DeepDroNeSt: AI-optimized Aerial Object Detection for IoT Applications

***Abstract*—Unmanned Aerial Vehicles (UAVs) are emerging as a powerful tool for various industrial and smart city applications. UAVs coupled with various sensors can perform many cognitive tasks such as object detection, surveillance, traffic management, urban planning, etc. These AI-driven tasks often rely on computationally expensive deep learning approaches, which cannot yield their full potential with embedded processors on a power-constrained battery-operated drone. Edge-AI has emerged as a popular alternative in such scenarios. This work proposes a novel deep learning approach which optimizes the 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 categorises 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. Finally, we reflect on what we have achieved and further propose several shining directions of future work to inspire further research and advancement in aerial object detection.**

***Index Terms*—**

I. Introduction

Intelligent UAVs have a major projected role for the development of 'Smart Cities'. The objective of smart cities is to provide robust infrastructure and services with minimal resource utilization. It facilitates integration of information and communication technologies (ICT) such as Internet of Things (IoT) and blockchain networks to provide holistic integrated services impacting living and security [29]. The UAVs with computer vision capabilities have a critical role in realizing these targets.

Object detection is an integral component in computer vision. Drones and Unmanned Aerial Vehicles (UAV) which are equipped with high resolution cameras can be employed in an extensive range of applications including surveillance, disaster response strategies and construction of transportation systems. Aerial image datasets like VisDrone and UAVDT differ from natural images such as those in ImageNet or COCO in a variety of aspects. We outline some of these distinctions in section IV.

Traditional object detection methods can be divided into three stages: the region proposal generation stage with selective search (SS) method, the feature extraction stage, and the classification stage. A large number of region proposals are generated with the help of SS. Various hand-crafted features such as the scale-invariant feature transform (SIFT) or histogram of oriented gradients (HOG) are extracted at a shallow level which are then fed into a classifier such as a support vector machine (SVM) [38]. Deep learning is an automatic feature extraction framework which overcomes several problems of the aforementioned approach such as generation of large number of redundant proposals by the time-consuming SS and manual feature design process.

A large number of deep learning approaches have been proposed for object detection in aerial images using various techniques like weakly supervised learning, special augmentation techniques and combining segmentation with detection. We present an analysis of these methods in section II.

In this paper we focus on a specific architecture, RetinaNet, a single stage end-to-end object detector. RetinaNet utilizes the Feature Pyramid Network (FPN) along with a focal loss that seeks to reduce class imbalance. The FPN generates feature maps that have both high spatial and semantic information which are then passed on to regression and classification submodels that predict the bounding box offsets and class scores respectively.

We specifically focus on the critical parameters of the model that influence detecting smaller objects such as those in the VisDrone 2019 dataset. We experimentally analyze the effect that different backbones have in detecting objects of various categories. We also show that the default anchors used in RetinaNet, although suitable for object detection in natural images are inadequate when it comes to aerial images.

II. Related Work

Generic Object Detection - With the rapid development of deep convolutional neural networks, several methods have been proposed for object detection. The current top deep learning based general purpose 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 category-agnostic region proposals are generated from an image. These contain most of the objects while filtering out majority of the negative locations. In the second stage, CNN features are extracted from these regions, and then category-specific classifiers are used to determine the category labels of the proposals. Representatives of region-based detectors include R-CNN [9], Faster-RCNN [30] and R-FCN [4].

Single stage detectors such as SSD [24] and RetinaNet [23] completely eliminate region proposal generation and are hence region-free. They directly predict class probabilities and bounding box offsets from full images with a single feed forward CNN network.

While generic object detectors achieve excellent performance on natural image datasets (e.g MS COCO [21] and PASCAL VOC [7]) these achieve only mediocre performance on high resolution aerial images (e.g VisDrone [42] and DOTA [36]).

