Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 19(8): 7920–7932.
DOI: 10.3934/mbe.2022370
Received: 18 April 2022
Revised: 12 May 2022
Accepted: 23 May 2022
Published: 27 May 2022

Research article

Enhancing object detection in aerial images

Vishal Pandey[1], Khushboo Anand[1], Anmol Kalra[2], Anmol Gupta[1], Partha Pratim Roy[1] and Byung-Gyu Kim[3,*]

[1] Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee, India

[2] Department of Computer Science and Engineering, College of Engineering, Roorkee, India

[3] Department of IT Engineering, Sookmyung Women's University, Seoul, South Korea

* **Correspondence:** Email: bg.kim@sookmyung.ac.kr.

Abstract: Unmanned Aerial Vehicles have proven to be helpful in domains like defence and agriculture and will play a vital role in implementing smart cities in the upcoming years. Object detection is an essential feature in any such application. This work addresses the challenges of object detection in aerial images like improving the accuracy of small and dense object detection, handling the class-imbalance problem, and using contextual information to boost the performance. We have used a density map-based approach on the drone dataset VisDrone-2019 accompanied with increased receptive field architecture such that it can detect small objects properly. Further, to address the class imbalance problem, we have picked out the images with classes occurring fewer times and augmented them back into the dataset with rotations. Subsequently, we have used RetinaNet with adjusted anchor parameters instead of other conventional detectors to detect aerial imagery objects accurately and efficiently. The performance of the proposed three step pipeline of implementing object detection in aerial images is a significant improvement over the existing methods. Future work may include improvement in the computations of the proposed method, and minimising the effect of perspective distortions and occlusions.

Keywords: object detection; aerial images; VisDrone-2019; drones; RetinaNet

Abbreviations: DMNet: Density-map guided object detection network; DOTA: Dataset of object deTection in aerial images; MCNN: Multi-column convolution neural network; MSCOCO: Microsoft common objects in context; R-CNN: Region-based convolution neural network; RPN: Region proposal network; SSD: Single-shot detector; VOC: Visual object classes; YOLO: You only look once

1. Introduction

In the upcoming years, the large scale infrastructural changes and development of smart cities will see a rise in the use of drones for various applications such as package delivery, traffic monitoring, policing, pollution management, etc. [1,2]. Object detection in such situations will be a crucial part of the solution ecosystem. Object detection in aerial images is necessary for various critical applications in computer vision, such as pedestrian detection, autonomous vehicles, surveillance, crowd counting and scene classification. Object detection approaches can broadly be divided into two categories, namely region-based methods and single-shot methods.

Region-based methods include Fast R-CNN [3], Mask R-CNN [4], Faster R-CNN [5]. Fast R-CNN [3] uses an exhaustive selective search to generate region proposals, and based on those proposals, it performs feature extraction followed by classification. Overcoming the exhaustive selective search, Faster R-CNN [5] uses convolution network to generate Region Proposal Network (RPN). Mask R-CNN [4] further extends Faster R-CNN [5] by introducing an additional branch of predicting segmentation masks in a pixel-to-pixel manner.

Single-shot methods include YOLO [6], SSD [7] and many other methods proposed in recent times. These methods view the problem of object detection as a regression problem. YOLO [6] views the entire image at once and simultaneously predicts bounding box and class probabilities. It makes use of objectness scores for prediction. It is not recommended for small objects that account for larger aerial datasets percentages. SSD [7], on the other hand, uses multi-sized convolutional feature maps and has a smaller trade-off between speed and accuracy.

Since the advent of deep learning architectures, general object detection has witnessed state-of-the-art performances by these methods on natural datasets MSCOCO [8] and Pascal VOC [9]. Zhang et al. [10] and Chu et al. [11] have implemented state-of-the-art object detection on MSCOCO [8] dataset and KITTI dataset [12] respectively which are natural datasets where the images have been taken from the ground in a first person view. However, these methods have appeared to be less than satisfactory on datasets of aerial or satellite images such as the VisDrone dataset [13], DOTA [14]. Despite the good results in other application areas, state-of-the-art techniques often fail in performance when applied to aerial images.

Aerial images are generally captured via satellites, drones, and aeroplanes equipped with cameras that capture images from the top-view and larger field view. Since the images contain a larger field of view and usually the top-view of the images, this contributes to several challenges in aerial images compared to general object detection. Some of these challenges are mentioned below.

a. Scale Variation: The angle of photos or camera viewpoint varies significantly, resulting in the variance of the scale of different objects per image and class.
b. Imbalanced Data: Highly imbalanced and non-uniform distribution of objects corresponding to various categories. Bigger size objects are easier to detect by the detector than small size objects.
c. Occlusion: Occlusion refers to one object getting hidden by another one. Object occlusion issues between objects are very common in aerial images.
d. Low Resolution: Since images are taken from a particular height, resolution may vary specifically for small objects.
e. Lightning variation: Nighttime pictures pose a difficulty in detecting the accurate coverage of an object.

f. The dominance of small objects: A large percentage of the dataset comprises small objects; therefore, it is necessary to give importance to small objects for better performance.

