International Journal of Computer Science, Engineering and Information Technology (IJCSEIT),
Vol.12, No.2/3/4, August 2022

MULTIPLE OBJECTS AND ROAD DETECTION IN UNMANNED AERIAL VEHICLE

M. Saranya, Kariketi Tharun Reddy, Madhumitha Raju and Manoj Kutala

Computer Science College of Engineering Guindy, Chennai, India

ABSTRACT

Unmanned Aerial Vehicles have greater potential to widely used in military and civil applications. Additionally equipped with the cameras can also be used in agriculture and surveillance. Aerial imagery has its own unique challenges that differ from the training set of modern-day object detectors, since it is made of images of larger areas compared to the regular datasets and the objects are very small on the contrary. These problems do not allow us to use common object detection models. Currently there are many computer vision algorithm that are designed using human centric photographs, But from the top view imagery taken vertically the objects of interest are small and fewer features mostly appearing flat and rectangular, certain objects closer to each other can also overlap. So detecting most of the objects from the birds eye view is a challenging task. Hence the work will be focusing on detecting multiple objects from those images using enhanced ResNet, FPN, FasterRCNN models thereby providing an effective surveillance for the UAV and extraction of road networks from aerial images has fundamental importance.

KEYWORDS

Unmanned Aerial Vehicle. ResNet, FPN, FasterRCNN, multiple object detection, aerial images, road detection.

1. INTRODUCTION

An unmanned Aerial Vehicle (UAV), popularly known as Drone, is an airborne system or an aircraft operated remotely by a human operator or autonomously by an onboard computer. Unmanned aerial vehicles (UAVs), because of their capabilities of offering ubiquitous computing resources, have been adopted to provide services in different application scenario. Camera based vision systems could be configured over UAV to perform image processing and video streaming. Vision assisted UAV system could be leveraged to deploy navigation autonomous control for itself. For example, relying on the camera involved sensory systems, vision assisted UAV could achieve path finding, balance control and landing maneuvering in a adaptively controlled manner. There are various applications when exploiting the wide-area images such automatic traffic control, military reconnaissance, suspicious vehicle/convoy detection, border security, surveillance of high-security areas, and intelligence.

The need for object detection systems is increasing due to the ever-growing number of digital images in both public and private collections. Object recognition systems are important for reaching higher level autonomy for robots [3]. Applying computer vision (CV) and machine learning (ML), it is a hot area of research in robotics. Drones are being used more and more as robotic platforms. The research in this article is to determine how to use existing object detection systems and models can be used on image data from a drone. One of the advantages of using a drone to detect objects in a scene may be that the drone can move close to objects compared to other robots [4], for example, a wheeled robot. However, there are difficulties with UAVs because of top-down view angels [5] and the issue to combine with a deep learning system for

DOI : 10.5121/ijcseit.2022.12401

compute intensive operations [6]. When a drone navigates a scene in search for objects, it is of interest for the drone to be able to view as much of its surroundings as possible [7, 8]. However, images taken by UAVs or drones are quite different from images taken by using a normal camera. For that reason, it cannot be assumed that object detection algorithms normally used on "normal" images perform well on taken by drone images. Previous works on this stress that the images captured by a drone often are different from those available for training, which are often taken by a hand-held camera. Difficulties in detecting objects in data from a drone may arise due to the positioning [9, 10] of the camera compared to images taken by a human, depending on what type of images is trained on. Additionally, a challenging problem is the strong weight and area constraint of embedded hardware that limits the drones to run with limited hardware resource. Currently there are many computer vision algorithm that are designed using human centric photographs, But from the top view imagery taken vertically the objects of interest are small and fewer features mostly appearing flat and rectangular, certain objects closer to each other can also overlap. So detecting most of the objects from the birds eye view is a challenging task.

In recent years, aerial imaging technology which is being used by UAV has been developing fast. Conventional aerial imaging techniques often fail to process high-resolution aerial images and the efficiency of algorithms is relatively poor. Object detection in aerial images is of vital importance to the subsequent analysis of aerial images. Conventional object detection methods cannot deal with the huge amount of data in the area of aerial image processing. Drone surveillance has higher mobility and large surveillance scope in contrast to fixed cameras. It has imperfections such as unstable background, low resolution, and illumination changes. There is a high demand for intelligent drones in real-world applications.

However, object detection in drone images or videos is different from traditional object detection. The size varies in aerial photos from object instances. Not only because of spatial sensor resolutions but also because of the size differences within the same type of object. Many tiny instances of objects are crowded in aerial photos. Besides, the frequencies of instances are unbalanced in aerial photos, Objects in aerial photographs also appear in arbitrary directions. There are also several instances of the large aspect ratio therefore, both the efficiency and the accuracy are low, and these algorithms cannot meet the requirements of aerial image processing in real applications. Enhanced object detection methods with superior performance are required to replace the conventional methods. Road detection which is an important part of Object detection in UAV has also attracted much attention and extensive research in remote sensing. It is used in many fields such as emergency rescue, autonomous driving, city planning, etc. However, it is difficult for the accurate results because of the complex road scene which includes shadows and occlusion caused by trees, vehicles and buildings.

The typical UAV operations are mapping and surveying, Person tracking, Agriculture, Path planning. The detection of objects from UAV view plays a major role in the surveillance even it can also act as medical aid. The detection of such small images for UAV equipped with vision technologies. Hence, this work aims to improve the object detection in drone images and make it useful for real-time applications by detecting multiple objects in the images and by detecting the roads.

1.1. Overall Objectives

The typical UAV operations are mapping and surveying, Person tracking, Agriculture, Path planning. The detection of objects from UAV view plays a major role in the surveillance even it can also act as medical aid. The detection of such small images for UAV equipped with vision technologies. The issues are

- Detection of Objects from the bird's eye view.
- To classify objects into various categories.
- To detect road in the aerial images.

2. RELATED WORK

The traditional target detection training methods are training the classifier based on the target feature such as scale invariant feature transform (SIFT) [1], histogram of oriented gradient [3] , local binary pattern [4]. Typical targeted acquisition methods have made significant progress in both acquisition accuracy and acquisition speed. However, much remains to be done. For example, the features are hand-made, and this requires researchers to have excellent prior knowledge. In addition, performance is often poor with background or lack of light. Performance is often worse when the structure of the object has a large change, the background is complex, or the light is insufficient.