Aerial Object Detection - We review some of the proposed deep learning approaches for detecting objects in aerial images. Among the earliest methods weakly supervised learning was one such approach. This approach was followed by Han et al.[11] , Fan Zhang et al. [40] and Dingwen Zhang et al. [39]. In weakly supervised learning the training set only includes binary labels indicating whether an object of a particular category is present or not and hence does not require manual annotation of the bounding boxes. [11] achieves performance comparable to the baseline supervised models in the Google Earth dataset. [40] uses a set of two convolutional neural network architectures that share features to detect aircrafts in satellite images. [39] trains a target detector through an iterative process after generating suitable training examples by self-adaptive segmentation and negative mining.

Deep Convolutional Neural Networks (DCNN) were used to extract orientation robust features by Zhu et al. [41] They experimentally show that the POOL 5 layer in Alexnet is able to model the rotation factor better than the FC6 and FC7 layers and hence the layer combinations which involved the POOL 5 layer achieved better recall. The Alexnet model was pre-trained on Imagenet and their evaluations were done on two datasets containing images obtained from Google Earth on vehicle and plane object detection.

Jiang et al. [13] used graph-based superpixel segmentation to generate proposals and then trained a CNN to classify these proposals for vehicle detection. A rotation-invariant layer based upon Alexnet was introduced by Cheng et al [2]. They used a novel optimization function which enforces training samples before and after rotation to share the same features. In [32], a simple CNN based approach is presented for automatic detection in aerial images and in [6], deep belief networks are employed which shows promising results.

III. Dataset

Large datasets are prerequisite for supervised deep learning. There are numerous datasets available for object detection like ImageNet [5], Open Images Dataset [17], CIFAR-10 [16], FashionMNIST [37] among many otheres. In comparision to general domain, the datasets for aerial object detection are less in number, size, and quality. In general, the aerial object datasets could be either captured by satellites like DOTA [36], COWC [27], NWPU VHR-10 [3] , or captured by drones like VisDrone [42], Inria Aerial Image Labelling Dataset [26], Stanford Drone Dataset [31].

For our experiments, we will be using Visdrone2019 object detection dataset [42]. The dataset is collected by Aisky-eye team and 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. (Figure - 1)

IV. Major Challenges in Aerial Object Detection

Aerial object detection has peculiar characteristics which make it a lot different and difficult than the natural object detection tasks. This section highlights some of the major challenges to aerial object detection.

A. Low spatial resolution

A typical object detection dataset is prepared from a drone flying at a high altitude from the ground. The pictures it captures 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 together.

As we can see from Figure - 4, the cars, as seen from drone, are very small in size in proportion to the image. This small size becomes a major challenge for object detection algorithms. Further notice the crowded cars in the image, which makes it difficult to classify each of them separately with crisp boundaries.

B. Variety of sizes of objects

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 trucks and buses to objects like bicycles 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 and have greater receptive field. This is in contrast to smaller objects where the resolution keeps on decreasing as we go deeper in the network and although the deeper layers are semantically strong, they become very spatially weak that makes detecting such objects very difficult. Hence, handling different varieties together becomes challenging for network designer.

Figure - 3 shows the variation of sizes of objects discussed in the above paragraphs. The bus at the front has substantially large spatial size than the cars present at the far end of the scene. These kinds of variation in object size makes it difficult for general object detection algorithms to detect and classify.

C. Discriminatory Loss function

The most common metric for performance in object detection is Intersection over Union (IoU). It measures the goodness of the predicted bounding box with respect to ground truth bounding box. But according to Lu et al. [25], IoU differs in its properties for small and large images. For the same decrease in the intersection between predicted and target boxes, for both small and larger image, the drop in IoU is more for smaller images as compared to larger images. This discrimination leads the model to focus on small images lesser, and hence their performance suffers.

Fig. 1: Sample images from Visdrone dataset. Ground truth bounding boxes are shown as green boxes. Notice the variation of scenes, variation of number of objects and their sizes.

