Vehicle Detection from Unmanned Aerial Images with Deep Mask R-CNN

Rıdvan Yayla, Emir Albayrak, Uğur Yüzgeç

Abstract

In this paper, a classification approach which is applied to Mask Region-based Convolutional Neural Network as deeper is proposed for vehicle detection on the images from UAV instead of the familiar methods. The different types of unmanned aerial vehicles are widely used for a lot of areas such as agricultural spraying, advertisement shooting, fire extinguishing, transportation and surveillance, exploration, destruction for the military. In recent years, deep learning techniques are progressively developed for object detection. Segmentation algorithms based on CNN architecture are especially widely used for extracting meaningful parts of an image. Additionally, Mask R-CNN based on CNN architecture rapidly detects the object with high-accuracy on an image. This study shows that the high-accuracy results are obtained when the Mask R-CNN is applied as deeper in vehicle detection on the images taken by UAV.

Keywords: Convolutional neural networks, Deep learning, Mask R-CNN, Vehicle detection.

1 Introduction

Nowadays, the images and the videos that are received from unmanned aerial vehicles (UAV) are widely used in a lot of areas such as commerce, agriculture, security, and the military. The vehicles or moving objects are mostly taken notice of for object detection due to the security and their variations. The vehicle detection is especially required when a moving target is detected for learning its coordinates in a military operation. Moreover, the incoming vehicles are directed to the free

doi:10.56415/csjm.v30.09

parking areas when a vehicle intensity rate is determined in an open-top car parking. Additionally, illegal structures can also be detected with deep learning algorithms by comparing them to the previous drone images.

Deep learning, which is a branch of machine learning, is basically an increased layer in the number of hidden layers of Multi-Layer Perceptron (MLP) by a large number [1]. In recent years, region-based segmentation model based on Convolutional Neural Network (CNN) architecture is an effective solution for object detection. Region-based Convolutional Neural Network (R-CNN), Fast R-CNN, Faster R-CNN and Mask R-CNN are the most used models for object detection. In this study, the accuracy of vehicle detection from a UAV image is evaluated by comparing Mask R-CNN models based on different backbone networks with pre-trained weights such as Resnet-101, Vgg16, (Microsoft) MS Coco. When a deeper Mask R-CNN algorithm is used for vehicle detection, the better results are obtained for the vehicle segmentation.

2 Methodology

Object recognition is a scientific field, which is related to computer vision and image processing, that deals in visual images and videos with the detection of instances of semantic objects of an exact type (such as traffic light, cars, or person). There are a lot of object detection applications that include image and video improvement in many fields of computer vision. Deep learning methods have been commonly used in the last decade for object detection. The most well-known and used techniques are R-CNN, Fast R-CNN, Faster R-CNN and Mask R-CNN for object detection. Additionally, YOLO algorithm based on CNN architecture that is also used for non-dynamic images is especially used for real-time video images. Li et al. developed a YOLO-Grape model for real-time detection of multiple varieties of table grapes in complex situations [2]. Lin and Li developed an integrated circuit board object Model based on YOLO for fast quality management process [3]. These studies are based on real-time object detection but the YOLO algorithm just uses a bounding box for object detection and the performance of the YOLO is compared by Fast R-CNN that is the previous

version of Mask R-CNN. The advantage of Mask R-CNN is to segment objects as a whole with masking.

Nie et al. realized an inshore ship detection based on Mask R-CNN by integrating to Soft-Non-Maximum Yang. They improved to ship segmentation with their proposed method by using instance segmentation on Mask R-CNN [4]. Li and Cheng developed a pedestrian detection application based on Mask R-CNN. In their study, they determined weak points of Mask R-CNN and they optimized their own network based on Mask R-CNN. By this way, they have provided a quality pedestrian gender segmentation with their own created dataset [5]. Moreover, Vemula and Frye proposed a study that is related to power-line detection by using UAV images. They improved Mask R-CNN detection by transfer learning and Mask R-CNN powerline detector is proposed based on UAV images [6]. Song and Zhao have used "Mask R-CNN" for the gastric cancer diagnosis from the medical images. Instead of the well-known object detection, they optimized different size medical images for the pathological experiments and obtained successful results on gastric cancer detection [7].

As it is seen in the actual examples, the Mask R-CNN can be effectively optimized with different methods and it is used for the different purposes from security to health. In our study, the vehicles on UAV images are segmented by using Mask R-CNN as deeper.

2.1 Convolutional Neural Network (CNN)

CNN is a powerful deep learning model for object detection and it provides high accuracy in recent years. CNNs consist of neurons with renewable weights and biases [8]. Each neuron takes some inputs and realizes a dot product and arbitrarily follows it with a non-linearity after these steps. There are six common layers in the CNN architecture, and these layers are as given below [9]:

- Image Input Layer
- Convolutional Layer
- Rectified Linear Unit (ReLu)
- Pooling Layer
- Fully Connected Layer

• Classification (Softmax) Layer

The image which consists of the pixels is converted to matrix format in the image input layer. In the convolutional layer, a field of the image is handled and a convolution operation is performed with a small part of the input matrix having the same dimension. An activation function is used to get rid of the negative values generated during the convolution process. Because the non-negative values are not generated, the Rectified Linear Unit (ReLU) is the basically used activation function in deep learning architecture. The ReLu function is given in the equation (1) [9].

$$\mathrm{ReLu}(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ if } x \geq 0 \end{cases}. \tag{1}$$

In the CNN architecture, the pooling layers are periodically added to the network. The pooling layer is independently executed in each depth slice of the input and it is resized by using max operation. According to this process, pooling layer is also called max pooling layer. The aim of the max operation is to decrease the number of parameters and calculations in the network.

