Vehicle Detection and Classification in Aerial Images using Convolutional Neural Networks

Chih-Yi Li[1] and Huei-Yung Lin[2]
[1]*Department of Electrical Engineering, National Chung Cheng University, Chiayi 621, Taiwan*
[2]*Department of Electrical Engineering and Advanced Institute of Manufacturing with High-tech Innovations, National Chung Cheng University, Chiayi 621, Taiwan*

Keywords: Aerial Image, Convolutional Neural Network, Vehicle Detection.

Abstract: Due to the popularity of unmanned aerial vehicles, the acquisition of aerial images has become widely available. The aerial images have been used in many applications such as the investigation of roads, buildings, agriculture distribution, and land utilization, etc. In this paper, we propose a technique for vehicle detection and classification from aerial images based on the modification of Faster R-CNN framework. A new dataset for vehicle detection, VAID (Vehicle Aerial Imaging from Drone), is also introduced for public use. The images in the dataset are annotated with 7 common vehicle categories, including sedan, minibus, truck, pickup truck, bus, cement truck and trailer, for network training and testing. We compare the results of vehicle detection in aerial images with widely used network architectures and training datasets. The experiments demonstrate that the proposed method and dataset can achieve high vehicle detection and classification rates under various road and traffic conditions.

1 INTRODUCTION

In recent years, due to the popularity of unmanned aerial vehicles (UAVs), the acquisition of aerial images has become more convenient. A large volume of aerial images can be obtained very quickly. The use of big data is a trend of the future research related to aerial image analysis. The techniques for aerial images has been adopted in many applications such as the investigation of roads, buildings, agriculture distribution, and land utilization, etc. One specific importance is the detection of vehicles from aerial imaging. This type of vehicle detection is suited for transportation related work including traffic monitoring, vehicle identification and tracking, parking analysis and planning. At the same time, the technology maturity of UAV with the characteristics of being lightweight, inexpensive and flexible makes the aerial photography easily apply to the traffic data collection and emergency response. Thus, it will be helpful and efficient in the applications if the analysis of aerial images can be accelerated.

The content of aerial images generally covers a large number of objects, including trees, lands, roads and buildings, etc. In the early research, satellite imagery was used to analyze the landscape, the distribution of forest, land usage, river and road areas. It has also been used for the detection of special buildings or large venues in the past few decades, especially for the military purposes such as aircraft and runway detection. Due to the advances of deep learning techniques in recent years, object detection can be achieved under complex backgrounds and a variety of application scenarios. Thus, it becomes more feasible to use aerial images for the detection and classification of vehicles.

The methods for vehicle detection in aerial images are generally divided into two categories, the traditional approaches and the machine learning techniques (Cheng and Han, 2016). In traditional approaches, the feature extraction is one important step in object detection, which consists of the use of texture, shape, color, and spatial information. The work presented by Kembhavi *et al.* relies on three features, histogram of oriented gradient (HoG), color probability maps (CPM) and pairs of pixels (PoP), to solve the regression problem using partial least squares (PLS) (Kembhavi et al., 2011). Lenhart *et al.* use the difference in color channels to detect vehicles with more significant color features and the grayscale images to extract the blob-like spots (Lenhart et al., 2008). In (Shao et al., 2012), a vehicle detection framework which combines different features including HoG and local binary patterns (LBP) is proposed. Furthermore,

Li, C. and Lin, H.
Vehicle Detection and Classification in Aerial Images using Convolutional Neural Networks.
DOI: 10.5220/0008941707750782
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 775-782
ISBN: 978-989-758-402-2; ISSN: 2184-4321

many traditional techniques such as frame difference and optical flow (Yalcin et al., 2005) are used to detect moving vehicles.

