VisDrone-DET2021: The Vision Meets Drone Object detection Challenge Results

Yaru Cao[1], Zhijian He[2], Lujia Wang[2], Wenguan Wang[3], Yixuan Yuan[4],
Dingwen Zhang[5], Jinglin Zhang[6], Pengfei Zhu[1], Luc Van Gool[3], Junwei Han[5], Steven Hoi[7],
Qinghua Hu[1], Ming Liu[2], Chong Cheng[17], Fanfan Liu[13], Guojin Cao[9],
Guozhen Li[14], Hongkai Wang[10], Jianye He[8], Junfeng Wan[12], Qi Wan[15],
Qi Zhao[11], Shuchang Lyu[11], Wenzhe Zhao[13], Xiaoqiang Lu[9], Xingkui Zhu[11], Yingjie Liu[16], Yixuan Lv[13],
Yujing Ma[11], Yuting Yang[9], Zhe Wang[8], Zhenyu Xu[8], Zhipeng Luo[8], Zhimin Zhang[8],
Zhiguang Zhang[8], Zihao Li[13], Zixiao Zhang[9],

[1]Tianjin University, Tianjin, China.
[2]The Hong Kong University of Science and Technology, Hong Kong, China.
[3]ETH Zurich, Zurich, Switzerland.
[4]City University of Hong Kong, Hong Kong, China.
[5]Northwestern Polytechnical University, Xian, China.
[6]Nanjing University of Information Science and Technology, Nanjing, China.
[7]Singapore Management University, Singapore.
[8]DeepBlue Technology(Shanghai) Co., Ltd, Shanghai, China.
[9]Xidian University, Xi'an, China.
[10]Xi'an University of Technology,Xi'an China.
[11]Beihang University, Beijing, China.
[12]Beijing University of Posts and Telecommunications, Beijing, China.
[13]University of Chinese Academy of Sciences, Beijing, P.R.China.
[14]Dalian University of Technology, Dalian, P.R.China.
[15]Shenzhen University, Shenzhen, P.R.China.
[16]Tianjin Polytechnic University, Tianjin, China.
[17]Xi'an University of Technology, Xi'an, China.

Abstract

Object detection on the drone faces a great diversity of challenges such as small object inference, background clutter and wide viewpoint. In contrast to traditional detection problem in computer vision, object detection in bird-like angle can not be transplanted directly from common-in-use methods due to special object texture in sky's view. However, due to the lack of a comprehensive data set, the number of algorithms that focus on object detection using data captured by drones is limited. So the VisDrone team gathered a massive data set and organized Vision Meets Drones: A Challenge (VisDrone2021) in conjunction with the IEEE International Conference on Computer Vision (ICCV 2021) to advance the field. The collected dataset is the same as the previous dataset object detection challenge. Specifically, the team needed to predict the bounding boxes of the objects of ten predefined classes. We received results from a number of teams using different approaches, and this article describes the 8 team's approach. We conducted a detailed analysis of the assessment results and summarized the challenges. More information can be found at: http://www.aiskyeye.com/.

1. Introduction

Object detection is developing very rapidly in the field of computer vision, driving a variety of industrial detection-

based applications such as autonomous driving, animal detection, face detection and activity recognition. Object detection technology is a method which deals with detecting instances of semantic object of a certain class inside a picture, indicating object position and object class via a boundingbox description. In order to finish this "what is the object and where is the object" task, object detection process develop deep-learning based frontend to extract feature, which requires training tons of groundtruth pictures to approximate realistic object class. The more groundtruth is similar to inference object, the more accuracy we can get in the result. However, object viewed on the drone is different from traditional ground-view object, which involve some unique challenging factors (*e.g.*, viewpoint change, scale variation, occlusion, and background clutter) in object detection. These studies have been severely limited by the lack of large-scale public benchmarks based on drones. Following VisDrone-DET2018 Challenge[29], VisDrone-DET2019 Challenge[7] and VisDrone-DET2020 Challenge[6], we held the 4th Vision Meets Drone Object Detection in Images Challenge (VisDrone-DET2021) on June 15, 2021, in conjunction with IEEE International Conference on Computer Vision (ICCV 2021).