These issues contribute to making aerial image object detection more challenging. However, addressing these challenges can help significantly in improving the existing methods and contributing to some very critical applications.

Multiple researchers like Jiang et al. [15], Yang et al. [16], Behera et al. [17], Santosh et al. [18], and Balamurugan et al. [19] focus on addressing the IOT challenges around drones applicable in smart cities like traffic monitoring and communication. Another important application of using drones in smart cities is scene text recognition [20, 21]. In the object detection domain, past works Li et al. [22] have focused on the strategy of simply using architectures for general objects with putting some emphasis on small objects and improving detection. Unel et al. [23] suggested tiling, which is the use of cropping images for boosting small object detection. Extending the tiling technique [23], Yang et al. [24] have proposed the idea of clusters to crop dense object regions and using ScaleNet to balance the shape of crops of image. Wang et al. [25] have addressed the issue of invariant viewpoint by focusing on leveraging the spatial information with the help of increased receptive fields, which can efficiently cover all objects. Li et al. [26] addressed the use of density maps for the tiling process. They used Multi-column CNN (MCNN) proposed by Zhang et al. [27] to generate the density maps and then use those maps to crop the image further. Yu et al. [28] have introduced the concept of dilated convolutions.

Aerial images contain objects of various scales; therefore, networks with better receptive fields are desired. Dilated convolutions help in expanding the receptive fields. Zhang et al. [27] and Shen at al. [29] in their work have addressed the effects of using dilated convolutions and how it can contribute to the detection of small objects. The class-imbalance problem refers to the scenario where the minority class has fewer examples as compared to the majority class. This leads to a drop in the overall performance as the detector tends to over classify the majority class. Hensman et al. and Masko et al. [30] showed that an imbalanced classification task could be handled by the data augmentation process in the training data. Nemoto et al. [31] in their work addressed the idea of data augmentation via rotation and inversion for minimising the class imbalance. In recent times, two-stage detectors under the category of R-CNN have been adopted for the task of object detection. They are preferred for their computation speed and accuracy.

However, considering the aerial images and the challenges like scale variation, occlusion, and small and densely located objects, these detectors often witness a performance drop when applied directly to the aerial datasets. Lin et al. [32] have proposed a method, RetinaNet, which is a single-stage detector, but it can perform similar results as compared to the two-stage detector approach by R-CNN. RetinaNet combines the features of both Region Proposal Networks and Feature Pyramids which are used for multi-scale object detection.

Aerial images often contain background as the objects to be detected are relatively smaller in size. Suppose the whole image is fed directly into the detector. In that case, the computation time is unnecessarily added without any results. The detector has to look for objects in the region proposals of the background where there might be no object present. Many recent approaches have adopted the methodology of first cropping out the sub-regions where the object is absent and then performing the final detection. Inspired by this methodology, we have adopted a similar approach. We view the problem as a composition of three modules. First, in order to address the class imbalance problem, we have augmented the rare, occurring class instances with rotation of 90, 180 and 270 degrees. Second, we have used an improved CNN to generate density maps which then help in cropping that image. Third, in order to accurately give better results on aerial images we have adopted RetinaNet [32] with improved anchor sizes. All the experiments have been performed on the VisDrone [13] dataset.

The contributions of this paper can be summarised as follows:

1) We have tried to address the class imbalance problem in the VisDrone dataset by rotation augmentation of infrequent classes. We have also demonstrated the impact of this augmentation in improving the final results.
2) We have proposed an improved version of MCNN with better receptive fields achieved using dilated filters considering the size variation of objects in the dataset.
3) We have suggested effective anchors for RetinaNet to detect smaller objects specific to aerial datasets accurately. We have also demonstrated the performance of anchor optimised RetinaNet on various backbone architectures.

The rest of the paper is organised as follows. Section 2 explains the dataset and the three-step methodology proposed for object detection in aerial images. Section 3 presents the analysis of the results obtained from the proposed methods as compared to existing methods. Section 4 discusses a summary of the results and their impact, along with limitations and possible future directions for the study.

2. Materials and methods

This section describes the dataset used in this study along with the three step proposed methodology and the procedure of the experiment employed by the study.

