Relation networks for few-shot video object detection

Daniel Cores[1], Lorenzo Seidenari[2], Alberto Del Bimbo[2], Víctor M. Brea[1], and Manuel Mucientes[1]

[1] Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain
[2] Media Integration and Communication Center (MICC), University of Florence, Italy
daniel.cores@usc.es, victor.brea@usc.es, manuel.mucientes@usc.es,lorenzo.seidenari@unifi.it,alberto.delbimbo@unifi.it

Abstract. This paper describes a new few-shot video object detection framework that leverages spatio-temporal information through a relation module with attention mechanisms to mine relationships among proposals in different frames. The output of the relation module feeds a spatio-temporal double head with a category-agnostic confidence predictor to decrease overfitting in order to address the issue of reduced training sets inherent to few-shot solutions. The predicted score is the input to a long-term object linking approach that provides object tubes across the whole video, which ensures spatio-temporal consistency. Our proposal establishes a new state-of-the-art in the FSVOD500 dataset.

Keywords: few-shot object detection · video object detection

1 Introduction

The paradigm of few-shot object detection aims to address the issue of large training sets in modern object detectors, which shows up mainly in high annotation costs, or, which is worse, eventually in useless deep learning models, simply because there might not be enough data to train them. This makes few-shot object detection a timely topic.

Video object detectors leverage spatio-temporal information to tackle challenges such as motion blur, out of focus, occlusions or high changes in an object appearance to increase detection precision [9, 6, 7], which is not straightforward for image object detectors working on isolated frames. Similarly, the issue of large training sets as a need for a high precision in video object detectors calls for few-shot video object detectors, which today are less abundant in the literature than their image few-shot object detector counterpart.

Few-shot becomes even more challenging when referred to attention mechanisms, which today have become widely accepted for modeling object proposal relationships in video object detection [24], increasing precision, but at the cost

of multiple video frames in each training iteration. On the contrary, a few-shot video object detector should learn from a limited number of labeled instances per object category while exploiting spatio-temporal information at inference time. In a proper few-shot framework, whole videos are not available for training. This leads to a data gap between train and test sets that must be addressed.

All the above leads us to design a new few-shot object detector to take advantage of spatio-temporal information in videos. Our solution comprises a relation model with attention mechanisms and a classification score optimization component for spatio-temporal feature aggregation and consistency.

The main contributions of this work are:

- A method to bridge the data gap between training and test sets through synthetic frame augmentation, which permits the relation module to match new objects with proposals from the current image in the training phase, and avoids the degradation of our solution in the single image setting.
- A new spatio-temporal double head for spatial and spatio-temporal information with a spatial branch for object location and classification in the current frame, and a spatio-temporal branch to improve the classification and predict the overlap among detections and ground truth as a detection precision metric.
- Object tube generation from the predicted overlap among detections and ground truth to link detections in successive frames. This benefits from the generally high spatio-temporal redundancy of videos to modify the classification confidence of object detections.

Our approach outperforms both single image and previous video object detectors in FSVOD500, a specific dataset for few-shot video object detection.

2 Related Work

General object detectors based on deep neural networks fall mainly into two main categories: one- and two-stage detectors. Two-stage detectors [21, 2] generate a set of object proposals with high probability of containing an object of interest, and then perform a bounding box refinement and object classification. In contrast, one-stage detectors [16, 20] directly calculate the final detection set by processing a dense grid of unfiltered candidate regions.

More recently, object detection frameworks specifically designed to work with videos [9, 6, 7] were introduced to take advantage of spatio-temporal information, improving detection precision. The main idea behind these methods is to aggregate per-frame features throughout time, achieving more robust feature maps. This aggregation can be done at pixel- [12] or at object-level [9, 6, 7], linking object instances through neighboring frames.

Recently, the few-shot object detection problem has drawn significant attention in order to replicate the success achieved in the image classification field. Learning new tasks from just a few training examples is very challenging for most machine learning algorithms, especially for deep neural networks. Current

few-shot object detectors follow two main approaches: meta-learning and fine-tuning.