As the research of feature extraction develop, researchers have found that convolutional neural networks can learn better features from large-scale data and overcome the shortcomings of hand-designed features. In 2014, Girshick et al. [5] proposed a region-based convolutional neural network (Region-based CNN, R- CNN) model, which became representative of target detection based on classification convolutional neural networks. First, the model uses the selective search algorithm to extract several proposal regions from the image. Then, the proposal regions are changed to a uniform size, and the feature is extracted using convolutional neural network. Finally, the features are classified by multiple support vector machines (SVM). In order to improve the speed and accuracy of the R-CNN model, fast R-CNN model is proposed [6]. Compared with the RCNN model which extracts the feature for each candidate region, the Fast R-CNN model only extracts once for the detected image, and then the feature corresponding to the proposal region is mapped to a feature vector with fixed length by spatial pyramid pooling. Both R-CNN and fast RCNN use selective search algorithm to extract proposal region, which takes a lot time. To overcome this issue, the Faster RCNN which replaces the selective search algorithm with region proposal networks (RPN) is proposed [7]. The RPN which connects to the last convolutional layer of Fast R- CNN is used to generated proposal region by predicting object bounds and objectness scores with a series of anchor boxes.

Real time deployments of autonomous drone/UAV in various fields require object detection methods embedded in them should be more robust and accurate so they have deep neural networks [8] to detect and classify the objects in image/frame of video. But this method requires more computational power and appropriate / sufficient amount of training samples to obtain accurate results. The dataset used does not contain images taken in different climatic conditions. Based on faster R-CNN [9] proposes the improved loss function in the positioning, which solves the problem that the L1 and L2 norm function cannot accurately reflect the overlap between the prediction region proposal and the ground truth. The proposed algorithm is better than other RCNN algorithms only on the specified data. The algorithm is an effective way to detect small objects. Faster R-CNN [10] merges RPN with Fast RCNN into a unified detection network by sharing feature maps. RPN is proposed to generate high quality region proposals, which simplifies fixed shape anchors and classifies each into foreground or background. The region proposal process effectively distinguishes the foreground and the background regions effectively and eliminates the interference of complex background information. However it also increases the computational complexity of detection.

There has been extensive work on road detection and classification from Unmanned Ground Vehicles (UGVs) (vanishing point detection [11], superpixel grouping [12], semi-supervised learning [13], [14]). All of them assume that the vehicle is detecting a single road, not multiple intersecting or bifurcating roads. These methods relies on the fact that the longest linear region in the image is the road. Although the algorithm works well on detecting straight roads, it will fail on images that have intersecting or bifurcating roads.

A histogram based thresholding method [15] was developed to select possible road regions from the images. Furthermore a probabilistic line detection algorithm combined with a clustering method was developed to further refine the road region. Mourad Bouziani et al. proposed an object-based classification approach for high resolution remote sensing imagery [16]. For classifying individual pixels, a maximum likelihood method is used initially. Then a multi-spectral segmentation based on classification algorithm is used to classify different objects in the image. Here, the region growing algorithm starts from seed points and groups the adjacent pixels based on homogeneity criterion. To identify the needed pixels, a rule based classification method is applied on the segmentation results. Object-based classification approach produces a better result than that of pixel-based classification. But the maximum likelihood classifier produces an error causing result for some class of values, and also the selection of seed points is sometimes difficult.

On the other hand, Detecting roads and other man-made features is useful in low-altitude imagery for developing geo-referenced mosaics, route planning and emergency management systems. A road detection and tracking system will vastly improve the utility of UAV acquired images sets and make a significant step towards the operational deployment of mini and micro-UAV systems for geographical information systems (GIS) purposes. While road detection has been extensively studied in satellite imagery, large number of images obtained and the limited resolution of onboard IMU and GPS systems make road-detection from UAVs a challenging problem.

Issues

- Performance is often poor with background or lack of light.
- Detecting objects in different scales also particularly for small objects.
- Many approaches were proposed for object detection such as models using RCNN and Fast RCNN but these approaches use selective search algorithm to extract proposal region which takes a lot of time.
- Some methods that use deep neural networks require more computational power and sufficient amount of training samples to obtain accurate results.

To address these problems, we propose a model using Faster R-CNN using Feature Pyramid Network for the detection of multiple objects taken from UAV, also recognizing and detecting roads in aerial images using image processing as there is no annotations for road in the dataset.

3. SYSTEM DESIGN

3.1. System Architecture

Figure 1 describes about the overall system architecture of the proposed multiple object detection using Faster RCNN. The text annotations with aerial images is converted to tf-records for the backbone network and encoding the classes, concatenating all the objects co-ordinates into one file which is used during training purpose. Then backbone network as RESNET is used for

extracting features so the extracted feature are high semantic but low resolution so we enhance it with feature pyramid network for good resolution with high semantic features and pass the feature maps. RPN is to output a set of proposals, each of which has a score of its probability of being an object and also the class/label of the object. RPN can take any sized input to achieve this task. These proposals are further refined by feeding to 2 sibling fully connected layers-one for bounding box regression and the other for box

