AI-enabled Object Detection in Unmanned Aerial Vehicles for Edge Computing Applications

Ayush Jain†, Rohit Ramaprasad†, Pratik Narang, Murari Mandal, Vinay Chamola, F. Richard Yu, *Fellow, IEEE* and Mohsen Guizani, *Fellow, IEEE*

***Abstract*—Unmanned Aerial Vehicles (UAVs) are emerging as a powerful tool for various industrial and smart city applications. The UAVs coupled with various sensors can perform many cognitive tasks such as object detection, surveillance, traffic management, and urban planning. These tasks often rely on computationally expensive deep learning approaches. Execution of the compute intensive algorithms are usually not feasible with the embedded processors on a power-constrained UAV. Therefore, the Edge-AI has emerged as a popular alternative in such scenarios by offloading the heavy-lifting tasks to the Edge devices. This work proposes a deep learning approach for detection of objects in aerial scenes captured by UAVs. In our setup, the power-constrained drone is used only for data collection, while the computationally intensive tasks are offloaded to a GPU edge server. Our work first categorize the current methods for aerial object detection using deep learning techniques and discusses how the task is different from general object detection scenarios. We delineate the specific challenges involved and experimentally demonstrate the key design decisions which significantly affect the accuracy and robustness of model. We further propose an optimized architecture which utilizes these optimal design choices along with the recent ResNeSt backbone in order to achieve superior performance in aerial object detection. Lastly, we propose several research directions to inspire further advancement in aerial object detection.**

***Index Terms*—Aerial computing, UAVs, object detection, artificial intelligence**

I. Introduction

Intelligent Unmanned Aerial Vehicles (UAVs) have a major projected role in intelligent transportation, surveillance, environmental monitoring, and security. It facilitates integration of information collected from various sensors and communication technologies (ICT) such as Internet of Things (IoT) and blockchain networks to provide holistic integrated services impacting living and security [1]. The UAVs with computer vision capabilities have a critical role in realizing these targets. Object detection is an integral component in numerous computer vision applications. The UAVs equipped with high resolution cameras can be employed in an extensive range of applications including surveillance, disaster response strategies and construction of transportation systems.

In the last decade, a lot of performance improvement has been obtained in generic object detection [2]. More specifically, the deep learning algorithms for object detection [3] have significantly outperformed the traditional approaches. The feature learning framework in deep learning addressed several problems in the previous approaches such as generation of large number of redundant proposals by the time-consuming selective search method and manual feature design process. However, the aerial images [4] differ from regular images [5], Some of the challenges in aerial view include low spatial resolution of objects, multitude of object sizes (small, medium, large), and complex backgrounds. This makes aerial object detection much more difficult than the general object detection.

In this paper we focus on a specific architecture, RetinaNet [3] which is a single stage end-to-end object detector. It is specially suitable for object detection in aerial view due to its focal loss, which helps it in focusing on tougher samples, and feature pyramid network, which helps it to adjust to wide variety of object sizes that can occur in images captured by drones. We specifically focus on the critical parameters of the model that influence the detection of smaller objects. We provide a detailed experimental analysis of these parameters in the VisDrone 2019 dataset. We perform experiments with the robust ResNeST50 backbone and compare the results with the traditional VGG16 and ResNet50 backbones. We also show that the default anchors used in RetinaNet are inadequate when it comes to aerial images, and needs to be carefully optimized to get good performance. Based on the experiments, quantitative and qualitative results, we list some possible future research directions for the readers.