Meta-learning based methods are one of the research lines to address the few-shot object detection problem. They learn a similarity metric, so a query object can be compared with support sets from different categories. FSRW [14] proposes a framework based on YOLOv2 [20] including a feature reweighting module and a meta-feature extractor. Alternatively, a two-stage approach based on Faster/Mask R-CNN is proposed on Meta R-CNN [26]. This work proposes a new network head that applies channel-wise soft attention between RoI features given by a Region Proposal Network (RPN), and category prototypes to calculate the final detection set. The authors in [11] also propose to modify the RPN of a two-stage architecture, including an attention RPN that takes advantage of the support information to increase the RPN recall. A key component in all meta-learning based object detectors is to compute category prototypes from a set of annotated objects that are general enough to represent each object category. DAnA [5] implements a background attenuation block to minimize the effect of background in the final category prototype and also proposes a new method to summarize the support information in query position aware (QPA) support vectors.

TFA [25] proves that a simple transfer-learning approach, in which only the last layers of existing detectors are fine-tuned with new scarce data, can achieve results comparable to those of the state-of-the-art of meta-learning based detectors. The fine-tuning approach is further developed in DeFRCN [19], including a gradient decoupled layer (GDL) that modulates the influence of the RPN and the network classification and localization head in the training process and a prototypical calibration layer (PCB) to decouple the classification and localization tasks. Few-shot detectors that follow a fine-tuning approach suffer from a non-exhaustive training process, in which objects from novel categories in the base images are not annotated. Therefore, the model learns to treat novel categories as background in the pre-training stage. This issue is addressed in [3] by mining annotations from novel categories in base images. An automatic annotation framework was also proposed in [15] to expand the annotations available for the novel categories through label mining in unlabeled sets.

To the best of our knowledge, there is only one attempt to consider spatio-temporal information for few-shot object detection [10]. They are also the first to propose a dataset specifically designed for few-shot video object detection. Their method is based on a meta-learning approach, including a tube proposal network to associate objects across frames and a Tube-based Matching Network to compare tube features with support prototypes. Alternatively, we propose a fine-tuning based video object detector that implements feature aggregation at object level to enhance per-frame proposal features.

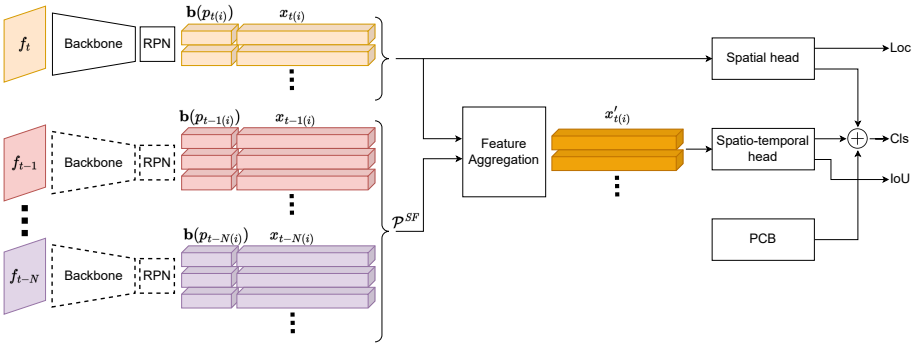

Fig. 1: FSVDet architecture overview, without the Confidence Score Optimization (CSO) module.

3 Proposed Method

3.1 Problem Definition

We follow the standard few-shot object detection setting established in previous works [14, 25, 11, 19]. The whole dataset is divided into $\mathcal{D}_{base}$, with several annotated objects of the base classes $\mathcal{C}_{base}$, and $\mathcal{D}_{novel}$, with only a few annotations of the novel classes $\mathcal{C}_{novel}$ for training, with $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$. The detection problem is defined as K-shot, where K is the number of objects annotated for each novel category $\mathcal{C}_{novel}$. Typically, the number of annotations K used to evaluate few-shot detectors in the literature ranges from 1 to 30 examples.