2.1. Dataset description

With the high demand of practical applications like surveillance, there have been various datasets available for analysis, comparison and setting up benchmarks to identify the best performance acquired so far. Since, natural images significantly differ from aerial images, special aerial image datasets have gained more focus. We have experimented our methodology and tested its performance against the VisDrone-2019 [13] dataset. Figure 1 shows some sample images from the dataset with ground-truth bounding boxes. VisDrone 2019 is one of the most popular datasets used for object detection in aerial images. The total 8599 images are divided into three sets where the training set includes 6471 images, validation set includes 548 images and the remaining 1580 images are being included into the test set. The dataset contains images with occlusion, lightning variation and densely packed objects encapsulating the various challenges as depicted in Figure 2. Since the images are taken from an altitude, the camera viewpoint changes. The dataset contains ten classes/categories which includes human beings and transport vehicles. Human being classes are ‘pedestrian’ which refer to those standing while class ‘person’ refer to human who is standing, walking, running or any other pose. Transport vehicle classes include ‘bicycle’ ‘car’ ‘van’ ‘truck’ ‘tricycle’ ‘awning tricycle’ ‘bus’ and ‘motor’.

2.2. Proposed method

We have adopted a three step methodology that utilises the advantages of these architectures. First step deals with the process of addressing the class-imbalance problem. In the second step, we have proposed an improved MCNN with dilated filters or convolutions in order to gain higher receptive fields and generate better image crops out of the original image. We have performed decomposition of original image followed by object detection in order to minimise the background noise. Third step performs the detection using RetinaNet and analyses the performance of RetinaNet with default anchors and also proposes adjusted anchor sizes in RetinaNet to boost the detection of small objects. The framework of our approach is illustrated in Figure 3.

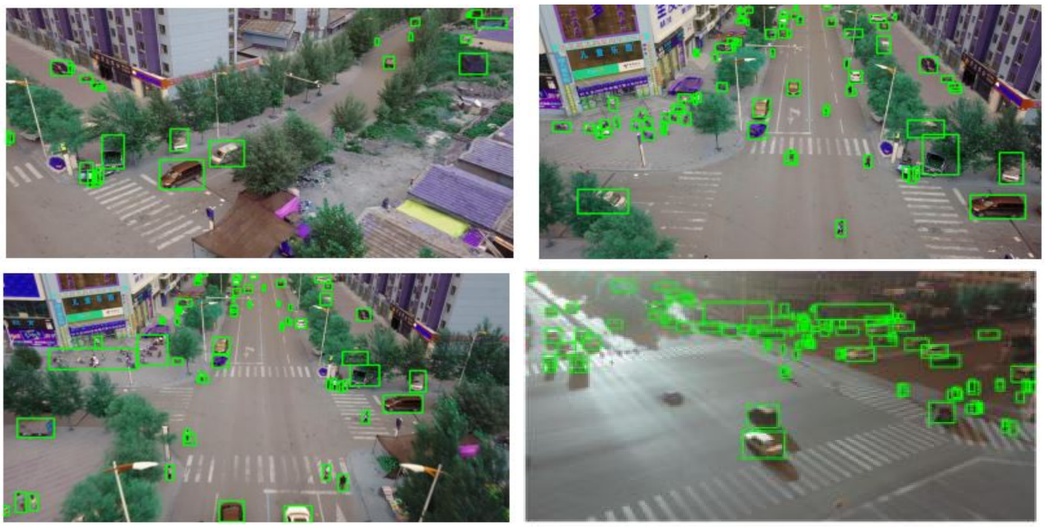

Figure 1. Sample images from VisDrone [13] dataset with ground-truth bounding boxes.

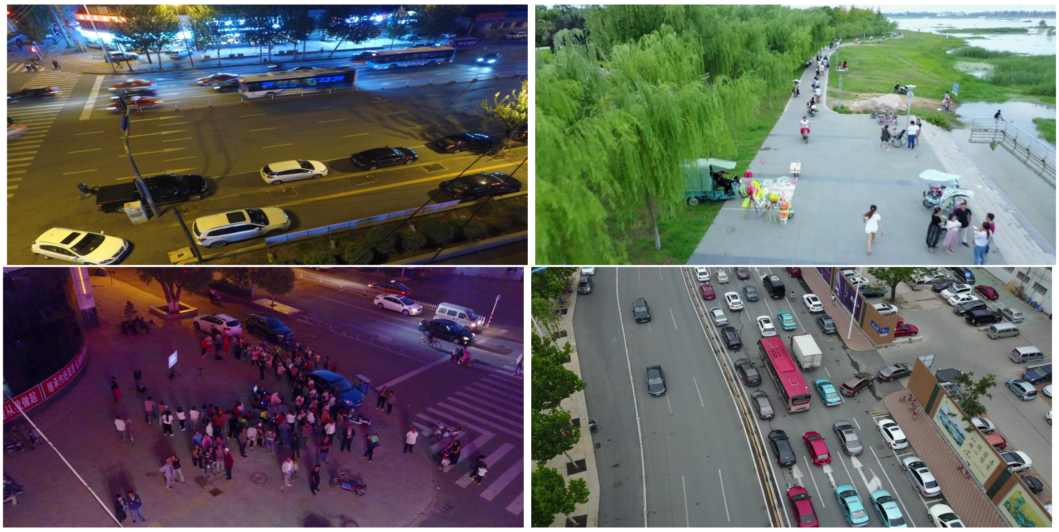

Figure 2. Images depicting challenges such as illumination variation, size variation, occlusion.