II. General Overview of Aerial Object Detection

The existing deep learning based object detection frameworks can be divided into two main categories: **region-based** and **region-free**. The region based frameworks involve two stages. In the first stage region proposals are generated from an image with no associated category information. These filter out majority of the negative locations, and generates the regions where the objects are most likely to be present. In the second stage, additional features are extracted from these regions, and then classifiers are used to determine the category labels of the proposals. Most common region-based detectors include R-CNN, Faster-RCNN and R-FCN. Region free detectors, also known as single-stage detectors such as YOLO, YOLOv2, YOLO3, SSD and RetinaNet [3] completely eliminate region proposal generation and are hence called region-free. It directly predicts the class probabilities and bounding box offsets from full images with a single feed forward CNN network. Mandal et al. [6], [7] designed

† The authors have equal contribution

Ayush Jain, Rohit Ramaprasad, Murari Mandal and Pratik Narang are with the Department of CSIS, BITS Pilani, India

Vinay Chamola is with the Department of EEE, BITS Pilani, India

F. Richard Yu is with Carleton University, Canada

Mohsen Guizani is with the Department of CSE, Qatar University, Qatar

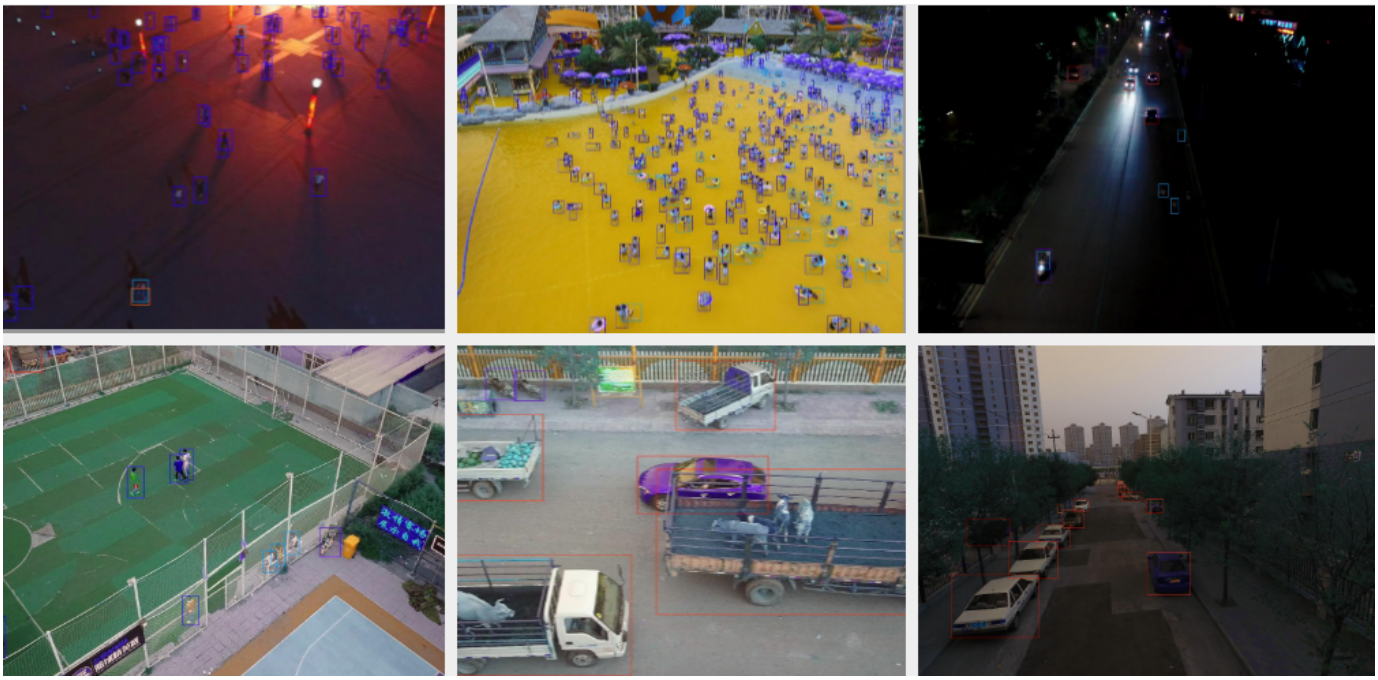


Fig. 1: **Challenges in aerial object detection**. Notice the variation of scenes, variation in the number of objects and their sizes. Images from Visdrone dataset [4].

one-stage detectors to efficiently learn the small-size object features from aerial views. They also present a new labeled aerial dataset for experimental analysis. Qin et al. [8] present the specially optimized one-stage network by combining the feature and semantic information of the objects. The network is constructed with the feature enhancement, multiscale detection, and feature fusion modules. The multiscale features have been frequently used with several improvisations in many recent works.