Video object detection aims to localize and classify objects in every frame of an input video. Object detection at each time step t is performed by considering information from the reference frame f_t and N previous frames $f_{t-N}, ..., f_{t-1}$. In the proposed framework, we also include a long-term optimization that considers information from the complete video to optimize the confidence of the detections. Although we propose a video object detector, annotations are object instances in single video frames. Thus, $\mathcal{D}_{novel}$ follows the same definition as for single-image approaches.

3.2 FSVDet: Few-shot video object detection

We propose FSVDet, a two-stage video object detection framework that can be trained with few labeled examples. The selected spatial baseline is DeFRCN [19], a modification of the original Faster R-CNN [21] able to perform a quick adaptation from a set of base categories $\mathcal{C}_{base}$ to a new domain $\mathcal{C}_{novel}$.

Fig. 1 shows an overview of the proposed network architecture. First, per-frame image features and object proposals are independently calculated. Then, proposal boxes $\mathbf{b}(p_{t(i)})$ and the corresponding features $\phi(p_{t(i)})$, extracted through RoI Align for proposal $p_{t(i)}$ in the current frame f_t, are fed to the spatial branch

of the double head. Object localization is exclusively performed with information extracted from the current frame, while the classification combines spatial and spatio-temporal information. Spatial classification and spatio-temporal classification scores for each category are combined as:

$$s = s_{tmp} + s_{spt}(1 - s_{tmp}) \tag{1}$$

being s_{spt} the spatial classification score and s_{tmp} the spatio-temporal score.

As our spatial baseline is DeFRCN [19], we also include the Prototypical Calibration Block (PCB) originally proposed in that work. This module applies a classification score refinement based on a similarity distance metric between network detections and category prototypes. For the category prototype calculation, annotated images from $\mathcal{D}_{novel}$ are fed to a CNN pre-trained on ImageNet to extract deep image features. Then, a RoI Align layer extracts object features with ground truth boxes, calculating K feature maps per each category in $\mathcal{C}_{novel}$. The final category prototype is calculated by an average pooling operator over the K feature maps of each category. The final classification score is calculated as follows:

$$s' = \beta \cdot s + (1 - \beta) \cdot s^{\cos} \tag{2}$$

being $s^{\cos}$ the cosine distance used as similarity metric between category prototypes and detection features extracted by the same CNN as the category prototypes. The hyperparameter β sets the tradeoff between the two confidence scores.

Finally, we define the Confidence Score Optimization method (CSO) that links object detections throughout the video and updates their classification scores. The linking method is based on the category agnostic scores and the overlap between detections in consecutive frames. The goal of this method is to modify the classification score of detections in each tube, ensuring spatio-temporal consistency. Further details of this method are given in Sec. 3.6.

3.3 Single image spatio-temporal training

Traditional video object detectors randomly select support frames in the input video sequence for each reference frame for training [9, 6, 7]. However, few-shot object detectors are trained with a limited number of annotated instances per category. Hence, the training set is composed of single images rather than fully annotated videos.

Our model overcomes this issue with the generation of a set $\mathcal{F}$ of synthetic support frames f_q for each training image I_t by inserting transformations of objects from I_t in different positions of f_q. In so doing, each annotated object $\alpha_{t(j)}$ from I_t is subject to horizontal random flipping and inserted L times in f_q in random positions within the image boundaries. The relative size of the object with respect to the whole image sets an upper bound for L. Undesirable artifacts from a naive insertion of a cropped object in a different position from