This paper summarizes 8 object detection algorithms that are better evaluated in this challenge and evaluates their comprehensive performance. These algorithms are based on state-of-the-art detectors, and most of them have recently been published in top computer vision conferences or journals,*e.g.*, Cascade R-CNN [2], YOLO[19] and CenterNet[28, 8]. The full experiment results can be found on our website `http://www.aiskyeye.com/`, which is helpful to further promote object detection research in drone-capture scenarios.

2. Related Work

2.1. Object detection Datasets

Recently, numerous datasets have been proposed to deal with the challenges in object detection, such as scale variations, background clutter, and illumination variation in the wild. The most frequently used object detection datasets include PASCAL VOC [10], MSCOCO[17], ImageNet[5], and DOTA[22]. PASCALVOC has two data sets, VOC2007 and VOC2012. It contains about $10,000$ images in 20 categories with bounding boxes for training and validation. MS COCO data set is a data set established by Microsoft. For target detection tasks, COCO covers 80 categories. Competitions based on this data set now cover detection, segmentation, key point recognition, annotation, and other central tasks of machine vision. The training and validation set for the annual contest contain $120,000$ images and over $40,000$ test images. ImageNet is a computer vision recognition project and is the world's largest image recognition database. It contains about 15 million images, 22000 categories, with strict manual screening and marking. DOTA is a common data set for remote sensing aerial object detection, which contains 2806 aerial images and $188,282$ instances in 15 categories. Aerial images are different from traditional datasets and have their own characteristics, such as larger-scale variation; Dense small object detection; Detect the uncertainty of the target.

2.2. Object detection Methods

The current mainstream target detection algorithms are mainly based on deep learning models. This algorithm divides target detection algorithms into two categories: two-stage algorithms and one-stage algorithms. The two stages are based on Region Proposal's R-CNN algorithm (R-CNN[13], Fast R-CNN[12], Faster R-CNN[20]). It is necessary to generate the target Region candidate frame, and then classify the candidate frame through the convolutional neural network. Another single-stage algorithm, taking the YOLO[19] and SSD[18] algorithms as examples, only uses convolutional neural networks to extract features, and directly predicts the categories and positions of different targets. The rapid development of various detection frameworks has led researchers to focus on improving detection performance by integrating complex models. Using data enhancement strategy to train the deep model can deal with the lack of training data.

In [15], several detection models are integrated on the 2016 COCO object detection challenge to achieve the most advanced performance. Xu *et al.* [23] uses SingleShot MultiBox Detector[18] as the backbone, combining integrated learning with context modeling and multi-scale feature representation. In addition, for the sake of reducing the false positives of the mined bounding boxes, Gao *et al.* [11] gathers the classification heads of the box predictor and the region proposal predictor. In order to reduce the computational cost, Chen and Shrivastava[3] developed the Group Ensemble Network, which integrates a ConvNets integration in a single ConvNet through a shared base and multi-head structure.

In terms of data enhancement, the most widely used strategies are random tailoring and multi-scale training. Some methods improve the accuracy of training by randomly deleting or adding objects to the image[27, 9]. Mosaic is a new data enhancement method that can detect objects outside the normal context by mixing 4 training images from different contexts. A comprehensive experiment has been conducted in [1], proving that CutMix[25] and Mosaic data enhancement are effective in YOLOv4 Breakthrough.

3. The VisDrone-DET2021 Challenge

The VisDrone-DET2021 Challenge aims to predict the bounding boxes of objects of predefined classes with a real-

valued confidence. Participants are required to submit their algorithm and evaluate the released VisDrone-DET2021 dataset. They are allowed to use external training data to improve the model. However, it is forbidden to submit different variants of the same algorithm. Meanwhile, the submission with the detailed algorithm description obtains the authorship in the ICCV 2021 workshop proceeding.

3.1. Dataset

Similar to the last three years[29, 7, 6], we use the same as the former three challenges of dataset. In particular, the challenge consists of $6,471$ for training images, 548 for verification images, and $3,190$ for test images. In the test set, we have 1580 images from the test-Challenge subset for the workshop competition, and 1610 images from the Test-dev subset for the public evaluation. In addition, ten object categories are predefined, *i.e.*, pedestrian, person, bus, car, van, truck, bicycle, awning tricycle, motorcycle, and tricycle. Some special vehicles (such as machine shop trucks, forklifts and tankers) that rarely occur are ignored in the assessment.

