Spatio-Temporal Object Detection From UAV On-Board Cameras

Daniel Cores[0000−0002−5548−4837], Víctor M. Brea[0000−0003−0078−0425], and Manuel Mucientes[0000−0003−1735−3585]

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) - Universidade de Santiago de Compostela, Santiago de Compostela, Spain
{daniel.cores,victor.brea,manuel.mucientes}@usc.es

Abstract. We propose a new two stage spatio-temporal object detector framework able to improve detection precision by taking into account temporal information. First, a short-term proposal linking and aggregation method improves box features. Then, we design a long-term attention module that further enhances short-term aggregated features adding long-term spatio-temporal information. This module takes into account object trajectories to effectively exploit long-term relationships between proposals in arbitrary distant frames. Many videos recorded from UAV on-board cameras have a high density of small objects, making the detection problem very challenging. Our method takes advantage of spatio-temporal information to address these issues increasing the detection robustness. We have compared our method with state-of-the-art video object detectors in two different publicly available datasets focused on UAV recorded videos. Our approach outperforms previous methods in both datasets.

Keywords: Object detection · Spatio-temporal features · CNN

1 Introduction

Object detectors precision has raised in recent years mainly fueled by the advances in Convolutional Neural Networks (CNNs). However, there are some scenarios that remain a huge challenge for state-of-the-art object detectors. Thus, videos recorded by on-board cameras mounted on Unmanned Aerial Vehicles (UAVs) are usually hard, mainly due to the high object density and the generally small object size. Moreover, camera movements might also increase the effect of motion blur that might degrade image quality at certain frames.

Traditional image object detectors are not designed to take into account temporal information available in videos. Therefore, the extended approach of applying a traditional object detector at frame level is suboptimal when it comes to video object detection. Spatio-temporal frameworks have been proposed to exploit spatio-temporal information to tackle occlusion and motion blur issues, generally increasing the detection precision. Still, most state-of-the-art video object detectors fail to effectively exploit spatio-temporal information when dealing with crowded images containing small objects.

This paper proposes a new spatio-temporal object detector architecture designed to overcome the main issues concerning object detection in videos recorded by cameras mounted on UAVs. Our implementation is based on a two stage object detector architecture. First, a short-term object aggregation method is implemented to exploit spatio-temporal information from the nearby frames. Then, shot-term enhanced box features are fed to an attention module that establishes long-term relations among object proposals in distant frames.

The main contributions of this work are:

- A new strategy to link object proposals in neighbouring frames. We avoid the use of short object tubelets to reduce the overhead of including spatio-temporal information. Instead, our Region Proposal Network (RPN) is fed with per frame anchor boxes as in the single image domain. Then, proposals associated with the same anchor in consecutive frames are linked.
- A new attention method to establish long-term proposal relationships. Our implementation takes into account object trajectories to update proposal positions. Therefore, at a given frame the attention module is fed with updated positions for each proposal instead of the original location in the corresponding frame. This makes possible to compare geometry features between proposals originally calculated in distant frames for the first time.
- We evaluate our method in two publicly available datasets. Video sequences in these datasets were recorded by UAVs with built-in cameras in different scenarios. We also compare our results with state-of-the-art video object detectors, proving that our approach achieves the best results.

2 Related work

Single image object detectors follow two main approaches: two stage and one stage architectures. Two-stage object detectors [14] first generate object proposals, which are defined as regions with high probability of containing objects of interest. Then, the network head refines and classifies these proposals. One stage approaches [17] try to solve the detection problem without any proposal generator.

Using feature maps at different pyramid levels was first popularized by Feature Pyramid Network (FPN) [12]. Feature maps with different resolutions make the network robust against a wide range of object sizes. This idea was further developed in PANet [13] and EfficientDet [15].

Recently, the success of attention mechanisms in the natural language processing domain modeling different element dependencies [18] was implemented in the single image object detection [7]. It allows to establish relationships between object proposals to enhance box features.