2.2.1. Data augmentation of infrequent classes

VisDrone dataset suffers from the class-imbalance problem. As illustrated in Figure 4, the classes ‘car’ and ‘pedestrian’ account for a larger portion of the dataset while classes ‘bus’ and ‘awning tricycle’ occur in small proportions. We performed Data Augmentation to minimise this problem. We performed augmentation of the infrequent categories by picking out images with rarely occurring objects present from the dataset and adding those entries back into the dataset. In order to avoid the case of overfitting of oversampled images upon direct augmentation, we augmented the images by rotating them at angles 90, 180, 270. This method helps in reducing the difference between minority classes and majority classes.

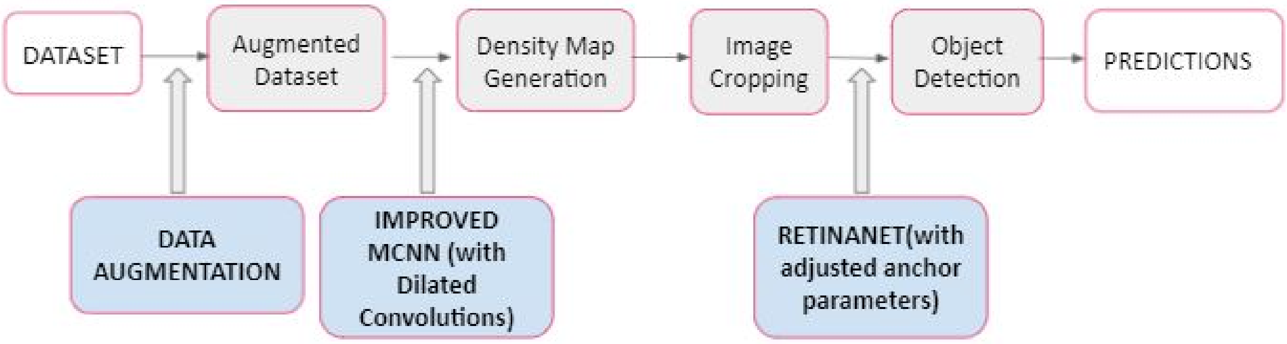

Figure 3. Proposed methodology.

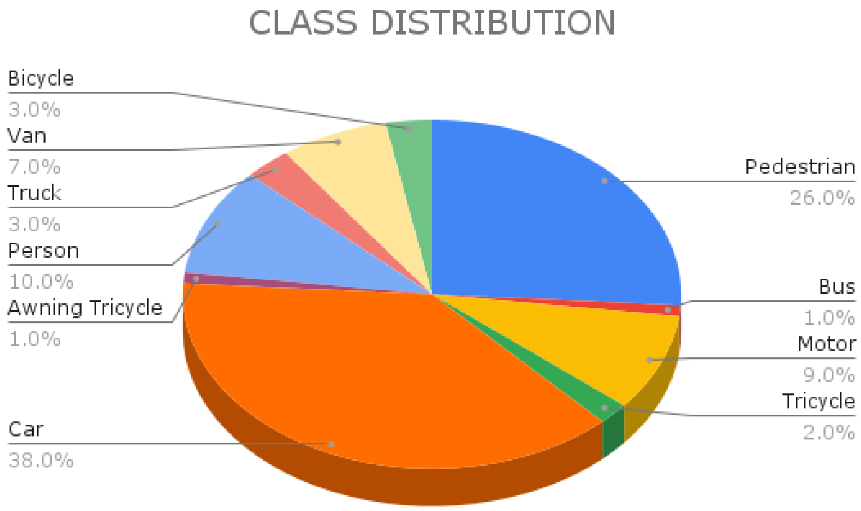

Figure 4. Visualization of class distribution in VisDrone dataset in terms of percentage.

2.2.2. Image cropping using improved MCNN with dilated convolutions

Before feeding images for the detection, we divide the image into multiple crops. We have adopted the methodology of image cropping where input image is decomposed into several crops or chips by leveraging the advantages of density maps. Segmentation followed by detection helps to minimise the background noise. We used the concept of density-based maps used in crowd counting which is one of the extensive practical applications, to the domain of aerial imagery object detection. For the purpose of density map generation, we have adopted improved MCNN which is a multi-column architecture where we have involved four columns in order to cover all the scales and capture the features properly. Conventional MCNN [27] used three columns for small, medium and large size objects but for an effective receptive field for smaller objects we have incorporated four columns with adjusted filters. In addition to it, to increase the receptive field of columns we have introduced dilated convolutions Instead of standard convolution-pooling layers. Dilation factor helps in looking at larger areas while preserving spatial resolution by expanding the receptive field of kernels. Figure 5 is the proposed architecture of improved MCNN.

