VEHICLE DETECTION IN HIGH RESOLUTION IMAGE BASED ON DEEP LEARNING

Haobo Gao [1, *], Xin Li [1]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 430079, China, gaohb_mail@163.com, xli2126@whu.edu.cn

KEY WORDS: Deep Learning, Vehicle Detection, SSD, High Resolution, Convolutional Neural Network

ABSTRACT:

Despite its high accuracy and fast speed in object detection, Single Shot Multi-Box Detector (SSD) tends to get undesirable results especially for small targets such as vehicles on high-resolution images. In this paper, we propose a new convolutional neural network based on SSD to detect vehicles on high-resolution images. In the proposed framework, the feature fusion module and detection module are incorporated. In the feature fusion module, feature maps of different scales are integrated into a fusion feature for object detection, which could improve the accuracy effectively. Besides, to prevent the network from overfitting and speed up the training, the batch normalization layer is embedded between the detection layers in the detection module. Some ablation experiments provide strong evidence for the effectiveness of these above structures. On the UCAS-High Resolution Aerial Object Detection Dataset, our network has the ability to achieve the 0.904 AP (average precision) with 0.094 AP higher than SSD512 but similar speed to it.

1. INTRODUCTION

With the development of society and economy, the number of vehicles is constantly increasing. Monitoring traffic conditions enables the related transportation department to better control traffic and plan road. Accurate monitoring can avoid traffic accidents and ease traffic jams. Compared with traditional sensors, remote sensing technology featuring rich information, low cost and wide coverage, which is widely used in traffic applications.(Sakai et al., 2019). As one of the important applications, vehicle detection can be applied in traffic monitoring, road planning, target tracking, etc.(Tang et al., 2017b). Therefore, vehicle detection in remote sensing images has attracted more and more attention in recent years. The existing approaches for vehicle detection in remote sensing images could be simply categorized into three types, including computer vision methods, traditional machine learning methods and deep learning.

Many computer vision methods of vehicle detection have been proposed in the literature. As the key part of these methods, the selection of features usually includes spectral features (Bar and Raboy, 2013), texture features(Chen et al., 2016), shape features(Zhang et al., 2017), etc. With a range of features, vehicle detection is achieved in different applications. When its geometric and radiometric feature, such as the shape and the spectral information, is obvious and distinctive, the vehicle can be detected with a better reliability, however, the detection result is relatively poor especially for complex background.

In the traditional machine learning methods, the Haar-like feature (Papageorgiou et al., 1998), artificial neural network (ANN) and Histogram of Oriented Gradient(HOG) and otherwise are extensively used in vehicle detection. An approach based on Haar-like features with adaptive enhancement technology was proposed (Leitloff et al., 2010). Taking the symmetry of vehicles in remote sensing images into consideration, Ren (Ren et al., 2016) presented a method of optimizing Haar-like features. Cao (Cao et al., 2011) combines the HOG features from the Adaboost classifier to build the feature vector and trains the SVM classifier for vehicle detection.

Although these methods are able to improve the accuracy of vehicle detection effectively, their features used for object detection should be designed manually. When we change the targets, features should be redesigned. Mentioned above indicate that the model has poor generalization performance, which restricts the further development of vehicle detection technology.

Deep learning techniques have shown their superior advantages in feature expression. Therefore, vehicle detection technology based on deep learning has been widely studied An oriented single shot multi-box detector was proposed for detecting vehicles with arbitrary orientations by Tang et al(Tang et al., 2017a). Hybrid depth convolution neural network (HDNN) extracts variable scales of features for detection(Chen et al., 2014). Sommer (Sommer et al., 2017) systematically studied the potential of Fast R-CNN and Faster R-CNN on aerial images. Overall, deep learning has made some breakthroughs in vehicle detection. However, there are still problems that need urgent solutions such as insufficient learning of the features due to the small size of the vehicle.

In this paper, we propose a convolutional neural network based on SSD for vehicle detection. By fusing feature maps of different scales, more information about vehicles could be available. The batch normalization layer is incorporated to prevent the network from overfitting and speed up the training. The main contribution of this paper is designing a multi-scale feature fusion network for vehicle detection in high resolution images, which can improve the accuracy remarkably.