A. Datasets

In general, the aerial object datasets could be either captured by satellites like DOTA [9], or captured by drones like VisDrone [4]. For our experiments, we use Visdrone 2019 object detection dataset [4] which consists of 8599 snapshot images taken from drone-mounted cameras. The dataset has ten varied classes (pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, motor) captured from different regions of China in different environment and lighting conditions. As is evident from the classes, the images have objects of different shapes and sizes with over 2.6 million hand annotated bounding boxes which makes it suitable for our experiments.

III. Major Challenges in Aerial Object Detection

Aerial object detection has some interesting characteristics which makes it much more difficult task than the regular view object detection. We highlight some of the major challenges to aerial object detection.

A. Low spatial resolution

A typical aerial image is captured from a drone flying at a high altitude from the ground which is a bird's eye view of the objects on the ground. Hence most of the objects like pedestrians, cars, bicycles etc, are very small in spatial size and are generally crowded. This small size becomes a major challenge for object detection algorithms. Further the crowded objects in the image makes it difficult to classify each of them separately with distinctive boundaries.

B. Multitude of object sizes

Due to the large field-view of drones, they typically capture a large number of objects which differ in their spatial dimensions. For instance, the Visdrone dataset has class labels for objects as big as trucks and buses to objects like bicycles and awning-tricycles which are comparatively very small. These large objects are generally easier to detect because the deeper layers of neural networks are able to form informative features for them due to their greater receptive field. This is in contrast to smaller objects where the resolution keeps on decreasing as we go deeper in the network. Although the deeper layers are semantically strong i.e. they encode more meaningful relationships, they become spatially very weak i.e. they loose out on a lot of fine-grained information which makes detecting such objects very difficult. Hence, handling different varieties together is a challenge to deep learning based models.

C. Discriminatory Loss function

The most common performance metric for object detection is Intersection over Union (IoU). Intersection over union means how much overlap is there between ground truth and predicted bounding boxes as a ratio of the total area covered by these boxes. It determines how accurate the predicted bounding box is with respect to the ground truth box. However, Lu

