Journal Pre-proof

Deep learning-based object detection in low-altitude UAV datasets: A survey

Payal Mittal, Akashdeep Sharma, Raman Singh

PII: S0262-8856(20)30178-5

DOI: https://doi.org/10.1016/j.imavis.2020.104046

Reference: IMAVIS 104046

To appear in: *Image and Vision Computing*

Received date: 26 September 2020

Accepted date: 9 October 2020

Please cite this article as: P. Mittal, A. Sharma and R. Singh, Deep learning-based object detection in low-altitude UAV datasets: A survey, *Image and Vision Computing* (2020), https://doi.org/10.1016/j.imavis.2020.104046

Deep learning-based object detection in low-altitude UAV datasets: A survey

Payal Mittal[1], Akashdeep Sharma[2], *Panjab University, Chandigarh* and Raman Singh[3], *Thapar Institute of Engineering & Technology, Patiala E-mail Addresses:* payalmittal6792@gmail.com[1], akashdeep@pu.ac.in[2], *raman.singh@thapar.edu*[3]

Abstract— Deep learning-based object detection solutions emerged from computer vision has captivated full attention in recent years. The growing UAV market trends and interest in potential applications such as surveillance, visual navigation, object detection, and sensors-based obstacle avoidance planning have been holding good promises in the area of deep learning. Object detection algorithms implemented in deep learning framework have rapidly became a method for processing of moving images captured from drones. The primary objective of the paper is to provide a comprehensive review of the state of the art deep learning based object detection algorithms and analyze recent contributions of these algorithms to low altitude UAV datasets. The core focus of the studies is low-altitude UAV datasets because relatively less contribution was seen in the literature when compared with standard or remote-sensing based datasets. The paper discusses the following algorithms: Faster RCNN, Cascade RCNN, R-FCN etc. into two-stage, YOLO and its variants, SSD, RetinaNet into one-stage and CornerNet, Objects as Point etc. under advanced stages in deep learning based detectors. Further, one-two and advanced stages of detectors are studied in detail focusing on low-altitude UAV datasets. The paper provides a broad summary of low altitude datasets along with their respective literature in detection algorithms for the potential use of researchers. Various research gaps and challenges for object detection and classification in UAV datasets that need to deal with for improving the performance are also listed.

***Index Terms*—Deep learning, Object detection, Unmanned Aerial Vehicles, Computer Vision, Low-Altitude Aerial Datasets**

I. Introduction

An Unmanned Aerial Vehicle (UAV), commonly known as drones is a flying device controlled either by a human operator or through autonomously operating onboard workstations [1]. Depending on several purposes, drones collect on-demand images from low altitude airspace which provide vast support for emergency item deliveries, border patrolling, emergency rescue in case of disaster and visual surveillance for crowd safety such tasks [2]. The market growth opportunity for vision processing in drones or consumer aerial vehicles expand the total number of vehicle owners. Further, different countries encourage existing drone users to upgrade their hardware for better computation is notable [3]. Recently, law enforcement agencies of countries have issued several guidelines to fly UAVs in a restricted manner so that they cannot harm intruder privacy [4]. The Drone flight regulations in several Countries' national legislation, especially in European Countries, request that the drones should not fly over crowds, and even more several laws define the minimum distance the drone can fly near a crowd [5]. The recent advent of UAVs in a wide range of applications such as visual surveillance, rescue, and entertainment, is accompanied by the demand for safety. A report by Goldman Sachs forecasts the global drone industry to reach $100 billion by 2020. Further, the world based retail or consumer drone market is depicted in fig. 1 a). The report also concludes that the drone market is ready for strong growth in the commercial space with the evolution of the regulations [6]. Most drone-enabled services rely on onboard imaging capabilities, and the significant applications include detections, classification, environmental monitoring, transportations systems, and aerial assessments such as disaster response and building inspection as depicted in fig. 1 b) [7]. UAV captured images and their post analysis are two major categories that fall in commercial applications of aerial vehicles. Applications in aerial images include landslide mapping, search and rescue, wildlife monitoring, the creation of digital elevation maps and utilization of mounted camera for a multitude of purposes. The technology behind innovation in aerial applications is responsible for digital video stabilization, autonomous navigation, and terrain analysis [8].

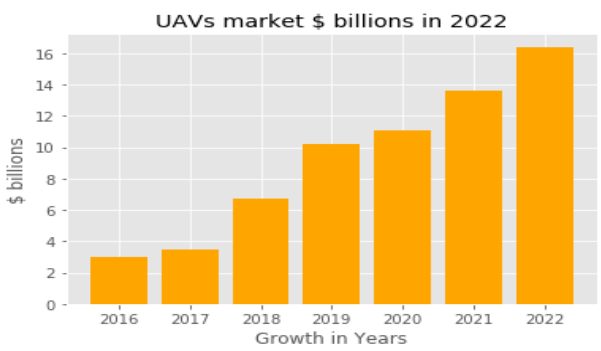

Fig. 1. a) World Retail/Consumer Drone market $ billions [6]

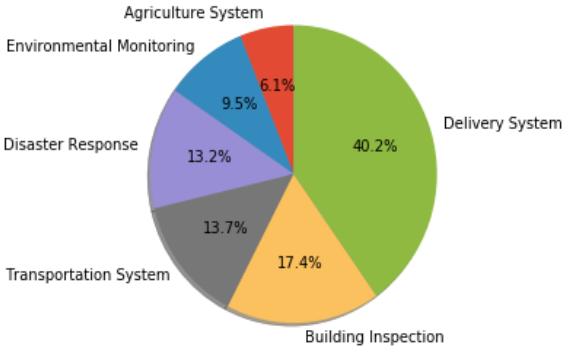

Fig. 1. b) Distribution of applications in UAV datasets [9]

The analysis utilizes post-flight data to produce fine-grained data and from crop quantity measure to water quality can be accessed in a fraction of time [10]. Fortunately, the characteristics of aerial vehicles such as the cost effectiveness, high performance and low power consumption made it possible to incorporate functional intact vision capabilities [11] into UAVs. The rapid proliferation of aerial vehicle technology is already well underway in a computing world. In computer vision, the most popular way to localize or detect an object in an image is to represent its location with the help of bounding boxes [12]. Object detection in low-altitude UAV datasets have been performed using deep learning and some detections examples have displayed in fig. 2. Object detection, a technique of identifying variable objects in a given image and inserting a boundary around them to provide localization The analysis utilizes post-flight data to produce fine-grained data and from crop quantity measure to water quality can be accessed in a fraction of time [10]. Fortunately, the characteristics of aerial vehicles such as the cost effectiveness, high performance and low power consumption made it possible to incorporate functional intact vision capabilities [11] into UAVs. The rapid proliferation of aerial vehicle technology is already well underway in a computing world. In computer vision, the most popular way to localize or detect an object in an image is to represent its location with the help of bounding boxes [12]. Object detection in low-altitude UAV datasets have been performed using deep learning and some detections examples have displayed in fig. 2. Object detection, a technique of identifying variable objects in a given image and

Fig. 2. Examples of object detection in UAV datasets [13]-[14]

inserting a boundary around them to provide localization coordinates. Object detection in aerial images has gained the attention of researchers working in this field as aerial vehicles provide stereoviews from a camera mounted on them. Deep learning based approaches for object detection is revolutionizing the capabilities of autonomous navigation vehicles robustly [15]. The work presented in paper is intended to offer a wide-ranging indication on the use of deep learning based object detection approaches specifically on low-altitude aerial datasets. It will serve as a repository of all current growth made in deep learning based object detection in low-altitude datasets and also help young researchers to consult research issues for further perusal in this field.

1.1 Motivation and Contribution

Our study is focused by the need to find several convolutional networks based methods for object detection in low altitude UAV datasets and group them together for the assistance of research society. The proposed survey is

different from already published surveys as the studies included are based only on low-altitude aerial datasets rather than focusing on standard [16]-[17] or remote-sensing based datasets [18]. Further, our study provides a complete state-of-the art object detection algorithms for low-altitude aerial images which includes single, two and advanced stages of detectors. The categorization of object detectors in the form of stages is discussed in next sections. Moreover, Table 1 lists out the major surveys related to general and UAV based object detection and points out the sharp disparities of the current survey to these published works. [12] analyzed a review of state of the art algorithms in application domains of deep learning based object detection. The generic and salient object detection were discussed for both face and pedestrian detection by adopting the methodology of multi- scale and multi-feature boosting forest. The paper focused on state-of-the-art deep learning based object detectors but did not include any information pertinent to low-altitude UAV datasets. [19] provided a detailed review of vision-based systems in UAVs that fostered the development of UAVs in advanced and modern tasks. This paper elaborated more concepts related to UAVs and less focus can be seen on specific object detection area. [20]-[21] presented survey of application based studies in UAVs specifically about disaster response and traffic surveillance respectively. Both the surveys lack in discussing deep learning based object detectors in a detail manner. [22] provided a comprehensive survey about the progress made in object detection since 2012. It also includes one, two and advanced stages of deep object detection but only for general image datasets. [23] avoid problem of inherent class imbalance and improve detection capability by introducing a novel deep network in low-altitude aerial images. However, it does not provide a survey for object detection and lists the current state-of-the-art for low-altitude aerial images.