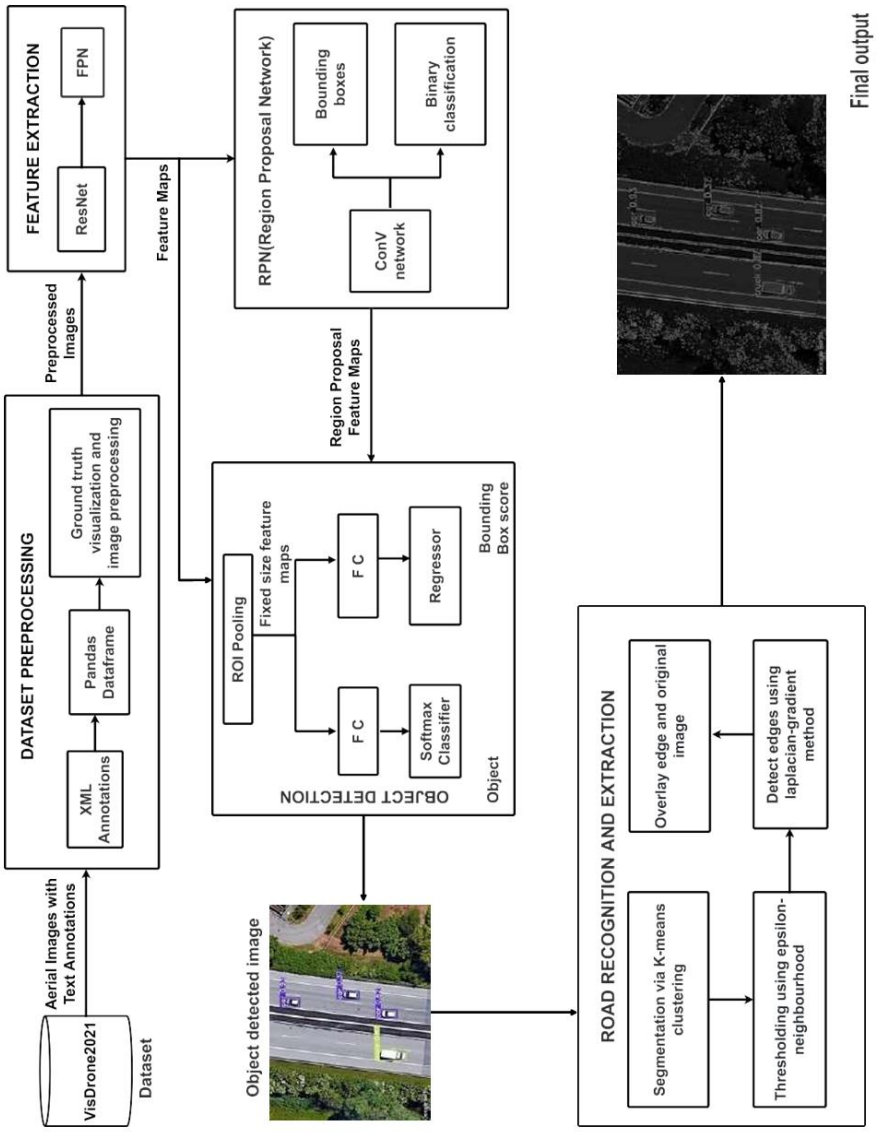

Figure 1. Overall System Architecture of Multiple Object Detection using Faster RCNN

Classification After RPN, we get proposed regions with different sizes. Different sized regions means different sized CNN feature maps. It's not easy to make an efficient structure to work on features with different sizes. Region of Interest Pooling can simplify the problem by reducing the feature maps into the same size and passing them through two individually fully connected layer one for classifying object and the other for bounding box regression. Then for road extraction as no annotations are available, by using image processing techniques and by overlaying them on the images roads are detected.

3.1.1. Data Pre-Processing

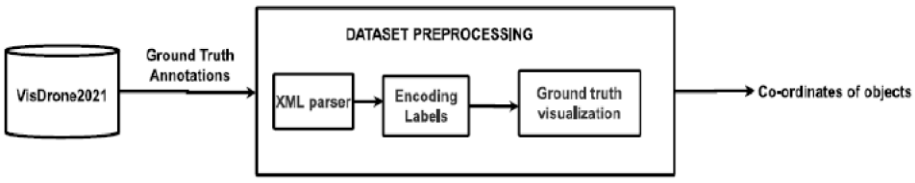

Figure 2. Data Preprocessing

Input : Visdrone dataset images and annotations
Output : Preprocessed images and Annotations

The figure 2 depicts that the pre-processing of the aerial image text annotations, in this module as a preliminary step we convert visdrone Dataset Annotations to XML Annotations. As object detection has developed, different file formats to describe object annotations have emerged. The most common annotation formats have emerged from challenges and amassed datasets so we convert most common formats: VOC XML and convert Xml annotations into pandas data frame for better representation and understanding.

3.1.2. Feature Extraction

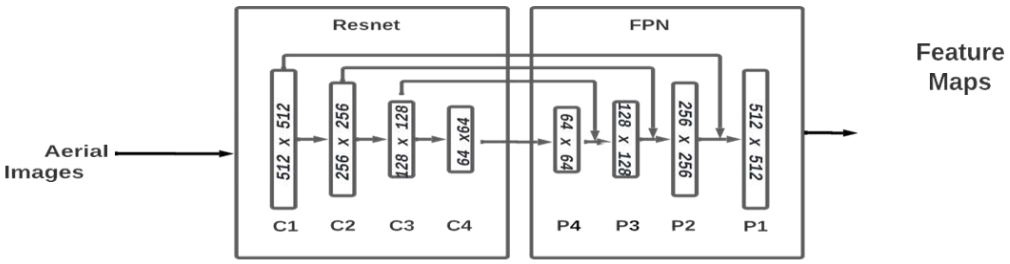

Figure 3. Feature Extraction

Input : Aerial images with ground truth values
Output : Feature maps images

Here figure 3 depicts that feature extraction in Resnet network with enhanced feature pyramid network by upscaling the feature extracted from the Resnet network and adding it to the next layer. Faster RCNN is proposed as one of the state-of-the-art generic object detection architectures. The first stage of Faster R-CNN is feature extraction. The RPN utilizes a convolution layer with kernel of 3*3 to slide on the feature map generated by the shared backbone network and then uses two convolution layers with kernel 1 to slide on 1 to classify the foreground and background. The last stage of Faster RCNN is a classification network.

To improve the performance of the backbone network, the neural network is always deepened or widened. With the increase of hyper-parameters, the ResNet is proposed, which is applied to solve the weakening issue of the very deep network. To improve Faster R-CNN and extract powerful information, we apply the ResNet in our method to improve the detection accuracy.

Considering the high-level semantic information and the low-level location information, the multi scale feature is extracted by constructing feature pyramid structure to make the network

more robust and accurate to the small object detection, especially for the task of aerial image object detection.

3.1.3. Region Proposal Network

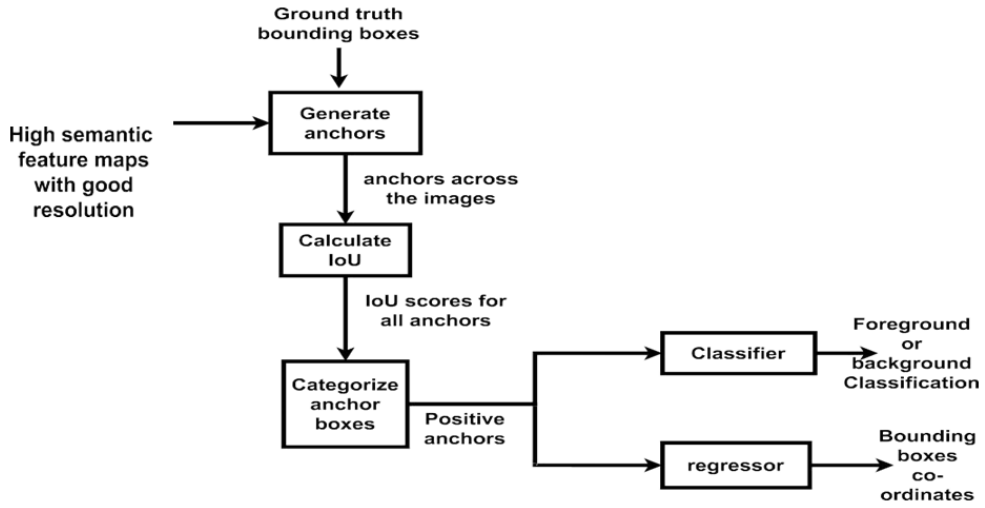

Figure 4. Region Proposal Network

Input : Feature Maps
Output : Region of interest (proposals)