2. RELATED WORK

2.1 Convolutional Neural Network

With the development of deep learning, convolutional neural network (CNN) has become a new research hotspot due to its powerful ability of feature expression. As one of the most important and successful neural networks in deep learning, CNN

* Corresponding author

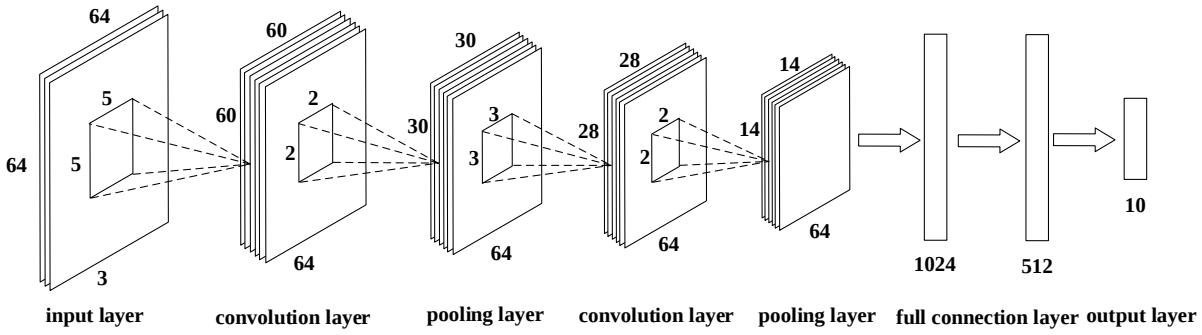


Figure 1. The structure of CNN

has been widely used in a variety of applications, such as classification, object detection, image registration, etc. With the deepening of research, many networks, such as Alexnet (Krizhevsky et al., 2012), VGG(Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015), Resnet (He et al., 2016) have been proposed. Weight sharing and local perceptron are CNN's characteristics, which not only makes the network structure easier to optimize but also reduces the risk of overfitting. In general, it is a multi-layer perceptron, which uses convolution to extract low-level features and combines low-level features into high-level features. CNN mainly includes input layer, hidden layer and output layer, and the hidden layer includes convolution layer, pooling layer and full connection layer. The structure of CNN is shown in Figure 1.

2.2 Batch Normalization

The stochastic gradient descent (SGD) is widely used to train a convolutional neural network, which features simplicity and efficiency (Bottou, 2010). However, due to the linear transformation and nonlinear activation mapping in each layer, small fluctuations of the parameters are amplified with the number of network layers increasing, and changes of parameters lead to their poor distribution in each layer, resulting in the gradient disappears. Therefore, aiming at these problems, the Batch Normalization(BN) (Ioffe and Szegedy, 2015) is proposed in deep neural network training. BN re-parametrizes the underlying optimization problem to make the loss landscape more stable and smooth. This implies that the direction of gradient descent is more predictive, which enables us to use a larger learning rate and faster network convergence (Santurkar et al., 2018). So, it is used in most deep learning models because it's practical success.

2.3 SSD Network

In terms of the design of network, the object detection network can be divided into two categories: region proposal and end-to-end. The former method employs a proposal network to extract the position of the object, and then determines the object categories, which mainly includes: R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), etc. The end-to-end method directly extracts and distinguishes the objects on the feature maps, which greatly improve the detection efficiency. YOLO (you only look once) (Redmon et al., 2016), SSD (Single Shot Multi-Box Detector) (Liu et al., 2016), YOLO v2 (Redmon and Farhadi, 2017), YOLO v3 (Redmon and Farhadi, 2018) ,etc. are the representative methods of this kind.

As a one-stage detection method, SSD holds better performance on the PASCAL VOC (Everingham et al., 2010), COCO (Lin et al., 2014), and ILSVRC (Russakovsky et al., 2015) datasets. It combines the advantage of Yolo and Faster R-CNN, which is a multi-scale object detection network. Thanks to the characteristics of multi-scale and multi-box, SSD could detect objects in different scales separately on each feature map. Considering the inaccurate positioning caused by the fixed size of boxes, SSD sets different scales and ratios of boxes for each feature map. The structure of SSD is shown in Figure 2.

Due to the better performance of detection, a range of improvements have been made in different settings, such as FSSD (Li and Zhou, 2017), DSSD (Fu et al., 2017). These improvements effectively increase the accuracy of object detection by extracting more abundant information from the feature maps.