Table 1. Comparison of past surveys and proposed work

STUDIES	DESCRIPTION	DEEP LEARNING BASED OBJECT DETECTION	LOW-ALTITUDE UAV DATASETS	MULTI-APPLICATION	ONE AND TWO STAGE DETECTORS	ADVANCED OBJECT DETECTION METHODS
[12]	REVIEWED OF OBJECT DETECTION IN A MULTI-APPLICATION ENVIRONMENT	√	×	√	√	×
[19]	ANALYSIS OF COMPUTER VISION BASED OBJECT RECOGNITION ALGORITHMS	×	×	√	×	×
[24]	DESCRIBED DEEP LEARNING BASED OBJECT DETECTION METHODS PARTICULARLY ON REMOTE SENSING DATASETS	√	×	√	×	×
[25]	HIGHLIGHTED DEEP LEARNING BASED OBJECT DETECTION BASED REGION PROPOSAL GENERATION	√	×	×	×	×
[26]	DISCUSSED DEEP LEARNING BASED OBJECT DETECTION PARTICULARLY FOR PEDESTRIANS	√	×	×	√	×
[27]	FOCUSED ON TYPICAL GENERIC OBJECT DETECTION ARCHITECTURES	√	×	√	√	×
[20]	SURVEY OF UAV IMAGERY ACQUISITION FOR DISASTER RESEARCH	×	√	×	×	×
[21]	SURVEY OF UAV FOR TRAFFIC SURVEILLANCE	×	×	√	×	×
[22]	PROVIDED A COMPREHENSIVE SURVEY OF OBJECT DETECTION SINCE 2012	√	×	×	√	√
[28]	ASPECTS OF GENERIC OBJECT DETECTION	√	×	×	√	√
[29]	SURVEY ABOUT MILESTONE DETECTORS, DETECTION DATASETS	√	×	√	√	×
[30]	PRESENTED COMPUTER VISION BASED UAV CONCEPTS OF OBJECT RECOGNITION	×	×	√	×	×
[23]	AVOID PROBLEM OF INHERENT CLASS IMBALANCE AND IMPROVE DETECTION CAPABILITY	√	√	×	√	×
[31]	RECENT ADVANCES IN DEEP LEARNING FOR OBJECT DETECTION	√	×	√	√	√
PROPOSED	SUMMARIZE MULTIPLE APPLICATION BASED DEEP LEARNING BASED	√	√	√	√	√

	OBJECT DETECTION ALGORITHMS IN LOW-ALTITUDE UAV DATASETS					

From these studies, it is quite evident that most survey papers presented summarized deep learning based object detection algorithms in a general manner or targeted at specific aerial applications. In our proposed work, a comprehensive survey based on deep learning-based object detection algorithms on low-altitude UAV datasets has been presented. There is a need to summarize the contents all under one umbrella for budding researchers, academician, industry and end users. This paper is aimed for researchers new to the field of object detection tasks related to low altitude-based UAV datasets.

The primary objectives of the research paper are as follows:

- To review the current taxonomy of deep learning-based object detection algorithms with respect to aerial data
- To provide a comprehensive list of low-altitude UAV datasets present in literature and analyze the current state of the art object detection algorithms in datasets
- Literature findings about why advanced deep detectors work better than popular one-stage and two-stage deep detectors
- Summarize their comparative performances by analyzing results on low-altitude benchmark datasets

The organization of the paper is as follows: Section II elaborates the work related to development of detectors with respect to low-altitude aerial datasets. Section III describes a comprehensive analysis of deep learning based taxonomy for object detection in low-altitude UAV datasets. A detailed review has been done on one-stage and two-stage based algorithms for object detection. Other recent advanced approaches for object detection based on current development in the deep learning area are also covered. Section IV discusses about available low-altitude UAV datasets for object detection algorithms and studies using benchmark datasets have also been considered. The performance issues in low-altitude aerial data which lead to research gaps and challenges in the field of object detection are also listed in section VI. The last section concludes the whole study and list some critical consequences for further investigation.

II. RELATED WORK

UAVs have been widely exploited in application areas such as search and rescue [32], security and monitoring [33], disaster management [20], crop management [34] and communications missions [35]. Aerial vehicles have ability to fly at different speeds to hover over a target, to perform flight outdoors and maneuvers in close proximity to objects over a point of interest [36]. These features make them fit to replace humans in operations where human intervention becomes difficult or exhaustive. There exist major challenges in low-altitude UAV based object detection when compared with standard images such as huge scale variations, densely distribution of objects, arbitrary orientations, object relative motion and turbulence of atmospheric conditions lead to blurring of objects [37]. All these challenges led to the development of object detection approaches in low-altitude aerial images that use low-level scene features as well as deep features for processing. There exist some other important critical issues in object detection on drone platforms due to which difference in mAP can be seen [38].

Small object detection: Objects are usually small in size in aerial scenes so, it is necessary to extract more contextual semantic information for discriminative representation of small objects.

Occlusion: Occlusion is another acute issue that limits the detection performance, especially in drone based scenes where objects usually are occluded by other objects or background obstacles. It is essential to handle occlusions by context or semantic information.

Large scale variations: The objects have a substantial difference in scales, even for the objects in the same category. Meanwhile, fusing multi-level convolutional features to integrate contextual semantic information is also effective to handle scale variations, just like the architecture in FPN. In addition, multi-scale testing and model ensemble are effective to deal with the scale variations.

Class imbalance: Class imbalance is another issue of object detection. The most straightforward approach is using the sampling strategy to balance the samples in different classes. Meanwhile, some methods integrate the weights of different object classes in the loss function to handle this issue such as focal loss [39]. How to solve the class imbalance issue is still an open problem.

It is quite evident in recent years, a boost in research publications happened due to emergence in the field of deep learning based object detection but high value of accuracy cannot be achieved in case of low-altitude UAVs. The domain of object detection is infinite in nature if we consider each and every development but we would strictly stick to algorithms which have scope in low-altitude aerial images. The literature of object detection in aerial images has been classified into two categories: classical and modern object detection approaches. The classical categorization includes conventional techniques which include vision based as well as machine classifiers based

approaches whereas modern deals with deep learning based algorithms which is our focus area. Classical approaches of object detection include all major developments made in the field of aerial images using handcrafted features based machine learning approaches. Classical approaches include vision technologies in aerial images such as inertial optical flow [40]-[41], shape-based descriptors [42]-[43], online boosting with features based on histogram of orient gradient (HOG) [44], deformable part based (DPM) descriptors [45], multiple trained cascaded Haar classifiers [46] and Markov random field descriptors [47]. Machine learning based approaches in aerial images make use of automated classifiers on handcrafted features to boost the performance of algorithms. These algorithms include Bayesian networks [48], graph cut methods [49], HOG with SVM classifier [50], hybrid of viola jones and SVM [51], multi-scale HOG [52], AdaBoost classifier [53], SIFT descriptor with SVM classifier [54],[35] and stochastic constraints based detection methods [55].

The inertial optical flow based evaluation to search motion objects and to differentiate in linear as well angular velocities. The shape-based feature extraction approach was used for detecting stationary and moving objects in cluttered scenes. Some detection systems were based on a bayesian network which combined several features of learning and makes use of gradient mask filters as features for the low-resolution and blurred noise in aerial data but failed miserably. The mentioned classical approaches of object detection such as inertial optical flow based evaluation, shape context feature descriptors and boosting framework with HOG features in aerial images were found to be struggling to learn discriminative object-specific features such as contour size and hierarchical shape dynamics. These methods were prone to produce more errors in practical conditions because these were not robust for moving objects, dynamic backgrounds or illumination variability which are inherent characteristics of aerial images. As a result, these failed to attain a considerable accuracy because these models were limited by the resolution of the captured images and the insufficiency of the existing feature descriptions. For example, HOG, SIFT, Markov Random field and other such feature descriptors are inefficient in feeding a vector with millions of numbers to an algorithm due to large amount of time and consider noisy information as well.

In the latest years, to reduce human efforts in processing and increase the efficiency of algorithms, deep learning based object detection in modern approaches came into existence. Deep object detection algorithms such as faster RCNN [56], Mask RCNN [57], FPN [58], R-FCN [59], YOLO [60], SSD [61], RetinaNet [39] use more complex and deep visual features extracted from the image to generate image regions. The modern approaches of object detection have better accuracy than classical approaches due to more computation processing capabilities. The modern approaches of object detection can further be divided into traditional deep and advanced object detectors. But in case of low-altitude aerial images, we will study why traditional deep learning based algorithms does not perform well so more recent advanced algorithms such as Cascade RCNN [62], CornerNet [63], CenterNet [64] and RefineDet [65] need to be implemented to achieve better results.

A. Modern Approaches of Object detection

Deep learning-based object detection algorithms are dominant and proven tools for allowing intelligent solutions for detection problems. Deep learning approaches automatically learn features of objects at multiple abstraction levels without depending on handcrafted features. The recent advancements in deep learning based models made object detection applications easier to develop than ever before. Besides, with current deep approaches focusing on full end-to-end detection and classification pipelines, performance has also been improved significantly. Generally, in object recognition process, a classifier takes an input image and produces a single output in the form of probability distribution in terms of class scores over multiple classes, but when the image has multiple objects of interest, classification produces fewer impressive results. A classifier might classify the image into less positive categories, but cannot locate objects in the picture but in object detection technique, it gives much more confident predictions for the likeliness of multiple objects. Deep architectures such as LeNet [66], AlexNet [67], Inception [68], ResNet [69], DenseNet [70], Inception-ResNet [71] etc. were successful in achieving classification accuracy but less efforts were seen in object detection. The central issue debated during the ILSVRC workshop did the CNN classification results inserted on detection problems. ImageNet [72] can simplify to detection results on the PASCAL VOC [16]. This debate motivated many researchers working in this direction for the imposition of pre-trained models to detection task. All existing object detection techniques in deep learning have been distributed into two broad categories: region proposal or two pass algorithms and non-region proposal or single-shot based detector algorithms. In region proposal methods, input images are pass into CNN after proposing category independent region through proposal techniques. Image region is the interest area of processing image which a user wants to detect and classify. These regions are generated automatically by using specific techniques such as edge box [73], selective search [74] or by deep Region Proposal Network (RPN) [56] without considering the image features. Among extracted regions, a feature vector was extracted using a deep network which was used by SVM for classification into a particular category. In contrast, single pass detectors do not generate multiple regions of an image and pass the whole image at once into a fixed grid based CNN. Boundary box coordinates were inserted around the specified region in an image

for detection. These methods are speedy and do not need a complex pipeline as compared with region-based approaches.

The modern approaches of object detection have practical uses in real-time world in an efficient manner through the ability to count people [75], detect faces [76] , indexing in visual search engines and aerial image analysis [77] . The advantages of modern object detection over classical approaches exist in enhancing accuracy due to large processing capabilities and automatically extracting features without performance overhead. But at the same instance, disadvantages such as more amount of labelled data, computation overhead and hardware requirements. The overall performances of modern approaches make it suitable for object detection in low-altitude aerial images. In the next section, a comprehensive analysis of several deep learning-based algorithms has been presented in reference to UAV datasets that evaluate object detection methods to find classified regions and predict class scores.