The figure 4 is RPN module which tells the Fast R-CNN where to search for objects. After processing the whole image with several convolutional and max pooling layers the network produces a convolutional feature map. The goal of RPN is to output a set of proposals, each of which has a score of its probability of being an object and also the class/label of the object. RPN can take any sized input to achieve this task. These proposals are further refined by feeding to 2 sibling fully connected layers one for bounding box regression and the other for box classification i.e, is the object foreground or background. The RPN that generates the proposals slides a small network over the output of the last layer of the feature map. This network uses an n*n spatial window as input from the feature map. Each sliding window is mapped to a lower dimensional feature. The position of the sliding window provides localization information with reference to the image while the regression provides finer localization information.

3.1.4. Object Detection

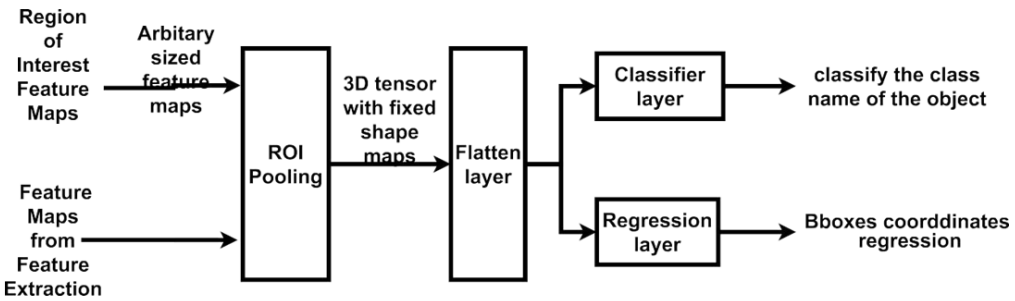

Figure 5. ROI pooling and object detection

We have obtained regions of interest for an arbitrarily-sized input image with different size. We get the proposals, defined in pixel-space coordinates, back to the feature maps. We get them all into a fixed size so they can later be fed into a fully-connected neural network as shown in figure 5. A technique called ROI pooling is used. ROI pooling is used for utilizing a single feature map for all the proposals generated by RPN in a single pass. ROI pooling solves the problem of fixed image size requirement for object detection networks. Now we perform the final classification, so we will first pass our proposals (which are cuts, although resized, of the original feature map) through the block Four of the ResNet.

Also, once we do this, since we've already extracted all the features we care about, we'll perform Global Average Pooling which means, essentially, to average out the spatial information, some feature was present all around. That means we'll be left with a single vector per proposal And, finally, pass this fixed-length vector through two fully-connected layers: one for the bounding box resizing and one for the classes.

3.1.5. Road Recognition and Extraction

Input: Object-classified Images
Output: Road extraction on object-classified images

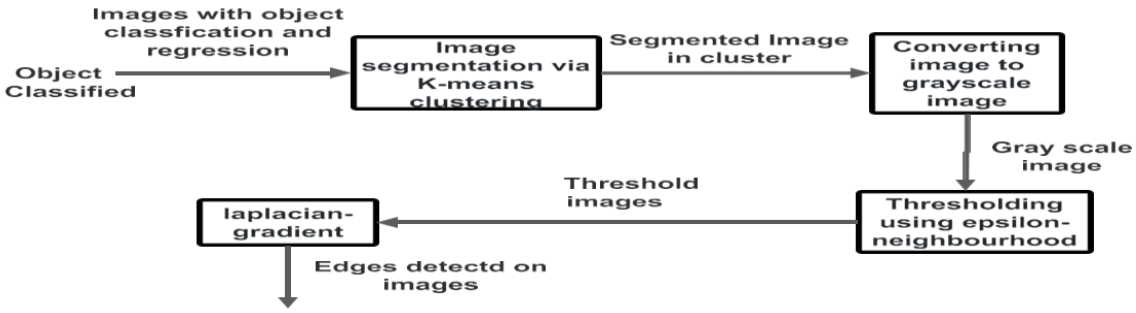

Figure 6. Road Extraction

Figure 6 show the steps involved in in extracting the roads. Here we have images with multi object classification we need to detect roads unfortunately the dataset does not have the annotations for the road. So we use the image processing techniques and extract the road and overlay it on the input image. This is done by first applying Image segmentation via K-means clustering. *K* -means clustering algorithm is an unsupervised algorithm and it is used to segment the interest area from the background. Then next we convert the image into grey scale and thresholding using epsilon-neighborhood. In final part we Detect edges using Laplacian-gradient method. The Laplacian edge detector uses only one kernel. It calculates second order derivatives in a single pass. In the end we overly the results on the input image.

4. IMPLEMENTATION DETAILS

4.1. Environmental Setup

The model was trained and tested on GPU enabled google colab. the following libraries were used-

- Numpy
- OpenCV

- Keras
- Tensorflow
- Pandas

4.2. Data Preprocessing

- It involves converting all the annotations file of text format to Pascal Vic format that is XML file format using a python script as shown in figure 7 below.
- The XML files are then used to convert the data of annotations into pandas data frame where the details of all the annotations are specified along with the path of image and annotations file.
- Removing all unwanted data in the data frame and performing label encoding.
- Ground truth visualization (figure 10) using the data frame obtained from the above steps.