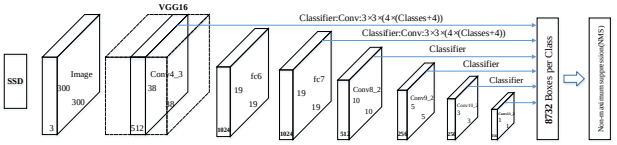


Figure 2. The structure of SSD

3. PROPOSED METHOD

Although SSD detects objects with several feature maps of different scales, these feature maps are irrelevant which tends to make the network unable to combine the overall information and local information for detection. And this weakens the capacity of object detection, especially for small objects like vehicles. At the same time, due to the small size of vehicles in the images, the low-level feature map contains less information. As a result, the network structure is redundant and detection speed decreases greatly. Therefore, fusion of feature maps and optimizing network structure will improve the accuracy and the speed of detection.

3.1 Network Structure

Our network is designed based on SSD, which mainly is divided into two parts: feature fusion module and vehicle detection module. In the feature fusion module, high-level and low-level feature are fused to generate a new fusion feature. Next, the fusion feature are inputted into the detection module for multi-scale detection. We use three levels of feature maps for fusion and introduce batch normalization to the detection module. Considering the small size of the vehicle in the image, we only use the four scales for detection. The method will be detailed discussed in Section 3.2 and 3.3. The overall structure of our network is shown in Figure 3.

3.2 Feature Fusion Module

What is the most distinctive between our method and SSD is the

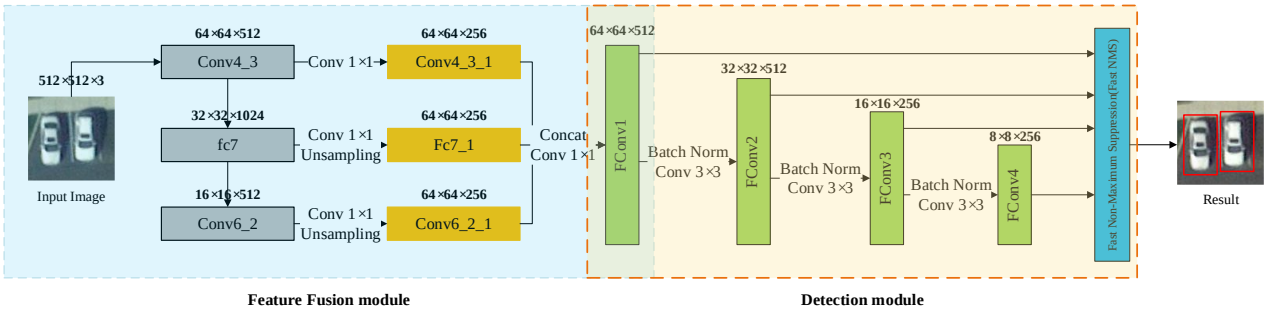


Figure 3. The structure of our network

detection base map. SSD directly uses the conv4_3 in VGG as the base map. Inspired by FSSD, our method fuses the low-level and high-level feature maps to a new feature map as the base map for detection. The structure of the module is shown in Figure 4. There are some factors to be considered when designing the feature fusion module, which will be introduced in the following.

Convolution Layer: In SSD512 based on VGG16, it chooses conv4_3, fc7, conv6_2, conv7_2, conv8_2, conv9_2 and conv10_2 as feature map to detect objects. Besides, the feature map is resized to 1/8, 1/16, 1/32, 1/64, 1/128, 1/256, 1/512 of the original image, respectively. Due to the small size of vehicles in the image, assuming that the size of feature map smaller than 1/32 of original one contains less information, we introduce conv6_2 for feature fusion instead of using conv7_2 as in FSSD. Conv7_2 has been tried, but the ablation experiment gains unsatisfactory results.

Concatenate layer: There are two methods of feature merging: concatenation and element-wise summation. Concatenation merges the channels of the features, and the value involved in each channel is unchanged. As two feature maps concatenate, the same number of channels is unnecessary. The element-wise summation is implemented by adding values of each corresponding channel in the feature map with equal number of channels. According to the analysis of the experiment the concatenation is selected as the method for feature merging.

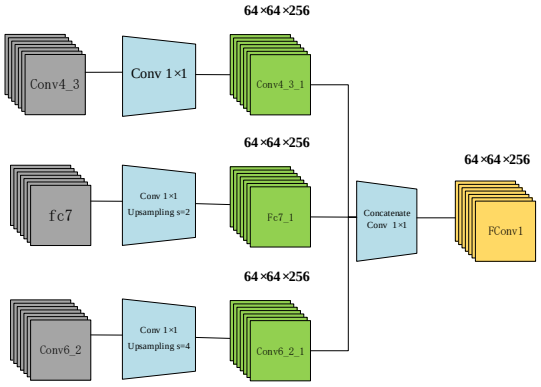


Figure 4. The feature fusion module

Up-sampling layer: To fuse feature maps more conveniently and effectively, the size and channel of feature map is needed to be adjusted to the same value. Firstly, 1×1 convolution is used to make conv4_2,fc7,conv6_2 share the same number of channels. Feature maps generated from fc7, conv6_2 are down-sampled with 2×2 max-pooling leading to different size with conv4_3. The feature maps are resized to equivalent size with conv4_3 by up-sampling. In this way, all the feature maps have the same size and channel.