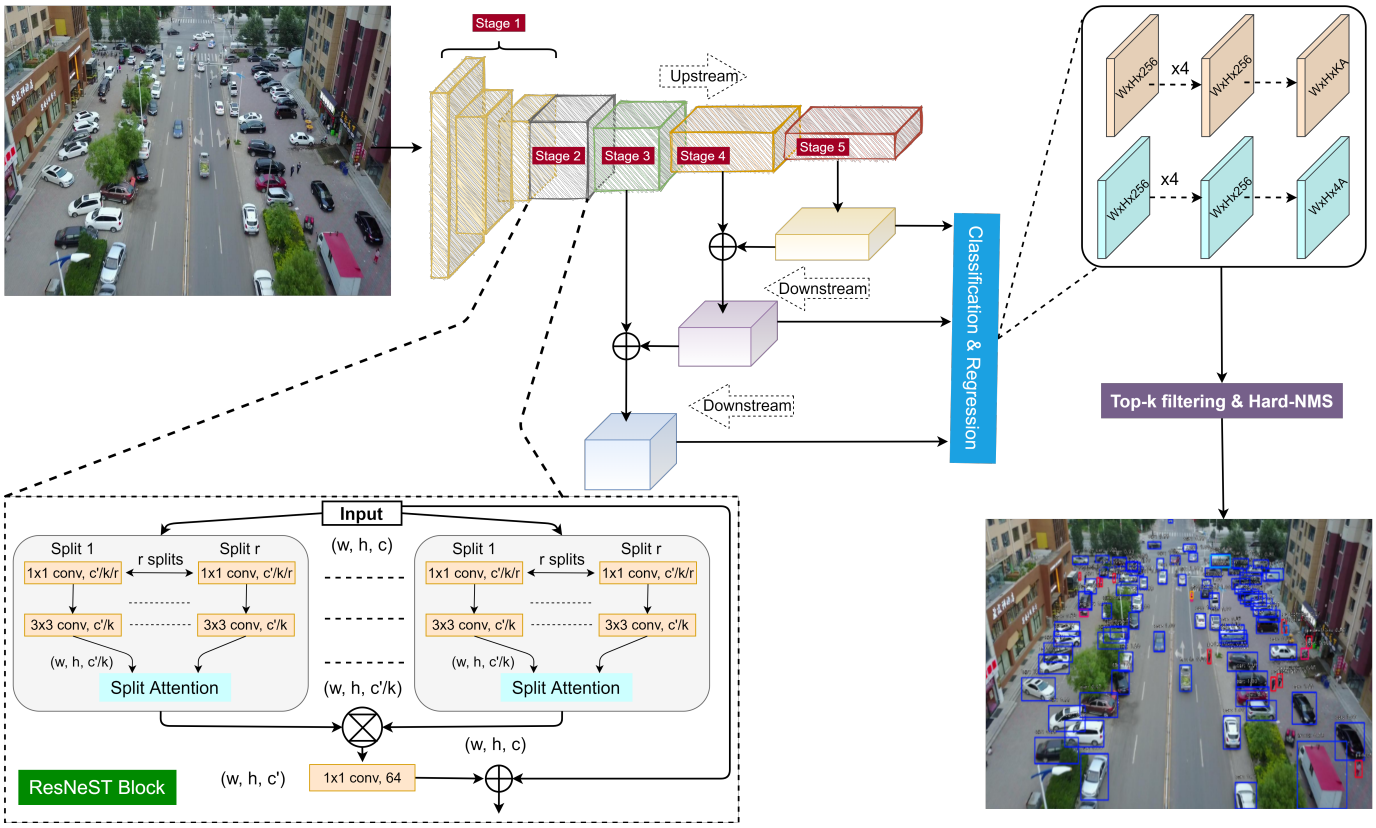


Fig. 2: **Architecture Diagram.** The input image taken from a drone is passed through ResNest and further downstream feature extraction layers followed by regression and classification subnets. We get the final object detection bounding boxes and labels as the output of regression and classification subnets. The blue bounding boxes in the output image show the predicted objects. Captions are omitted for clarity.

et al. [10] shows how IoU has an inherent bias to favour larger objects. For the same decrease in the intersection between predicted and target boxes for both small and large objects, the drop in IoU is larger for smaller objects compared to larger objects. This discrimination forces the model to focus less on small objects thereby reducing its performance.

D. Occlusion and variation in surrounding illumination

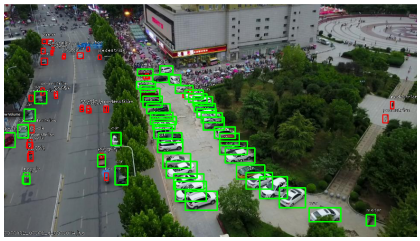
Occlusion is a general problem in object detection but it's effect is more prominent when dealing with aerial images. Occlusion along with the shadowing of objects by large buildings and trees makes detection even more difficult. Due to the small size of the objects, even partial occlusion can make detection of such objects almost impossible. Another difficulty is variation in surrounding illumination. For example some of the cars might lie in illuminated area, while some might lie in dark region. For instance, the cars in dark region with dark shade bonnets are hardest to detect for any object detector, and hence a lot of potential work needs to be done to tackle this problem of low and varying illumination.