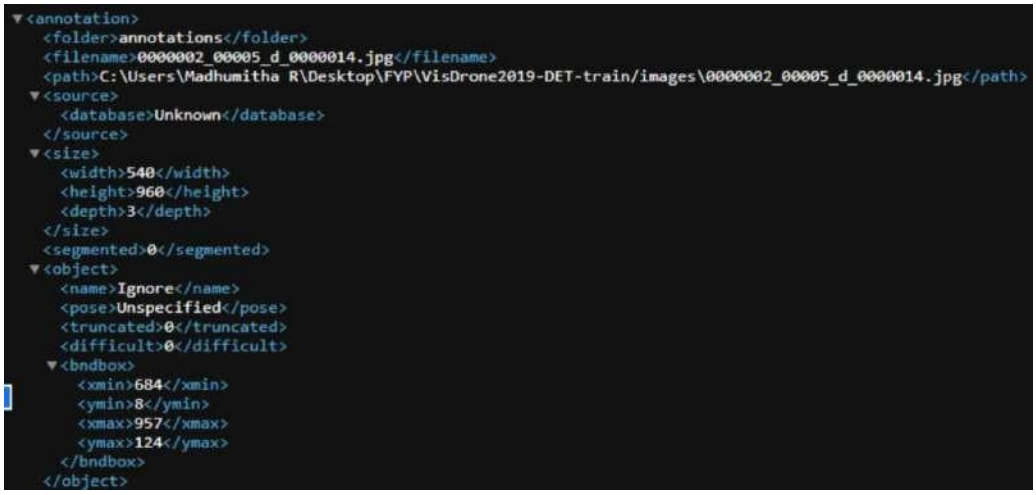

Figure 7. XML annotations

In figure 7 we can see the XML files for each image in the dataset which can be converted from text annotations to XML.

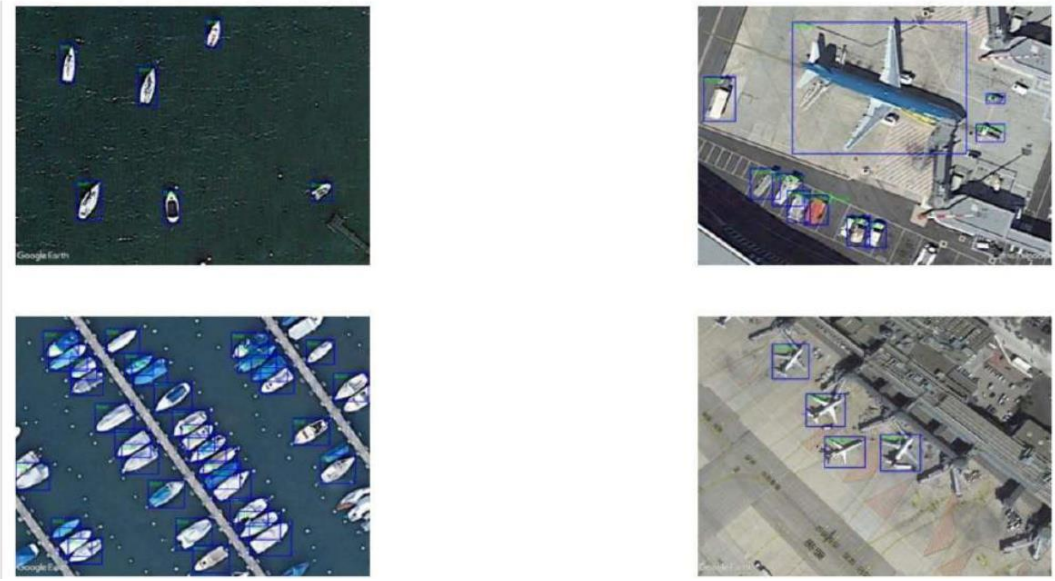

Figure 8. Label encoding

Here figure 8 describes the labelling of the object and overall count of each object in the aerial images training dataset.

```
df_train.head()
```

	xml_path	img_path	img_id	labels	xmin	ymin	xmax	ymax
0	D:\Satellite Images\Annotations\0001.xml	D:\Satellite Images\Images\0001	0001	4	384.0	436.0	424.0	454.0
1	D:\Satellite Images\Annotations\0001.xml	D:\Satellite Images\Images\0001	0001	4	879.0	423.0	922.0	440.0
2	D:\Satellite Images\Annotations\0001.xml	D:\Satellite Images\Images\0001	0001	4	95.0	444.0	142.0	459.0
3	D:\Satellite Images\Annotations\0001.xml	D:\Satellite Images\Images\0001	0001	4	58.0	443.0	95.0	460.0
4	D:\Satellite Images\Annotations\0001.xml	D:\Satellite Images\Images\0001	0001	4	150.0	441.0	194.0	457.0

Figure 9. Dataset in data frames

Figure 9 depicts each object in data frames which contain XML file of a particular image and their corresponding object coordinates.

Fig 10: Ground truth visualization boxes

Here figure 10 describes the ground truth visualization boxes of each object in an aerial image dataset.

4.3. Feature Extraction

For Feature extraction we have used a pretrained resnet50 model on ImageNet dataset out high 50 layers in the model only the last two layers are discarded so that the extracted features are high. The input shape is 224,224,3 the output features are 7,7,2048 as we can see the depth of the features extracted are high.