The main approach to address the video object detection problem is to aggregate spatial features through time getting more robust feature maps. Several works have proposed to perform this aggregation at pixel level. Optical flow was first used by these methods to link features in the nearby frames [23, 19]. As an alternative, the use of deformable convolutions was explored to identify

these relationships [1]. Recurrent Neural Networks (RNN) have also been successfully applied to perform the pixel level aggregation in [20] by defining a new memory module that aggregates the spatio-temporal information. Motivated by the success achieved by attention mechanisms in the single image object detection domain, there have also been attempts to implement pixel level aggregation methods applying the same ideas [6].

As an alternative to pixel level aggregation, object level aggregation methods have also been implemented to effectively aggregate spatial information throughout time. These methods focus on areas of high probability of containing an object instead of aggregating spatial information from the whole image. Some spatio-temporal object detectors propose to link the per frame detection sets applying tracking algorithms [10, 9]. As a follow on, [8] includes a Tubelet Proposal Network (TPN) that links object proposals instead of final detections. It exploits the generally large receptive field of CNNs to propagate static proposals throughout nearby frames, and adapts each proposal in the corresponding frame to the actual object position. This idea of object linking by means of short tubelet generation is further developed in [16]. In this case, authors designed a Cuboid Proposal Network (CPN) in which object tubelets are initialized as anchor cuboids, a spatio-temporal extension of the concept of anchor boxes defined in the single image domain. Anchor cuboids where also used in [3] as the first step for short-term feature aggregation. Moreover, this framework also includes a long-term object linking algorithm that reuses short-term tubelets to increase the robustness of the association process. Instead of relying on anchor cuboids to link proposals, we propose a new method based on anchor box linking that reduces the overhead in comparison with a single image object detector baseline.

Attention mechanisms have also proved to be useful to establish relationships between object proposals in different frames. Authors in [4] successfully extended the method described in [7] for single images. However, this spatio-temporal approach only exploits short-term information. Long-term information is added in [2], implementing a location free attention mechanism that only uses appearance features, ignoring geometry information such as object position and shape. Comparing object positions in distant frames is not meaningful to establish object relationships and adds noise to the linking process. Alternatively, we keep track of object trajectories in order to update proposal positions throughout time making possible to also exploit geometry features in the long-term.

3 Method

We propose a short- and long-term aggregation method that can be applied to two stage object detectors in order to take advantage of spatio-temporal information available in videos. Both short- and long-term aggregation stages take as input per frame object proposals calculated following the same strategy as the single image baseline. This reduces the overhead of including these techniques on traditional object detectors.

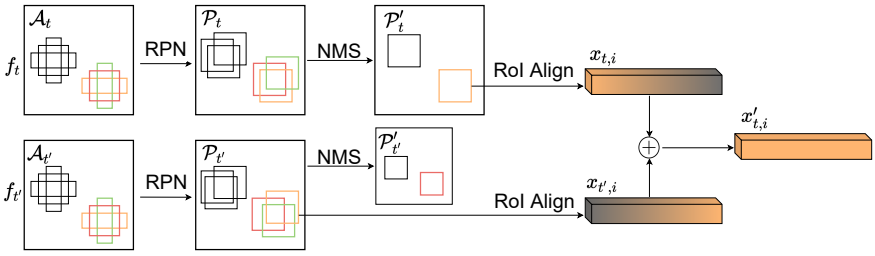

Fig. 1: Short-term aggregation process with one support frame $f_{t'}$.

Box proposal features at each reference frame f_t are enhanced with features from nearby support frames $f_{t-N}, ..., f_{t-1}, f_t, f_{t+1}, ..., f_{t+N}$ by an object level aggregation method. Sec. 3.1 describes the short-term linking and aggregation strategy. A long-term spatio-temporal module (Sec. 3.2) is fed with short-term aggregated box features to establish long-term relationships and further improve object features.

Most previous works use spatio-temporal features to localize and classify the object. In contrast, in our implementation, spatio-temporal information is only used to boost the classification precision, as we argue that the most valuable information to localize the object comes from the current frame. Therefore, we use spatial information to localize the object and spatial and spatio-temporal information to classify each object. The final classification score is calculated as:

$$p = p_{tmp} + p_{spt}(1 - p_{tmp}) \tag{1}$$

being p_{tmp} the classification score calculated with spatio-temporal features and p_{spt} the score of the classification in the reference frame with just spatial information.