IV. Experiments and Analysis

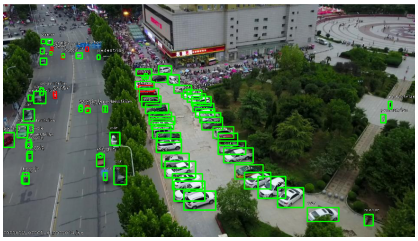
To understand and solve the multi-facet problems of object detection in aerial view as described in the previous Section, we conduct numerous experiments to analyze the effect of the backbone and anchor configuration in RetinaNet. We vary the number of scales and aspect ratios and perform anchor optimization using the optimization algorithm of Zlocha et.al. [11] We use the VGG16 and ResNet50 backbones which have been pre-trained on ImageNet. We also use ResNeSt50 backbone [12] which is pre-trained on the COCO-2017 dataset. Based on the conducted experiments, we propose a novel detector, DeepDroneSt which is optimized for aerial object detection. DeepDroneSt utilizes a superior ResNeSt50 features for improved performance. The extracted features are passed to the feature pyramid network to ensure spatially and semantically strong features. The regression and classification sub-networks take as inputs the features from each level of the pyramid and predict the offsets and confidence scores for the bounding boxes respectively. The final bounding boxes are then obtained from these predictions using top-k filtering and a technique called hard non-max suppression, which tries to eliminate the duplicate bounding boxes that might be predicted by the model.

A. Implementation Details

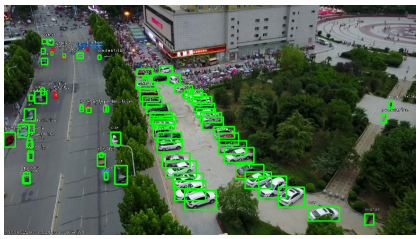
All architectures in Table I are trained end-to-end using a single Tesla V100 GPU. The images are resized so that their minimum side is equal to 800 pixels and maximum side is less than 1333 pixels. We use stochastic gradient descent with a



(a) The performance of default anchors on an image from VisDrone dataset. The green ones have the anchors available with the default anchors while the red ones don't have any anchor matching



(b) The performance of optimized anchors with 3 ratios and 5 different scales. Notice the stark decrease in number of red boxes as compared to Figure 3a



(c) The performance of optimized anchors with 5 ratios and 3 different scales. Notice the further decrease in number of red boxes as compared to Figure 3a and Figure 3b

Fig. 3: **Effect of anchor optimization.** The figures shows better object recognition from left to right with the help of more diverse and optimized anchors

#ratios	#scales	backbone	AP[%]	AP50[%]	AP75[%]	AR1[%]	AR10[%]	AR100[%]	AR500[%]
3	**3**	VGG16	15.09	23.62	16.99	1.17	10.37	19.54	19.55
		ResNet50	13.30	20.82	14.97	1.28	9.99	17.87	17.88
		ResNeSt50	**17.56**	**27.84**	**19.45**	**1.69**	**11.63**	**22.23**	**22.23**
5	**3**	VGG16	14.68	26.13	15.06	0.82	8.33	20.30	20.34
		ResNet50	12.69	23.51	12.65	0.81	7.84	18.33	18.37
		ResNeSt50	**16.89**	**29.56**	**17.59**	**0.94**	**9.44**	**22.16**	**22.16**
5	**5**	VGG16	16.40	28.15	17.33	0.88	9.23	22.36	22.4
		ResNet50	14.13	25.50	14.28	0.84	8.72	20.06	20.09
		ResNeSt50	**19.41**	**35.93**	**19.21**	**0.94**	**9.59**	**27.54**	**27.54**

TABLE I: AP values from different anchor configurations and backbones

mini-batch consisting of a single training sample. *Smooth L1* loss is used for the regression submodel and focal loss with $\alpha = 0.25$ and $\gamma = 2$ for the classification submodel. Initial learning rate is 10^{-5} and adam optimizer is employed. We experiment with 3 sets of anchor ratio and scale combinations. The first anchor set consists of 3 ratios and 3 scales with their default values used by retinanet [3]. The second and third anchor sets consist of 5 ratios and 3 scales and 5 ratios and 5 scales respectively which are optimized using an algorithm proposed by [11]. Each of these sets are trained separately with VGG16, ResNet50 and ResNeSt50 backbones. The models using the first combination of 3 ratios and 3 scales are trained for 50 epochs whereas the models using the other combinations are trained for 60 epochs to compensate for the increase in number of anchor parameters. Additionally we train another RetinaNet model with VGG16 as backbone with 5 ratios and 3 scales on the VisDrone dataset with the "large object" categories removed. Specifically we remove the car, van, truck and bus annotations from the training set and analyze the reasons behind the performance difference with and without these larger object classes. All experiments use a confidence threshold of 0.5 and NMS threshold of 0.5.