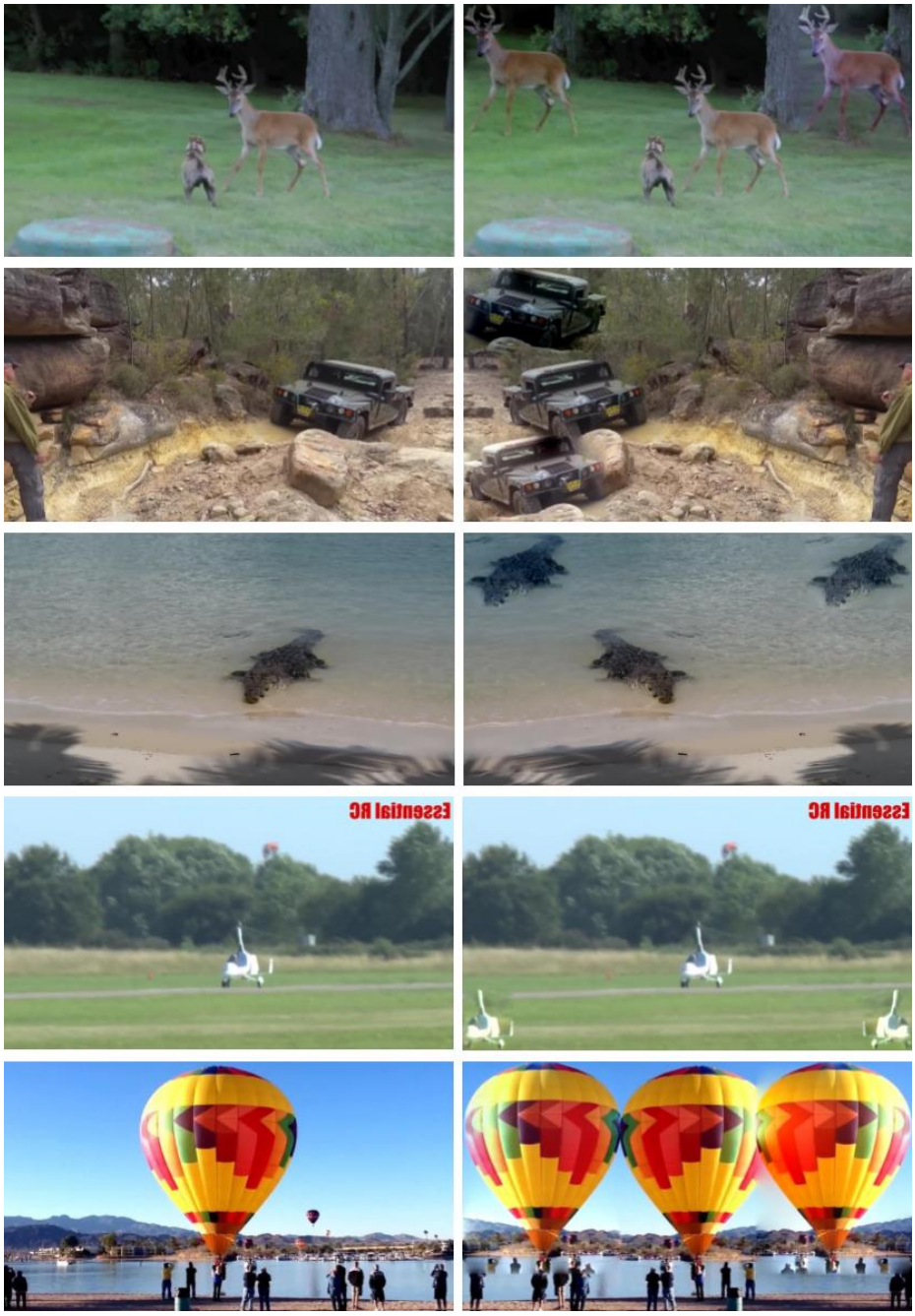

Fig. 2: Examples of real reference frames (left) and their corresponding synthetic support frames (right). The synthetic support frames might appear flipped.

the original object in an image is addressed with a seamless cloning operator [18]. Fig. 2 shows a set of synthetically generated support frames.

Reusing the original image background to place new objects decreases the probability of inconsistencies between background and foreground features. Reducing these background mismatches is crucial as the detector learns not only the object features but also the context features [4, 1]. Previous methods for position selection to insert new objects in video frames rely on spatio-temporal consistency and many annotated objects to select valid positions [4, 1]. However, the data availability restrictions of few-shot training makes the application of these methods infeasible.

