$$

To be specific, MFSA leverages the overall relation obtained by Eq. (4) to guide feature augmentation for frames at a batch level (see the lower part of Figure 2). Empowered by the **Multi-Focus** strategy, repeated and huge computation for building the relation can be avoided when performing detection on another $K - 1$ supporting frames of the batch. This is helpful to address *focus inefficiency*. Thus, the aggregation and detection efficiency can be improved largely.

3. EXPERIMENTS

3.1. Dataset and Evaluation Metrics

We evaluate the proposed method on the ImageNet VID dataset [13]. The ImageNet VID dataset consists of 3862 videos in the training set and 555 videos in the validation set, respectively. There are 30 categories of fully-annotated video snippets with bounding boxes and tracking IDs. Following the protocols in [8, 20, 21, 3, 10], we perform the evaluation on the validation set and use the mean average precision (mAP@IoU=0.5) metric.

3.2. Implementation Details

We choose Faster-RCNN [1] with ResNet-101 [19] pretrained on ImageNet as our basic detector. RPN is appended after conv4 stage During training and inference, the NMS threshold of IoU is set as 0.7, through which 300 candidate boxes are generated for each frame. We train our model on both ImageNet VID and DET datasets. Each training mini-batch contains three frames. The sampling stride is set to $[-l, -0.5l]$ and $[0.5l, l]$. l is the sampling stride. The ratio of frames sampled from the DET and VID are approximately set as $1 : 1$. 220K iterations of SGD training are carried out in an

Table 1. Ablation study on the ImageNet VID validation set. The results are obtained by three variants of our method. The best mAP results are highlighted by bold.

Methods	(a)	(b)	(c)
Single-Focus	✓	✓	
Multi-Focus			✓
RPM		✓	✓
mAP(%)(overall)	73.6	$\mathbf{83.7}_{\uparrow\mathbf{10.1}}$	$83.6_{\uparrow 10.0}$
mAP(%)(slow)	82.1	$90.4_{\uparrow 8.3}$	$\mathbf{90.6}_{\uparrow\mathbf{8.5}}$
mAP(%)(medium)	71.0	$\mathbf{82.5}_{\uparrow\mathbf{11.5}}$	$82.4_{\uparrow 11.4}$
mAP(%)(fast)	52.5	$\mathbf{67.3}_{\uparrow\mathbf{14.8}}$	$67.2_{\uparrow 14.7}$
Runtime(ms)	63.6	$344.6_{\uparrow 281.0}$	$66.2_{\uparrow 2.6}$

end-to-end fashion. For inference, 21 frames are contained in each batch. We only utilized simple data augmentation (only random flipping operation) in our experiments. All the experiments are performed on the NVIDIA GTX 2080Ti.

3.3. Ablation Study

Table 1 details the performance comparisons across different variants of our method. Faster R-CNN (method (a) in Table 1) simply performs object detection on a single frame, which achieves 73.6% mAP. Method (b) (SFSA in Figure 2) achieves a significant boost of 10.1% mAP (from 73.6% to 83.7%). The result verifies that with RPM to build the overall semantic relation among frames, a significant step has been achieved towards solving the problem of *focus ineffectiveness*. Furthermore, we apply RPM with the **Multi-Focus** strategy (MFSA in Figure 2), denoted as method (c) in Table 1. Obviously, method (c) still obtains an absolute boost of 10.0% mAP to method (a). Moreover, method (b) and (c) obtain superior results to the baseline on the slow-, medium- and fast-motion groups. However, method (b) encounters huge runtime consumption (5.4 times to method (a)). This is because great and repeated consumption of aggregation is unavoidable when adopting the **Single-Focus** strategy. Notice that with the **Multi-Focus** strategy towards addressing *focus inefficiency*, method (c) runs at a comparable speed to method (a) (+2.6 ms) but only introduces little degradation of accuracy (-0.1% mAP) to method (b). That verifies the effectiveness of the **Multi-Focus** strategy to enhance the aggregation and detection efficiency. Therefore, we choose method (c) as our formal method due to the better overall performance.