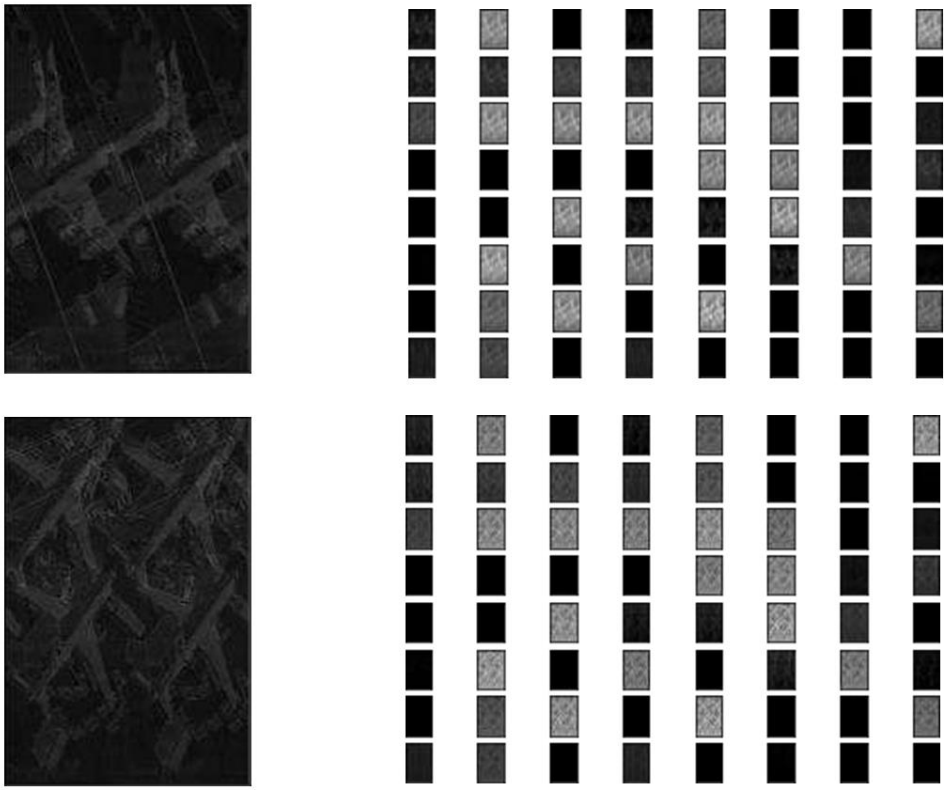

Figure 10. Feature maps with 64 filters applied

Here figure 10 depicts the feature maps that are extracted from the Resnet layer upscaled with FPN in aerial images and applied with filters.

4.4. Region Proposal Networks

- Here we first define all the functions and implement them at the last together.
- The initial status for each anchor is ‘negative’. Then, we set the anchor to positive if the IOU is >0.7.
- If the IOU is >0.3 and <0.7, it is ambiguous and not included in the objective.
- One issue is that the RPN has many more negative than positive regions, so we turn off some of the negative regions.
- We also limit the total number of positive regions and negative regions to
- 256. y_is_box_valid represents if this anchor has an object. y_rpn_overlap represents if this anchor overlaps with the ground-truth bounding box.

Positive anchors and their respective positions:

Figure 11. Positive anchor boxes

In figure 11 we can see the positive anchor boxes that are generated from RPN along with the actual ground truth boxes.

4.5. Roi Pooling and Classification

ROI Pooling layer is the function to process the ROI to a specific size output by max pooling. Every input ROI is divided into some sub-cells, and we applied max pooling to each sub-cell. The number of sub-cells should be the dimension of the output shape. Classifier layer is the final layer of the whole model and just behind the ROI Pooling layer. It's used to predict the class name for each input anchor and the regression of their bounding box.

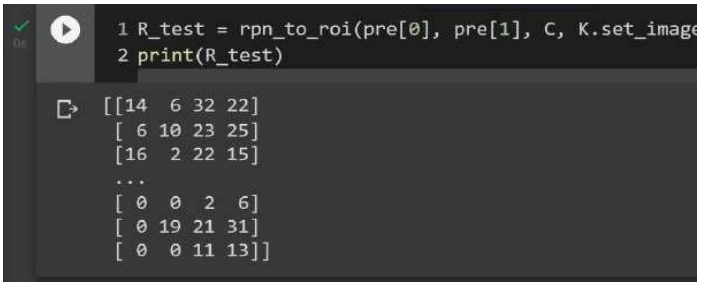

Figure 12. ROI Pooling maps

Figure 12 depicts the equal size feature maps which initially are of different arbitrary size feature maps and in Figure 13 describe the classification of object into respective class along with its bounding box.

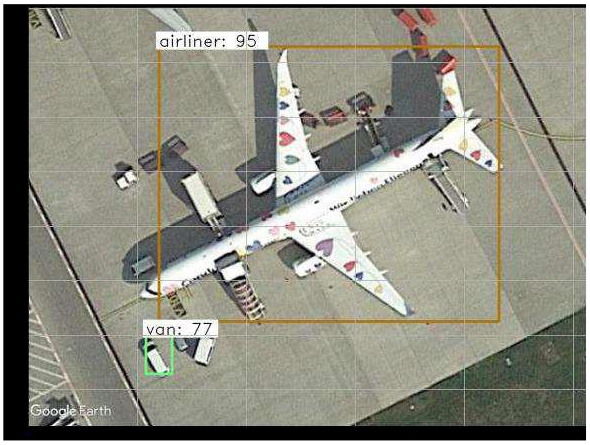

Figure 13. object classification and regression

4.6. Road Recognition and Extraction

Here first we read the object classified input image and then apply image segmentation via K-means clustering. Converting the images to 0 to 1 intensity values and reshape 3 channel image into 2D space the segment image into 4 clusters then we reshape the image to its original shape. And converting the image back to 0-255 intensity values. Now we convert into gray scale image by converting color scale of image and can find the intensity with maximum probability in histogram. Then convert the image into epsilon thresholding and then to edge detection by overlaying the extracted road image over original Black and White image. In figure 14, figure (a) is Gray scale image, figure (b) is Threshold image, figure (c) is edge detection and figure (d) is overlay image which is the output image.

4.7. Dataset

We evaluated our proposed model on VisDrone 2019 data. VisDrone is a large-scale benchmark with carefully annotated ground-truth for various important computer vision tasks, to make vision meet drones. For Object Detection we have used the VisDrone2021 Dataset. The VisDrone2021 dataset is collected by the AISKYEYE team at Lab of Machine Learning and Data Mining, Tianjin University, China.

Figure 14. road extraction

The benchmark dataset consists of **400** video clips formed by **265,228** frames and **10,209** static images, captured by various drone-mounted cameras, covering a wide range of aspects including location (taken from 14 different cities separated by thousands of kilometers in China), environment (urban and country), objects (pedestrian, vehicles, bicycles, etc.), and density (sparse and crowded scenes). Note that the dataset was collected using various drone platforms (i.e., drones with different models), in different scenarios, and under various weather and lighting conditions. These frames are manually annotated with more than **2.6million** bounding boxes or points of targets of frequent interests, such as pedestrians, cars, bicycles, and tricycles. Some important attributes including scene visibility, object class and occlusion, are also provided for better data utilization.

4.8. Performance Metrics

CLASSIFIER LOSS:

The training loss for the RPN is also a multi-task loss, given by:

Here i is the index of the anchor in the mini-batch. The classification loss $L_{cls}(p_i,p_i)$ is the log loss over

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

Two classes (object vs not object). p_i is the output score from the classification branch for anchor i, and p_i is the groundtruth label (1 or 0).