3.4 Proposal feature aggregation

Relations among objects in the same image were successfully explored in [13], defining an object relation module based on the multi-head attention introduced in [24]. Similar approaches were also successfully applied in video object detection, mining relations between objects in different frames [13, 9, 7].

The goal of the feature aggregation module is to compute M relation features $\mathbf{r}_R^m$ for each object proposal $p_{t(i)} \in \mathcal{P}_t$ in the reference frame f_t:

$$\mathbf{r}_R^m(p_{t(i)}, \mathcal{P}^{SF}) = \sum_{r=1}^{R} \sum_{j=1}^{|\mathcal{P}_r|} w_{t(i),r(j)}^m \left(W_V \, \phi(p_{r(j)})\right), \quad m = 1, ..., M \tag{3}$$

where $\mathcal{P}^{SF} = \{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_R\}$ contains object proposals in the R support frames. In training, $\mathcal{P}_t$ contains the proposals extracted from the input image I_t while $\mathcal{P}^{SF}$ contains proposals from the synthetic frames set $\mathcal{F}$ (Sec. 3.3). $w_{t(i),r(j)}^m$ is a pairwise relation weight between $p_{t(i)}$ and each proposal $p_{r(j)}$ in the support frame f_r based on appearance and geometry similarities [13].

Following previous work [9, 7], we implement a multi-stage relation module with a basic stage and an advanced stage. The basic stage enhances proposal features in the current frame $\mathcal{P}_t$ by mining relationships with the proposals in the support frames $\mathcal{P}^{SF}$, generating $\mathcal{P}'_t$. In the advanced stage, the proposals in $\mathcal{P}^{SF}$ are first improved by aggregating them with the top-ϵ% scoring proposals in $\mathcal{P}^{SF}$. This enhanced support proposal features are finally used in this advanced stage to aggregate with proposals in $\mathcal{P}'_t$, generating the final proposal set.

3.5 Loss function

We propose a double objective optimization loss for the training of the spatio-temporal branch of the network double head: a classification loss and an overlap prediction. On the one hand, the classification loss function is implemented as a cross-entropy loss, following the standard approach to train a multi-class classifier. On the other hand, the overlap prediction is based on estimating the overlap of each detection with the actual objects in the image. This loss function is implemented as a binary cross-entropy loss. Hence, the final loss for the spatio-temporal branch is defined as:

$$\mathcal{L} = \mathcal{L}_{CLS} + \mathcal{L}_{IoU} \tag{4}$$

3.6 Confidence Score Optimization (CSO)

Confidence prediction plays a fundamental role in object detection. First, redundant detections are removed by means of Non-Maximum Suppression (NMS). Then, a confidence threshold is usually applied to discard low confidence detections. Commonly, the classification score is used as detection confidence.

Long tube generation has been successfully applied to optimize detection confidence, leveraging long-term spatio-temporal consistency [9, 8]. This idea is the basis of our Confidence Score Optimization (CSO) method, which first builds long tubes from detections of one object and, then, increases the classification scores based on the detections of the tube. We propose to calculate the link score *ls* between a detection $d_{t(i)}$ in frame f_t and a detection $d_{t'(j)}$ in $f_{t'}$ as follows:

$$ls(d_{t(i)}, d_{t'(j)}) = \hat{\psi}(d_{t(i)}) + \hat{\psi}(d_{t'(j)}) + 2 \cdot IoU(d_{t(i)}, d_{t'(j)}) \tag{5}$$

being $\hat{\psi}(d_{t(i)})$ the predicted IoU score described in Sec. 3.5. This increases the probability of linking detections with greater predicted overlap with the ground truth.

To build the object tubes we apply the Viterbi algorithm, using as association scores the values generated by Eq. 5 [9, 6, 7]. Then, detections belonging to each tube are updated, setting their classification score to the mean classification score of the top-20% detections of the tube.