3.4. Comparisons with state-of-the-art methods

As shown in Table 2, we compare our method with several state-of-the-art methods. To the best of our knowledge, our proposed MFSA achieves 83.6% mAP, which is the best performance among methods without complex operations (such as adding DCN [24], complex data augmentations [3, 25] or offline post-processing techniques [26, 27, 28]). Meanwhile,

Table 2. Comparisons with several state-of-the-art methods. The results are from their papers, where GPUs are different. K, X, XP and ti mean K40, TITAN X, TITAN XP and GTX 1080 Ti, respectively. Ti, adopted in our method, represents RTX 2080Ti. The best results are highlighted by bold. (1), (2) and (3) mean per-frame, propagation-based and aggregation-based methods, respectively.

Types	Methods	mAP (%)	Model Size (params)	runtime (fps)
(1)	Faster-RCNN [1]	73.6	59.3M	15.7(Ti)
(2)	DFF [8]	73.1	96.6M	20.3 (K)
	LWDN [10]	76.3	–	20.0 (X)
	PLSA [11]	77.1	**63.7M**	**30.8 (V)**
	LSTS [12]	77.2	–	23 (V)
(3)	FGFA [20]	76.3	100.4M	1.4 (K)
	MANet [21]	78.1	–	5.0 (XP)
	SELSA [3]	82.7	–	2.9 (Ti)
	LRTR [5]	79.3	–	10.0 (X)
	OGEMN [22]	79.3	–	8.9 (ti)
	MEGA [4]	82.9	–	8.7 (Ti)
	HVR-Net [23]	83.2	–	–
	MFSA (Ours)	**83.6**	65.3M	15.1 (Ti)

MFSA achieves a superior tradeoff among accuracy, inference speed and model size. Among propagation-based methods ((2) in Table 2), PSLA and LSTS run at a faster speed but suffer from inferior accuracy than MFSA ($\downarrow 6.5\%$ and $\downarrow 6.4\%$ mAP, respectively). For the model size, MFSA only introduces $\uparrow 6M$ to the baseline and $\uparrow 1.6M$ to PLSA. Among aggregation-based methods ((3) in Table 2), MFSA outperforms HVR-Net and MEGA with better accuracy ($\uparrow 0.4\%$ and $\uparrow 0.7\%$ mAP) with better accuracy performance. Notice that LRTR [5] presents a light-weight modification to a per-frame detector and shares a similar motivation with MFSA. Obeviously, MFSA achieves better overall performance.

4. CONCLUSION

In this work, we propose a **Multi-Focus** guided Semantic Aggregation (MFSA) method for video object detection. In MFSA, we introduce a Relation Propagation Module (RPM) to consider overall semantic relation among a batch of frames to capture and propagate valuable semantic information. Moreover, we propose a simple yet effective **Multi-Focus** strategy to leverage the relation to provide semantic guidance for feature alignment at the batch level. Therefore, a significant step towards solving the huge aggregation computation burden has been achieved. Therefore, our method can address the problems of *focus ineffectiveness* and *inefficiency* significantly. Compared with several competing methods, our method achieves promising performance and an outperforming speed-accuracy tradeoff on the ImageNet VID dataset.

5. REFERENCES

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and et al., "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. of NIPS*, 2015, pp. 91–99.

[2] Jifeng Dai, Yi Li, Kaiming He, and et al., "R-fcn: Object detection via region-based fully convolutional networks," in *Proc. of NIPS*, 2016, pp. 379–387.

[3] Haiping Wu, Yuntao Chen, Naiyan Wang, and et al., "Sequence level semantics aggregation for video object detection," in *Proc. of ICCV*, 2019, pp. 9217–9225.

[4] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang, "Memory enhanced global-local aggregation for video object detection," in *Proc. of CVPR*, 2020, pp. 10337–10346.