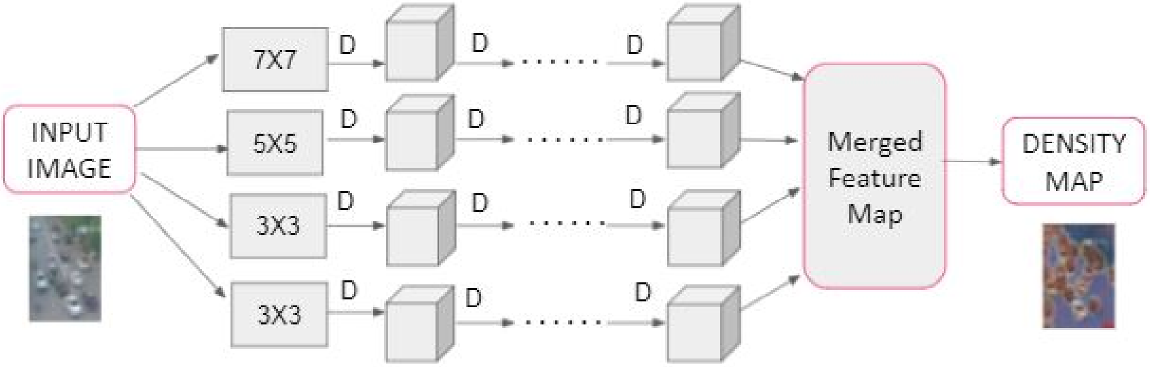

Figure 5. Improved MCNN with 4 columns for various scales. ‘D’ represents Dilation factor.

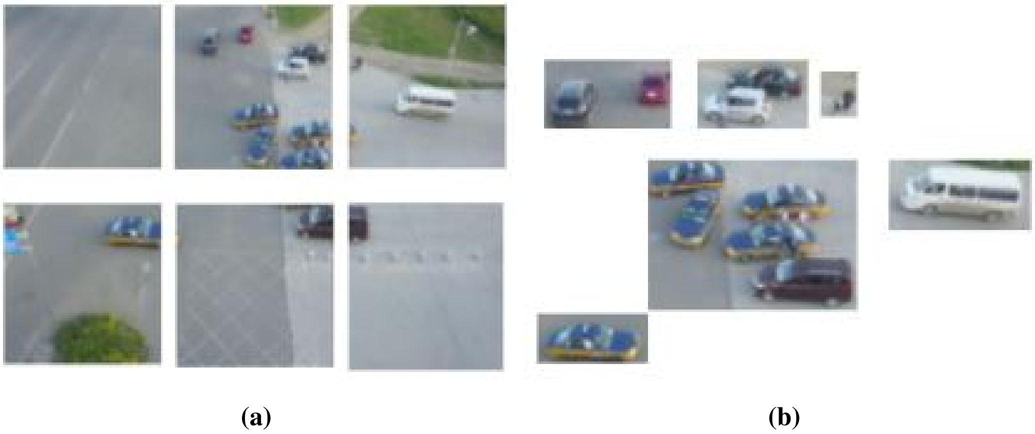

Figure 6. Visualization of (a) uniform cropping vs. (b) density crops achieved by our methodology. (a) corresponds to region proposals/ crops suggested by uniform cropping and (b) corresponds to cropped chips corresponding to MCNN with dilated convolutions. The uniform crops contain mostly background pixels. The crops made by density based methodology are more useful with respect to our task.

This way features are adaptive and robust towards perspective distortion, varying resolution, and scale variation. In this case, MCNN consists of four columns or four parallel CNNs with different filter sizes to capture features corresponding to different scales. After density maps, the cropping module generates image crops for the forward process of detection. Cropping module is able to produce better crops, as can be seen in Figure 6, due to distribution of objects as background area can be filtered out effectively and the number of objects belonging to each crop can also be reduced. The cropping module has been used similarly as mentioned in [26]. The images are cropped by accumulating the pixel intensities and comparing their sum with a given threshold value. This is achieved with the use of a sliding window that slides over the density map and gathers the information to generate a density mask. The windows with intensities higher than threshold are combined together to break the image into crops. Then the final detection is performed. This facilitates the effective recognition of small objects.