The participants are asked a detection result of a specific algorithm, together with the detailed description submitted to an evaluation server does not exceed 10 times. The best of the ten submissions is the final result. We encourage participants to use the provided training data, but also allow them to use additional training data. If external data is used, the usage of the external data must be stated when submitting. In order to make a fair comparison, we will rank the algorithms trained on the external VisDrone test-dev on the rankings. In this challenge, the results submitted by different accounts are strictly prohibited from using the same algorithm. The top few teams that provide better performance are the co-authors of the result paper, and the teams that do not provide algorithm descriptions will not appear in the paper.

For assessment, we use the MS COCO assessment protocol[17] on the detection algorithms for ranking, *i.e.*, AP, AP50. AP75, AR1, AR10, AR100, and AR50 indicators. In particular, AP is the main indicator, and the calculation method is to average the total step size0.05 of all 10 object categories to the Intersection over Union (loU) threshold in the range $[0.50, 0.95]$. AP50 and AP75 are the average accuracy when loU threshold is 0.50 and 0.75, respectively. In addition, we calculate the sum of 1,10,100, given by the average recall. Under all object categories and loU thresholds, each image is detected 500 times.

3.2. Submission

We received 113 submissions from all over the world in the VisDrone-DET2021 Challenge. In the following, we briefly overview the submitted algorithms included in the crowd counting task of VisDrone2020 Challenge and pro-

vide the corresponding descriptions in Appendix A. Among all the submissions, several methods use ensemble model to improve the accuracy, *e.g.*, DNEFS(A.7. Two algorithms are based on YOLO[19] with various effective modules, including SOLOER(A.2), TPH-YOLOv5(A.4). Fourth methods are derived from R-CNN[13], *e.g.*, VistrongerDet(A.5), Cascade++(A.6) and DBNet(A.1). Swin-T(A.3)is a transformer with layered design including sliding window operation, which can limit attention calculation within a window. On the one hand, it can introduce the locality of CNN convolution operation, and on the other hand, it can save calculation.EfficientNet(A.8) propose a weighted bidirectional feature pyramid structure BiFPN for fast feature fusion work and a new joint scaling method, which is a high-efficiency and fast detection network suitable for different calculation conditions.

3.3. Overall Evaluation

The overall results of the submissions are presented in Table 1. Compared with the winner detectors HAL-RetinaNet in the VisDrone-DET2018 Challenge[29], DPNet-ensemble in the VisDrone-DET2019 Challenge[7] and DroneEye2020 in the VisDrone-DET2020 Challenge[6]. There are top 10 methods in the VisDrone-DET2021 Challenge achieving a better mAP score of more than 38.00. By using test-dev dataset in the training phase, the top performer DBNet(A.1)in Table 1 performs slightly better than the former top performer DroneEye2020, *i.e.*, 39.43 vs 37.37. As discussed above, the ensemble of several networks is effective to improve the accuracy of object detection. In Table 1, DNEFS(A.7) adopts FPN[16], Cascade R-CNN[2] and YOLOv5(`https://github.com/ultralytics/yolov5`) as the baseline network, based on that, we apply some effective strategies to get better accuracy, such as attention mechanism, double head. On the other hand, the use of the Cascade-RCNN[2] framework has become wide-spread recently (*e.g.*, (A.1), (A.6)) The top model is based on Cascade R-CNN[2] for Object Detection which follow the setting get better localization performance through cascading refine boxes. Besides, Cascade++(A.6) basing on the existing object detector Cascade R-CNN[2] adopts an multi-stage object detection architecture. As the same, Cascade++ composes of a sequence of detectors that is trained with increasing IoU thresholds and auto label-assigning, and the network is an anchor-free method. Compared with the baseline Cascade-RCNN[2] Method with the mAP score of 16.09, the submitted variants largely improve the performance by combining several effective modules. In Table 1, TPH-YOLOv5(A.4) bases on the original architecture of YOLO-v5(`https://github.com/ultralytics/yolov5`), and replace the convolutional blocks with Transformers. Then it adds one more prediction head to ease the negative influence from the large vari-