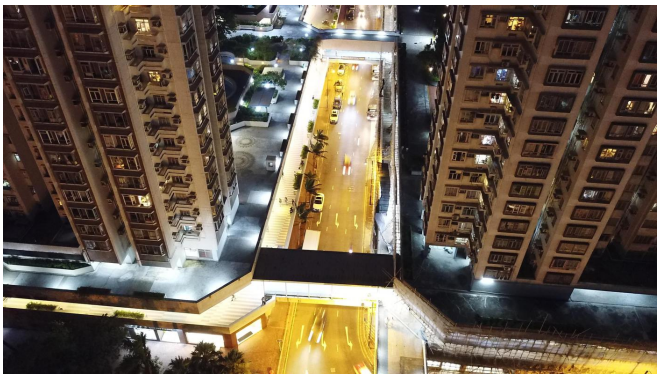


Fig. 2: A typical image from Visdrone dataset for object detection. Note carefully, that the objects are crowded and very small in size, which makes the object detection difficult.

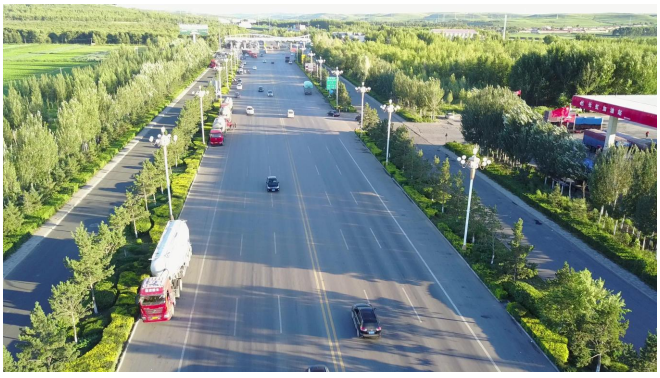


Fig. 3: Notice the difference in sizes of various objects. The bus is considerably bigger than the cars present at the far side of the image.

D. Occlusion and variation in surrounding illumination

Occlusion is a common problem in object detection in general. But it's effect becomes more profound when dealing with aerial images. Not only occlusion, but shadowing of the objects by large buildings and trees become common. Due to their small sizes, even partial occlusion makes detection difficult.

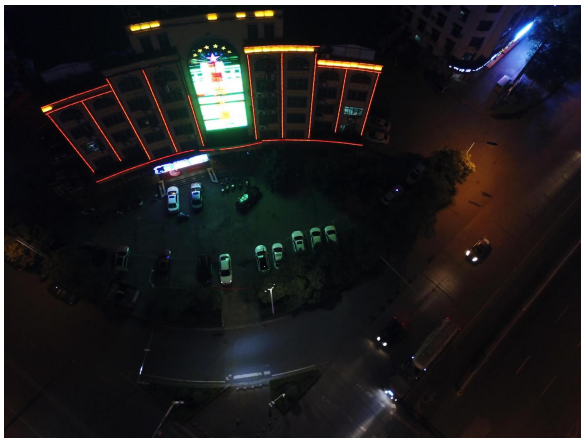
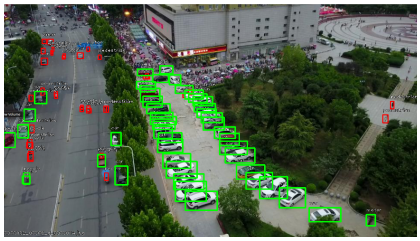


Fig. 4: Notice in the figure the lighting of the building illuminates some of the cars, while the others remain in dark. Further observe that the cars with white bonnets are more easier to spot than black ones in dim lighting.

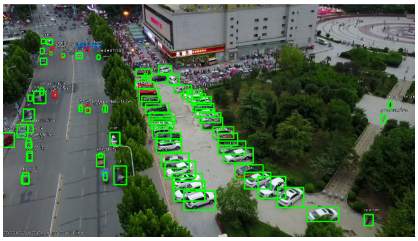
Another issue is variation in surrounding illumination. We can observe from Figure -4 that some of the cars lie in illuminated area, while some of the other cars lie in dark region. The cars in dark region with dark shade bonnets are hardest to detect for any object detector, and hence a lot of potential work needs to be done to tackle this problem of low and varying illumination.