III. DEEP LEARNING-BASED OBJECT DETECTION ALGORITHMS

Aerial imaging through UAVs is used in numerous applications such as entertainment, detections and classifications studies, wildlife observation, and other intriguing purposes. In recent era, unlike aircraft, UAVs are affordable to end users looking for aerial imaging systems within a confined budget. The advanced approaches of deep learning based object detection have bright future in an efficient manner. Amongst deep learning based detectors, several innovations of object detection algorithms in low altitude UAVs have been witnessed in the recent years. The viewpoint variation is one of the biggest challenges in images captured from drones, since the dataset distribution contains images captured in top view angle, while other images might be captured from a lower view angle. The features learned from the object in different angles are not transferable. So, it becomes mandatory to detect aerial based objects from powerful detectors. Deep learning-based object detection algorithms have been categorized into two stage, one stage and advanced methods for aerial images as highlight in fig. 3. The algorithms such as faster RCNN [56], Mask RCNN [57], Cascade RCNN [62], FPN [58] and R-FCN [59] fall under the taxonomy of two stage detectors whereas YOLO [78], SSD [61], RefineDet [65] and RetinaNet [39] under one stage detectors. The recent advancements in object detection which are also quite popular among aerial data such as CornerNet [63], Objects as points [79] and Foveabox [80] which are based on anchorless methodology are also listed under the fig. 3. Moreover, the brief description of each deep learning based detector, the category in which it belongs, backbone network, input sizes, GitHub code repositories and losses description is given in table II. The description of loss function contains classification and localization loss with respect to each detector. The overall objective loss function is a weighted sum of the localization loss and the confidence loss where localization loss is the mismatch between the predicted boundary box and the ground truth box and classification loss is the loss in assigning class labels to predicted boxes. Moreover, a combination of L1 and L2 loss is known as smooth L1 loss which is suboptimal for accurate object localization. In the next section, we briefly discuss all the developments made in object detectors with respect to low-altitude aerial images.

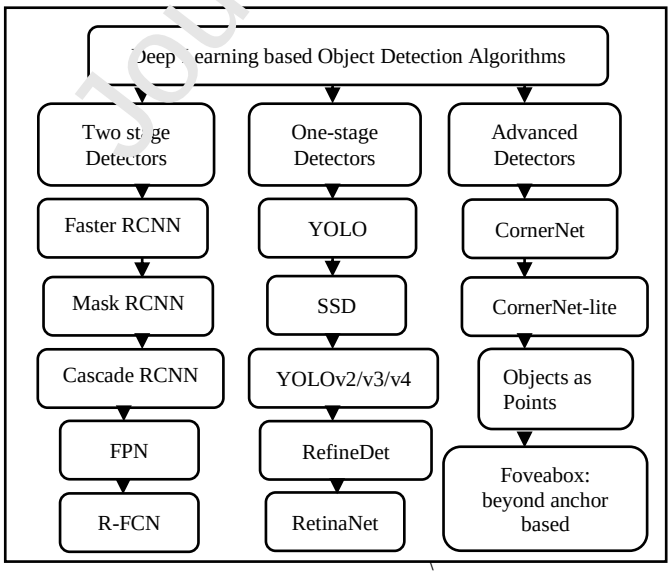

Fig. 3. Taxonomy of deep learning based object detection methods

A. Two stage Object Detection Algorithms

The two stage object detection algorithms mean detecting objects in two passes. The different stages generate a sparse set of regions of interest (RoIs) and classify each of them by a network. Early object detection models such as OverFeat [81] showed that different tasks of localizing, classifying and predicting bounding boxes could be learned using a unified shared deep network. These approaches work inside a combined framework by using convolutional networks for detection. The multiscale sliding window approach in OverFeat algorithm can be efficiently implemented. One of the first advanced algorithm using deep learning for object detection named as RCNN [82] was published in 2014 which presented an almost 50% improvement on the object detection challenge [72]. RCNN computes object location from a large set of region candidates, crops them, and classifies each using a deep network. Meanwhile, [83] proposed a deep CNN based on multi-scale spatial pyramid pooling to sample vehicle detection from aerial imagery with different sizes to learn multi-scale characteristics of objects. This advanced technique restores the edges of detection objects disturbed by environment clutter, improving detections by avoiding the cropping induced deformation of input images of different sizes. To reduce expensive process of consumed training time in object detection through RCNN, fast RCNN algorithm was proposed in [84] based on box regressing approach that comprised of end-to-end training algorithm which performs classification of object proposals and identifies spatial locations to obtain bounding box

TABLE II. A BRIEF DESCRIPTION OF DEEP LEARNING BASED DETECTORS

Category	**Backbone**	**Input Size**	**Object Detection Algorithm**	**GitHub Code Repositories**	**Classification Loss**	**Localisation Loss**
2-stage	VGG16	1000*600	Faster RCNN (2015)	https://github.com/smallcorgi/Faster-RCNN_TF	Log loss over 2 classes	Smooth L1 Loss
2-stage	Inc-Res-v2	1000*600	Deformable R-FCN (2017)	https://github.com/msracver/Deformable-ConvNets/tree/master/rfcn	Cross entropy	Smooth L1 Loss
2-stage	ResNeXt-101	1280*800	Mask RCNN (2017)	https://github.com/matterport/Mask_RCNN	Categorical cross entropy	Smooth L1 Loss
2-stage	Res101-FPN	1280*800	Cascade RCNN (2018)	https://github.com/zhaoweicai/Detectron-Cascade-RCNN	Categorical cross entropy	Smooth L1 Loss
1-stage	VGG16	300*300	SSD (2016)	https://github.com/balancap/SSD-Tensorflow	softmax_cross_entropy_with_logits	Smooth L1 Loss
1-stage	ResNet-101	640*640	RetinaNet (2017)	https://github.com/fizyr/keras-retinanet	focal loss, where α=0.25	Smooth L1 Loss
1-stage	DarkNet-53	608*608	YOLO V3 (2018)	https://github.com/pjreddie/darknet	Binary Cross Entropy	Sum of squared error
1-stage	VGG16	320*320	RefineDet (2018)	https://github.com/sfzhang15/RefineDet	Cross entropy/log loss Softmax loss	Smooth L1 Loss
1-stage	Hourglass	512*512	CornerNet (2018)	https://github.com/princeton-vl/CornerNet-Lite https://github.com/princeton-vl/CornerNet	focal loss, where α=2, b=4	Smooth L1 Loss
1-stage	VGG16	512*512	M2Det (2019)	https://github.com/qijiezhao/M2Det	softmax_cross_entropy_with_logits	Smooth L1 Loss

1-stage	Hourglass	512*512	CenterNet (2019)	https://github.com/xingyizhou/CenterNet https://github.com/Duankaiwen/CenterNet	focal loss, where α=2	L1 Loss

coordinates. Fast RCNN algorithm had several advantages when compared with RCNN:

- Quality of detection in terms of performance metric mean area precision (mAP) was higher than RCNN
- Training was done in a single stage by implementing multi-task (classification as well as regression) loss
- Training process simultaneously filters all deep layers
- No memory storage required for feature extraction

Another major advancement in case of fast RCNN is that the whole network can be trained with multi-task losses that improve significant accuracy. [56] again updated fast RCNN by introducing faster RCNN named algorithm by incorporating the following advancements:

- Designing almost cost-free regions from RPN instead of explicitly techniques for region proposals used in RCNN and fast RCNN, produced a unified pipeline of fast RCNN and RPN as a single network.
- This method enabled an integrated object detection system to run at real-time frame rates.
- RPN takes the output feature maps from the same deep network used in RCNN and slide filters over them to form region proposals, resulting in 4*k coordinates and 2*k scores per location in output.
- It predicted offsets relative to the corner of some reference boxes called anchors. These anchors are pre-selected with multi-scales and aspect ratios at each location.
- The learned RPN also enhanced region proposal quality and the cumulative object detection accuracy.

The detection of objects in an image at multiple scales is a fundamental challenge in computer vision and scale invariant Feature Pyramid Networks (FPN) [58] seems to be a standard solution [58]. The main objective of FPN is to produce a multi-scale feature representation at high-resolution levels. The principle advantage of featuring each level of an image pyramid is that it produces a multi-scale feature representation in which all levels are semantically strong, including the high-resolution levels. A remarkable increase can be seen for average precision in COCO dataset [17] by 2.3 points and PASCAL dataset [16] by 3.8 points over baseline of faster RCNN on ResNets [56]. FPN, easily extended to mask proposals and further improves average recall and speed significantly for object detection tasks and even in semantic segmentation methods [85]. FPNs can be utilized in many applications rather than object detection tasks such as generating segmentation proposals. SharpMask [86], used FPNs to generate proposals as they were trained on image crops for predicting instance segments and respective class scores. Based on FPN, Mask R-CNN [57] further extends a mask predictor by adding an extra branch in parallel with the bounding box recognition. Moreover, Cascade R-CNN [62] trains multi-stage R-CNNs with increasing IoU thresholds stage-by-stage and thus the multi-stage R-CNNs are sequentially more powerful for accurate localization. As a result, the last stage R-CNN can produce detections with the most accurate localization accuracy. Lastly, R-FCN [59] have been proposed for accurate and efficient object detection. In contrast to previous two pass detectors such as fast and faster RCNN, which applied a costly per-region subnetwork, R-FCN detector is fully convolutional with almost all computation shared on the entire image. Region-based feature maps or positive-sensitive score maps were proposed in R-FCN to address a contradiction between translation-invariance in classification and translation-variance in detection. This method adopts fully convolutional image classifier backbones such as ResNets [69] for object detection. Recently, deep learning algorithms, two-stage detectors (R-CNNs), have achieved state-of-the-art detection performance in computer vision. But our focus is on detecting low-altitude aerial objects, significant object detection accuracy cannot be achieved from the above discussed two-stage methods as they are based on sliding-window search and shallow-learning-based features with heavy computational costs and limited representation power. However, several challenges limit the applications of R-CNNs in object detection from low-altitude aerial images [87]:

- The vehicles in large-scale aerial images are relatively small in size, and R-CNNs have poor localization performance with small objects;
- R-CNNs are particularly designed for detecting the bounding box of the targets without extracting attributes;
- The manual annotation is generally expensive and the available manual annotation of vehicles for training R-CNNs are not sufficient in number.
- Faster R-CNN involves two fully connected layers for RoI recognition, while R-FCN produces a large score maps.

Thus, the speed of these networks is slow due to the heavy-head design in the architecture. Even if we significantly reduce the base model, the computation cost cannot be largely decreased accordingly.

Recently, a new two-stage detector, light-head R-CNN [88] address the shortcomings present in faster RCNN by making the head of network as light as possible, by using a thin feature map and a cheap R-CNN subnet. Further, [89] propose Deformable Convolutional Networks to model geometric transformations by learning additional offsets without supervision.

The recent methods such as Cascade RCNN, light-head RCNN make advancement than existing deep learning based object detectors in case of low-altitude aerial images but still some modifications have to be done such as use of attention mechanisms [90] in deep networks to detect objects of interests. The aerial images are of higher resolution in nature so a larger size of receptive field is needed.