4 Experiments

We have evaluated our proposal on the FSVOD-500 dataset [10]. It contains 2,553 annotated videos with 320 different object categories for $\mathcal{D}_{base}$, and 949 videos with 100 object categories for $\mathcal{D}_{novel}$ [3]. Object categories in $\mathcal{C}_{base}$ and $\mathcal{C}_{novel}$ are completely different. Following [10], we experiment with different partitions of $\mathcal{D}^{train}_{novel}$ and $\mathcal{D}^{test}_{novel}$ for fine-tuning. Thus, we randomly divide $\mathcal{D}_{novel}$ into two subsets ($\mathcal{D}^{train}_{novel}$ and $\mathcal{D}^{test}_{novel}$), keeping the same distribution of videos per object category. The subsets are interchanged, so each video is in $\mathcal{D}^{test}_{novel}$ once. We repeat this whole process 5 times —with different random splits—, and the reported results include the mean and standard deviation of these 5 executions.

We use ResNet-101 pretrained on ImageNet [22] as backbone. For training, we first learn the spatial part of FSVDet on $\mathcal{D}_{base}$. Then, we fine-tune both the spatial and spatio-temporal parts on $\mathcal{D}^{train}_{novel}$ —the spatio-temporal weights are randomly initialized. Training of the spatial part is done with a batch size of 16, with a learning rate of 2×10^{-2} for the first 20K iterations, reducing it to 2×10^{-3} for the next 5K iterations, and to 2×10^{-4} for the last 5K iterations. For

[3] FSVOD500 also contains a validation set with 770 annotated videos with 80 object categories. We do not use this set in the experimentation.

Table 1: Results on the few-shot video object detection FSVOD-500 dataset.

Type	Method	AP_{50}
Obj. Det.	Faster R-CNN [21]	$26.4_{\pm 0.4}$
Few-shot Obj. Det.	TFA [25]	$31.0_{\pm 0.8}$
	FSOD [11]	$31.3_{\pm 0.5}$
	DeFRCN [19]	$37.6_{\pm 0.5}$
Vid. Obj. Det.	MEGA [6]	$26.4_{\pm 0.5}$
	RDN [9]	$27.9_{\pm 0.4}$
Mult. Obj. Track.	CTracker [17]	$30.6_{\pm 0.7}$
	FairMOT [27]	$31.0_{\pm 1.0}$
	CenterTrack [28]	$30.5_{\pm 0.9}$
Few-shot Vid. Obj. Det.	FSVOD [10]	$38.7_{\pm 0.7}$
	FSVDet	$\mathbf{41.9_{\pm 2.0}}$

fine-tuning the learning rate is 1×10^{-2} for the first 9K iterations, reducing it to 1×10^{-3} for the last 1K iterations. The spatio-temporal training is performed for 40K iterations with 1 image per batch and an initial learning rate of 2.5×10^{-4}, reducing it to 2.5×10^{-5} after the first 30K iterations. The loss function is the cross entropy with label smoothing regularization [23]. For training, 2 synthetic support frames are generated for each input image with a maximum number of $\gamma = 5$ new objects into each support frame. For test, 15 support frames are used for each video frame to mine object relations. The hyperparameter β that modulates the influence of the Prototypical Calibration Block on the classification score is set to 0.5. For the relation module, the ratio of proposals selected for the advanced stage ϵ is 20%.

Tab. 1 shows the results of several state-of-the-art approaches on the FSVOD-500 dataset for a shot size of $K = 5$ —the one tested in [10]. As the few-shot video object detection problem remains almost unexplored with only one previous work, for comparison purposes we also include single image few-shot object detectors, traditional video object detectors and methods based on multiple object tracking (MOT). The results for traditional video object detectors and MOT-based methods were originally reported in [10]. FSVDet outperforms previous approaches, improving our single image baseline (DeFRCN) by 4.3 points, and previous few-shot video object detectors by 3.2 points. Traditional video object detectors that include attention mechanisms for mining proposal relationships [9, 6] fail to perform few-shot object detection, falling behind single image few-shot detectors. This proves the need for algorithms specifically designed for few-shot video object detection.