V. Experiments

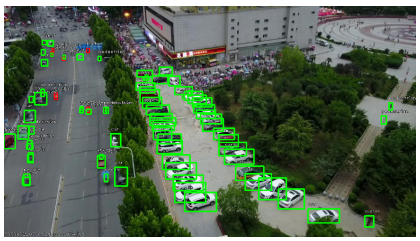
We run numerous experiments to analyze the effect of the backbone and anchor configuration in RetinaNet. We vary the number of scales and aspect ratios and perform anchor



(a) The performance of default anchors on an image from VisDrone dataset. The boxes are ground truth annotation boxes. The green ones have the anchors available with the default anchors while the red ones don't have any anchor matching, hence it won't contribute during training.



(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 Fig - 5a



(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 Fig - 5a and Fig - 5b

Fig. 5: Effect of Anchor Optimization

Number of ratios	Ratios	Number of scales	Scales
3 (default)	[0.5, 1, 2]	3 (default)	[1.0, 1.26, 1.587]
5 (optimized)	[0.405, 0.64, 1.0, 1.562, 2.469]	3 (optimized)	[0.4, 0.5, 0.629]
5 (optimized)	[0.377, 0.628, 1.0, 1.592, 2.652]	5 (optimized)	[0.4, 0.504, 0.596, 0.633, 0.868]

TABLE I: Different combinations of ratios and scales.

#ratios	#scales	backbone	set	AP[%]	AP50[%]	AP75[%]	AR1[%]	AR10[%]	AR100[%]	AR500[%]
3	3	**VGG16**	**Test-Dev**	15.09	23.62	16.99	1.17	10.37	19.54	19.55
			Validation	14.55	23.21	15.54	0.52	6.01	18.76	18.82
		ResNet50	**Test-Dev**	13.30	20.82	14.97	1.28	9.99	17.87	17.88
			Validation	13.73	22.03	14.80	0.56	6.08	18.18	18.21
5	3	**VGG16**	**Test-Dev**	14.68	26.13	15.06	0.82	8.33	20.30	20.34
			Validation	14.22	25.67	13.80	0.44	4.61	18.82	18.95
		ResNet50	**Test-Dev**	12.69	23.51	12.65	0.81	7.84	18.33	18.37
			Validation	12.67	23.75	12.08	0.44	4.49	17.67	17.78
5	5	**VGG16**	**Test-Dev**	16.40	28.15	17.33	0.88	9.23	22.36	22.4
			Validation	16.2	28.22	16.59	0.48	5.37	21.45	21.57
		ResNet50	**Test-Dev**	14.13	25.50	14.28	0.84	8.72	20.06	20.09
			Validation	14.31	25.78	13.91	0.45	4.87	19.46	19.55

TABLE II: AP scores of various configurations on VisDrone 2019 dataset

optimization using the optimization algorithm of Zlocha et.al. [43] As is common practice, the backbones are pre-trained on ImageNet and then fine-tuned on the detection dataset.

Implementation Details - All architectures in Table II are trained end-to-end using a single GPU. Similar to [22] the images are are resized so that their minimum side is equal to 800 pixels. We use stochastic gradient descent with a 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 as per [23]. Initial learning rate is 10^{-5} and Adam optimizer is employed. We experiment with 3 sets of anchor ratio and scale combinations as given in Table I. Each of these sets are trained separately with VGG16 and ResNet50 backbone. The models using the first combination of 3 ratios and 3 scales are trained for 50 epochs whereas the models using the other two 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 categories.

A. Effect of Anchors

A lot of popular object detection algorithms like Fast RCNN [35] depend on region proposals approach of generating the probable image locations for object detection. Though they have been very successful, due to high time complexity of region proposals, another popular approach, anchors, is being used by object detection models like Faster RCNN [30] and RetinaNet [30]. 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. Usually the default sizes and scales work well for most of the objects, but we found out that they struggle a lot with small images captured from drones.

The large number of red boxes in figure - 5a clearly implies 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 level (usually 50%).

Table I shows the combinations of ratios and scales used. The first combination uses the default values for 3 ratios and 3 scales. The remaining two have been optimized using [43]. From Table II we notice that 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 both 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 for both VGG16 and ResNet50 backbones. 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.