3.1 Short-term feature aggregation

Our short-term aggregation method links proposals throughout the neighbouring frames and aggregates box features accordingly. Per frame object proposals are initialized as anchor boxes $\mathcal{A}_t = \{a_{t,i}\}_{i=1}^{A}$. Then, the proposal set $\mathcal{P}_t = \{p_{t,i}\}_{i=1}^{A}$ is calculated by an RPN modifying each anchor box to better fit the objects of interest. Each proposal $p_{t,i}$ consists of a bounding box $\mathbf{b}(p_i)$ and an objectness score $\mathbf{s}(p_i)$. Finally, spatially redundant proposals are removed applying Non-Maximum Suppression (NMS), getting the final proposal set $\mathcal{P}'_t$. This process is shown in Fig. 1.

For each reference frame f_t, we link proposals that come from the same anchor box in the same position for every supporting frame $f_{t'}$ in $\{f_{t-N}, ..., f_{t-1}, f_{t+1}, ..., f_{t+N}\}$. The high overlap in consecutive frames for the same object and the high field of view of CNNs make this lightweight linking strategy suitable for the short-term. However, since $|\mathcal{P}'_t| \leq |\mathcal{P}_t|$ due to the removed proposals, we link proposals in $\mathcal{P}'_t$ in the reference frame with proposals

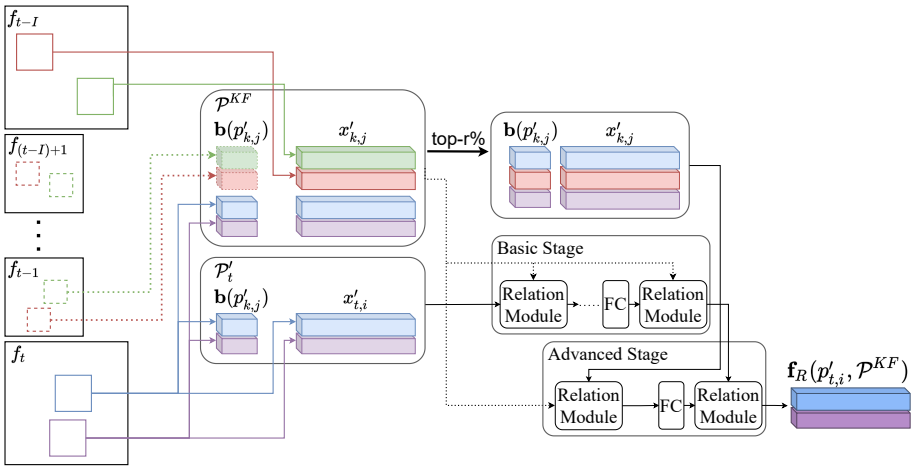

Fig. 2: Long-term aggregation strategy with one key frame f_{t-I}.

in $\mathcal{P}_{t'}$ in the supporting frames. Otherwise, proposals in $\mathcal{P}'_t$ and $\mathcal{P}_{t'}$ might not share the same anchors —see proposal associated with orange anchor and proposal associated with red anchor in Fig. 1 after NMS.

Box features are extracted for each proposal by the RoI Align method. Then box features that came from the same anchor box are aggregated (Fig. 1):

$$x'_{t,i} = \sum_{l=-N}^{N} \omega^s_{t+l,i}\, x_{t+l,i} \tag{2}$$

being $x'_{t,i}$ the output aggregated feature map calculated as a weighted sum of the per frame box feature maps $x_{t+l,i}$. The short-term aggregation weights $\omega^s_{t+l,i}$ are calculated as:

$$\omega^s_{t+l,i} = exp\left(\frac{x_{t,i}\, x_{t+l,i}}{|x_{t,i}||x_{t+l,i}|}\right) \tag{3}$$

normalizing the $\omega^s_{t+l,i}$ using a Softmax function to ensure that $\sum_{l=-N}^{N} \omega^s_{t+l,i} = 1$.