Method	AP[%]	AP50[%]	AP75[%]	AR1[%]	AR10[%]	AR100[%]	AR500[%]
DBNet(A.1)	39.43	65.34	41.07	0.29	2.03	12.13	55.36
SOLOer (A.2)	39.42	63.91	40.87	1.75	10.94	44.69	55.91
Swin-T(A.3)	39.40	63.91	40.87	1.76	10.96	44.65	56, 83
TPH-YOLOv5(A.4)	39.18	62.83	41.34	2.61	13.63	45.62	56.88
VistrongerDet(A.5)	38.77	64.28	40.24	0.77	8.10	43.23	55.12
cascade++(A.6)	38.72	62.92	41.05	1.04	6.69	43.36	43.36
DNEFS(A.7)	38.53	62.86	40.19	1.42	9.38	43.10	54.87
EfficientDet(A.8)	38.51	63.25	39.54	1.82	11.12	43.89	55.12
DPNet-ensemble	37.37	62.05	39.10	0.85	7.96	42.03	53.78
DroneEye2020	34.57	58.21	35.74	0.28	1.92	6.93	52.37
Cascade R-CNN	16.09	31.91	15.01	0.28	2.79	21.37	28.43

Table 1. Object detection results in the VisDrone-DET2021 Challenge.

ance of object scales. Swin-T(A.3) is a Swin Transformer with a hierarchical design including sliding window operation, which limited the attention calculation to a window can introduce the locality of the CNN convolution operation on the one hand, and save the amount of calculation on the other hand. And it achieve comparable Performance with 39.41 mAP, which performs similarly as DBNet(A.1).

4. Conclusions

In this paper, we present the results of the VisDrone-DET2021 Challenge. It is the fourth annual object detector benchmarking activity, following the very successful VisDrone-DET2018, VisDrone-DET2019 and VisDrone-DET2020 challenges[29, 7, 6]. Many of the submitted object detection methods set a new state-of-the-art technique by being evaluated on the same data set. Specifically, the top performer is DBNet(A.1), with the overall mAP score of 39.43. Besides,there are four ways to obtain the overall mAP score of more than 39.00: DBNet(A.1), SOLOER(A.2), Swin-T(A.3), TPH-YOLOv5(A.4). The experimental results indicate that ensemble learning of a few powerful detectors can largely boost the detection performance. Besides. Cascade R-CNN and YOLOv5 are other popular detection frameworks. It is worth mentioning that the best detector DBNet(A.1) improves the mAP score by over 5% than before, which shows the development of object detection in the past year. However, the best mAP score is still below 40% and is still a long way from achieving good performance in practical applications. In addition, due to the limited resources on the UAV platform, the computational complexity of the submitted algorithm requires further consideration. We hope that we will continue to provide a platform for advancing target detection methods on drone-captured scenarios.

A. Submitted Object Detection Algorithms

In this appendix, we provide a short summary of all crowd counting algorithms that were considered in the VisDrone-DET2021 Challenge.

A.1. DBNet

Zhe Wang, Jianye He, Zhenyu Xu, Zhimin Zhang, Zhiguang Zhang, Zhipeng Luo
{wangzh,hejianye,xuzy,zhangzhm,zhangzhg,luozp}@deepblueai.com

Our model is based on Cascade R-CNN[2] for Object Detection . In detail, we follow the setting get better localization performance through cascading refine boxes. Deformable Convolution(DCN)[4] with bottleneck ratio of 4 is applied both on the layer2,layer3 and layer4 of ResNext-101, and We use data augmentation(RandomFlip, ShiftScaleRotate, Multi-Scale, CenterCrop) and image process algorithm Gaussian noise, RandomBrightnessContrast, Cutout to expand the dataset. After analyzing the dataset and prediction results, we mark the easily confused annotation as is crowd and other classes as is crowd. In this way, we can only do loss without back propagation, so as not to have a negative impact on the model.In addition, we also improved the crop function. If we cut the images directly, it will have a great impact on the boundary, we only keep the box witch IOU greater than 0.8 with original box.