B. Effect of Backbone

One important section of RetinaNet model or any other general object detector models are their backbone architecture which are used for extracting features from input images. The way it works is that the earlier layers extract simple features like edges while the deeper layers combine the features of previous layers to extract more concrete and distinctive features from the image. Also, in general, due to the convolutions and max pool layers, the resolution of the image keeps on decreasing as we go deeper in the architecture. This means that while the earlier layers are spatially stronger, they are semantically weaker. The deeper layers, on the other hand, are spatially weaker (due to low resolution), while semantically stronger. Hence, most of the architectures tend to use deeper feature extractors to perform computer vision tasks like object detection.

In all the three sets of anchors we observe that VGG16 performs better than the corresponding ResNet50 model in precision and recall (Table - II). We can thus conclude that VGG16 is able to extract semantic and spatially stronger features for smaller objects which are predominant in aerial images compared to ResNet50. This means that the deeper layers, which are supposed to be semantically stronger, becomes syntactically too weak for very small images, due to wider and less precise receptive field.

C. Object size modalities

Variety of object sizes is one of the major characteristic of aerial object detection 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.

Figure-6b and Table-III 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 as compared to its performance on larger objects like car, van, truck and bus.

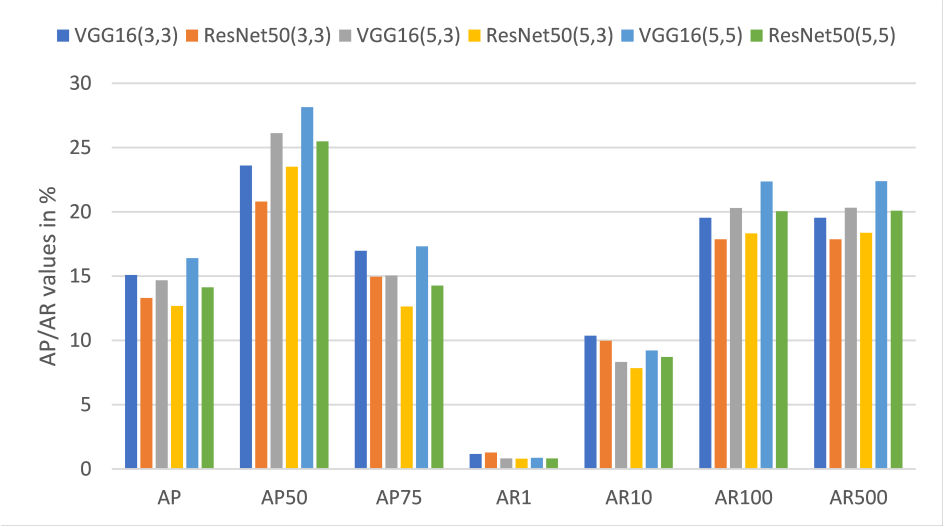
To further investigate the above results, we masked the classes of larger objects namely the car, van, truck and bus, and retrained the model on remaining classes. The results of the experiment is summarised in Figure 7. Notice the huge increase in detection accuracy of the smaller classes. Except for the awning tricycle class, all other classes have witnessed a considerable increase in accuracy with the largest increase present in Pedestrian and Tricycle classes. 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. Surprisingly the accuracy of awning tricycle class has dropped. We attribute this drop to the extremely small size of the awning tricycle objects making it almost insignificant compared to even the other small object categories.