3.2 Long-term feature aggregation

Although our anchor-based strategy provides a lightweight effective linking method in the short-term, it is not suitable for the long-term. The high overlap between nearby frames cannot be assumed in the long-term. Therefore, we design a new method based on attention mechanisms to link proposals from distant supporting key frames selected at a fixed interval I.

The attention module calculates M relation features $\mathbf{f}^m_R$ given the support proposal set $\mathcal{P}^{KF}$ and proposals in the current frame $p'_{t,i}$:

$$\mathbf{f}_R^m(p'_{t,i}, \mathcal{P}^{KF}) = \sum_{k=1}^{K} \sum_{j=1}^{|\mathcal{P}'_k|} w^m_{t(i),k(j)} \, (W_V \, x'_{k,j}), \quad m = 1, ..., M \tag{4}$$

being W_V a transformation matrix optimised through backpropagation. $\mathcal{P}^{KF}$ includes the per key frame proposals, being K the number of key frames. We also include in $\mathcal{P}^{KF}$ proposals from current frame allowing to establish relationships also with current proposals. Each proposal $p'_{t,i}$ has an associated appearance feature $x'_{t,i}$ and geometry features $\mathbf{b}(p'_{t,i})$. Therefore, we use the short-term enhanced box features instead of the weaker RoI Align output of previous works [4, 2]. The relation weight $w^m_{t(i),k(j)}$ is calculated as:

$$w^m_{t(i),k(j)} = \frac{g^m_{t(i),k(j)} \, exp(a^m_{t(i),k(j)})}{\sum_q g^m_{t(i),q} \, exp(a^m_{t(i),q})} \tag{5}$$

being $a^m_{t(i),k(j)}$ the appearance weight and $g^m_{t(i),k(j)}$ the geometry weight. The appearance weight is calculated as a normalized dot product:

$$a^m_{t(i),k(j)} = \frac{\langle W_H \, x'_{t,i}, W_Q \, x'_{k,j} \rangle}{\sqrt{d_h}} \tag{6}$$

where W_H and W_Q of Eq. 6, as well as W_G in Eq. 7, are also learnt in the training process as W_V (Eq. 4). W_H and W_Q project the appearance features in the reference frame and supporting key frames respectively, being d_H the projected dimension.

Geometry weights are computed as:

$$g^m_{t(i),k(j)} = max\{0, W_G \, \mathcal{E}(\mathbf{b}(p'_{t,i}), \mathbf{b}'(p'_{k,j}))\} \tag{7}$$

where function $\mathcal{E}$ takes proposals bounding box definitions $\mathbf{b}(p')$ and embeds the vector $\left(log\left(\frac{|x_i - x_j|}{w_i}\right), log\left(\frac{|y_i - y_j|}{h_i}\right), log\left(\frac{w_j}{w_i}\right), log\left(\frac{h_j}{h_i}\right)\right)$ in a high dimensional representation following the method outlined in [18]. For features from key frames we do not use the original geometry features $\mathbf{b}(p'_{k,j})$, but a modified version $\mathbf{b}'(p'_{k,j})$ taking into account object trajectories —proposals with the same color in Fig. 2 belong to the same trajectory. We link proposals in consecutive frames reusing previously computed relation weights defining the association score between a proposal i in one frame and a proposal j in the next frame as:

$$\mathcal{S}_{i,j} = \mathbf{s}(p'_{\tau,i}) \, exp(\overline{w}_{\tau(i),k(j)}) \tag{8}$$

where $\overline{w}_{\tau(i),k(j)}$ is the average relation weight between proposal i in frame f_τ and proposal j in f_k. As we have previous relation weights already calculated, we can establish object trajectories from key frames to the previous frame f_{t-1} maximising the association score $\mathcal{S}_{i,j}$ applying the Hungarian method [11]. Using this updated geometry features allows to compare proposal positions as if they were in consecutive frames rather than in arbitrary distant frames.