3.3 Batch Normalization Module

Batch normalization mentioned in Section 2.2 shows excellent performance, which can speed up training and improve accuracy. L2Normalization is involved in SSD's object detection, while batch normalization is used to generate the fusion feature maps of FSSD. However, neither of these two normalization methods are engaged in scaling feature maps, which lowers the accuracy. Therefore, batch normalization is introduced in our network. When generating different scales of feature maps, normalization is added before inputting to the next layer, to avoid overfitting and low accuracy. At the same time, considering the tininess of vehicles in the subsequent feature maps, we abandoned some large-scale feature maps to improve detection efficiency. The module structure is shown in Figure 5.

4. EXPERIMENT

UCAS-High Resolution Aerial Object Detection Dataset (UCAS-AOD) (Zhu et al., 2015) is selected to train and test our network. 269 images (3240 vehicles) with about 1300×700 pixels in size are used in our experiment. Among these images, 215 is used to train and validate, and 54 to test. The predicted box will be correct if its intersection over union (IoU) with the ground truth is higher than 0.5. We choose average precision (AP) as the metric of detection. Faster R-CNN, SSD300, SSD512 and YOLOv3 models are selected as comparison. To make the comparison more reasonable, all models are trained on single Nvidia 1050Ti GPU. Our and other modules are implemented based on Keras package.

4.1 Ablation Study

In this section, some important factors are considered in network design. We compare the results deriving from different settings to verify the effectiveness of the module. All the models are trained with UCAS-AOD and the results are summarized in Table 1.

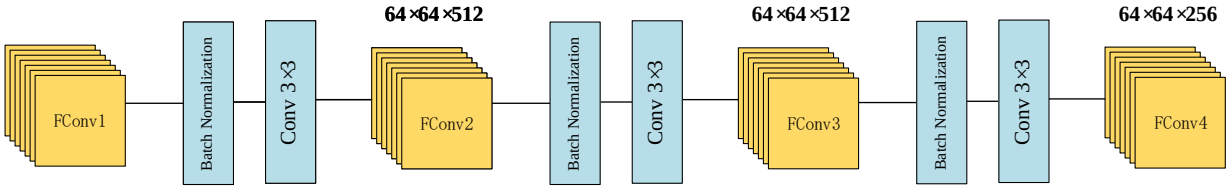


Figure 5. The detection module

data	BN	fusion method	fusion layers	detection layers	AP
UCAS-AOD	√	concat	conv4_3-conv6_2	4	**0.904**
UCAS-AOD	√	concat	conv4_3-fc7	4	0.878
UCAS-AOD	√	concat	conv4_3-conv7_2	4	0.901
UCAS-AOD	×	concat	conv4_3-conv6_2	4	0.876
UCAS-AOD	√	concat	conv4_3-conv6_2	3	0.895
UCAS-AOD	√	concat	conv4_3-conv6_2	5	0.888
UCAS-AOD	√	ele-sum	conv4_3-conv6_2	4	0.881

Table 1. Results of the ablation study on UCAS-ADC. **BN** means that batch normalization layer is added in the detection layer. The options of layers we can fuse include conv4_3, fc 7, conv6 2. The **fusion layers** represents the layers we choose to merge. The **detection layers** represents the number of detection layers. The **AP** is measured on UCAS-AOD test set.

4.1.1 The Fusion Layers

We make a comparison with networks of different fusion features. Considering the complexity of the network, we fuse four feature maps (conv4_2, fc7, conv6_2, conv7_2). The AP on UCAS-AOD is 0.907. However, when we remove conv7_2, the AP is similar to that with four feature maps, which shows that conv7_2 is useless for detection. Then we continue to remove conv6_2, the AP is decreased to 0.878, which proves conv6_2 could improve the accuracy of detection. So we choose conv4_2, fc7 and conv6_2 to fuse our feature.