2.2.3. Object detection by RetinaNet with better anchors

The final step is to perform detection on the crops obtained in the previous step. For the detection, any object detection architecture can be used. Algorithms such as YOLO [6], Fast R-CNN [3], Faster R-CNN [5] etc. can be chosen as the baseline architecture keeping in mind the trade-offs and challenges. We have adopted RetinaNet for this purpose due to its recent promising results in detecting small objects. RetinaNet which is a single-stage detector but it can perform similar results as compared to the two-stage detector approach by R-CNN. RetinaNet combines the features of both Region Proposal Networks and Feature Pyramids which are used for multi-scale object detection. However, we analyse that the default configuration of RetinaNet anchors is inadequate for detection in aerial images. The default anchor sizes are 32, 64, 128, 512 which work fine for normal objects but often miss out the smaller objects. We adjusted the anchor sizes to 16, 32, 64, 128 which improves the detection as there is dominance of small objects and it is then able to capture the instances that it missed earlier. Further, we demonstrate the effect of adjusted anchors with respect to different backbone architectures. We have used ResNet50, ResNet101 to analyse the effect of RetinaNet.

2.3. Experiment implementation

We have first tried to implement and replicate the results of [26]. In order to understand their approach and look for scope of improvements, we have followed their procedure of implementation. Their work uses MCNN for generating the crops and Faster R-CNN [5] for detection. As suggested, we have used MMDetection toolbox [33] for this purpose. MMDetection toolbox is a rich toolbox which is open-source and contains a number of popular detector algorithms. It also provides the flexibility of hyperparameter tuning and provides inference and ability to train on standard datasets or customizable datasets. Before applying final detection, there is a need to convert text annotations of the images into annotations of MSCOCO [8] format as the detector algorithms are designed to run on these formats containing natural images. For the backbone of the network, the ResNet-50 and ResNet-101 architectures were used.

Inspired by [26], we have implemented our suggested methodology on the VisDrone-2019 dataset [13], with data augmentation as the first step. The classes with total number of instances less than 10,000 in the training set were 'bus' 'tricycle' 'awning tricycle' with 5926, 8058, and 3246 instances respectively. We picked out the images containing these classes and augmented their 90, 180, and 270 degree rotations back to the dataset. We have implemented improved MCNN in PyTorch as the underlying framework for creating density maps. Our MCNN uses four parallel CNNs with local receptive fields of varying sizes 7, 5, 3, and 3. Filters of varying sizes enable looking at objects of different scales, for example, filters with small receptive fields will help in creating maps for smaller objects. In addition to it, we have used dilated convolutions instead of normal (conv-pooling-conv-pooling) in the structure. The dilation filter values (1,2,4,8) vary for each column corresponding to the kernel size. The resultant from each column are then stacked together to generate the density map. We have used the Stochastic Gradient Descent optimizer and Mean Absolute Error for the generation of density maps similar to the setting in [26].

Image crops are generated by cropping the connected neighbouring region as specified by the density maps. The cropping module implementation has been referred from [26]. For the purpose of detection, any state-of-the-art object detector can be used, here in this case we have implemented RetinaNet using Keras as per [32]. Before applying final detection, we converted annotations into the format of RetinaNet. RetinaNet has default anchor configurations in which we included anchor size 16 as some of the objects span over very less pixel area in the image and omitted the highest size 512. In addition to it, we increased the scales from 3 to 5 to increase the number of anchor boxes. As per the standard setting in [32], we have used Adam optimiser, Smooth L1 loss and focal loss with standard hyperparameters. We trained the model with default anchor sizes and then adjusted the size to boost the performance. For the backbone of the network, we have used ResNet-50 and ResNet-101 architectures and trained them separately to observe the effect of depth of backbone architectures.

3. Results and discussion

Average Precision has been used to evaluate and compare performance. It refers to precision across various threshold values and evaluates the overlap between the ground-truth bounding boxes and the detected boxes.

Initially, we have implemented the approach presented in [26]. It is important to note that those experiments were performed on only half of the data which resulted in decrease of performance. The results are summarised in Table 1.

Table 1. Results of standard vs. our implementation of [26].

Method	Backbone Architecture	AP
DMNet [26]	ResNet-50	28.2
DMNet [26]	ResNeXt-101	29.4
DMNet (Our implementation of [26])	ResNet-50	27.1
DMNet (Our implementation of [26])	ResNeXt-101	27.9

Further we have run experiments and evaluated the performance of our methodology on VisDrone [13] Test-Dev dataset. We have tried to analyse the impact of every step by including them in the methodology in a step-wise manner. We have introduced data augmentation keeping rest of the structure in their default configurations, that is, MCNN and RetinaNet. Then we introduced improved MCNN keeping RetinaNet still same with default parameters. Afterwards, we added the improved RetinaNet to assess the overall performance. The results are illustrated in Table 2.

We can infer from the table that our results outperform the standard [26] across both the backbone architectures ResNet 50 and ResNet 101 respectively. The significance of each step can be clearly inferred from the results. Improved MCNN helps in increasing receptive fields of the filters, RetinaNet with adjusted anchor parameters helped in detecting smaller objects efficiently. Further, experiments done with ResNet-101 gave better results as compared to ResNet-50. Instances belonging to classes such as ‘people’ ‘bicycle’ ‘awning tricycle’ often go undetected which leads to performance drop.