B. One stage Object Detection Algorithms

Early success in deep learning based object detection field was achieved through two-stage detectors but speed was a real challenge in former approaches. The higher efficiency attribute of one-stage detectors over two-stage detectors makes them deployable in low-altitude object detection scenarios. The researchers are eventually shifted towards one stage detectors due to adaptability towards meeting challenges like providing high speed and less memory requirements. Single stage algorithms have a different concept than two stage detectors in which the whole image is passed at once into a fixed grid based CNN rather than in patches. In the initial days of one stage based detection algorithms, [60] suggested a real-time single pass based detection algorithm named YOLO which produced better results i.e. mAP higher than two stage detectors in short time. The key idea was to look at an image to predict number of objects and identify location of objects. YOLO approach trained on complete images and directly boosted detection performance. This integrated model had numerous benefits over established methods of object detection:

- Fast speed of base network which run at 45 fps on high performance CPU.
- Learned generalizable representations of objects means less chances of breaking down when pertained to new domains

The one-object rule of YOLO limits close detected objects and high localization errors and lower recall value forced to develop YOLOv2, the second version of YOLO, aims at improving accuracy significantly while consistent at maintaining speed. YOLOv2 pushes mAP by adding batch normalization, high resolution classifier and convolutional with anchor boxes. YOLO applied a softmax activation function for conversion of class scores into probabilities. YOLOv3 [78] replaced softmax function with logistic classifiers to use multi-label classification and used binary cross-entropy loss for reducing computational complexity. YOLOv3 made 3 predictions at 3 different scales per location and to determine anchors, it applied k-means clustering process. This efficient development, YOLOv3, achieves relatable results than previous versions on low-altitude aerial datasets. But, this advanced version made significant improvement in detecting small size object but still higher localization errors exist. There is also the development of YOLOv4 [91], just few days back, provide efficient results with optimal speed. YOLOv4 consists of CSPDarknet53 [92] as the backbone in which CutMix [93] and Mosaic data augmentation, DropBlock regularization [94] and class label smoothing [95] methods utilized for the functioning of backbone network. It achieved state-of-the-art with 43.0% average precision on the MS-COCO dataset. SSD algorithm [61], a more advanced single shot detector proved more accurate when compared with two stage detectors that performed region proposals. The feature detection of SSD provided significant improvement over previous detectors which were calculated by running a convolutional network on input only once. Further, it utilized anchor boxes concept for learning coordinates of bounding boxes. The detection results of SSD proved significant improvement with a mAP of 31.2% on MS-COCO test dataset as compared to 21.6% in YOLO. The amendment in speed comes from improvements such as eradicating bounding box proposals, feature resampling stage, using separate filters for suggested aspect ratios and small filter for predicting class scores and offsets in bounding box locations.

RefineDet [9] improves one-stage detector by two-step cascade regression. These two inter-connected modules imitate the two-stage structure to produce accurate detection results with high efficiency. RefineDet achieves the current state-of-the-art results on generic object detection (i.e., PASCAL VOC 2007 , PASCAL VOC 2012 and MS COCO [17]). Some literature work [96] had introduced the attention mechanism in RefineDet to further improve the performance specifically for aerial images.

RetinaNet [39] is another FPN based single stage detector, which involves Focal-Loss to address class imbalance issue caused by extreme foreground-background ratio.. The loss function for a large number of background examples resulted in the degenerative model in RetinaNet solved by introducing focal loss [39], a new dynamic loss function used to alter weights between positive and negative examples of training data. Through novel focal loss, reshaping of cross entropy loss has been made towards correctly classified training examples. It also prevents a large number of easy negative examples from flooding the object detector during the training process. To evaluate the effectiveness of focal loss, a simple dense detector was designed and trained. RetinaNet is able to match or we can say that achieved better result of nearly 39.1% AP on MS-COCO test dataset when compared with YOLO or SSD.

The highest accuracy object detectors are based on region proposal methods i.e. RCNN series, where a classifier is applied to a sparse set of candidate object locations. In contrast, one pass detectors have the potential to be faster and simpler but have trailed the accuracy of two pass detectors. The central cause for lacking in accuracy is extreme foreground background class imbalance encountered during training of dense detectors.

C. *Advanced Approaches in Deep learning based Object Detection*

In previous sections, a detailed analysis has been presented for the one stage and two stage deep learning based object detectors in low-altitude UAV images. If we consider the inherent characteristics of aerial images which are definitely challenging than standard images, then, technologically powerful detectors will be needed. In recent times, researchers made significant contribution towards building detectors which are powerful yet efficient. The current state-of-the-art utilized anchorless concept in one stage detectors which will help aerial datasets also to obtain better results than one stage and two stage detectors. Although anchor based detectors in which we have discussed one stage and two stage detectors have achieved much progress in object detection, it is still difficult to select optimal parameters of anchors. Few drawbacks of using anchor boxes in one-stage detector are:

- One-stage detector places anchor boxes densely over an image and generate final box predictions by scoring anchor boxes and refining their coordinates through regression
- Anchor boxes for training will creates a huge imbalance between positive and negative anchor boxes and slows down training [39]
- Become very complicated when combining with multiscale architectures where a single network makes separate predictions at multiple resolutions, with each scale using its own set of anchor boxes. [61][39][97]

To guarantee high recall, more anchors are essential but will introduce high computational complexity; moreover, different datasets correspond to different optimal anchors. To solve these issues, anchor-free detectors attract much research in recent times and have achieved significant advances with complex backbone networks. CornerNet [63] performs object detection by detecting objects as paired key points to eliminate the need for designing a set of anchor boxes commonly used in prior single-stage detectors. In addition, a new type of corner pooling layer was introduced that helped the network to better localize corners. It achieves a 42.2% average precision on MS COCO, outperforming all existing one-stage detectors. To decrease the high processing cost, CornerNet-Lite [98] introduced which is a combination of two efficient variants of CornerNet: CornerNet-Saccade with an attention mechanism and CornerNet-Squeeze with a new compact backbone architecture. Moreover, [64] detected the object as a triplet, rather than a pair, of key points, which improves both precision and recall. [79] models the detecting object as the center point of its bounding box, and localizes all other object properties, such as size, 3D location, orientation, and even pose. On the other hand, [80] proposed an accurate, flexible and completely anchor-free framework. It predicts category-sensitive semantic maps for the object and category-agnostic bounding box for each position that potentially contains an object. The related literature of advanced detectors with respect to low-altitude images is in development stage as only few of them were tested on only VisDrone aerial dataset [99]. The maximum value of mAP in case of 10 classes is around 23% which needs more inputs from researchers around the globe to achieve high accuracy in low-altitude aerial images.

Now, after knowing the taxonomy of deep learning based object detection algorithms for low-altitude aerial images, we can say that the traditional detectors include faster RCNN, YOLOv2, SSD detectors which lack in achieving large mAP, particularly when trained on challenging low-altitude aerial datasets. But, if have a look on some of the aerial literature studies, a jump in accuracies can be seen. The detection accuracy values in aerial images are relatable good when detectors trained on real-time UAV captured images, older aerial datasets such as VEDAI, Munich Vehicle and number of training images in aerial datasets are large. The categorization of two-stage and one-stage detectors is done to analyze in a better way. [100] focused on developing hard-mining strategies to tackle the detection of small objects. It detects vehicles using faster RCNN algorithm in infrared images of VEDAI dataset and obtained an average precision of 77.8 and recall of 31.04 in detection results. These results inspired by LeNet-5 network and were trained using SGD with a dropout of 0.5 to get 256*256 heatmaps from 1064*1064 pixels images. [101] proposed an enhanced detection technique based on faster RCNN to tackle challenges imposed by baseline RPNs i.e., faster RCNN has poor localization performance especially for small sized objects due to course feature maps. To improvise the recall metric, a hyper RPN (HRPN) was employed to classify small sized objects with a grouping of hierarchical feature maps on Munich Vehicle dataset. The classifier was also altered by boosting classifiers to validate candidate regions. The mAP under 1 scale and 1 ratio setting obtained was 0.7624, for 1 scale and 3 ratios were 0.7950, for 3 scale and 1 ratio was 0.7624 and 3 scale and 3 ratios were 0.7954. [102] introduced a deep network named rotatable residual network based on region proposal to find multi-oriented objects in aerial images. This deep network used a rotatable RPN to generate rotatable RoIs from feature maps and a strategy of batch averaging rotatable anchor was employed to initialize the shape of vehicle objects. Further, a rotatable position

sensitive type pooling layer was also intended to keep the orientation as well as position information on DLR 3K Munich 3K and VEDAI dataset. Recall values were provided with variable loss weights in ResNet-101 under Intersection-over-Union (IOU) of 0.5 such as on DLR, it was 0.733 and on VEDAI was 0.528. [103] trained faster RCNN for strong performance in small unmanned aerial systems. To provide robust training data to train a CNN, a combination of the publicly available dataset such as VEDAI, DLR 3K, SUSEX Avon Park, SUSEX Camp Atterbury and DARPA TAILWIND dataset was done for computing performance. VGG16 network achieved confidence threshold of 0.627 for 200 iterations at NMS threshold of 0.4 whereas ZF network obtained 0.686 threshold for 100 iterations at the same threshold. [104] applied faster RCNN algorithm on MIT and Caltech car dataset to detect different types of vehicles which are common in the traffic scene. For car model, average accuracy was found to be 79.9% for ZF network and 82.3% for VGG16 whereas for minibus model, 73.9% and 74.8% respectively and SUV model, 68.3% and 70.1% accuracy using improved faster RCNN network. [105] proposed an innovative bird detection framework in low-resolution aerial data images using a deep network. The processing of aerial data images was done through converting low-resolution to high-resolution images by super-resolution CNN (SRCNN) and very deep super-resolution (VDSR) techniques and then implemented faster RCNN algorithm to localize birds. Faster RCNN achieved mAP of 94.81% on BIRD-50 dataset and 95.51% on CUB-200 dataset whereas YOLO obtained 96.77% and 97.71% respectively. The above discussed studies achieved significant mAP due to the aerial datasets which have utilized are older and not in current use by researchers.