A.2. Scaled-yolov4 with transformer and BiFPN (SOLOER)

Xiaoqiang Lu, Guojin Cao, Zixiao Zhang, Yuting Yang
xqlu@stu.xidian.edu.cn,caoguojincn@163.com,
{Zhangzx1999,Ytyang_1}@stu.xidian.edu.cn

Transformers based on the self-attention mechanism have already shined in major computer vision tasks. Therefore, when we notice that the data set needs more attention

to guide the model, we choose to migrate the transformer to the backbone of the scaled-yolov4[1] structure. We get a very good result. In addition, although the original FPN structure can improve the detection ability of small targets, it has shortcomings. Its fusion mechanism is too simple, so we introduce a complex two-way feature fusion network BiFPN into the baseline. Progress has been made in local verification.

A.3. Swin-transformer object detection with coarse segmentation (Swin-T)

Hongkai Wang
2200421251@stu.xaut.edu.cn

There are two major challenges in applying Transformer to the image field: The visual entity changes greatly, and the performance of the visual Transformer may not be very good in different scenarios. That is, the target size is variable. Unlike the token size in the NLP task is basically the same, the target size in the target detection is different, and it is difficult to have a good effect with a single-level model.

The image has high resolution and many pixels, and Transformer's calculation based on global self-attention leads to a large amount of calculation. That is, the high resolution of the picture. Especially in the segmentation task, high resolution will cause the computational complexity to show a quadratic increase in the size of the input image, which is obviously unacceptable.

In response to the above two problems, we propose a Swin Transformer with hierarchical design including sliding window operation. The sliding window operation includes non-overlapping local window and overlapping cross-window. Limiting the attention calculation to a window can introduce the locality of the CNN convolution operation on the one hand, and save the amount of calculation on the other hand.

A.4. Improved YOLOv5 Based on Transformer Prediction Head (TPH-YOLOv5)

Xingkui Zhu,Shuchang Lyu,Yujing Ma,Qi Zhao
{adlith, lyushuchang, zhaoqi, zy1902407}@buaa.edu.cn

Our algorithm is based on YOLO-v5(https://github.com/ultralytics/yolov5), which is one of the most efficient DET models to solve dense object detection task. In Drone-captured images, there are two main problems: 1) The object scale varies violently, which will burden the regression. 2) The motion blur of images caused by drone will make the object hard to distinguish. To solve the above two main issues, we propose TPH-YOLOv5.

Based on the original architecture of YOLO-v5, we re-place the convolutional blocks with Transformers to explore the representation potential. We then add one more prediction head to ease the negative influence from large variance of object scales. Moreover, we adopt common-used multi-scale testing strategy and multi-model (TPH-YOLOv5x and TPH-YOLOv5x) ensemble method. To further improve the overall performance, we visualize the failure cases of testing samples and find that our proposed architecture has excellent localization ability but poor classification ability, especially on some categories like "bicycle" and "awning-tricycle". On this issue, we train an extra classifier(ResNet18) using the instance cropping from training data as classification training set. With an extra classifier, ourmethod get around 0.8%1.0% improvement on AP value.

A.5. Stronger Visual Information for Tiny Object Detection. (VistrongerDet)

MCPRL Sugar
wanjunfeng@bupt.edu.cn

For purpose better performance, the framework is based on two-stage R-CNN framework[12, 20, 16, 2]. Among them, Cascade R-CNN[2] is trained stage by stage, leveraging the observation that the output of a detector is a good distribution for training the next higher quality detector.

According to the characteristic of DET dataset, our framework integrates the following novel components.

1)FPN level enhancement. Inspired by[14], we adopted effective fusion factors to solve the above problems. Meanwhile, in order to make FPN features of objects more expressive, we introduced mask supervision on each layer during the model training. We refer to these two strategies as FPN level Enhancement.

2)ROI level enhancement. In FPN structure, the feature of each ROI is obtained by pooling on the features of one certain level. We propose Soft RoI Selection algorithm, which learns to fusion ROI features from adjacent levels by parameterizing the procedure of ROI pooling.

3)HEAD level enhancement. We exploit a novel Dual Sampler and Head Network(DSHNet)[24] to handle head classes and tail classes separately. And we cleverly add two supervisors to the classification head, multi-label prediction and grouping softmax, thereby indirectly avoiding changing the original structure.