In the fully connected (dense) layer, all neurons are fully connected to each other with all activations in the previous layer. The aim of the fully connected layers is basically feature extraction and classification. The predictions in dense layer are ready for the classification layer. The classification layer is a standard fully-connected (dense) layer that uses the softmax activation function. In this final layer, a prediction of classes that are created in the training model is made for the desired field of an image via the softmax function. The softmax function is given in the equation (2) [9].

$$s\left(x_i\right) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}. \tag{2}$$

Most of the application which is made by using Mask R-CNN is focused on medical, radar image, or object detection. In our study, the vehicle images that are taken from a UAV are segmented by using a deeper Mask R-CNN algorithm.

2.2 Region-based Segmentation

2.2.1 R-CNN

Girschick et al. developed an object detection method that can be implemented from high-capacity CNNs to region proposals so as to situate and segment objects in 2014 [10]. They proposed a method with using the selective search algorithm that takes out only 2000 regions from the image. These regions are called region proposals. The CNN is set to region proposals of the image, and a prediction is generated by the CNN. This method is called region-based convolutional neural network (R-CNN). The working principle of the R-CNN is summarized as follows:

- A region proposal in an image is sent to CNN architecture and the correct region is found.
- The region proposal is classified by using an operation that is called a selective search algorithm. The algorithm provides the analysis of the image by different-size windows that try to combine neighbor pixels by texture or color intensity.
- Selective search is executed on the image for finding the desired region.
- The obtained regions from the selective search algorithm are utilized for classification and feature extraction by using a pretrained CNN architecture.

R-CNN obtains perfect object detection correctness by using CNN for image classification but has mostly disadvantages. Because all the proposed regions are applied to CNN, the running time of the algorithm is long and more memory is required. The Fast R-CNN model based on R-CNN is developed for removing these disadvantages.

2.2.2 Fast R-CNN

R-CNN works very well for detecting an object but it is slow because of the fact that the Selective Search algorithm is used for determining

region proposals. In the R-CNN model, all the region proposals are delivered to ConvNet architecture, and it runs for all region proposals. Instead of that 2000 regions are to feed to CNN architecture, the convolution operation is made only once per image in Fast R-CNN and a feature map is extracted from it. In the convolutional feature map, the desired regions are recognised and are scanned into squares by using a kind of pooling layer that performs max pooling on inputs [11]. By this way, the desired regions are resized into a fixed size for setting into a fully connected layer. Each region proposal pursues on ConvNet architecture, and this method is called Region of Interest (ROI) Pooling; this kind of pooling layer is called Region of Interest (RoI) layer. By the RoI pooling, the region is divided into subregions. The RoI Pooling extracts a fixed-size frame from the feature map and uses the attributes for gaining the final class label and bounding box. [12]. The output attributes from the RoI Pooling layer are delivered into the sequential Fully Connected layers, the softmax and bounding box (BB) regression branches. Each RoI expresses the number of categories and a single background category. The softmax classification produces probability values of each RoI. The BB regression output is used for making bounding boxes from the region proposal algorithm more susceptible.

2.2.3 Faster R-CNN

Faster-CNN was developed for eliminating computing and running time in 2016. Shaoqing et al. developed an object detection algorithm that provides the network learning of region proposals [13]. The selective search algorithm is not used in the new approach due to the object detection algorithm. In this model, a discrete network is used to predict the region proposals in place of using the selective search algorithm on the feature map to recognize the region proposals. The network is called as Region Proposal Network (RPN).

Faster R-CNN consists of two modules. The first module proposes the regions by using deep fully ConvNets, and the second one uses region proposals via the Fast R-CNN detector [14]. The Faster R-CNN architecture is shown in Figure 1. Faster R-CNN uses the same convolutional network for both region proposal production and object detec-

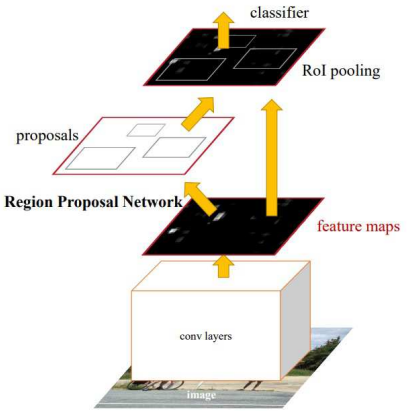

Figure 1. Architecture of the Faster R-CNN [13]

tion. Unlike the selective search algorithm, it is divided into ConvNets and the Region Proposal Network (RPN) which depends on the determined attributes of the image [15]. This method provides a decrease in the computing process and higher accuracy results.