In recent years, the convolutional neural networks (CNNs) have achieved good results in target detection and classification. Among them, Tang *et al.* use DLR Vehicle Aerial dataset and Vehicle Detection in Aerial Imagery (VEDAI) dataset to train the Single Shot MultiBox Detector (SSD) (Liu et al., 2016), and adopt the model as the backbone to perform network tuning (Tang et al., 2017). Sommer *et al.* evaluate the performance comparison between Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren et al., 2017) network architectures using public aerial image datasets (Sommer et al., 2017). Deng *et al.* propose a fast vehicle detection system based on R-CNN. It combines AVPN (Accurate Vehicle Proposal Network) and VALN (Vehicle Attributes Learning Network), and provides the results superior to Fast R-CNN (Deng et al., 2017). Lu *et al.* analyze the differences among YOLO (Redmon et al., 2016), YOLOv2 (Redmon and Farhadi, 2017) and YOLOv3 (Redmon and Farhadi, 2018) networks for vehicle detection using VEDAI, COWC and DOTA datasets (Xia et al., 2018) for training and testing (Lu et al., 2018). Similarly, Benjdira *et al.* present a small public vehicle dataset to compare the pros and cons of Faster R-CNN and YOLOv3 for vehicle detection (Benjdira et al., 2019).

Compared to the general object detection and recognition research and applications, the datasets for vehicle detection in aerial imagery are fairly limited. The VEDAI dataset used in this work is made available by Razakarivony and Jurie (Razakarivony and Jurie, 2016), and originated from the public Utah AGRC database. VEDAI contains a total of 1,250 images, and is manually annotated with nine classes of objects (plane, boat, camping car, car, pickup truck, tractor, truck, van, and others) and a total of 2,950 samples. The annotation of each sample includes the sample class, the center point coordinates, direction and the four corner point coordinates of the groundtruth. However, the targets in VEDAI are relatively easy to identify. Most of the vehicles in the images are sparsely distributed with simple backgrounds, and the vehicles in the densely distributed places such as parking lots are excluded.

Compared to the object detection in ground view images, vehicle detection in aerial images has several challenges. The targets are usually much smaller with monotonic appearance, and easily affected by the illumination changes. In this paper, we propose a technique for vehicle detection and classification in aerial images based on a modified Faster R-CNN framework. We compare the advantages, disadvantages and results of vehicle detection in aerial images with widely used network architectures. A new aerial image dataset for vehicle detection is introduced with the annotation of 7 common vehicle categories including sedan, minibus, truck, pickup truck, bus, cement truck and trailer. Figure 1 shows the schematic diagram of the proposed method for vehicle detection and classification in aerial images. It consists of the creation of our own aerial image dataset and the development of network architecture for vehicle detection.

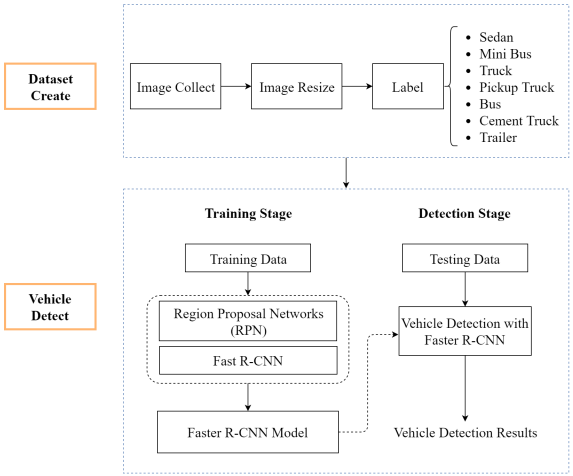


Figure 1: A schematic diagram of the proposed method for vehicle detection and classification in aerial images. It consists of the creation of our own aerial image dataset and the development of network architecture for vehicle detection.