[5] Mykhailo Shvets, Wei Liu, and Alexander C Berg, "Leveraging long-range temporal relationships between proposals for video object detection," in *Proc. of ICCV*, 2019, pp. 9756–9764.

[6] Fei He, Naiyu Gao, Qiaozhe Li, and et al., "Temporal context enhanced feature aggregation for video object detection.," in *Proc. of AAAI*, 2020, pp. 10941–10948.

[7] Zhu Chen, Weihai Li, Chi Fei, and et al., "Spatial-temporal feature aggregation network for video object detection," in *Proc. of ICASSP*, 2020, pp. 1858–1862.

[8] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, and et al., "Deep feature flow for video recognition," in *Proc. of CVPR*, 2017, pp. 2349–2358.

[9] Xizhou Zhu, Jifeng Dai, Lu Yuan, and et al., "Towards high performance video object detection," in *Proc. of CVPR*, 2018, pp. 7210–7218.

[10] Zhengkai Jiang, Peng Gao, Chaoxu Guo, and et al., "Video object detection with locally-weighted deformable neighbors," in *Proc. of AAAI*, 2019, pp. 8529–8536.

[11] Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, and et al., "Progressive sparse local attention for video object detection," in *Proc. of ICCV*, 2019, pp. 3909–3918.

[12] Zhengkai Jiang, Yu Liu, Ceyuan Yang, and et al., "Learning where to focus for efficient video object detection," in *Proc. of ECCV*. Springer, 2020, pp. 18–34.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, and et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[14] Hao Zhou, Chongyang Zhang, and Chuanping Hu, "Visual relationship detection with relative location mining," in *Proc. of ACM MM*, 2019, pp. 30–38.

[15] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu, "Video visual relation detection via multi-modal feature fusion," in *Proc. of ACM MM*, 2019, pp. 2657–2661.

[16] Han Hu, Jiayuan Gu, Zheng Zhang, and et al., "Relation networks for object detection," in *Proc. of CVPR*, 2018, pp. 3588–3597.

[17] Xiankai Lu, Wenguan Wang, Chao Ma, and et al., "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proc. of CVPR*, 2019, pp. 3623–3632.

[18] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, and et al., "End-to-end memory networks," in *Proc. of NIPS*, 2015, pp. 2440–2448.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and et al., "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016, pp. 770–778.

[20] Xizhou Zhu, Yujie Wang, Jifeng Dai, and et al., "Flow-guided feature aggregation for video object detection," in *Proc. of ICCV*, 2017, pp. 408–417.

[21] Shiyao Wang, Yucong Zhou, Junjie Yan, and et al., "Fully motion-aware network for video object detection," in *Proc. of ECCV*, 2018, pp. 542–557.

[22] Hanming Deng, Yang Hua, Tao Song, and et al., "Object guided external memory network for video object detection," in *Proc. of ICCV*, 2019, pp. 6678–6687.

[23] Mingfei Han, Yali Wang, Xiaojun Chang, and et al., "Mining inter-video proposal relations for video object detection," in *Proc. of ECCV*, 2020, pp. 431–446.

[24] Jifeng Dai, Haozhi Qi, Yuwen Xiong, and et al., "Deformable convolutional networks," in *Proc. of ICCV*, 2017, pp. 764–773.

[25] Lijian Lin, Haosheng Chen, Honglun Zhang, and et al., "Dual semantic fusion network for video object detection," in *Proc. of ACM MM*, 2020, pp. 1855–1863.

[26] Wei Han, Pooya Khorrami, Tom Le Paine, and et al., "Seq-nms for video object detection," *arXiv preprint arXiv:1602.08465*, 2016.

[27] Kai Kang, Hongsheng Li, Junjie Yan, and et al., "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE TCSVT*, vol. 28, no. 10, pp. 2896–2907, 2017.

[28] Jiajun Deng, Yingwei Pan, Ting Yao, and et al., "Relation distillation networks for video object detection," in *Proc. of ICCV*, 2019, pp. 7023–7032.