REGRESSION LOSS:

The regression loss $L_{re}(t_i, t_i)$ is activated only if the anchor actually contains an object i.e., the groundtruth p_i is 1. The term t_i is the output prediction of the regression layer and consists of 4 variables $[t_x, t_\gamma, tw, t_h]$. The regression target t_i* is calculated as

$$t_x^* = (x^* - x_a)/w_a, \quad t_y^* = (y^* - y_a)/h_a, \quad t_w^* = \log(w^*/w_a), \quad t_h^* = \log(h^*/h_a)$$

Here x, y, w, and h correspond to the (x, y) coordinates of the box centre and the height h and width w of the box. x_a, x* stand for the coordinates of the anchor box and its corresponding ground truth bounding box.

Table 3. Loss values

Models	Epoch - 1	Epoch - 25	Epoch - 50
classifier Accuracy	0.516	0.541	0.541
Loss RPN classifier	2.645	2.01	2.849
Loss RPN regression	1.114	1.006	0.939
Loss Detector classifier	1.546	1.484	1.534
Loss Detector Regressio n	0.037	0.077	0.067
Total Loss	5.343	4.579	4.827

The evaluation is done using the standard Mean Average Precision (mAP) at some specific IoU threshold (0.7). mAP is a metric that comes from information retrieval, and is commonly used for calculating the error in ranking problems and for evaluating object detection problems.

Table 4. Accuracy Scores

	Validation (AP)	Testing (AP)
Car	**0.84**	**0.80**
Longvehicle	**0.84**	**0.85**
Van	**0.80**	**0.80**
Bus	**0.83**	**0.80**
Stairtruck	**0.83**	**0.81**
Airliner	**0.82**	**0.86**
Pushbacktruck	**0.82**	**0.81**
Truck	**0.87**	**0.87**
Chartered	**0.85**	**0.84**
Propeller	**0.84**	**0.85**
Trainer	**0.80**	**0.84**
Boat	**0.81**	**0.83**
Helicopter	**0.80**	**0.82**
Mean AP	**0.8403**	**0.8400**

Precision is the degree of exactness of the model in identifying only relevant objects. It is the ration of TPs over all detections made by the model.

$$P = \frac{TP}{TP + FP} = \frac{TP}{all\ detections}$$

Recall measures the ability of the model to detect all ground truths— proposition of TPs among all ground truths.

$$R = \frac{TP}{TP + FN} = \frac{TP}{all\ ground\ truths}$$

F1 score is a weighted average of precision and recall. The value ranges from 0 to 1 where 1 means the highest accuracy.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.9. Performance Analysis

There are two loss functions we applied to both the RPN model and Classifier model. As we mentioned before, RPN model has two output. One is for classifying whether it's an object and the other one is for bounding boxes' coordinates regression. From the figure 16 below, we can see that it learned very fast at the first 20 epochs. Then, it became slower for classifier layer while the regression layer still keeps going down. The reason for this might be that the accuracy for objectness is already high for the early stage of our training, but at the same time, the accuracy of bounding boxes' coordinates is still low and needs more time to learn.

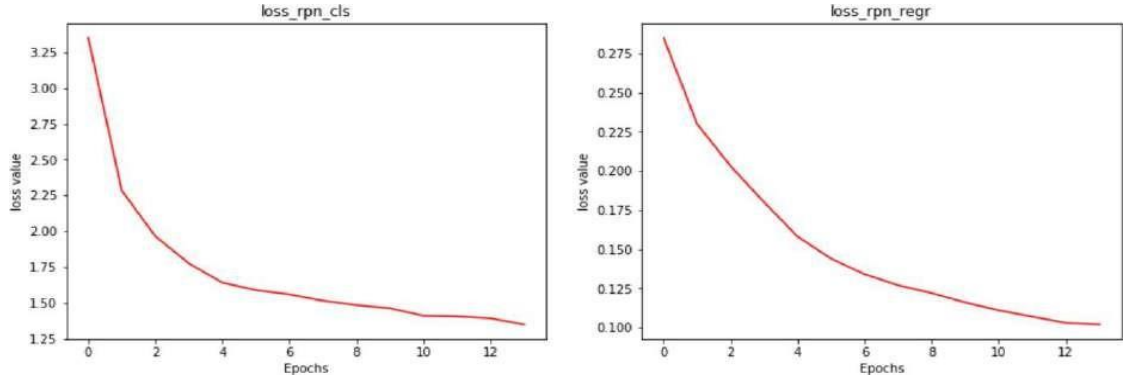

Figure 15. RPN classification and regression loss

*T*he similar learning process is shown in Classifier model, figure 15. Compared with the two plots for bboxes' regression, they show a similar tendency and even similar loss value. I think it's because they are predicting the quite similar value with a little difference of their layer structure. Compared with two plots for classifying, we can see that predicting objectness is easier than predicting the class name of a bbox.

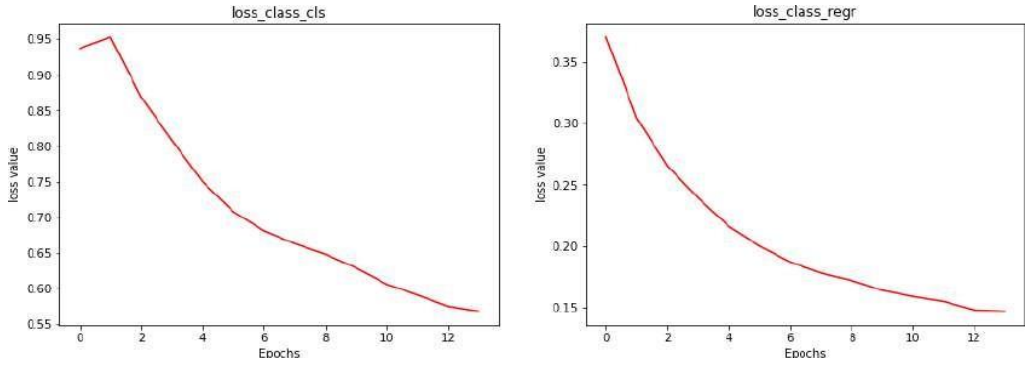

Figure 18. classifier classification and regression loss

This total loss figure 18 is the sum of four losses above. It has a decreasing tendency.