2 VAID DATASET

As mentioned previously, there are not many public datasets available specifically for vehicle detection in aerial images. Even for the existing datasets, only a very limited number of places and traffic scenes are covered. Several popular datasets for vehicle detection in aerial images include VEDIA, COWC, DLR-MVDA and KIT AIS. The description of these datasets are shown in Table 1. This paper introduces a new vehicle detection dataset, VAID (Vehicle Aerial Imaging from Drone), with the aerial images captured by a drone.[1] We collect about 6,000 aerial images under different illumination conditions, viewing angles from different places in Taiwan. The images are taken with the resolution of 1137×640 pixels in JPG format. Our VAID dataset contains seven classes of vehicles, namely 'sedan', 'minibus', 'truck', 'pickup truck', 'bus', 'cement truck' and 'trailer'. Figure 2

[1] VAID Dataset: http://vision.ee.ccu.edu.tw/aerialimage/

Table 1: Summary of the existing datasets for vehicle detection in aerial images.

Database	Image	Image Size	Resolution	Vehicle Size
VEDIA	1250	512×512	25cm	10×20
		1024×1024	12.5cm	20×40
COWC	53	2000×2000, 19000×19000	15cm	24×48
DLR-MVDA	20	5616×3744	13cm	20×40
KIT AIS Dataset	241	300-1800	12.5cm-18cm	15×25, 20×40

Figure 2: The common vehicles are classified to 7 categories, namely (a) sedan, (b) minibus, (c) truck, (d) pickup truck, (e) bus, (f) cement truck and (g) trailer, from the left to the right. The sample images are shown in the figure.

shows several cropped vehicle images from different categories.

Although the vehicles are divided into the above seven categories according to the popularity in Taiwan's road scenes, it is sometimes very tricky to annotate. The characteristics of small sedans viewing from the above are less obvious, and the types are more diverse, including two-door and four-door sedans, five-door hatchbacks, recreational vehicles and nine-seat vans. There are a few differences in the definition of a truck and a pickup truck for annotation. A truck is defined as a vehicle with a shelter in the cargo area or a vehicle with its own cargo area as a container, and the body and the front of the vehicle are completely disconnected. However, a pickup truck is not covered by the canopy. A minibus is a 21-seat medium size bus, while a bus includes passenger and big buses. The trailer category includes tank trucks, gravel trucks, tow trucks, container trucks with detachable tailgates. The images in the dataset are annotated using the labeling tool LabelImg in the format of PASCAL VOC, including the names of the classes and the bounding box coordinates.

The images in the dataset are taken by a drone (DJI's Mavic Pro). To keep the sizes of the vehicles consistent in all images, the altitude of the drone is maintained at about 90 – 95 meters from the ground during video recording. The output resolution is 2720×1530 at 2.7K and the frame rate is about 23.98 fps. For an average sedan with the length of 5 meters and the width of 2.6 meters, the apparent size in the image is about 110×45 pixels. In the VAID dataset, the images are scaled to the resolution of 1137×640, and a sedan in the images is about the size of 40×20 pixels.

The dataset covers ten geographic locations in southern Taiwan, and contains various traffic and road conditions. The images are taken on the sunny days when the light is sufficient, the interference caused by the shadow of the house in the afternoon, and the darker imaging condition in the evening. Figure 3 shows some of the dataset images with various road and traffic scenes.

3 PROPOSED FRAMEWORK

An overview of the proposed framework is illustrated in Figure 4. It is modified based on Faster R-CNN (Ren et al., 2017) and uses ResNet as the backbone structure for feature learning. This provides high efficiency, robustness and effectiveness during training.

3.1 Feature Learning

In the original Faster R-CNN architecture, the authors use VGG16 and ZF Net as the feature extraction networks. There exist many network architectures, such as AlexNet, ResNet, and Inception, which also provide good results in feature extraction. ResNet is one popular feature extraction network. On the evaluation using PASCAL VOC 2007, the mAP is increased from 73.2% to 76.4% if VGG16 is replaced by ResNet101. The mAP is also increased from 70.4% to 73.8% on PASCAL VOC 2012. The experiments in (Ren et al., 2018) show that, among the feature extraction networks for Faster R-CNN, the results obtained by ResNet50 are better than VGG and Inception.