4.1.2 Concatenation or Element-wise Summation

From Table 1, we can see that concatenation can achieve 0.904 AP while the element-wise summation only achieves 0.881 AP. The result shows that concatenation is 0.023 AP higher than element-wise summation when implementing vehicle detection in UCAS-AOD.

4.1.3 Batch Normalization Layer or Not

In SSD512, the network only uses L2Normalization in the conv4_2, which makes the network easier to overfitting in detection. To find a simple and efficient way to avoid this problem, we add batch normalization between low-level and high-level feature layers. As can be seen from Table 1, the additional batch normalization layer brings 0.028 AP improvement, which proves the effectiveness of batch normalization in the network.

4.1.4 The Detection Layers

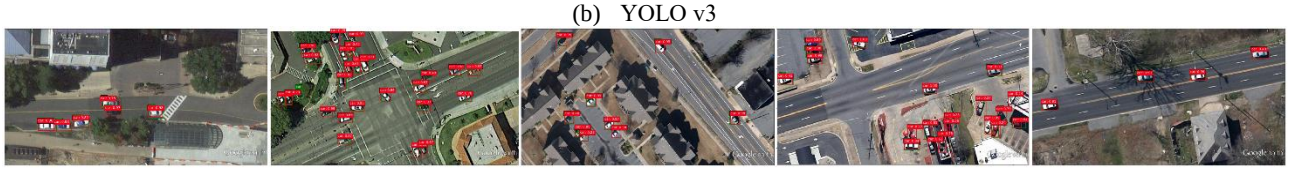
The author uses conv4_3, fc7, conv6_2, conv7_2, conv8_2, conv9_2, and conv10_2 to detect different scales objects in SSD512. However, given that the tininess of the vehicle, the high-level feature map contains less information. Therefore, we choose fconv1, fconv2, fconv3, and fconv4 as the detection layers in our network. The result can achieve 0.904 AP, but when we decrease or increase the number of detection layers, their AP are dropped to 0.895 and 0.888, respectively. The results show that the four detection layers are better than other numbers of layers.

4.2 Results on UCAS-AOD

According to the ablation experiments, the network structure is designed as follows: VGG16 with 512×512 pixels as input is the backbone network. Conv4_3, fc7 and conv6_2 are converted to 256 channels with 1×1 convolution layer, and then fc7 and conv6_2 up-sample to 64×64. Afterward, conv4_3, fc7 and conv6_2 are concatenated together to a fusion feature. Finally, several detection blocks (including one batch normalization layer, one 3×3 convolutional layer with stride 2 and one ReLU layer) are applied to detect vehicles.

(a) SSD512

(b) YOLO v3



(c) Ours

Figure 6.The result of (a) SSD512, (b) YOLO v3 and (c) Ours. From the figure (a) (b), there are many missing and false detection in the results of SSD and Yolo, and the detection result is poor in the detection of some smaller vehicles. Although our method has some problems in some areas with dense vehicles, the overall performance is better, and the smaller vehicles can also be accurately detected.

Method	data	base network	speed(FPS)	GPU	input size	AP
Faster R-CNN	UCAS-AOD	Resnet50	0.82	Nvidia 1050Ti	600×600	0.733
SSD300	UCAS-AOD	VGG16	12.30	Nvidia 1050Ti	300×300	0.652
SSD512	UCAS-AOD	VGG16	6.91	Nvidia 1050Ti	512×512	0.81
YOLO v3	UCAS-AOD	Darknet-53	10.95	Nvidia 1050Ti	416×416	0.828
Ours	UCAS-AOD	VGG16	**7.41**	Nvidia 1050Ti	512×512	**0.904**

Table 2. UCAS-AOD test detection results. The speed of all models are tested on a single **Nvidia 1050Ti GPU**. The metric of speed is Frame per Second (**FPS**).

We train our models on Nvidia 1050Ti GPUs for 14k iterations. The learning rate starts at 0.002 and decreases to 0.0002 after 11k iterations. The anchor box's size has been modified according to the size of vehicles, which mainly range between 30 to 80 pixels. We adopt Adam with beta1 0.9 and beta2 0.999 (Kingma and Ba, 2014) to optimize our pre-train network. We compare the proposed method with several detection models, including Faster R-CNN, SSD300, SSD512 and YOLOv3. The default parameter settings are used in our experiments. Figure 6 shows the detection results of different methods. Table 2 lists the AP of different methods on the test dataset. Our network achieves 0.904 AP and scores the highest AP among those methods mentioned above, which increases by 0.091 compared with the SSD512. From Figure 6, we can see that our network has fewer problems of false and missed detection than other methods. Moreover, most of the results using our network have higher confidence. Therefore, our network has better performance.