A.6. Cascade++

Fanfan Liu, Guozhen Li, Qi Wan, Zihao Li, Yixuan Lv, Wenzhe Zhao
liufanfan19@mails.ucas.ac.cn,lgzh@mail.dlut.edu.cn,
wanqi2019@email.szu.edu.cn,lizihao191@mails.ucas.ac.cn,
lvyixuan19@mails.ucas.ac.cn,zwz@mail.ie.ac.cn

We design our method basing on the existing object detector Cascade R-CNN[2], which adopts an multi-stage object detection architecture. As the same, our Cascade++ composes of a sequence of detectors which is trained with increasing IoU thresholds and auto label-assigning. The Cascade++ network is an anchor-free method.

As we adopt anchor-free detectors, the original label-assigning policy in the original Cascade R-CNN is not applicable. So we proposed auto label-assigning, changing the relatively hard assigning policy to a soft and adaptive one. In this process we adaptively select proposals as positive ones according to the cost of each proposal.

A.7. Drone Networks with Effective Fusion Strategy(DNEFS)

Yingjie Liu
ezsifjbq@gmail.com

DNEFS adopts FPN[16], Cascade R-CNN[2] and YOLOv5(`https://github.com/ultralytics/yolov5`) as the baseline network, based on that, we apply some effective strategies to get better accuracy, such as attention mechanism, double head. In the training stage, we use different image scales to enhance the multi-scale feature extraction ability, specifically, we set no limit for the long side of the input image, and its short side is randomly picked from 600, 800, 1000 and 1200, after the training process, we continue to apply SWA[26] strategy to train additional 12 epochs. The backbone network is Resnets-vs based on mxnet framework. We also train YOLOv5m and YOLOv5l with a large image size of 1280, besides, each image is split into four pieces for small objects, no additional hyper-parameters are modified. In the testing stage, we implement multi-scale testing tricks and fuse all detection results by using WBF[21] method.

A.8. EfficientDet

Chong Cheng
2200421362@stu.xaut.edu.cn

We Propose a weighted bidirectional feature pyramid structure BiFPN for fast feature fusion work;In addition, we also proposed a new joint scaling method, which unified the backbone, feature space network, head, image space resolution backbone, feature network, head, image resolution of backbone, feature network, head, imagine solution. We use EfficientNet as the backbone, combined with the above two points, a high-efficiency and fast detection network suitable for different calculation conditions is generated.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6154–6162. IEEE Computer Society, 2018.

[3] Hao Chen and Abhinav Shrivastava. Group ensemble: Learning an ensemble of convnets in a single convnet. *CoRR*, abs/2007.00649, 2020.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 764–773. IEEE Computer Society, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[6] Dawei Du, Longyin Wen, Pengfei Zhu, Heng Fan, Qinghua Hu, Haibin Ling, Mubarak Shah, Junwen Pan, Apostolos Axenopoulos, Arne Schumann, Athanasios Psaltis, Ayush Jain, Bin Dong, Changlin Li, Chen Chen, Chengzhen Duan, Chongyang Zhang, Daniel Stadler, Dheeraj Reddy Pailla, Dong Yin, Faizan Khan, Fanman Meng, Guangyu Gao, Guosheng Zhang, Hansheng Chen, Hao Zhou, Haonian Xie, Heqian Qiu, Hongliang Li, Ioannis Athanasiadis, Jincai Cui, Jingkai Zhou, Jong Hwan Ko, Joochan Lee, Jun Yu, Jungyeop Yoo, Lars Wilko Sommer, Lu Xiong, Michael Schleiss, Ming-Hsuan Yang, Mingyu Liu, Minjian Zhang, Murari Mandal, Petros Daras, Pratik Narang, Qiong Liu, Qiu Shi, Qizhang Lin, Rohit Ramaprasad, Sai Wang, Sarvesh Mehta, Shuai Li, Shuqin Huang, Sungtae Moon, Taijin Zhao, Ting Sun, Wei Guo, Wei Tian, Weida Qin, Weiping Yu, Wenxiang Lin, Xi Zhao, Xiaogang Jia, Xin He, Xingjie Zhao, Xuanxin Liu, Yan Ding, Yan Luo, Yang Xiao, Yi Wang, Yingjie Liu, Yongwoo Kim, Yu Sun, Yuehan Yao, Yuyao Huang, Zehui Gong, Zhenyu Xu, Zhipeng Luo, Zhiguo Cao, Zhiwei Wei, Zhongjie Fan, Zichen Song, and Ziming Liu. Visdrone-det2020: The vision meets drone object detection in image challenge results. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12538 of *Lecture Notes in Computer Science*, pages 692–712. Springer, 2020.