3.2 Region Proposal Network

In the feature extraction process of Faster R-CNN, the RPN shares the convolutional layers, which can reduce a large amount of cost compared to the selective search method. When using the sliding window, the RPN generates multiple anchor boxes for matching,

Figure 3: Some aerial images in our VAID dataset captured using a drone. It consists of different road and traffic scenes.

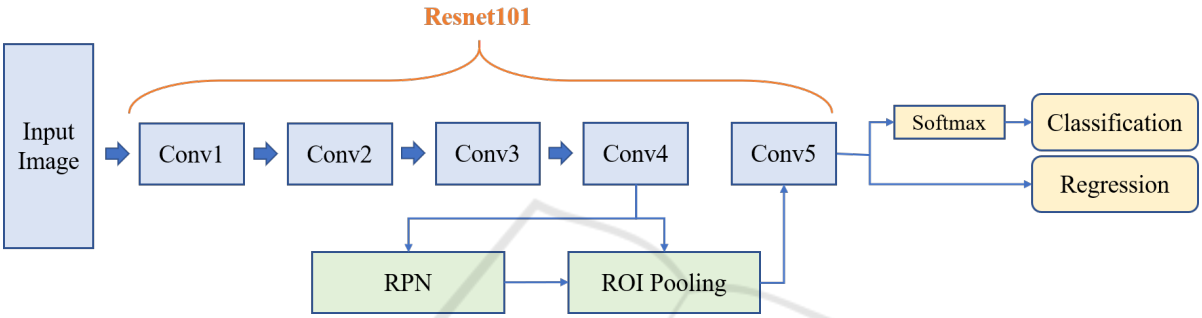


Figure 4: The proposed modified Faster R-CNN architecture for vehicle detection and classification.

which are used to effectively predict the aspect ratio and scale settings of larger objects in PASCAL VOC. Compared to the general object detection, the vehicles in aerial images are clearly smaller than the common objects of interest. Thus, if the original magnification of the anchor box is used, the target might be too small to be detected.

In the original implementation, the anchor boxes are built on multiple scales and aspect ratios. The three aspect ratios used are 0.5, 1 and 2, and the three bounding box scales are 128^2, 256^2 and 512^2. The aspect ratios are changed to [0.2, 0.5, 1, 1.2, 2] to handle to the small target size problem while maintaining the recognition rate of large vehicles.

In the early work, the most commonly used activation function for DNN is the sigmoid function. To solve the deficiencies of the sigmoid function, Rectified Linear Unit (ReLU) is increasingly used recently. It is able to converge faster, and the computation is less. The vanishing gradient problem can be effectively solved with the observation whether the input is greater than zero. ReLU gives part of the output zero, which makes the neural network sparse and reduces the overfitting problem. Faster R-CNN has two activation functions in the RPN, one for the classification and the other for the bounding box prediction. In the experiments, it is found that the accuracy can be improved if the activation function of the classification prediction is changed.

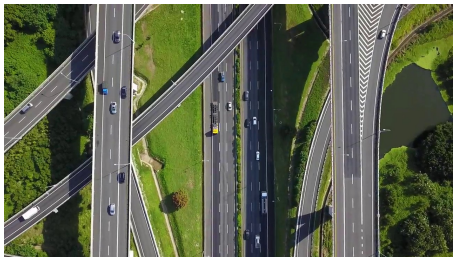
4 EXPERIMENTAL RESULTS

In the experiments, we introduce the VAID dataset presented in this paper, and compare it with the existing aerial image datasets on the processes and results of network training for vehicle detection and classification.