B. *Effect of Anchors*

A lot of popular object detection algorithms like Fast RCNN [13] depend on region proposals approach of generating the probable image locations for object detection. Though they have been very successful, generating these image locations is computationally expensive. To tackle this problem, another popular approach based on anchors, is being used by object detection models like Faster RCNN and RetinaNet [3]. The anchors are crude bounding boxes, whose adjustments are trained by the network to fit the objects of focus in the image. Anchors can be of different sizes and aspect ratios depending upon the properties of objects in the image. Although in most cases, the default sizes of the anchor boxes work well in localizing the objects, we found out that these are highly inefficient when it comes to localizing smaller objects as those captured from drones.

The large number of red boxes in Figure 3a clearly indicates that the default anchors sizes perform poorly on aerial images especially on smaller objects like pedestrians. The presence of these red boxes on small objects means that these objects will not be able to contribute to the training as most of them do not have an IoU overlap above the threshold (usually 50%). Hence we optimize these anchors to adjust to the sizes of aerial objects. Also the number of ratios and scales need to be decided by the network designer. Ideally, we can have a lot of anchors with various ratios and scales. But the problem is that as we increase the number of anchors, the training time increases drastically. Also just increasing the number of ratios and scales of anchors do not guarantee better performance. As we can see from Table I, despite the change from the non-

optimized 3 ratios and 3 scales anchors to the optimized 5 ratios and 3 scales anchors both the AP and AR 1-100 scores have decreased for all the backbones. However increasing both the ratios and scales to 5 and 5 and optimizing them increased both the precision and recall to a large extent. The results indicate that detection of smaller objects is more sensitive to an increase in scales than to an increase in ratios of anchor boxes. We hypothesize that this is because more number of anchor boxes will be assigned to ground truth boxes during training because of an increase in IoU due to the larger variety of scales. The chance of an increase in IoU with increase in scale is more than that due to an alternate ratio as the larger scale covers more area bringing more number of ground truth objects under it.

C. Effect of Backbone

One vital component of the RetinaNet model or any other general object detector is their backbone architecture which is used for extracting features from the high dimensional input images. The earlier layers extract simple features like edges while the deeper layers combine the features of previous layers to extract more concrete and distinct features from the image. Generally, due to the convolutional and max pool layers, the resolution of the image keeps on decreasing as we go deeper in the architecture. Consequently, the earlier layers are spatially stronger but semantically weaker while the deeper layers are spatially weaker (due to low resolution) but semantically stronger. Most of the architectures tend to use features from the deeper layers to perform computer vision tasks like object detection.

We experimented with most common backbones like VGG16 and ResNet50 which have different design and number of layers. We observe that in all the three sets of anchors, VGG16 outperforms ResNet50 in both precision and recall (Table I). We conclude VGG16 is able to extract semantically and spatially stronger features for smaller objects which are predominant in aerial images compared to ResNet50. Due to more number of layers in ResNet50, the deeper layers become spatially too weak for very small images, due to wider and less precise receptive field. Though ResNet50 uses skip connections which could theoretically skip many layers and learn a less deeper feature extractor, in our experiments we observed that the model typically does not learn to skip layers.

We also tried a recently proposed ResNest50 backbone. We observed a consistently better performance with ResNest50 as compared to other backbone layers. The superior performance of ResNest50 can be attributed to its split-attention block which learns the weights for different features and is thus able to learn stronger representations for the different object classes. While residual connections make the model easier to learn the identity function to prevent the problem of vanishing/exploding gradients, aggregate transformations help the model to learn better feature representations.