Finally, box features are calculated as the concatenation of M relational features adding the result to the original appearance features:

$$\mathbf{f}_R(p'_{t,i}, \mathcal{P}^{KF}) = \mathbf{f}_R(p'_{t,i}, \mathcal{P}^{KF}) + concat[\{\mathbf{f}_R^m(p'_{t,i}, \mathcal{P}^{KF})\}_{m=1}^{M}] \quad (9)$$

We stack a set of relation modules following an approach similar to [4]. Fig. 2 shows this pipeline organized in *Basic* and *Advanced Stages*. The inputs to the *Basic Stage* are the key frame supporting proposals and the reference frame proposals $\mathcal{P}'_t$ that are iteratively improved. Then, the top r% proposals in $\mathcal{P}^{KF}$ are enhanced with the whole set $\mathcal{P}^{KF}$ as supporting proposals in the *Advanced Stage*. Finally, the second step of the *Advanced Stage* takes *Basic Stage* output and these enhanced proposals to calculate the final box features.

4 Experimental Results

4.1 Datasets

We evaluate our method in two publicly available datasets: Unmanned Aerial Vehicle Benchmark (UAVDT) [5] and VisDrone [21]. Both datasets are focused on videos recorded from on-board cameras mounted on UAVs. The UAVDT dataset contains 30 training videos and 20 testing videos recorded in urban areas with annotated objects belonging to one category: vehicles. The VisDrone dataset provides 56 training and 17 testing videos with annotations of 11 different object categories. Compared to UAVDT, the number of objects per frame is significantly higher in VisDrone with 25 and 85 objects per frame on average respectively.

4.2 Implementation details

In our implementation, per frame features are extracted at different FPN levels using ResNeXt-101 as backbone with deformable convolutions [22] on *conv3*, *conv4* and *conv5*. The backbone is pre-trained in the ImageNet classification dataset.

To train our spatio-temporal network, we first train the single frame baseline, setting the base learning rate to 1.25×10^{-3} for 45K iterations, and reducing it by $*0.1$ at 30K and 40K iterations. Then, we keep its weights frozen only training the spatio-temporal double head and the attention modules. For this spatio-temporal training, the initial learning rate is also set to 1.25×10^{-3} for 15K iterations, reducing it at 12K iterations. Input images are resized keeping the shortest dimension bellow 540px for UAVDT and 720px for VisDrone.

Short-term support frames and long-term key frames are selected following different strategies in the training and testing stages. In the short-term case, instead of selecting $2N + 1$ consecutive frames $\{f_{t-N}, ..., f_{t-1}, f_t, f_{t+1}, ..., f_{t+N}\}$ for training, we randomly select two support frames in $(t - N, ..., t + N)$ for each reference frame. In testing, every video frame is processed sequentially making possible to reuse previous backbone calculations. However, in training

Table 1: Comparison with state-of-the-art spatio-temporal frameworks.

(a) UAVDT

Method	$AP^{@.5}$	$AP^{@.75}$	$AP^{@.5-.95}$
FGFA [23]	57.6	25.6	28.9
RDN [4]	60.4	32.0	32.5
MEGA [2]	59.4	30.7	31.7
Ours	**61.0**	**37.1**	**34.9**

(b) Visdrone.

Method	$AP^{@.5}$	$AP^{@.75}$	$AP^{@.5-.95}$
FGFA [23]	30.7	11.8	14.1
RDN [4]	31.5	11.7	14.4
MEGA [2]	31.8	11.7	14.5
Ours	**32.1**	**12.9**	**15.4**

we randomly select a fixed sized subsample of frames to reduce the effect of very large videos. Therefore, this optimization cannot be applied, increasing the training time when working with large N. In the long-term case, we follow a similar strategy in the training stage, randomly selecting two key frames from the complete video rather than evenly spaced key frames from previous frames. In our experiments we set $N = 5$.

We also report the performance of state-of-the-art video object detection frameworks in the same datasets. We use the implementations provided in [2]. To ensure a fair comparison, we keep the same parameters —apart from input image resolution— unchanged for both datasets.

4.3 Results

In this section we compare our framework with the state-of-the-art spatio-temporal object detectors in the selected datasets. The spatio-temporal methods included in the comparison are FGFA [23], RDN [4] and MEGA [2]. We report the Average Precision at different IoU levels for every dataset.