5 Conclusions

We have proposed FSVDet, a new few-shot video object detection framework that, first, applies attention mechanisms to mine proposals relationships between different frames. Then, a spatio-temporal double head classifies object proposals leveraging spatio-temporal information, and it also predicts the overlap of each proposal with the ground truth. Finally, overlapped predictions are used in an object linking method to create long tubes and optimize classification scores. Moreover, we have defined a new training strategy to learn from single images while considering a group of input frames at inference time. FSVDet outperforms previous solutions by a large margin, and establishes a new state-of-the-art result for the FSVOD-500 dataset for few-shot video object detection.

Acknowledgment

This research was partially funded by the Spanish Ministerio de Ciencia e Innovación (grant number PID2020-112623GB-I00), and the Galician Consellería de Cultura, Educación e Universidade (grant numbers ED431C 2018/29, ED431C 2021/048, ED431G 2019/04). These grants are co-funded by the European Regional Development Fund (ERDF).

References

1. Bosquet, B., Cores, D., Seidenari, L., Brea, V.M., Mucientes, M., Bimbo, A.D.: A full data augmentation pipeline for small object detection based on generative adversarial networks. Pattern Recognition p. 108998 (2022)
2. Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
3. Cao, Y., Wang, J., Lin, Y., Lin, D.: Mini: Mining implicit novel instances for few-shot object detection. arXiv preprint arXiv:2205.03381 (2022)
4. Chen, C., Zhang, Y., Lv, Q., Wei, S., Wang, X., Sun, X., Dong, J.: Rrnet: A hybrid detector for object detection in drone-captured images. In: IEEE International Conference on Computer Vision Workshops (ICCV) (2019)
5. Chen, T.I., Liu, Y.C., Su, H.T., Chang, Y.C., Lin, Y.H., Yeh, J.F., Chen, W.C., Hsu, W.: Dual-awareness attention for few-shot object detection. IEEE Transactions on Multimedia (2021)
6. 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)
7. Cores, D., Brea, V.M., Mucientes, M.: Short-term anchor linking and long-term self-guided attention for video object detection. Image and Vision Computing **110**, 104179 (2021)
8. Cores, D., Brea, V.M., Mucientes, M.: Spatiotemporal tubelet feature aggregation and object linking for small object detection in videos. Applied Intelligence pp. 1–13 (2022)

9. 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)
10. Fan, Q., Tang, C.K., Tai, Y.W.: Few-shot video object detection. In: European Conference on Computer Vision (ECCV) (2022)
11. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-RPN and multi-relation detector. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4013–4022 (2020)
12. 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)
13. 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)
14. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: IEEE International Conference on Computer Vision (ICCV). pp. 8420–8429 (2019)
15. Kaul, P., Xie, W., Zisserman, A.: Label, verify, correct: A simple few shot object detection method. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14237–14247 (2022)
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
17. Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: European Conference on Computer Vision (ECCV). pp. 145–161 (2020)
18. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003 Papers, pp. 313–318 (2003)
19. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: DeFRCN: Decoupled Faster R-CNN for few-shot object detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 8681–8690 (2021)
20. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7263–7271 (2017)
21. 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)
22. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
24. 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 (NIPS). pp. 5998–6008 (2017)
25. Wang, X., Huang, T., Gonzalez, J., Darrell, T., Yu, F.: Frustratingly simple few-shot object detection. In: International Conference on Machine Learning (ICML). pp. 9919–9928 (2020)

26. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta R-CNN: Towards general solver for instance-level low-shot learning. In: IEEE International Conference on Computer Vision (ICCV). pp. 9577–9586 (2019)
27. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: A simple baseline for multi-object tracking. arXiv preprint arXiv:2004.01888 p. 6 (2020)
28. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision (ECCV). pp. 474–490. Springer (2020)