Meanwhile, [106] proposed new dataset to localize waste plastic bottles in the wild which has 25,407 UAV captured images with diverse backgrounds. The oriented bounding boxes were used to annotate for achieving detailed information and several other object detection algorithms were evaluated on UAV-BD dataset such as faster RCNN, SSD, YOLOv2, and RRPN. The average precision values of RRPN (88.3%), SSD (87.6%), faster RCNN (86.4%) and YOLOv2 (67.3%) were obtained. In this case, the dataset obtained was of only single class and very huge in nature. Further, [107] used faster RCNN based on Caffe framework for vehicle target detection and drawing the moving trace of each vehicle. The crowdsourcing marking platform was used to mark vehicle targets in frames by adding class and location label. The accuracy rate achieved was 96.5% on vehicle detection, and traffic flow statistics showed an average of 92.7% on total traffic conditions. [108] presented a novel dataset from Microsoft opensource simulator namely as AirSim based on wildlife monitoring and applied faster RCNN algorithm into 70 thermal infrared videos captured from a simulator. Efficient performance metrics such as precision 0.4690 and recall 0.0925 were obtained while tested into models from popular SPOT animal poacher dataset. [109] implemented two main modules of detection by faster RCNN and Hungarian method based tracking method using deep CNN in UAV images. The training was performed on VIVID and CAVIAR dataset while testing on real-time drone images and achieved precision of 0.87. [110] designed concentric circles and pentagons shaped UAV landing signs and employed faster RCNN to identify landing marks for UAV. The experiments demonstrated that a speed of nearly 81 milliseconds each frame and 97.8% accuracy was achieved by using faster RCNN for classification and detection. [111] investigated the use of regression based single CNN named YOLOv2 for detection of vehicles in UAV captured images as well as CSK tracking method as novel data annotation method for real-time UAV feed and Stanford drone dataset. IoU was used for performance score and value bigger than 0.7 overlapped with ground truth box considered as positive samples for training. All implementations were based on Caffe framework and achieved higher mAP of 77.12%, 67.09% and 72.95% for SSD, YOLO and YOLOv2 respectively from real-time dataset as compared with standard drone dataset [112] with 56.25%, 42.31% and 68.0% values respectively. [113] presented a holistic approach for designing CNN networks for UAV purposes utilizing low-power based embedded processors. Real-time images were collected for training of single pass based detector and the structure of Tiny-YOLO model used as a criterion with 9 convolutional layers and max- pooling layers ranging from 4 to 6. Further, a comprehensive assessment of proposed TinyYoloVoc, TinyYoloNet, SmallYolov3 and DroNet from baseline network was performed and DroNet achieved significant detection accuracy with 5 to 18 fps. [114] examined state-of-the-art convolutional based detectors on 9525 labeled images captured from 11 multi-rotor drones using a Pan-Tilt-Zoom (PTZ) camera. The PR curves and speed of detection models mentioned as fps in case of SSD MobileNet was 20.8, 12.0 for SSD Inception V2, 3.1 for RFCN ResNet 101, 2.4 for FRCNN ResNet 101, 0.7 for FRCNN Inception ResNet and 13.0 for YOLOv2. The highest accuracy obtained from faster RCNN but was slowest in detection and training while YOLOv2 was much faster than other models considering speed-accuracy trade-offs. [115] evaluated YOLOv2 algorithm on a custom dataset of 500 manually labelled images retrieved from Parrot A.R Drone. The main goal was to detect falling of persons and further tracked by Kalman filter using vision based drone control. The training of algorithm was done in mini-batches of 2000 with learning rate of 0.001 and momentum of 0.99 at 2000 epochs. The number of false positives and false negatives was 5.97% and 13.57% out of 86.24% positive detection results. At last, the real-time captured UAV images manually also achieved high detection accuracy. The large artificial aerial dataset also produced good results as [116] proposed end-to-end object detection

model YOLOv2 by creating artificial dataset of background subtracted real images. The dataset contained around 676,534 images from diverse backgrounds by proposing an algorithm for preparing the dataset. The network was trained and fine-tuned for 10k iterations with batch size of 128 and positioning batch normalization following convolutional layers. [117] employed SSD, a deep object detector to produce object region of interests from low-altitude aerial images. An alternative deep network capable of pedestrian action labels offered by human sources was also used to acquire common sub-space. The two-step framework of object detection and action recognition was proposed to extract object coordinates and blend high resolution person objects with distinctive possible actions by reinitialized VGG16 network. The experiment was performed on the Okutama-Action dataset which contains a total of 33 high-resolution videos containing the person class. The mAP for action detection at 0.5 IoU was found to be 15.39% for SSD512, 18.80% for SSD960 and 28.30% for proposed approach. [118] introduced an aerial dataset from drone and a processing method to categorize pose estimation of human as normal or abnormal through presence of perspective projection in in-flight images due to which people look tilted. Our literature findings state that majorly deep learning based detectors when trained on challenging low-altitude UAV datasets does not provide significant detection accuracy. The viewpoint variation is one of the biggest challenges in images captured from UAVs, since the dataset distribution contains images captured in top view angle, while other images might be captured from a lower view angle. The features learned from the object in different angles are not transferable. Moreover, a study of detectors when trained on challenging low-altitude UAV datasets is presented in table III. The description of deep detector approach, dataset, training details and performance metric. The mAP values are less and needs sincere improvements by paying attention to detection of aerial objects.

TABLE III. A SUMMARY FOR OBJECT DETECTION MAP VALUES WHEN TRAINED ON LOW-ALTITUDE AERIAL DATASETS

Studies	Description	Dataset	Training	mAP Values
[118]	introduced an aerial dataset from drone and a processing method to categorize pose estimation of human as normal or abnormal through presence of perspective projection in in-flight images due to which people look tilted.	aerial dataset of around 1350 images was formed using DJI Phantom4 drone	different resolution of 416*416, 480*480 and 544*544 sizes were considered for 45k iterations with 0.9 momentum and weight decay of 0.0005.	38.72
[117]	The proposed two-step framework to generate object proposals and fuse resolution proposals with different possible actions by reinitialized VGG16 network.	Okutama-Action dataset	containing a total of 33 and selected 24 videos for the training and remaining for test phase	18.80
[119]	A scale-aware network is proposed to determine the scale of predefined anchors, which can effectively reduce the scale search range, reduce the risk of overfitting, and improve the detection accuracy and speed in aerial images.	VisDrone dataset	open-source code of mask RCNN with Pytorch	33.9
[120]	The whole network structure consists of an input image will be input to the ResNet50 backbone, which is implemented with deformable convolution layers. The feature maps further refine with FPN then RPN extract some Region of Interest (RoI).	VisDrone dataset	The image is segmented into 4x4 blocks on average and merged into the training set with the original images. The training set is increased 5x as much as the original one	22.61
[77]	The Clustered object Detection (ClusDet) network consists of three key components: (1) a cluster proposal subnet (CPNet); (2) a scale estimation subnet (ScaleNet); and (3) a dedicated detection network (DetecNet).	VisDrone dataset	Each image is uniformly divided into 6 and 4 chips without overlap	32.4
[23]	introduce a Deep Feature Pyramid Network (DFPN) architecture. Similar to FPN, our goal is to leverage a ConvNet's pyramidal feature hierarchy, which has semantics from low to high levels, and build deep feature pyramids with high-level semantics throughout.	VisDrone dataset	training is performed using SGD, using an initial learning rate of 0.0001.	30.6

[96]	Different from ClusDet, this method is to consider the regions where the difficult targets are concentrated, and we abandoned ScaleNet of ClusDet to streamline the entire process.	VisDrone dataset	cropped images are generated for each image and the entire training dataset is four times larger than the original dataset	30.3
[121]	propose the novel PENet structure to detect objects in aerial images. PENet has three components: Mask Re-sampling Module (MRM), Coarse-PENet (CPEN), and Fine-PENet (FPEN).	VisDrone dataset	added three additional classes: human, non-vehicles, and vehicles into existing 10 classes of VisDrone dataset	41.1
[37]	ResNeXt-101 based Multi-scale inference and bounding box voting	VisDrone dataset	Train networks on 8 NVIDIA GTX 1080Ti GPUs, using mini-batch SGD as the optimization method	35.69

4. LOW-ALTITUDE UAV DATASETS

Most of the deep learning based object detection algorithms have been trained on PASCAL VOC dataset to detect different objects in dynamic environment. The dataset consists of 20 catalogues closely related to human life, including human and animal, vehicle, and indoor item. From the mentioned object categories, one can figure that the actual size of most objects in the dataset is large. Therefore, the detection model based on the dataset composed of large objects will not be effectively detected for the small objects in reality. To achieve better detection accuracy, aerial datasets should be utilized in an effective manner. In the latest times, an indicative number of low-altitude UAV datasets have been made open source for researchers and developers to analyze performance of deep learning-based object detection algorithms. The truth acceptance of an object detection algorithm decided from choosing a benchmark dataset to solve a specific problem. We have collected datasets from heterogeneous resources to form a list of all available low-altitude UAV datasets for evaluation of detection algorithms as depicted in Table IV. We have considered aerial images which are captured by drones, approximately 120 m or less flying above the ground. The snapshots related to standard UAV datasets such as CARPK [75], UAVBD [106], Okutama [122] and VEDAI [123] with varying scales and orientations are presented in fig. 4. The low-altitude UAV datasets are inherently different from ground video based detection datasets such as VOT2015[124]. The small scale and multiple orientations of objects due to different elevations when recording from a high perspective, making it difficult to detect all the objects in aerial images. UAV123 has 123 video sequences and 110k frames. It can be used for object detections of multiple classes such as a car, bike, person, ball, bird etc. and can easily be embedded with visual tracker benchmark by merging the respective configuration and sequence files. The main aim of developing this benchmark dataset is to solve trajectory forecasting in object tracking problems. Campus dataset [112], compiled around 50 videos in the real-time outdoor environment of a university campus that follow social etiquette-based ground rules for moving inside campus and further it not only includes pedestrians object classes but bicyclist, skateboarders, cart, car, and bus type of objects. A comprehensive list of specifically low-altitude UAV datasets for estimation of deep learning-based object detection and tracking has been shown in Table IV. Some datasets managed to provide source code functionality in their implemented methods so that users get an idea for problem assessment. The presented Table IV mentions the year in which dataset was made public, format and classes from where all required information can be achieved about their application areas. As observed from Table IV, Okutama dataset [122] is specifically dedicated to human action detection between different humans and objects. CARPK dataset [75] provides localization and counting of cars object in parking lot to gather free space information for new entrants. The UAVBD dataset [106] is dedicated to procuring waste plastic bottles from mountains and wild grasses for recycling from drone's view. VIRAT dataset [125] only confined to detection of vehicles and pedestrians in complex visual events. The recent and challenging aerial datasets include UAV- Gesture, VisDrone in which UAV-Gesture is dedicated to mainly recognizing gestures of humans captured by a low-altitude flying drone. The VisDrone dataset was collected using various drone platforms and under various weather and lighting conditions. These frames are manually, annotated with more than 2.6 million bounding boxes of targets of frequent interests, such as pedestrians, cars, bicycles, and tricycles. Some important attributes including scene visibility, object class and occlusion, are also provided for better data utilization. The majority of datasets have refined