VI. Proposed Method

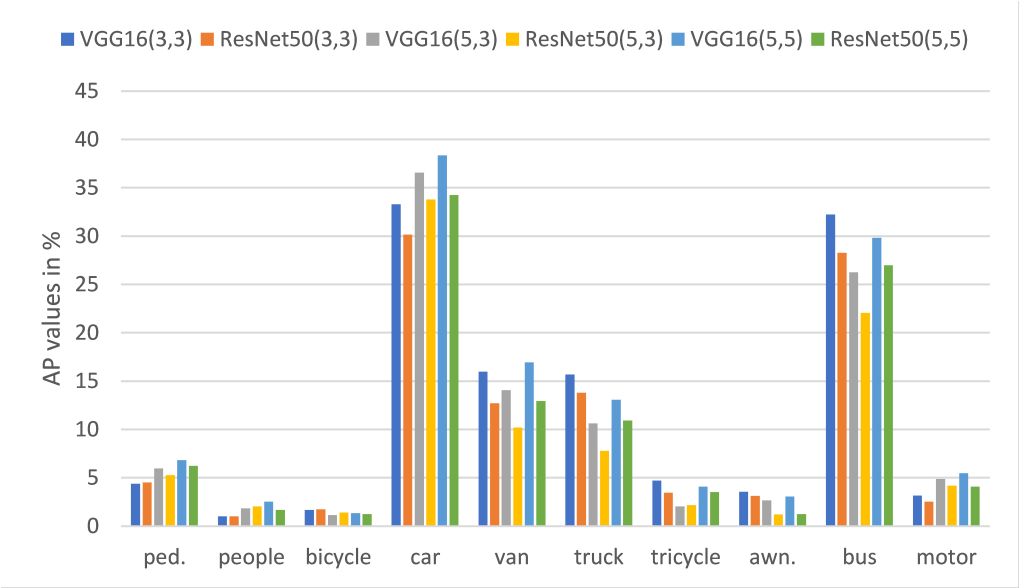
VII. Discussion

VIII. Future Directions

Specialised loss function Loss functions control the 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, easy 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 heavily as compared to larger images. One trivial loss function could be made by adding a term to loss function inversely proportional to the area of ground truth bounding box. 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.



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



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

Fig. 6: AP/AR scores for our various models

#ratios	#scales	backbone	pedestrian	people	bicycle	car	van	truck	tricycle	awn. tricycle	bus	motor
3	**3**	**VGG16**	4.4	1.02	1.67	33.31	15.99	15.67	4.71	3.55	32.24	3.15
		ResNet50	4.5	1.02	1.74	30.15	12.7	13.8	3.45	3.13	28.29	2.54
5	**3**	**VGG16**	5.97	1.83	1.13	36.57	14.05	10.64	2.03	2.66	26.27	4.87
		ResNet50	5.26	2.05	1.4	33.79	10.21	7.8	2.17	1.2	22.07	4.2
5	**5**	**VGG16**	6.82	2.54	1.36	38.36	16.93	13.07	4.09	3.07	29.83	5.49
		ResNet50	6.23	1.69	1.25	34.26	12.94	10.93	3.53	1.25	27	4.08

TABLE III: class-wise AP scores on the Test-Dev set of VisDrone 2019

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. The results from top performing layers can be ensembled, to provide the final results.

Visualisation of Feature Maps Anchor-less approaches While anchors play a major role in object detectors like

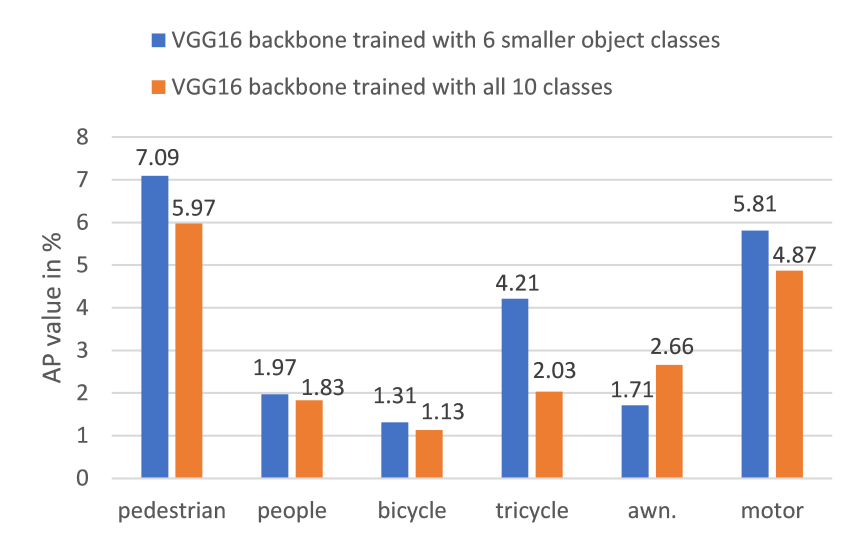


Fig. 7: comparison between AP values after training RetinaNet with VGG16 backbone with only 6 smaller classes and with all 10 classes