Table 2. Quantitative results of our methodology on VisDrone [13] dataset.

Method	Backbone	AP
DMNet [26]	ResNet-50	28.2
DMNet [26]	ResNet-101	28.5
Data Augmentation(Ours) + MCNN + RetinaNet	ResNet-50	28.1
Data Augmentation(Ours) + Improved MCNN(Ours) + RetinaNet	ResNet-50	28.7
Data Augmentation(Ours) + Improved MCNN(Ours) + RetinaNet	ResNet-101	28.9
Data Augmentation(Ours) + Improved MCNN(Ours) + Improved RetinaNet(Ours)	ResNet-50	29.6
Data Augmentation(Ours) + Improved MCNN(Ours) + Improved RetinaNet(Ours)	ResNet-101	29.9

Based on the results, we were able to gain understanding of the effect of various modules. Our methodology focuses on density map based cropping of images using improved version of MCNN that generates better image crops for detection. We can see the performance gain of 0.6 points when improved MCNN was used along with the default RetinaNet architecture over ResNet-50. This increase occurs due to dilated filters used in MCNN that are able to expand the receptive fields at various scales. This performance further improves by a factor of 0.2 when ResNet-101 is used.

Further, it also suggests an improved RetinaNet with adjusted anchors and ratios that increase the detection of smaller objects. After replacing the default RetinaNet parameters by our suggested parameters, we were able to witness a further increase of 0.7 with ResNet-50. This can be accounted for by the fact that anchors were now able to capture the small objects with smaller pixel coverage which went undetected earlier. Similar to earlier experiments, ResNet-101 boosted it to more by a factor of 0.3. We can also infer that the deeper backbone architectures deliver better results.

Additionally, we also adopted data augmentation of infrequent classes as a pre-processing step that helped in increasing the rare-occurring instances. This pipeline is able to outperform the other density based approaches in terms of AP due to the robust properties of the improved architectures.

4. Conclusions

Object Detection in aerial images has become an interesting and demanding area of research particularly due to its practical applications in the upcoming development of smart cities. It has emerged as an area with a wider range of possibilities and wide spectrum of approaches that have been adapted for improving the performance. This article discusses the challenges due to which detectors used in general object detection designed for natural images fail in achieving state-of-the-art performance in this domain. We have also discussed some of the architectures and methodologies proposed in recent times considering the challenges of aerial images. We have adopted data augmentation of infrequent classes as a part of pre-processing to increase the instances of less occurring classes. Further, expanding the ideology of density based image cropping, our approach generates density maps with improved MCNN with dilation factor, then crops images and finally performs detection with RetinaNet with default as well as adjusted anchor parameters across different backbone architectures. We have emphasised on the importance of every contribution suggested in this work with observations. The performance of the proposed methodology is a significant improvement over the existing state-of-the-art performance. However, due to the three step pipeline, our methodology becomes slightly computationally expensive. Future work can be done towards improvement in the computations. Further, Ensemble Learning methodology has always been proven to exhibit better performances as compared to the stand-alone detectors. Therefore, ensemble methodology also provides a good scope to try different combinations of models on aerial images. In the field of aerial imagery object detection, the scenario of heavy overlapping and extremely thin objects demands new approaches for robust object detection. Additionally, minimizing the effect of perspective distortion, occlusion are open avenues for further work.

Acknowledgments

This research was conducted under and supported by a project from the Life Sciences Research Board, Defence Research and Development Organization titled 'Automated Cognitive Metrics Classification Framework to Improve Human Machine Collaboration' (Project number: DRD-1659-CSE) to Prof. Partha Pratim Roy.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. H. Alsamhi, O. Ma, M. S. Ansari, F. A. Almalki, Survey on collaborative smart drones and internet of things for improving smartness of smart cities, *IEEE Access*, **7** (2019), 128125–128152. https://doi.org/10.1109/ACCESS.2019.2934998

2. M. A. Khan, B. A. Alvi, A. Safi, I. U. Khan, Drones for good in smart cities: A review, in *International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC)*, (2018), 1–6.

3. R. B. Girshick, Fast R-CNN, in *2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 1440–1448. https://doi.org/10.1109/ICCV.2015.169

4. K. He, G. Gkioxari, P. Dollár, R. B. Girshick, Mask R-CNN, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 2980–2988. https://doi.org/10.1109/ICCV.2017.322

5. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39** (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

6. J. Redmon, S. Divvala, R. B. Girshick, A. Farhadi, You Only Look Once: Unified, real-time object detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 779–788. https://doi.org/10.1109/CVPR.2016.91

7. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Y. Fu, et al., SSD: Single Shot MultiBox Detector, in *European Conference on Computer Vision*, (2016), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

8. T. Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft COCO: Common Objects in Context, in *European Conference on Computer Vision*, (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

9. M. Everingham, S. Eslami, L. Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, *Int. J. Comput. Vision*, **111** (2014), 98–136. https://doi.org/10.1007/s11263-014-0733-5