D. Object Size Modalities

The wide variety of object sizes is one of the distinctive characteristics of aerial image datasets. The object sizes can vary from as large as truck, aeroplanes, to as small as pedestrians and motor-cycles. The aiskyeye dataset also has varied object categories which makes object detection on it very challenging. To study this effect, we evaluate the model's performance on various kinds of object categories. Figure 4b shows the class wise performance of RetinaNet on Aiskyeye dataset. We observe that the model performs very poorly on small objects like pedestrain, people and bicycle when compared to larger objects like car, van, truck and bus. To further investigate the above results, we masked the classes of larger objects namely the car, van, truck and bus and retrained the model on remaining classes. Except for the awning tricycle, all other classes witnessed a considerable increase in accuracy with the largest increase present in Pedestrian and Tricycle classes. The pedestrian class witnessed an increase in AP from 5.97% to 7.09% and the tricycle class from 2.03% to 4.21%. The experiment demonstrates the dominance of the larger object categories on the loss function revealing that further improvements to the loss function can be made to improve detection on these smaller objects.

V. Promising Future Research Directions

A. Specialised Loss Function

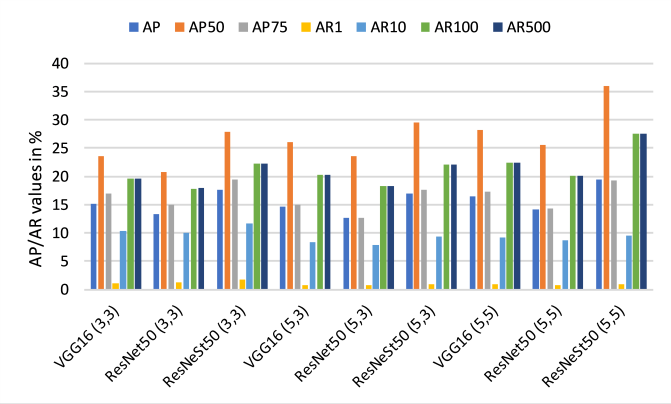
Loss functions are used to tell the model how good or bad its performance is on a particular task. It is one of the most critical parameters in learning process of machine learning models and its efficient design can lead to substantial increase in model performances. As shown in Experiments section C, larger objects are easier to detect as compared to smaller objects. Also, we showed that larger objects tend to dominate the existing loss functions over the small images, which hampers the learning for smaller images. Hence, a lot of work needs to be done in the design of efficient loss functions which can penalise the mis-detection of smaller images more heavily as compared to larger images. One such simple loss function could be made by simply scaling the loss function by a term which is inversely proportional to the area of ground truth bounding box. Thus the model will have larger loss for smaller objects as compared to larger objects. More complex loss functions can include contributions corresponding to increasing degree of complexity with features like occlusion percentage, truncation ratio, crowding and many others depending upon the attributes available for the respective datasets.

B. Experiments with Different Layer Depths

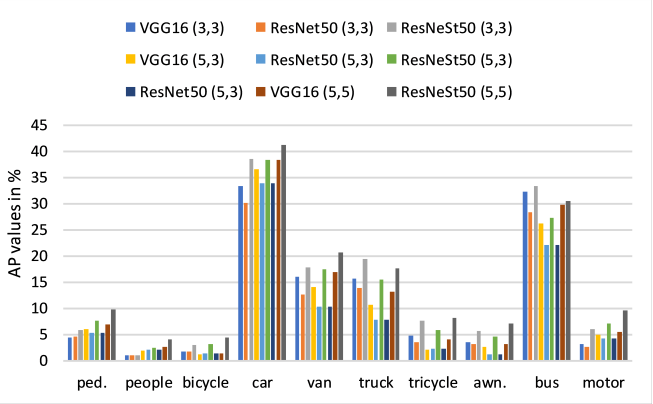
Due to small sizes of objects captured by drones, deeper layers sometimes becomes spatially too weak. Hence, using shallower layers might boost the performance significantly. Hence quantitative comparison of results on performing detection using different activation layers of backbone and FPN should be performed. Moreover, the results from top performing layers can be ensembled, to provide the final results.

C. Anchor-free Approaches

While anchors play a major role in object detectors like RetinaNet, it requires careful tuning for its effective use. In



(a) AP/AR scores on VisDrone2020 Test-Dev set. Values in bracket denote combination of ratios and scales.