backbone	#ratios	#scales	set	score threshold	AP[%]	AP50[%]	AP75[%]	AR1[%]	AR10[%]	AR100[%]	AR500[%]
ResNeSt50	3	3	Test-Dev	**0.5**	17.56	27.84	19.45	1.69	11.63	22.23	22.23
				0.4	19.32	31.93	20.74	1.69	11.75	25.5	25.5
				0.25	20.76	35.39	21.77	1.69	11.78	28.94	28.94
	5	3	Test-Dev	**0.5**	16.89	29.56	17.59	0.94	9.44	22.16	22.16
				0.4	18	32.28	18.32	0.94	9.51	24.27	24.27
				0.25	19.41	35.93	19.21	0.94	9.59	27.54	27.54

TABLE IV: AP scores for RetinaNet with ResNeSt50 bakbone at various score thresholds

RetinaNet, it requires careful tuning for its effective use. In cases of varying object sizes, it becomes a bottleneck to performance both qualitatively and computationally. Hence anchorless object detectors ([33], [15], [19]) might provide better flexibility and performance for aerial object detection.

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 debunk the noise in the image, after which the object detection could be done. Recently Fu et al., combined Mask RCNN [12] and RetinaNet to build RetinaMask [8]. Similar architectures could be developed with special focus for aerial object detection and segmentation.

Self-Supervised and Unsupervised Learning A lot of research has been done on developing supervised learning algorithms for object detection in general. 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 design for learning visual features that can be further used to solve the actual downstream task [14]. Some of the commonly used pretext tasks include minimizing reconstruction error in autoencoders [10], image inpainting [28], greyscale colorisation [18] among many others. Recently self-supervised and unsupervised learning have seen a upsurge of research interest and several new architectures ([20], [1], [34]) have been developed using limited or self-supervision. Similar research in aerial object detection can significantly improve the state of the art with low focus on data size.

IX. Conclusions

References

[1] Elad Amrani et al. “Learning to Detect and Retrieve Objects From Unlabeled Videos”. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019, pp. 3713–3717.

[2] G. Cheng, P. Zhou, and J. Han. “Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 54.12 (2016), pp. 7405–7415.

[3] Gong Cheng et al. “Multi-class geospatial object detection and geographic image classification based on collection of part detectors”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 98 (2014), pp. 119–132.

[4] Jifeng Dai et al. “R-FCN: Object Detection via Region-based Fully Convolutional Networks”. In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 379–387. URL: http://papers.nips.cc/paper/6465-r-fcn-object-detection-via-region-based-fully-convolutional-networks.pdf.

[5] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[6] W. Diao et al. “Efficient Saliency-Based Object Detection in Remote Sensing Images Using Deep Belief Networks”. In: *IEEE Geoscience and Remote Sensing Letters* 13.2 (2016), pp. 137–141.

[7] Mark Everingham et al. “The Pascal Visual Object Classes Challenge: A Retrospective”. In: *International Journal of Computer Vision* 111 (2014), pp. 98–136.

[8] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C Berg. “RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free”. In: *arXiv preprint arXiv:1901.03353* (2019).

[9] R. Girshick et al. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 580–587.

[10] Anupriya Gogna and Angshul Majumdar. “Semi supervised autoencoder”. In: *International Conference on Neural Information Processing*. Springer. 2016, pp. 82–89.

[11] J. Han et al. “Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning”. In: *IEEE Transactions on Geoscience and Remote Sensing* 53.6 (2015), pp. 3325–3337.

[12] Kaiming He et al. “Mask R-CNN”. In: *CoRR* abs/1703.06870 (2017). arXiv: 1703.06870. URL: http://arxiv.org/abs/1703.06870.

[13] Q. Jiang et al. “Deep neural networks-based vehicle detection in satellite images”. In: *2015 International Symposium on Bioelectronics and Bioinformatics (ISBB)*. 2015, pp. 184–187.