Table IV. Comprehensive list of low-altitude UAV datasets

Year	Dataset	Height	Format	Description	Annotation	classes	Task	Reference Link
2011	VIRAT	Not specified	25 videos	event recognition in surveillance	Horizontal BB	Multi class	Detection Tracking	https://www.crcv.ucf.edu/data/VIRAT.php
2016	UAV123	10-50m	110k frames	tracking objects in diverse scenes	Horizontal BB	Multi class	Tracking	https://ivul.kaust.edu.sa/Pages/Dataset-UAV123.aspx
2016	Stanford Campus	Not specified	50 videos	trajectory forecasting	Horizontal BB	Multi class	Tracking	http://cvgl.stanford.edu/projects/uav_data/
2016	VEDAI	Not specified	1200 images	small vehicle detection	Oriented BB	Multi class	Detection	https://downloads.greyc.fr/vedai/
2017	CARPK	40m	90K images	counts and localize target cars in videos	Horizontal BB	single	Detection Counting	https://lafi.github.io/LPN/
2017	UAVDT	10-70m	80k frames	detection of vehicles in complex backgrounds	Horizontal BB	Multi class	Detection Tracking	https://sites.google.com/site/daviddo0323/projects/uavdt
2017	Okutama	10-45m	77.4k frames	concurrent human action detection	Horizontal BB	Single class Multi actions	Detection	http://okutama-action.org
2018	UAV-BD	10-30m	25k images	aiming to find and localize plastic bottles in the wild.	Oriented BB	single	Detection	http://jwwangchn.cn/UAV-BD/
2018	UMCD	6-15m	50 videos	UAV mosaicking and change detection	Attributes (Object type, shape, wrt bg)	Multi class	Object and change detection	http://www.umcd-dataset.net
2018	UAV-GESTURE	3-5m	119 videos	gestures of humans	Horizontal BB	Single class Multi action	Detection Tracking	https://github.com/asankagp/UAV-GESTURE
2018	VisDrone	Not specified	179,264 frames	detect and track multiple object categories	Horizontal BB	Multi class	Detection Tracking	http://www.aiskyeye.com

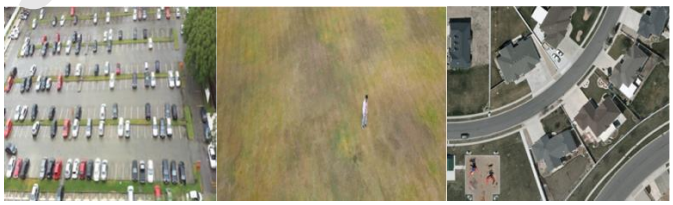

Fig. 4. Snapshots of CARPK, Okutama, VEDAI Aerial dataset [75][122][123]

ground truth annotations of images such as the inclusion of region of interest coordinates, class in which image belongs to, track ids etc. Although very little work has been done in all low-altitude UAV datasets provided but it can be observed that the detection accuracy achieved by aerial datasets are lower when compared with standard image datasets. Additionally, the use of low-altitude UAV datasets by various deep detectors has been listed in fig. 5. The cumulative number of publications under the categories such as one pass, two pass, advanced approaches have been discussed. In years 2013-2020, a number of publications published in various categorizations of object detection algorithms. The cumulative sum of each algorithm in brackets has listed as well as some datasets such as CARPK, VEDAI, Stanford Drone, VisDrone displayed greater contribution than other datasets. It can be concluded

that in emerging years, one pass based and other advanced detectors aim better mAP when trained on above discussed aerial datasets.

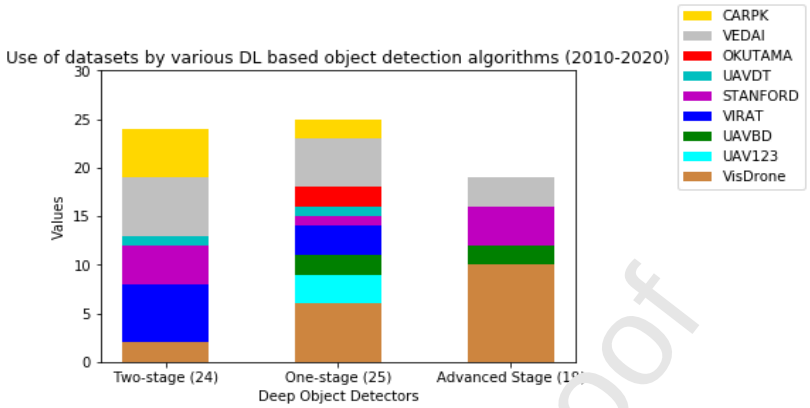

Fig. 5. Use of standard UAV datasets by various object detection algorithms

5. DISCUSSIONS

Deep learning based aerial object detection has been proven successful in ensuring public safety in terms of motor vehicle accidents, ship collisions, border and power line surveillance and solar farms energy inspection such real-time crucial applications [126]. We have discussed broadly two categories of object detection methods in low-altitude aerial images i.e. one pass and two pass detectors. The two-stage algorithms achieved significant results but the speed is slower whereas the accuracy of the recent advanced approaches such as CenterNet, RefineDet, CornerNet is better in case of aerial images with faster speed. The current state-of-the-art two-stage methods, such as Faster R-CNN [56], R-FCN [59], and FPN [58], Cascade RCNN [62] have three benefits over the one-stage detectors which are as follows:

(1) applying two-stage structure with sampling heuristics to handle class imbalance;
(2) utilizing two-step cascade to regress the object box parameters;
(3) describing two-stage features to describe the objects

The performances of various discussed algorithms on low-altitude UAV datasets can be evaluated. This section provides few useful comprehensions about the growth of various object detection techniques for the assistance of researchers, academicians and end users. Different findings have been identified through a literature survey that supports further research in the low altitude UAV processing capabilities. The anticipation of effective deep learning techniques such as one-pass and two-pass based detectors lags behind when compared with advanced techniques as the success achieved by recent detectors in low-altitude based UAV object detection is mind-boggling. In our observation, substantial proliferation in the number of deep learning approaches for object detection in low-altitude UAV datasets have been observed in recent publications as shown in fig. 6. It can be inferred after 2016, one and two stage detectors shown publication interest among researchers and recently, advanced detectors shown more progress in low-altitude UAV images.

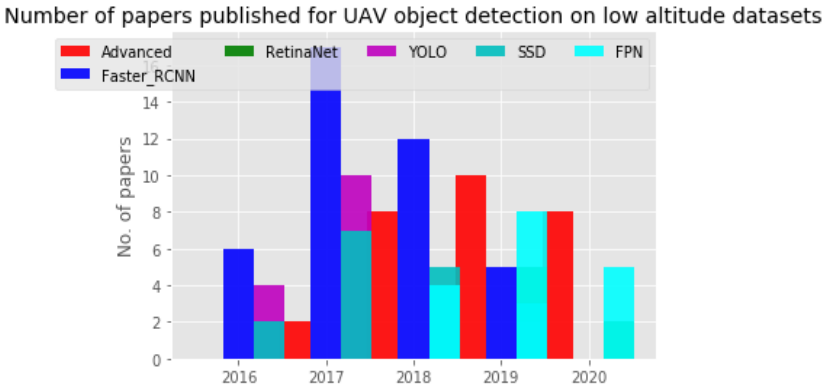

Fig. 6. A summary of papers published in deep learning-based UAV object detection on low altitude datasets

A. Performance based Comparisons in Object Detection Methods

It is very important to compare performance accuracies as well as computational complexities on standard datasets of image recognition as well as on low-altitude aerial images. A good algorithm balances a tradeoff between accuracy and inference time. The accuracy metric used is called Mean Average Precision or mAP. Average Precision (AP) is calculated for each class from the area under precision-recall curve. Now predictions are sorted by their confidence score from highest to lowest. Then 11 different confidence thresholds called ranks are chosen such that the recall at those confidence values have 11 values ranging from 0 to 1 by 0.1 interval. The thresholds should be such that the Recall at those confidence values is 0, 0.1, 0.2, 0.3 to 0.9 and 1.0. AP is now computed as the average of maximum precision values at these chosen 11 recall values. We use the evaluation protocol in MS COCO [17] to evaluate the results of detection algorithms, including AP, AP50, AP75 metrics. Specifically, AP is computed by averaging over all 10 Intersection over Union (IoU) thresholds (i.e., in the range [0.50 : 0.95] with the uniform step size 0.05) of all categories, which is used as the primary metric for ranking. AP50 and AP75 are computed at the single IoU thresholds 0.5 and 0.75 over all categories. able on challenging MS-COCO dataset and detectors attained much lower accuracy in this challenging dataset of 80 classes. The recent anchorless concept based detector i.e. CornerNet outperforms existing detectors which is based on a novel hourglass-104 backbone. The mentioned table V have been divided into

Table V. Detection Results on MS-COCO test dataset [63]

Method	Backbone	AP	AP50	AP75
Two-pass detectors				
Faster RCNN	ResNet101	36.2	59.1	39.0
Mask-RCNN	ResNeXt101	39.8	62.3	43.4
Cascade RCNN	ResNet101	42.8	62.1	46.3
One-pass detectors				
Yolov2	DarkNet-19	21.6	44.0	19.2
SSD513	ResNet101	31.2	50.4	33.3
RetinaNet	ResNet101	39.1	59.1	42.3
RefineDet	ResNet101	41.8	62.9	45.7
CornerNet	Hourglass-104	40.6	56.4	43.2

two-pass and one-pass detectors based on their respective AP scores and it is quite clear that recent approaches such as RetinaNet, Cascade RCNN, RefineDet and CornerNet perform better than faster RCNN, YOLO and SSD

Table VI. Detection Results on Test dataset on VisDrone 2019 [99]

Method	AP	AP50	AP75
Faster RCNN	3.55	8.75	2.43
R-FCN	7.20	15.17	6.38
Cascade RCNN	16.09	31.91	15.01
Yolov3	10.25	21.56	8.70
SSD	2.52	4.78	2.47

RetinaNet	11.81	21.37	11.62
RefineDet	14.90	28.76	14.08
CornerNet	17.41	34.12	15.78

algorithms. But in case of challenging aerial VisDrone dataset, even the highest performing detector in MS-COCO test-dev dataset i.e. CornerNet only achieves an average mAP of 17.41 which is very less when compared with 40.6 as depicted in table VI. The aerial images are in high-resolution i.e. about 2, 000 × 1, 500 pixels, while most of the images are less than 500 × 500 pixels in standard datasets such as MS-COCO.