[7] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, Liefeng Bo, Hailin Shi, Rui Zhu, Aashish Kumar, Aijin Li, Almaz Zinollayev, Anuar Askergaliyev, Arne Schumann, Binjie Mao, Byeongwon Lee, Chang Liu, Changrui Chen, Chunhong Pan, Chunlei Huo, Da Yu, DeChun Cong, Dening Zeng, Dheeraj Reddy Pailla, Di Li, Dong Wang, Donghyeon Cho, Dongyu Zhang, Furui Bai, George Jose,

Guangyu Gao, Guizhong Liu, Haitao Xiong, Hao Qi, Hao-ran Wang, Heqian Qiu, HongLiang Li, Huchuan Lu, Il-doo Kim, Jaekyum Kim, Jane Shen, Jihoon Lee, Jing Ge, Jingjing Xu, Jingkai Zhou, Jonas Meier, Jun Won Choi, Junhao Hu, Junyi Zhang, Junying Huang, Kaiqi Huang, Keyang Wang, Lars Sommer, Lei Jin, Lei Zhang, Lianghua Huang, Lin Sun, Lucas Steinmann, Meixia Jia, Nuo Xu, Pengyi Zhang, Qiang Chen, Qingxuan Lv, Qiong Liu, Qishang Cheng, Sai Saketh Chennamsetty, Shuhao Chen, Shuo Wei, Srinivas S S Kruthiventi, Sungeun Hong, Sungil Kang, Tong Wu, Tuo Feng, Varghese Alex Kollerathu, Wanqi Li, Wei Dai, Weida Qin, Weiyang Wang, Xiaorui Wang, Xiaoyu Chen, Xin Chen, Xin Sun, Xin Zhang, Xin Zhao, Xindi Zhang, Xinyu Zhang, Xuankun Chen, Xudong Wei, Xuzhang Zhang, Yanchao Li, Yifu Chen, Yu Heng Toh, Yu Zhang, Yu Zhu, Yunxin Zhong, Zexin Wang, Zhikang Wang, Zichen Song, and Ziming Liu. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 213–226, 2019.

[8] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6568–6577. IEEE, 2019.

[9] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1310–1319. IEEE Computer Society, 2017.

[10] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

[11] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. NOTE-RCNN: noise tolerant ensemble RCNN for semi-supervised object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9507–9516. IEEE, 2019.

[12] Ross B. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society, 2015.

[13] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society, 2014.

[14] Yuqi Gong, Xuehui Yu, Yao Ding, Xiaoke Peng, Jian Zhao, and Zhenjun Han. Effective fusion factor in FPN for tiny object detection. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1159–1167. IEEE, 2021.

[15] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wo-

jna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3296–3297. IEEE Computer Society, 2017.

[16] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944. IEEE Computer Society, 2017.

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 21–37. Springer, 2016.

[19] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016.

[20] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015.

[21] Roman A. Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image Vis. Comput.*, 107:104117, 2021.

[22] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge J. Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3974–3983. IEEE Computer Society, 2018.

[23] Jie Xu, Wei Wang, Hanyuan Wang, and Jinhong Guo. Multi-model ensemble with rich spatial information for object detection. *Pattern Recognit.*, 99, 2020.

[24] Weiping Yu, Taojiannan Yang, and Chen Chen. Towards resolving the challenge of long-tail distribution in UAV images

for object detection. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 3257–3266. IEEE, 2021.

[25] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6022–6031. IEEE, 2019.

[26] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. SWA object detection. *CoRR*, abs/2012.12645, 2020.

[27] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13001–13008. AAAI Press, 2020.

[28] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019.

[29] P. Zhu, L. Wen, N. Mamgain, N. K. Vedurupaka, and E. Al. Visdrone-det2018: The vision meets drone object detection in image challenge results. In *European Conference on Computer Vision*, 2018.