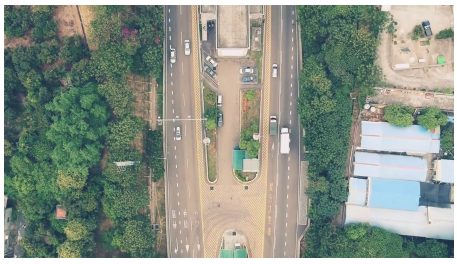
4.1 Dataset Comparison

We evaluate our method on the public VEDAI dataset and our VAID dataset, and test with the aerial images acquired from different places. The VEDAI dataset is the most common dataset for vehicle detection in aerial image. To focus on the targets for vehicle detection, the training data of the VEDAI dataset are processed as follows. (a) The objects in the categories with the labels ‘Boat’, ‘Plane’ and ‘Others’ are removed. (b) The remaining vehicles are labeled by their types to two categories. The first category contains the general cars, which is labeled as ‘Car’ in the original dataset. The second category consists of the combination of other types of vehicles which

(a) Scene A: A sedan is about 20×40 pixels.



(b) Scene B: A sedan is about 10×20 pixels.



(c) Scene C: A sedan is about 20×45 pixels.

(d) Scene D: A sedan is about 40×80 pixels.

Figure 5: The image data used for testing. (a) Scene A consists of the images recorded from two locations in a city. (b) Scene B contains a YouTube video recorded with a highway. (c) Scene C contains a YouTube video recorded with an expressway. (d) Scene D is a YouTube video recorded with a crossroad in Belarus.

includes 'Camping Car', 'Pickup Truck', 'Tractor', 'Truck' and 'Van'.

To compare with the VEDAI dataset, the original seven vehicle categories in our dataset are also adjusted to two categories. The first category is the general car, which is the vehicle classification of the original label 'Sedan', and the second category is the other six original categories of vehicles labeled as 'Minibus', 'Truck', 'Pickup Truck', 'Bus', 'Cement truck' and 'Trailer'. Our testing data are selected from four different scenes and image acquisition scenarios. Scene A consists of the self-recorded aerial images from two locations in a city. Scenes B – D are the videos obtained from YouTube, which record two highways and one expressway in Taiwan, and a crossroad in Belarus. The details of the testing data are shown in Figure 5 and tabulated in Table 2.

Table 3 and Figure 6 show the comparison results of using VEDAI and VAID as the training datasets. In the four different scenarios, training with the VEDAI dataset is worse than using our VAID dataset, especially for Scenes A and D in the testing environment. In VEDAI images, the vehicle only occupies a small region, so the classification error is high when the vehicle size is relatively large in the testing images such as Scene D. Scenes A – C are the road data acquired in Taiwan, and the vehicles such as trucks and trailers are rare in the VEDAI dataset. This also causes the classification problem for certain types of vehicles. Using our VAID dataset for training, high accuracy results are obtained for Scenes A, C and D. Especially, note that Scene D is the road data acquired in Belarus with the captured vehicle size larger than those in our training set. The low precision results of Scene B are mainly caused by the vehicle size in the scene (about 20×10) much smaller compared to the size of about 40×20 used for training.

4.2 Detection and Classification Comparison

In our VAID dataset, there are totally 5,937 aerial images with vehicles classified into 7 categories. It includes 4,456 images for training, 502 images for verification, and 979 images for testing. The detailed information of each class and the number of images for training, validation and testing are shown in Table 4. It can be seen that the number of training samples is not balanced for different classes, so it is very important to achieve better results with fewer samples when training the classification network.

To test the effect of different activation functions, we evaluate the network using softplus, ELU and ReLU, and the results are tabulated in Table 5. Among these activation functions, softplus is the worst, while ELU and ReLU converge faster than the original architecture and the mAP is also improved by about 0.4%. Thus, ReLU is adopted as the activation function of RPN for Faster R-CNN classification.