4.3 Speed

Testing speed is another essential part of detection methods. The inference speed is shown in Table 2. Our network can run at 7.41 FPS with the input size 512×512 on a single Nvidia1050Ti GPU. We also test other models in the same environment. Although our network adds a concatenation layer and several batch normalization layers, there is no reduction in speed compared with SSD512 due to the decrease of detection layers. In Table 2, it is clear that our network is similar in running speed to SSD512 while having the highest accuracy.

5. CONCLUSION

In this paper, we proposed a new convolutional neural network based on SSD for vehicle detection, which applies feature fusion and batch normalization layers together on it. Since the small size of vehicles in high-resolution images, it is hard to detect by simply using the same-level features. The fusion layer is designed for getting more information by concatenating features at different levels. Then we generate the detection layer based on the fusion feature layer and batch normalization is added into the detection layer to prevent the network from overfitting and to speed up training. The ablation experiments demonstrate the effectiveness of these structures. Similarly, the results on UCAS-AOD show that our network can improve the accuracy effectively without loss of speed, which could get 0.904 AP and achieve a speed of 7.41FPS. In the future, we will use more robust backbone networks such as Resnet to get better performance on other datasets.

REFERENCES

Bar, D.E., Raboy, S., 2013. Moving car detection and spectral restoration in a single satellite WorldView-2 imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6(5), 2077-2087.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent, Proceedings of COMPSTAT'2010. Springer, pp. 177-186.

Cao, X., Wu, C., Yan, P., Li, X., 2011. Linear SVM classification using boosting HOG features for vehicle detection in low-altitude airborne videos, 2011 18th IEEE International Conference on Image Processing. IEEE, pp. 2421-2424.

Chen, X., Xiang, S., Liu, C.-L., Pan, C.-H., 2014. Vehicle detection in satellite images by hybrid deep convolutional neural networks. IEEE Geoscience and remote sensing letters 11(10), 1797-1801.

Chen, Z., Wang, C., Luo, H., Wang, H., Chen, Y., Wen, C., Yu, Y., Cao, L., Li, J., 2016. Vehicle detection in high-resolution aerial images based on fast sparse representation classification and multiorder feature. IEEE transactions on intelligent transportation systems 17(8), 2296-2309.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International journal of computer vision 88(2), 303-338.

Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C., 2017. Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659.

Girshick, R., 2015. Fast r-cnn, Proceedings of the IEEE international conference on computer vision, pp. 1440-1448.

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

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, pp. 1097-1105.

Leitloff, J., Hinz, S., Stilla, U., 2010. Vehicle detection in very high resolution satellite images of city areas. IEEE transactions

on Geoscience and remote sensing 48(7), 2795-2806.

Li, Z., Zhou, F., 2017. FSSD: feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, European conference on computer vision. Springer, pp. 740-755.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, European conference on computer vision. Springer, pp. 21-37.

Papageorgiou, C.P., Oren, M., Poggio, T., 1998. A general framework for object detection, Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271), pp. 555-562.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788.

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, C., Huixian, H., Yuan, T., Chengxiao, W., 2016. Vehicle identification from remote sensing image based on image symmetry. Remote Sensing for Land & Resources 28(4), 135-140.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems, pp. 91-99.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision 115(3), 211-252.

Sakai, K., Seo, T., Fuse, T., 2019. Traffic density estimation method from small satellite imagery: Towards frequent remote sensing of car traffic, 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 1776-1781.

Santurkar, S., Tsipras, D., Ilyas, A., Madry, A., 2018. How does batch normalization help optimization?, Advances in Neural Information Processing Systems, pp. 2483-2493.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sommer, L.W., Schuchert, T., Beyerer, J., 2017. Fast deep vehicle detection in aerial images, 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 311-319.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9.

Tang, T., Zhou, S., Deng, Z., Lei, L., Zou, H., 2017a. Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks. Remote Sensing 9(11), 1170.

Tang, Y., Zhang, C., Gu, R., Li, P., Yang, B., 2017b. Vehicle detection and recognition for intelligent traffic surveillance system. Multimedia tools and applications 76(4), 5817-5832.

Zhang, X., Xu, W., Dong, C., Dolan, J.M., 2017. Efficient L-shape fitting for vehicle detection using laser scanners, 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 54-59.

Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., Jiao, J., 2015. Orientation robust object detection in aerial images using deep convolutional neural network, 2015 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 3735-3739.