Multiple deep learning detectors for object detection tasks have been studied for low-altitude UAVs in previous sections. The relative share of major deep learning-based object detection algorithms in low altitude UAVs have been depicted in fig. 7. It is presented to create awareness among readers about the growth of detection techniques with respect to each other in terms of publication share. There exist several objects detection algorithms based on advanced

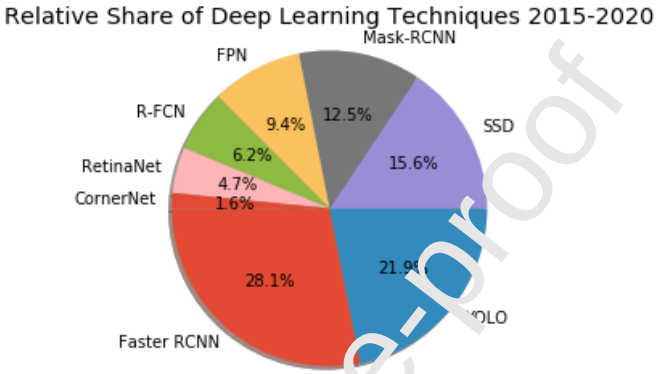

Fig. 7. Relative Share of Various deep learning based object detection algorithms

deep learning and making a choice for the right method has been crucial and depends on the specific problem chosen by user. It can be seen that one-stage and two stage have relatively larger share than advanced detectors as these have more support in implementation than recent advanced detectors. But, advanced detectors are more successful in detecting low-altitude based aerial objects. The current values of mAP can be improved if research focus is drawn on right direction. Additionally, few recommendations for future research have also been suggested by emphasizing some of the open issues prevalent in the domain of object detection. Some recommended solutions include:

a. To prevent large-scale features from covering small scale features of aerial images, the feature tensor that is outputted from different RoI pooling should be normalized before those tensors concatenated.
b. In order to get abstract object features, be ensure to have enough pixels to describe small objects so that a combination of the features of different scales represent the local details of the object.

The technological interventions and boost of applications influenced the aerial imaging market which is expected to expand at a rate of 14.2% in the coming years. Altogether, the manner of technological advancements indicates that aerial imaging techniques will truly evolve in the coming years and sincere efforts need to be done in object detection field of low-altitude aerial images.

Conclusion and Future Scope

Object detection has always been a fundamental but challenging issue in computer vision. To the best of our knowledge, this is the first survey in the literature which focuses on object detection using deep learning in low-altitude UAV datasets. The given studies based on low-altitude UAV datasets for object detection algorithms shows that inherent characteristics of aerial images possess serious challenges to algorithms performances. In the present study, a comprehensive survey on deep learning-based object detection algorithms has been done particularly on aerial datasets. Our work analyzes that in case of aerial datasets, recent advanced deep learning based detectors such as RetinaNet, Cascade RCNN, CornerNet achieves better mAP than previous state-of-the-art detectors among which faster RCNN, YOLO or SSD listed. We have also highlighted the pros and cons of one-pass and two-pass based detection algorithms with respect to aerial images and this study will be helpful for the researchers who are interested in exploring object detection from low-altitude aerial images. For instance, in case of challenging aerial VisDrone dataset, even the highest performing detector in MS-COCO test dataset i.e. CornerNet only achieves an average mAP of 17.41 which is very less when compared with 40.6 for standard images. The aerial objects of interest are often too small and too dense relative to the images. In addition, objects of interests are often in different relative sizes which makes them difficult for detect through standard algorithms. It is evident that sincere research focus needed on implementation of deep learning based object detection algorithms to low-altitude aerial images.

References

[1] E. Semsch, M. Jakob, D. Pavlicek, and M. Pechoucek, "Autonomous UAV surveillance in complex urban environments," in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 02*, 2009, pp. 82–85.

[2] M. Tzelepi and A. Tefas, "Human crowd detection for drone flight safety using convolutional neural networks," in *Signal Processing Conference (EUSIPCO), 2017 25th European*, 2017, pp. 743–747.

[3] Brian Dipert, "Vision Processing Opportunities in drones" 2017 [Online] Available: https://www.embedded-vision.com/platinum-members/embedded-vision-alliance/embedded-visiion training/documents/pages/drones. [Accessed: 24-Jan-2019].

[4] G. McNeal, "Drones and aerial surveillance: Considerations for legislatures," 2014. [Online]. Available: https://www.brookings.edu/research/drones-and-aerial-surveillance-considerations-for-legislatures/. [Accessed: 16-Dec-2018].

[5] L. Ruth, "Regulation of Drones: Comparative Analysis," 2019.

[6] K. Haye, "Drones-Reporting for Work," 2019.

[7] P. Cohn, A. Green, M. Langstaff, and M. Roller, "Commercial drones are here: The future of unmanned aerial systems," 2017. [Online]. Available: https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/commercial-drones-are-here-the-future-of-unmanned-aerial-systems. [Accessed: 16-Feb-2019].

[8] I. Colomina and P. Molina, "Unmanned aerial systems for photogrammetry and remote sensing: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 92, pp. 79–97, 2014.

[9] "Interact Analysis-A new version of intelligent automation." [Online]. Available: https://www.interactanalysis.com/drones-market-2022-predictions/.

[10] Z. Zhou, J. Irizarry, and Y. Lu, "A multidimensional framework for unmanned aerial system applications in construction project management," *J. Manag. Eng.*, vol. 34, no. 3, p. 4018004, 2018.

[11] S. Todorovic and M. C. Nechyba, "A vision system for intelligent mission profiles of micro air vehicles," *IEEE Trans. Veh. Technol.*, vol. 53, no. 6, pp. 1713–1725, 2004.

[12] Z.-Q. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *arXiv Prepr. arXiv1807.05511*, 2018.

[13] K. Mok, "Deep Learning Drone Detects Fights, Bombs, Shootings in Crowds." [Online]. Available: https://thenewstack.io/deep-learning-drone-detects-fights-bombs-shootings-in-crowds/. [Accessed: 13-Jan-2019].

[14] "Detection and counting of Arabian Oryx from aerial image." [Online]. Available: https://blogs.flytbase.com/ai-drones/. [Accessed: 18-Sep-2018].

[15] A. R. Pathak, M. Pandey, and S. Rautaray, "Application of Deep Learning for Object Detection," *Procedia Comput. Sci.*, vol. 132, pp. 1706–1717, 2018.

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[17] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.

[18] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. CVPR*, 2018.

[19] A. Al-Kaff, D. Martin, F. Garcia, A. de la Escalera, and J. M. Armingol, "Survey of computer vision algorithms and applications for unmanned aerial vehicles," *Expert Syst. Appl.*, 2017.

[20] S. M. Adams and C. J. Friedland, "A survey of unmanned aerial vehicle (UAV) usage for imagery collection in disaster research and management," in *9th International Workshop on Remote Sensing for Disaster Response*, 2011, vol. 8.

[21] A. Puri, "A survey of unmanned aerial vehicles (UAV) for traffic surveillance," *Dep. Comput. Sci. Eng. Univ. South Florida*, pp. 1–29, 2005.

[22] S. Agarwal, J. O. Du Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," *arXiv Prepr. arXiv1809.03193*, 2018.

[23] S. Vaddi, C. Kumar, and A. Jannesari, "Efficient object detection model for real-time UAV applications," *arXiv Prepr. arXiv1906.00786*, 2019.

[24] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, 2016.

[25] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?," *IEEE*

Trans. Pattern Anal. Mach. Intell., vol. 38, no. 4, pp. 814–830, 2016.

[26] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018.

[27] Z.-Q. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. neural networks Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, 2019.

[28] L. Liu *et al.*, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.

[29] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv Prepr. arXiv1905.05055*, 2019.

[30] C. Kanellakis and G. Nikolakopoulos, "Survey on computer vision for UAVs: Current developments and trends," *J. Intell. Robot. Syst.*, vol. 87, no. 1, pp. 141–168, 2017.

[31] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, 2020.

[32] S. Rathinam *et al.*, "Autonomous searching and tracking of a river using an UAV," in *2007 American control conference*, 2007, pp. 359–364.

[33] P.-M. Olsson, J. Kvarnström, P. Doherty, O. Burdakov, and K. Holmberg, "Generating UAV communication networks for monitoring and surveillance," in *2010 11th International Conference on Control Automation Robotics & Vision*, 2010, pp. 1070–1077.

[34] E. A. George, G. Tiwari, R. N. Yadav, E. Peters, and S. Sadana, "UAV systems for parameter identification in agriculture," in *2013 IEEE Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS)*, 2013, pp. 270–273.

[35] B. Xu, X. Xu, and C.-M. Own, "On the feature detection of nonconforming objects with automated drone surveillance," in *Proceedings of the 3rd International Conference on Communication and Information Processing*, 2017, pp. 484–489.

[36] S. Campbell, W. Naeem, and G. W. Irwin, "A review on improving the autonomy of unmanned surface vehicles through intelligent collision avoidance manoeuvres," *Annu. Rev. Control*, vol. 36, no. 2, pp. 267–283, 2012.

[37] J. Zhou, C.-M. Vong, Q. Liu, and Z. Wang, "Scale adaptive image cropping for UAV object detection," *Neurocomputing*, vol. 366, pp. 305–313, 2019.

[38] P. Zhu, L. Wen, X. Bian, L. Haibin, and Q. Hu, "Vision Meets Drones: A Challenge," *arXiv Prepr. arXiv1804.07437*, 2018.

[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *arXiv Prepr. arXiv1708.02002*, 2017.

[40] A. Miller, P. Babenko, M. Hu, and M. Shah, "Person tracking in UAV video," in *Multimodal Technologies for Perception of Humans*, Springer, 2008, pp. 215–220.

[41] D. Meier, R. Brockers, L. Matthies, R. Siegwart, and S. Weiss, "Detection and characterization of moving objects with aerial vehicles using inertial-optical flow," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 2473–2480.

[42] S. Wang, "Vehicle detection on aerial images by extracting corner features for rotational invariant shape matching," in *2011 IEEE 11th International Conference on Computer and Information Technology*, 2011, pp. 171–175.

[43] S. M. Thornton, M. Hoffelder, and D. D. Morris, "Multi-sensor detection and tracking of humans for safe operations with unmanned ground vehicles," in *Proceedings 1st. Workshop on Human Detection from Mobile Robot Platforms, IEEE ICRA. IEEE*, 2008.

[44] A. Su, X. Sun, H. Liu, X. Zhang, and Q. Yu, "Online cascaded boosting with histogram of orient gradient features for car detection from unmanned aerial vehicle images," *J. Appl. Remote Sens.*, vol. 9, no. 1, p. 96063, 2015.