[14] Longlong Jing and Yingli Tian. “Self-supervised visual feature learning with deep neural networks: A survey”. In: *arXiv preprint arXiv:1902.06162* (2019).

[15] Tao Kong et al. “Foveabox: Beyond anchor-based object detector”. In: *arXiv preprint arXiv:1904.03797* (2019).

[16] Alex Krizhevsky and Geoff Hinton. “Convolutional deep belief networks on cifar-10”. In: *Unpublished manuscript* 40.7 (2010), pp. 1–9.

[17] Alina Kuznetsova et al. “The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale”. In: *arXiv preprint arXiv:1811.00982* (2018).

[18] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. “Colorization as a proxy task for visual understanding”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6874–6883.

[19] Hei Law and Jia Deng. “Cornernet: Detecting objects as paired keypoints”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 734–750.

[20] Wonhee Lee, Joonil Na, and Gunhee Kim. “Multi-task Self-supervised Object Detection via Recycling of Bounding Box Annotations”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4984–4993.

[21] Michael Lin Tsung-Yiand Maire et al. “Microsoft COCO: Common Objects in Context”. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 740–755.

[22] T. Lin et al. “Feature Pyramid Networks for Object Detection”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 936–944.

[23] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[24] Dragomir Liu Weiand Anguelov et al. “SSD: Single Shot MultiBox Detector”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 21–37.

[25] X. Lu et al. “Gated and Axis-Concentrated Localization Network for Remote Sensing Object Detection”. In: *IEEE Transactions on Geoscience and Remote Sensing* 58.1 (2020), pp. 179–192.

[26] Emmanuel Maggiori et al. “Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark”. In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2017, pp. 3226–3229.

[27] T Nathan Mundhenk et al. “A large contextual dataset for classification, detection and counting of cars with deep learning”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 785–800.

[28] Deepak Pathak et al. “Context encoders: Feature learning by inpainting”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544.

[29] Fei Qi et al. “UAV network and IoT in the sky for future smart cities”. In: *IEEE Network* 33.2 (2019), pp. 96–101.

[30] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems*. 2015, pp. 91–99.

[31] Alexandre Robicquet et al. “Learning social etiquette: Human trajectory understanding in crowded scenes”. In: *European conference on computer vision*. Springer. 2016, pp. 549–565.

[32] I. Ševo and A. Avramović. “Convolutional Neural Network Based Automatic Object Detection on Aerial Images”. In: *IEEE Geoscience and Remote Sensing Letters* 13.5 (2016), pp. 740–744.

[33] Lachlan Tychsen-Smith and Lars 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.

[34] Keze Wang et al. "Towards human-machine cooperation: Self-supervised sample mining for object detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1605–1613.

[35] Xiaolong Wang, Abhinav Shrivastava, and Abhinav 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.

[36] G. Xia et al. "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.

[37] Han Xiao, Kashif Rasul, and Roland Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: *arXiv preprint arXiv:1708.07747* (2017).

[38] S. Xu et al. "Object Classification of Aerial Images With Bag-of-Visual Words". In: *IEEE Geoscience and Remote Sensing Letters* 7.2 (2010), pp. 366–370.

[39] D. Zhang et al. "Weakly Supervised Learning for Target Detection in Remote Sensing Images". In: *IEEE Geoscience and Remote Sensing Letters* 12.4 (2015), pp. 701–705.

[40] F. Zhang et al. "Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection". In: *IEEE Transactions on Geoscience and Remote Sensing* 54.9 (2016), pp. 5553–5563.

[41] H. Zhu et al. "Orientation robust object detection in aerial images using deep convolutional neural network". In: *2015 IEEE International Conference on Image Processing (ICIP)*. 2015, pp. 3735–3739.

[42] Pengfei Zhu et al. "VisDrone-DET2018: The Vision Meets Drone Object Detection in Image Challenge Results". In: *The European Conference on Computer Vision (ECCV) Workshops*. Sept. 2018.

[43] Martin Zlocha, Qi Dou, and Ben Glocker. "Improving RetinaNet for CT Lesion Detection with Dense Masks from Weak RECIST Labels". In: *arXiv preprint arXiv:1906.02283* (2019).