(b) class-wise AP scores on VisDrone2020 Test-Dev set. Values in bracket denote combination of ratios and scales.

Fig. 4: **AP/AR Test-Set Scores.** Fig. 4 (a) shows the performance on VisDrone2020 by various models. ResNeSt50 with 5 ratios and 5 scales anchor achieves the best accuracy. Fig. 4 (b) indicates the class-wise performance of the different algorithms. Notice the stark difference in performance between cars, vans and people, bicycle.

cases of varying object sizes, anchors become a bottleneck to both qualitative and computational performance. Hence anchorless object detectors [14], [15] might provide better flexibility and performance for aerial object detection.

D. Combined Segmentation and Detection

Background noise is one of the major challenges to object detection. Hence, first performing an instance segmentation on the image could be a useful step to reduce the noise in the image, after which the object detection could be done. Recently, the Mask-RCNN and RetinaNet were combined to build RetinaMask in the literature. Similar architectures could be developed with special focus for aerial object detection and segmentation. The results of our experiments can help in deciding the various design factors of these combined architectures.

E. Self-Supervised and Unsupervised Learning

A major drawback of supervised learning methods is the need for enormous amount of dataset for training the models. Self-supervised learning uses pretext tasks, which are designed for learning visual features that can be further used to solve the actual downstream task. Some of the commonly used pretext tasks include minimizing reconstruction error in autoencoders, image inpainting, greyscale colorisation among many others. Similar research in aerial object detection can significantly improve the state of the art by decreasing the dependence on very large training datasets.

VI. Conclusion

Aerial object detection is a very important field of research with huge potentials for practical applications in developing smart cities, and surveillance systems powered by the recent advances in IoT. In this paper we first discuss the existing deep learning based approaches towards detecting objects in aerial images. We then present a detailed analysis of the challenges and difficulties in detecting objects that are distinctive to aerial image datasets. Several experiments are carried out to understand the impact of many critical parameters that affect detection accuracy such as the anchor box configuration and backbone architecture used for feature extraction in RetinaNet. We also propose DeepDronest, an object detection model that outperforms current RetinaNet variants with the help of split-attention network of ResNest. We then present possible future areas of research where further improvements can be made such as designing a more efficient loss function or combining segmentation with detection.

References

[1] F. Qi, X. Zhu, G. Mang, M. Kadoch, and W. Li, "Uav network and iot in the sky for future smart cities," *IEEE Network*, vol. 33, no. 2, pp. 96–101, 2019.

[2] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[3] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb 2020.

[4] P. Zhu and W. et al., "Visdrone-det2018: The vision meets drone object detection in image challenge results," in *The European Conference on Computer Vision (ECCV) Workshops*, 9 2018.

[5] M. Lin, Tsung-Yiand Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[6] M. Mandal, M. Shah, P. Meena, S. Devi, and S. K. Vipparthi, "Avdnet: A small-sized vehicle detection network for aerial visual data," *IEEE Geoscience and Remote Sensing Letters*, 2019.

[7] M. Mandal, M. Shah, P. Meena, and S. K. Vipparthi, "Sssdet: Simple short and shallow network for resource efficient vehicle detection in aerial scenes," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3098–3102.

[8] H. Qin, Y. Li, J. Lei, W. Xie, and Z. Wang, "A specially optimized one-stage network for object detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[9] G. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.

[10] X. Lu, Y. Zhang, Y. Yuan, and Y. Feng, "Gated and axis-concentrated localization network for remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 179–192, 2020.
[11] M. Zlocha, Q. Dou, and B. Glocker, "Improving retinanet for ct lesion detection with dense masks from weak recist labels," *arXiv preprint arXiv:1906.02283*, 2019.
[12] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.
[13] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2606–2615.
[14] L. Tychsen-Smith and L. Petersson, "Denet: Scalable real-time object detection with directed sparse sampling," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 428–436.
[15] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Foveabox: Beyond anchor-based object detector," *arXiv preprint arXiv:1904.03797*, 2019.