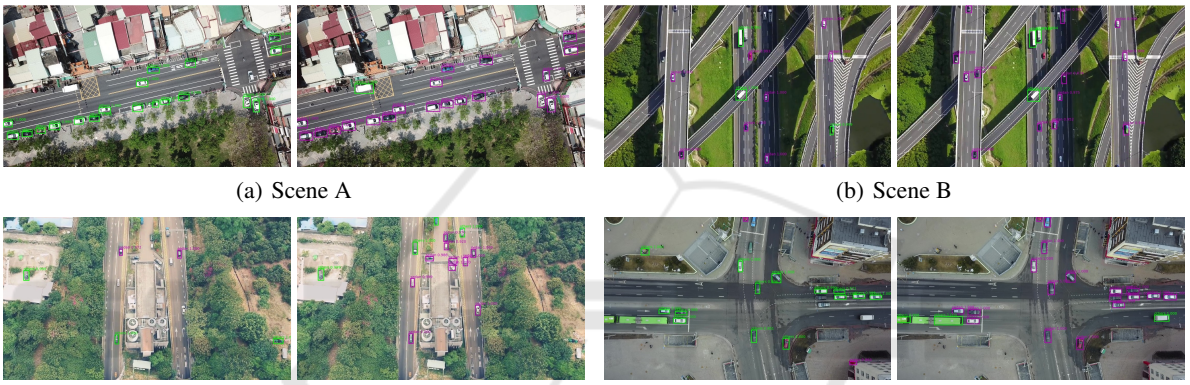
In this paper, Faster R-CNN with ResNet101 as

Table 2: The description of the training and testing sets.

	Training Set		Testing Set	
	Number of image	Image size	Number of image	Image size
VEDAI Dataset	1211	1024 × 1024	A: 99	1137×640
			B. 17	1280 ×720
VAID Dataset	4456	1137×640	C: 31	1280×720
			D: 35	1920×1080

Table 3: The result comparison (mAP) of using the VEDAI dataset and our VAID dataset for network training.

Training Data	VEDAI Dataset			VAID Dataset		
Class	Sedan	Others	mAP	Sedan	Others	mAP
A	18.8%	0%	9.5%	90.4%	62.6%	76.5%
B	59.4%	35.8%	45.4%	73%	55.2%	64.1%
C	31.1%	23.3%	27.2%	89.1%	86.8%	87.9%
D	12.6%	0%	6.7%	89.1%	90.7%	89.9%



(a) Scene A (b) Scene B

(c) Scene C (d) Scene D

Figure 6: The results of Scenes A – D using the proposed vehicle detection approach. Left: Training with the VEDAI dataset. Right: Training with our VAID dataset.

the feature extraction network is modified. It includes the data enhancement of the training images, the adjustment of the anchor box size, the change of the test image input size, and the modification of the ResNet101 activation function. Table 6 shows the results of the original Faster R-CNN, our modified Faster R-CNN, and the modification with the ReLU activation function for RPN. Using the same training data, the improvement with the pre-adjustment is about 1.5% in mAP, and the result of each category is also slightly improved. Figure 7 shows the vehicle detection and classification results in various common road scenes in Taiwan. The bounding boxes with different colors are used to represent different types of vehicles. Successful detection and classification with all kinds of vehicles are shown in the figures.

5 CONCLUSIONS

In this paper, we propose a technique for vehicle detection and classification from aerial images based on the modification of Faster R-CNN framework. A new dataset with images collected from a drone is further introduced for vehicle detection and available for public access. We compare the results of vehicle detection in aerial images with widely used network architectures and training datasets. The experiments demonstrate that the proposed method and dataset can achieve high vehicle detection and classification rates under various road and traffic conditions. In the future work, the vehicle detection will be incorporated with traffic management system for parking lot and traffic flow monitoring.

Table 4: The detailed information of each class and the number of images used for training, validation and testing.

Class Name	Sedan	Minibus	Truck	Pickup Truck	Bus	Cement Truck	Trailer
Training	29774	392	2382	2411	447	144	595
Validation	3483	39	267	266	49	21	62
Testing	6774	70	531	322	84	22	56

Table 5: Different activation functions are used in Faster R-CNN for comparison (mAP).