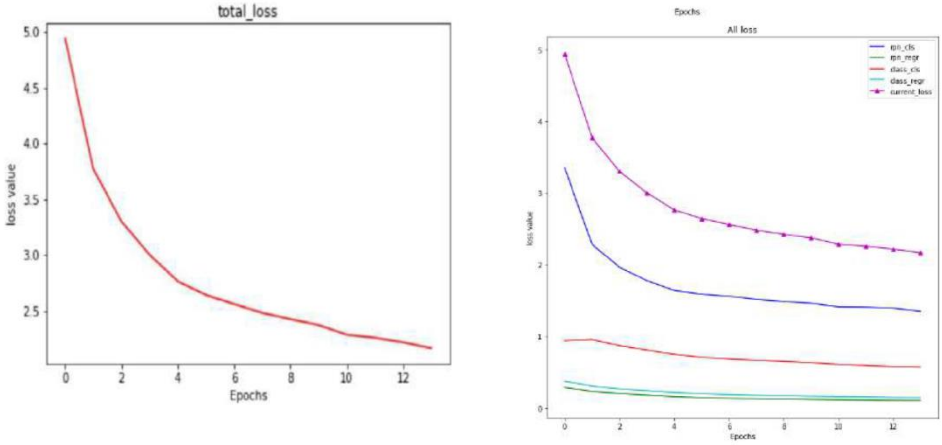

Figure 19. total loss of all 4 classes

4.10. Comparative Analysis

Table 5. Compaision of different state-of- the-art methods

Algorithm	**mAP test**	**mAP validation**
SSD	78.24	82.87
R-CNN	80.46	85.26
Fast R-CNN	81.18	82.23
Faster R-CNN	84.78	86.12
YOLO v3	87.12	85.51
Faster R-CNN with FPN	**84.02**	**84.03**

Table 5 represents the object detection results for the Dota v1.5 and xView 2018 datasets. As previously mentioned, the Dota v1.5 and xView 2018 datasets also contain aerial images, here we have used visdrone images. Taking into account the rate of methods convergence,

generalization ability, and speediness, the best result on detection of the objects with similar patterns has been shown by our model faster rcnn-with fpn.

5. CONCLUSION

Here we have presented multi object detection in aerial images using faster RCNN method and have used RPNs for efficient and accurate region proposal generation. By sharing convolutional features with the down-stream detection network, the region proposal step is nearly cost-free and recognised the roads using image processing and extracted them on the original image.

Faster R-CNN is one of the models that proved that it is possible to solve complex computer vision problems with the same principles that showed such amazing results at the start of this new deep learning revolution. New models are currently being built, not only for object detection, but for semantic segmentation, 3D-object detection, and more, that are based on this original model. Some borrow the RPN, some borrow the R-CNN, others just build on top of both.

Although, the mean average precision result is not high. The foundation is laid for the follow-up to the direction of further research. Increasing the performance and even moving the trained model on Nvidia Jetson TX2 is a direction for further research. Also a annotated dataset which also includes road. Here we also presented an approach for road recognition from aerial images. The proposed approach has a minimum of manually selectable parameters, so the process of road detection is automatically at all stages. Our approach is by using few classical image processing. Experimental results recognise the road but it needs to be improved.

6. FUTURE WORK

Object detection is breaking into a wide range of industries, with use cases ranging from personal security to productivity in the workplace. Object detection and recognition is applied in many areas of computer vision, including image retrieval, security, surveillance, automated vehicle systems and machine inspection. Significant challenges stay on the field of object recognition. The possibilities are endless when it comes to future use cases for object detection. Some of them are, automated cctv Surveillance is an integral part of security and patrol. Recent advances in computer vision technology have lead to the development of various automatic surveillance systems, however their effectiveness is adversely affected by many factors and they are not completely reliable. This study investigated the potential of automated surveillance system to reduce the CCTV operator workload in both detection and tracking activities, Object counting in object detection system can also be used for counting the number of objects in the image or real time video. The Object Detection and Recognition system In Images is web based application which mainly aims to detect the multiple objects from various types of images. It also recognizes the images after performing the detection. Apart from these object detection can be used for classifying images found online. Obscene images are usually filtered out using object detection and for road recognition in future research, we would like also to use for segmentation additional features such as texture and shape to separate roads from buildings cars and other objects in aerial images.

REFERENCES

[1] Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), pp.91-110.

[2] Bay, H., Tuytelaars, T. and Gool, L.V., 2006, May. Surf: Speeded up robust features. In European conference on computer vision (pp. 404-417). Springer, Berlin, Heidelberg.

[3] Dalal, N. and Triggs, B., 2005, June. Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE.

[4] Ojala, T., Pietikainen, M. and Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, 24(7), pp.971-987.

[5] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[6] Girshick, R., 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[7] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards realtime object detection with region proposal networks. Advances in neural information processing systems, 28.

[8] Subash, K.V.V., Srinu, M.V., Siddhartha, M.R.V., Harsha, N.S. and Akkala, P., 2020, March. Object detection using Ryze Tello drone with help of mask- RCNN. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 484-490). IEEE.

[9] Cao, C., Wang, B., Zhang, W., Zeng, X., Yan, X., Feng, Z., Liu, Y. and Wu, Z., 2019. An improved faster R-CNN for small object detection. IEEE Access, 7, pp.106838-106846.

[10] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards realtime object detection with region proposal networks. Advances in neural information processing systems, 28.

[11] Kong, H., Audibert, J.Y. and Ponce, J., 2009, June. Vanishing point detection for road detection. In 2009 ieee conference on computer vision and pattern recognition (pp. 96-103). IEEE.

[12] Rasmussen, C. and Scott, D., 2008, September. Shape-guided superpixel grouping for trail detection and tracking. In 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 4092-4097). IEEE.

[13] Dahlkamp, H., Kaehler, A., Stavens, D., Thrun, S. and Bradski, G.R., 2006, August. Selfsupervised monocular road detection in desert terrain. In Robotics: science and systems (Vol. 38).

[14] Lieb, D., Lookingbill, A. and Thrun, S., 2005, June. Adaptive Road Following using SelfSupervised Learning and Reverse Optical Flow. In Robotics: science and systems (Vol. 1, pp. 273280).

[15] Lin, Y. and Saripalli, S., 2012, May. Road detection from aerial imagery. In 2012 IEEE International Conference on Robotics and Automation (pp. 3588- 3593). IEEE.

[16] Bouziani, M., Goita, K. and He, D.C., 2010. Rule-based classification of a very high resolution image in an urban environment using multispectral segmentation guided by cartographic data. IEEE Transactions on Geoscience and Remote Sensing, 48(8), pp.3198-3211.

[17] Rao, Y., Liu, W., Pu, J., Deng, J. and Wang, Q., 2018, December. Roads detection of aerial image with FCN-CRF model. In 2018 IEEE Visual Communications and Image Processing (VCIP) (pp. 1-4). IEEE.