10. Y. Zhang, J. Chu, L. Leng, J. Miao, Mask-refined r-cnn: A network for refining object details in instance segmentation, *Sensors*, **20** (2020), 1010. https://doi.org/10.3390/s20041010

11. J. Chu, Z. Guo, L. Leng, Object detection based on multi-layer convolution feature fusion and online hard example mining, *IEEE Access*, **6** (2018), 19959–19967. https://doi.org/10.1109/ACCESS.2018.2815149

12. A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, *Int. J. Rob. Res.*, **32** (2013), 1231–1237. https://doi.org/10.1177%2F0278364913491297

13. D. Du, Y. Zhang, Z. Wang, Z. Wang, Z. Song, Z. Liu, et al., VisDrone-DET2019: The vision meets drone object detection in image challenge results, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, (2019), 213–226. https://doi.org/10.1109/ICCVW.2019.00030

14. G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. J. Belongie, J. Luo, 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), 3974–3983. https://doi.org/10.1109/CVPR.2018.00418

15. J. Jiang, F. Liu, W. W. Ng, Q. Tang, W. Wang, Q. V. Pham, Dynamic incremental ensemble fuzzy classifier for data streams in green internet of things, *IEEE Trans. Green Commun. Networking*, (2022), 1. https://doi.org/10.1109/TGCN.2022.3151716

16. Y. Yang, W. Wang, L. Liu, K. Dev, N. M. F. Qureshi, AoI optimization in the UAV-aided traffic monitoring network under attack: A stackelberg game viewpoint, *IEEE Trans. Intell. Transp. Syst.*, (2022), 1–10. https://doi.org/10.1109/TITS.2022.3157394

17. S. Behera, D. P. Dogra, M. K. Bandyopadhyay, P. P. Roy, Crowd characterization in surveillance videos using deep-graph convolutional neural network, *IEEE Trans. Cybern.*, (2021), 1–12. https://doi.org/10.1109/TCYB.2021.3126434

18. K. K. Santhosh, D. P. Dogra, P. P. Roy, Anomaly detection in road traffic using visual surveillance: A survey, *ACM Comput. Surv.*, **53** (2020), 1–26. https://doi.org/10.1145/3417989

19. N. M. Balamurugan, S. Mohan, M. Adimoolam, A. John, W. Wang, DOA tracking for seamless connectivity in beamformed iot-based drones, *Comput. Stand. Interfaces*, **79** (2022), 103564. https://doi.org/10.1016/j.csi.2021.103564

20. P. Keserwani, P. P. Roy, Text region conditional generative adversarial network for text concealment in the wild, in *IEEE Transactions on Circuits and Systems for Video Technology*, **32** (2022), 3152–3163. https://doi.org/10.1109/TCSVT.2021.3103922

21. P. Keserwani, A. Dhankhar, R. Saini, P. P. Roy, Quadbox: Quadrilateral bounding box based scene text detection using vector regression, in *IEEE Access*, **9** (2021), 36802–36818. https://doi.org/10.1109/ACCESS.2021.3063030

22. J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan, Perceptual generative adversarial networks for small object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1951–1959. https://doi.org/10.1109/CVPR.2017.211
23. F. O. Unel, B. Özkalayci, C. Çigla, The power of tiling for small object detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2019), 582–591. https://doi.org/10.1109/CVPRW.2019.00084
24. F. Yang, H. Fan, P. Chu, E. Blasch, H. Ling, Clustered object detection in aerial images, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 8310–8319. https://doi.org/10.1109/ICCV.2019.00840
25. H. Wang, Z. Wang, M. Jia, A. Li, T. Feng, W. Zhang, et al., Spatial attention for multi-scale feature refinement for object detection, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, (2019), 64–72. https://doi.org/10.1109/ICCVW.2019.00014
26. C. Li, T. Yang, S. Zhu, C. Chen, S. Guan, Density map guided object detection in aerial images, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2020), 737–746. https://doi.org/10.1109/CVPRW50498.2020.00103
27. Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016) 589–597. https://doi.org/10.1109/CVPR.2016.70
28. F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, preprint, arXiv:1511.07122.
29. Y. Zhang, T. Shen, Small object detection with multiple receptive fields, in *IOP Conference Series: Earth and Environmental Science*, **440** (2020), 32093. https://doi.org/10.1088/1755-1315/440/3/032093
30. D. Masko, P. Hensman, The impact of imbalanced training data for convolutional neural networks, *Degree Project in Computer Science, KTH Royal Institute of Technology*, 2015.
31. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.
32. T. Y. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, (2017), 2999–3007. https://doi.org/10.1109/ICCV.2017.324
33. K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, et al., MMDetection: Open MMLab detection toolbox and benchmark, preprint, arXiv:1906.07155.

AIMS Press