Although our method uses frames in advance in the short-term, long-term key frames are selected from previous frames. Therefore, our implementation can give the detection set with just a few frames of delay. That is the case for all the spatio-temporal approaches in the comparison except for MEGA [2]. In this case, key frames are randomly selected from the complete video. Thus, this method might not be suitable for certain applications in which using so many frames in advance is not possible.

Table 1 shows the results in the UAVDT and VisDrone datasets. Our method outperforms all the other methods in the UAVDT dataset (Table 2a) at every IoU level. It also shows that our approach not only gets better results but the difference is bigger as we set a more demanding IoU. Thus, the difference with the best spatio-temporal object detector is of 0.6% in $AP_{@.5}$ while in $AP_{@.75}$ it is of 5.1%. In the VisDrone dataset (Table 2b) our framework also improves the other methods in all the metrics. As in the previous case, the difference is more significant in $AP_{@.75}$ and $AP_{@.5-.95}$ with 1.2% and 0.9% over MEGA, the best spatio-temporal framework in this dataset.

5 Conclusions

We have proposed a new framework for spatio-temporal object detection that effectively exploits both short- and long-term information in videos recorded from UAVs on-board cameras. First, proposals are linked in the nearby frames allowing to aggregate short-term spatio-temporal information. Then, enhanced box features are further enriched by an attention stage that takes into account object trajectories to exploit geometry features.

Our framework outperforms state-of-the-art spatio-temporal object detectors in two publicly available datasets focused on videos recorded from UAVs. This proves the suitability of our method for this challenging real application.

Acknowledgements

This research was partially funded by the Spanish Ministry of Science, Innovation and Universities under grants TIN2017-84796-C2-1-R and RTI2018-097088-B-C32, and the Galician Ministry of Education, Culture and Universities under grants ED431C 2018/29, ED431C 2017/69 and accreditation 2016-2019, ED431G/08. These grants are co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

1. Bertasius, G., Torresani, L., Shi, J.: Object detection in video with spatiotemporal sampling networks. In: IEEE International Conference on Computer Vision (ICCV) (2018)
2. Chen, Y., Cao, Y., Hu, H., Wang, L.: Memory enhanced global-local aggregation for video object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10337–10346 (2020)
3. Cores, D., Mucientes, M., Brea, V.M.: RoI feature propagation for video object detection. In: European Conference on Artificial Intelligence (ECAI) (2020)
4. Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., Mei, T.: Relation distillation networks for video object detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 7023–7032 (2019)
5. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: Object detection and tracking. In: European Conference on Computer Vision (ECCV). pp. 370–386 (2018)
6. Guo, C., Fan, B., Gu, J., Zhang, Q., Xiang, S., Prinet, V., Pan, C.: Progressive sparse local attention for video object detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 3909–3918 (2019)
7. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3588–3597 (2018)
8. Kang, K., Li, H., Xiao, T., Ouyang, W., Yan, J., Liu, X., Wang, X.: Object detection in videos with tubelet proposal networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

9. Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X., et al.: T-CNN: Tubelets with convolutional neural networks for object detection from videos. IEEE Transactions on Circuits and Systems for Video Technology **28**(10), 2896–2907 (2017)
10. Kang, K., Ouyang, W., Li, H., Wang, X.: Object detection from video tubelets with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
11. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
13. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8759–8768 (2018)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
15. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 10781–10790 (2020)
16. Tang, P., Wang, C., Wang, X., Liu, W., Zeng, W., Wang, J.: Object detection in videos by high quality object linking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
17. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 9627–9636 (2019)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
19. Wang, S., Zhou, Y., Yan, J., Deng, Z.: Fully motion-aware network for video object detection. In: IEEE International Conference on Computer Vision (ICCV) (2018)
20. Xiao, F., Jae Lee, Y.: Video object detection with an aligned spatial-temporal memory. In: European Conference on Computer Vision (ECCV) (2018)
21. Zhu, P., Wen, L., Bian, X., Ling, H., Hu, Q.: Vision meets drones: A challenge. arXiv preprint arXiv:1804.07437 (2018)
22. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9308–9316 (2019)
23. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: IEEE International Conference on Computer Vision (ICCV) (2017)