[45] M. Andriluka *et al.*, "Vision based victim detection from unmanned aerial vehicles," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1740–1747.

[46] A. Gaszczak, T. P. Breckon, and J. Han, "Real-time people and vehicle detection from UAV imagery," in *Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, 2011, vol. 7878, p. 78780B.

[47] Z. Li, W. Shi, P. Lu, L. Yan, Q. Wang, and Z. Miao, "Landslide mapping from aerial photographs using change detection-based Markov random field," *Remote Sens. Environ.*, vol. 187, pp. 76–90, 2016.

[48] T. Zhao and R. Nevatia, "Car detection in low resolution aerial images," *Image Vis. Comput.*, vol. 21, no. 8, pp. 693–703, 2003.

[49] S. Sun and C. Salvaggio, "Aerial 3D building detection and modeling from airborne LiDAR point clouds,"

IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 6, no. 3, pp. 1440–1449, 2013.

[50] X. Cao, C. Wu, P. Yan, and X. Li, "Linear SVM classification using boosting HOG features for vehicle detection in low-altitude airborne videos," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 2421–2424.

[51] Y. Xu, G. Yu, Y. Wang, X. Wu, and Y. Ma, "A hybrid vehicle detection method based on viola-jones and HOG+ SVM from UAV images," *Sensors*, vol. 16, no. 8, p. 1325, 2016.

[52] D. Sugimura, T. Fujimura, and T. Hamamoto, "Enhanced cascading classifier using multi-scale HOG for pedestrian detection from aerial images," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 30, no. 03, p. 1655009, 2016.

[53] M. Teutsch, W. Krüger, and J. Beyerer, "Evaluation of object segmentation to improve moving vehicle detection in aerial videos," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2014, pp. 265–270.

[54] T. Moranduzzo and F. Melgani, "Automatic car counting method for unmanned aerial vehicle images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1635–1647, 2014.

[55] C. Huang, P. Chen, X. Yang, and K.-T. T. Cheng, "REDBEE: A visual-inertial drone system for real-time moving object detection," in *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, 2017, pp. 1725–1731.

[56] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[57] K. H. G. G. P. Dollár and R. Girshick, "Mask R-CNN."

[58] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature Pyramid Networks for Object Detection.," in *CVPR*, 2017, vol. 1, no. 2, p. 4.

[59] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[60] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[61] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *European conference on computer vision*, 2016, pp. 21–37.

[62] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[63] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[64] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6569–6578.

[65] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4203–4212.

[66] Y. LeCun and others, "LeNet-5, convolutional neural networks," *URL http//yann. lecun. com/exdb/lenet*, p. 20, 2015.

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[68] W. Ouyang *et al.*, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 346–361.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[70] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv Prepr. arXiv1404.1869*, 2014.

[71] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[72] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[73] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*, 2014, pp. 391–405.

[74] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object

recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.
[75] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017, vol. 1.
[76] H.-J. Hsu and K.-T. Chen, "Face recognition on drones: Issues and limitations," in *Proceedings of the First Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use*, 2015, pp. 39–44.
[77] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8311–8320.
[78] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv Prepr. arXiv1804.02767*, 2018.
[79] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv Prepr. arXiv1904.07850*, 2019.
[80] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Foveabox: Beyond anchor-based object detector," *arXiv Prepr. arXiv1904.03797*, 2019.
[81] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv Prepr. arXiv1312.6229*, 2013.
[82] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
[83] T. Qu, Q. Zhang, and S. Sun, "Vehicle detection from high-resolution aerial images using spatial pyramid pooling-based deep convolutional neural networks," *Multimed. Tools Appl.*, vol. 76, no. 20, pp. 21651–21663, 2017.
[84] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
[85] K. Chen, K. Fu, M. Yan, X. Gao, X. Sun, and X. Wei, "Semantic segmentation of aerial images with shuffling convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 173–177, 2018.
[86] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *European Conference on Computer Vision*, 2016, pp. 75–91.
[87] C.-J. Seo, "Vehicle Detection using Images taken by Low-Altitude Unmanned Aerial Vehicles (UAVs)," *Indian J. Sci. Technol.*, vol. 9, 2016.
[88] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head r-cnn: In defense of two-stage object detector," *arXiv Prepr. arXiv1711.07264*, 2017.
[89] W. Ouyang *et al.*, "Deepid-net: Deformable deep convolutional neural networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2403–2412.
[90] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[91] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv Prepr. arXiv2004.10934*, 2020.
[92] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," *arXiv Prepr. arXiv1911.11929*, 2019.
[93] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6023–6032.
[94] G. Ghiasi, T.-Y. Lin, and Q. V Le, "Dropblock: A regularization method for convolutional networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 10727–10737.
[95] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," in *Advances in Neural Information Processing Systems*, 2019, pp. 4696–4705.
[96] J. Zhang, J. Huang, X. Chen, and D. Zhang, "How to Fully Exploit The Abilities of Aerial Image Detectors," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, p. 0.
[97] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," *arXiv Prepr. arXiv1701.06659*, 2017.
[98] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "Cornernet-lite: Efficient keypoint based object detection," *arXiv Prepr. arXiv1904.08900*, 2019.
[99] D. R. Pailla, "VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results," 2019.
[100] J. O. Du Terrail and F. Jurie, "On the use of deep neural networks for the detection of small vehicles in ortho-images," in *Image Processing (ICIP), 2017 IEEE International Conference on*, 2017, pp. 4212–4216.

[101] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," *Sensors*, vol. 17, no. 2, p. 336, 2017.

[102] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R $^ 3$-Net: A Deep Network for Multi-oriented Vehicle Detection in Aerial Images and Videos," *arXiv Prepr. arXiv1808.05560*, 2018.

[103] J. Kaster, J. Patrick, and H. S. Clouse, "Convolutional neural networks on small unmanned aerial systems," in *Aerospace and Electronics Conference (NAECON), 2017 IEEE National*, 2017, pp. 149–154.

[104] L. Suhao, L. Jinzhao, L. Guoquan, B. Tong, W. Huiqian, and P. Yu, "Vehicle type detection based on deep learning in traffic scene," *Procedia Comput. Sci.*, vol. 131, pp. 564–572, 2018.

[105] C. Li, B. Zhang, H. Hu, and J. Dai, "Enhanced Bird Detection from Low-Resolution Aerial Image Using Deep Neural Networks," *Neural Process. Lett.*, pp. 1–19, 2018.

[106] J. Wang, W. Guo, T. Pan, H. Yu, L. Duan, and W. Yang, "Bottle Detection in the Wild Using Low-Altitude Unmanned Aerial Vehicles," in *2018 21st International Conference on Information Fusion (FUSION)*, 2018, pp. 439–444.

[107] J.-S. Zhang, J. Cao, and B. Mao, "Application of deep learning and unmanned aerial vehicle technology in traffic flow monitoring," in *Machine Learning and Cybernetics (ICMLC), 2017 International Conference on*, 2017, vol. 1, pp. 189–194.

[108] E. Bondi *et al.*, "AirSim-W: A Simulation Environment for Wildlife Conservation with UAVs," in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 2018, p. 40.

[109] H. D. Nguyen, I. S. Na, S. H. Kim, G. S. Lee, H. J. Yang, and J. J. Choi, "Multiple human tracking in drone image," *Multimed. Tools Appl.*, pp. 1–15, 2018.

[110] J. Chen, X. Miao, H. Jiang, J. Chen, and X. Liu, "Identification of autonomous landing sign for unmanned aerial vehicle based on faster regions with convolutional neural network," in *Chinese Automation Congress (CAC), 2017*, 2017, pp. 2109–2114.

[111] T. Tang, Z. Deng, S. Zhou, L. Lei, and H. Zou, "Fast vehicle detection in UAV images," in *Remote Sensing with Intelligent Processing (RSIP), 2017 International Workshop on*, 2017, pp. 1–5.

[112] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European conference on computer vision*, 2016, pp. 549–565.

[113] C. Kyrkou, G. Plastiras, T. Theocharides, S. I. Venieris, and C.-S. Bouganis, "DroNet: Efficient convolutional neural network detector for real-time UAV applications," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2018*, 2018, pp. 967–972.

[114] J. Park, D. H. Kim, Y. S. Shin, and S. Lee, "A comparison of convolutional object detectors for real-time drone tracking using a PTZ camera," in *Control, Automation and Systems (ICCAS), 2017 17th International Conference on*, 2017, pp. 696–699.

[115] C. Iuga, P. Dr\uagan, and L. Buşoniu, "Fall monitoring and detection for at-risk persons using a UAV," *IFAC-PapersOnLine*, vol. 51, no. 10, pp. 199–204, 2018.

[116] C. Aker and S. Kalkan, "Using deep networks for drone detection," *arXiv Prepr. arXiv1706.05726*, 2017.

[117] A. Soleimani and N. M. Nasrabadi, "Convolutional Neural Networks for Aerial Multi-Label Pedestrian Detection," in *2018 21st International Conference on Information Fusion (FUSION)*, 2018, pp. 1005–1010.

[118] H.-Y. Wang, Y.-C. Chang, Y.-Y. Hsieh, H.-T. Chen, and J.-H. Chuang, "Deep learning-based human activity analysis for aerial images," in *Intelligent Signal Processing and Communication Systems (ISPACS), 2017 International Symposium on*, 2017, pp. 713–718.

[119] R. Jin and D. Lin, "Adaptive Anchor for Fast Object Detection in Aerial Image," *IEEE Geosci. Remote Sens. Lett.*, 2019.

[120] X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and Small Object Detection in UAV Vision Based on Cascade Network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, p. 0.

[121] Z. Tang, X. Liu, G. Shen, and B. Yang, "PENet: Object Detection using Points Estimation in Aerial Images," *arXiv Prepr. arXiv2001.08247*, 2020.

[122] M. Barekatain *et al.*, "Okutama-Action: An aerial view video dataset for concurrent human action detection," in *1st Joint BMTT-PETS Workshop on Tracking and Surveillance, CVPR*, 2017, pp. 1–8.

[123] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, 2016.

[124] M. Kristan *et al.*, "The visual object tracking vot2015 challenge results," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 1–23.

[125] S. Oh *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, 2011, pp. 3153–3160.

[126] O. Gusikhin, D. Filev, and N. Rychtyckyj, “Intelligent vehicle systems: applications and new trends,” in *Informatics in Control Automation and Robotics*, Springer, 2008, pp. 3–14.