	Modified Faster R-CNN	Modified Faster R-CNN (softplus)	Modified Faster R-CNN (ELU)	Modified Faster R-CNN (ReLU)
Sedan	90.2%	90.2%	90.2%	90.2%
Minibus	97.9%	92.2%	96.6%	95.6%
Truck	84.9%	87.5%	87.1%	86.8%
Pickup Truck	78.7%	79.3%	78.6%	79.6%
Bus	89.4%	89.6%	89.8%	90.3%
Cement Truck	97.6%	93.8%	97.6%	98.1%
Trailer	83.8%	81.5%	84.7%	84.5%
Avg.	88.9%	87.7%	89.2%	89.3%

Table 6: The results from our modified Faster R-CNN (mAP).

	Original Faster R-CNN	Modified Faster R-CNN	Modified Faster R-CNN (ReLU)
Sedan	90.0%	90.2%	90.2%
Minibus	95.0%	97.9%	95.6%
Truck	83.4%	84.9%	86.8%
Pickup Truck	76.8%	78.7%	79.6%
Bus	88.9%	89.4%	90.3%
Cement Truck	94.2%	97.6%	98.1%
Trailer	85.6%	83.8%	84.5%
Avg.	87.7%	88.9%	89.3%

Figure 7: The results from our dataset and the modified Faster R-CNN with ReLU.

ACKNOWLEDGMENTS

The support of this work in part by the Ministry of Science and Technology of Taiwan under Grant MOST 106-2221-E-194-004 and the Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan is gratefully acknowledged.

REFERENCES

Benjdira, B., Khursheed, T., Koubaa, A., Ammar, A., and Ouni, K. (2019). Car Detection using Unmanned Aerial Vehicles: Comparison between Faster R-CNN and YOLOv3Benjdira, B., Khursheed, T., Koubaa, A., Ammar, A., &Ouni, K. (2019). Car Detection using Unmanned Aerial Vehicles: Comparison between Faster R-CNN and YOLOv3. 2019 1s. In *2019 1st International Conference on Unmanned Vehicle Systems-Oman, UVS 2019*.

Cheng, G. and Han, J. (2016). A survey on object detection in optical remote sensing images.

Deng, Z., Sun, H., Zhou, S., Zhao, J., and Zou, H. (2017). Toward Fast and Accurate Vehicle Detection in Aerial Images Using Coupled Region-Based Convolutional Neural Networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*.

Kembhavi, A., Harwood, D., and Davis, L. S. (2011). Vehicle Detection using Partial Least Squares. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 33(6):1250–1265.

Lenhart, D., Hinz, S., Leitloff, J., and Stilla, U. (2008). Automatic traffic monitoring based on aerial image sequences. *Pattern Recognition and Image Analysis*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C. (2016). SSD: Single shot multibox detector. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Lu, J., Ma, C., Li, L., Xing, X., Zhang, Y., Wang, Z., and Xu, J. (2018). A Vehicle Detection Method for Aerial Image Based on YOLO. *Journal of Computer and Communications*.

Razakarivony, S. and Jurie, F. (2016). Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Redmon, J. and Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*.

Redmon, J. and Farhadi, A. (2018). YOLO v.3. *Tech report*.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ren, Y., Zhu, C., and Xiao, S. (2018). Object Detection Based on Fast/Faster RCNN Employing Fully Convolutional Architectures. *Mathematical Problems in Engineering*.

Shao, W., Yang, W., Liu, G., and Liu, J. (2012). Car detection from high-resolution aerial imagery using multiple features. In *International Geoscience and Remote Sensing Symposium (IGARSS)*.

Sommer, L. W., Schuchert, T., and Beyerer, J. (2017). Fast deep vehicle detection in aerial images. In *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*.

Tang, T., Zhou, S., Deng, Z., Lei, L., and Zou, H. (2017). Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks. *Remote Sensing*.

Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. (2018). DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Yalcin, H., Hebert, M., Collins, R., and Black, M. J. (2005). A flow-based approach to vehicle detection and background mosaicking in airborne video. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*.