2.2.4 Mask R-CNN

Mask Region-based Convolutional Neural Network (Mask R-CNN) is a segmentation model based on Faster R-CNN model. Mask R-CNN provides that the network not only performs object detection but pixel-wise instance segmentation based on semantic segmentation. While the model predicts a bounding box recognition by this method, it adds a mask to object at the same time. Mask R-CNN is performed on some principles. These principles are listed as follows [16]:

- It works on Faster R-CNN model by inserting a parallel branch.
- It predicts segmentation mask by using a small (Fully Connected Layer) FCN.
- It performs better than state-of-art models in person keypoint detection, segmentation and bounding box detection.
- It changes RoIs in Faster R-CNN to a quantization-free layer

called RoI Align. RoI Align removes the quantization which causes the misalignment. 4 locations are sampled and bilinear interpolation is used.

• The separated networks run in parallel, and in this way, the prediction system runs at a higher speed.

• It uses an instance segmentation method that is based on semantic segmentation. The difference between these segmentation methods is as follows: while the semantic segmentation segments (classifies) the same objects into a single one with one label (person, car, etc.), the instance segmentation segments (classifies) the same objects as similar instances with different labels such as person1, person2, or car1, car2.

The Mask R-CNN is different from prior systems, where classification depends on mask prediction. It also uses the lost function for fixing weight errors. The lost function (L) for each sampled RoI is calculated as follows [16].

$$L = L_{cls} + L_{box} + L_{mask}. \tag{3}$$

In equation (3), L_{cls} expresses the classification loss, L_{box} is the bounding box loss, and L_{mask} is the masking loss. RoIPool is a useful process for exposing a small feature map such as 7×7 from each RoI [17]. RoI Align abolishes the harsh quantization of RoI Pool by aligning the determined features with the input. The bilinear interpolation which computes the exact values of the input features at four regularly sampled locations in each RoI bin and aggregates the result (using max or average) is used in this model [16][17]. The Mask R-CNN differs from the Faster R-CNN with two features.

• Masking: Faster R-CNN consists of two common features which are bounding box and prediction of object detection. The masking feature is added in the Mask R-CNN algorithm.

• Instance segmentation: If there are two same objects or same meaningful parts in an image, these objects or parts are divided into different masking or colourization [17]. Mask R-CNN uses a fully connected network to predict the mask.

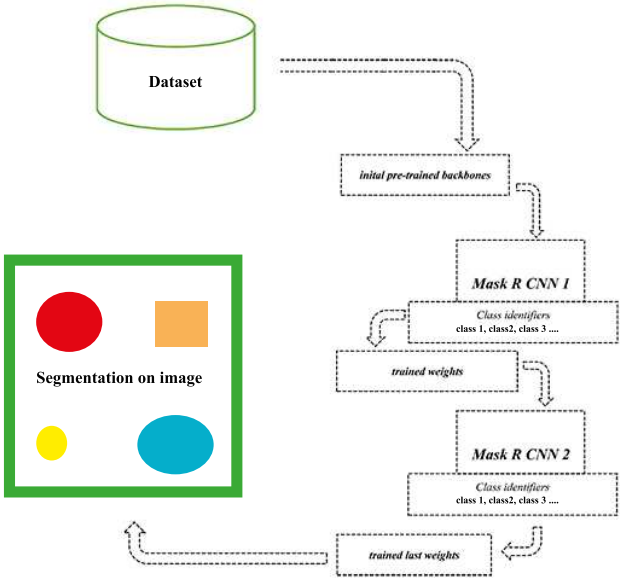

Figure 2. Working principle of Deep Mask R-CNN [18]

2.3 Deep Mask R-CNN

In [18], a multi-class hybrid classification method is proposed by using deep Mask R-CNN for smoothed Synthetic Aperture Radar (SAR) images. The method is trained with different backbone networks that are integrated with Deep Mask R-CNN, and these networks are compared to each other. When the Mask R-CNN is trained with default (pre-trained) backbone weights such as MS coco, inception v3, vgg16 resnet101 and resnet50, a new trained weights are obtained by Mask R-CNN at the end of the training process. The working principle of deep Mask R-CNN is shown in Figure 2. The segmentation of the meaningful parts is observed in the image after the pre-trained backbone weights are given first convolutional inputs in the initial running time. If the segmentation loss values are not sufficiently low, the obtained trained weights are repeatedly sent to Mask R-CNN within a new training process. The deep convolutional process is repeated until

the segmentation losses reach a constant low value.

The deep Mask R-CNN aims to decrease Mask R-CNN algorithm losses and to obtain a quality segmentation that is properly separated from each other [19]. According to the related work, it is shown that when the Mask R-CNN applied deeper to an image, the Mask R-CNN losses are decreased and more high-quality segmentation is obtained. In our study, the Deep Mask R-CNN, which is the second hybrid study of the related work, is used for a quality vehicle segmentation by decreasing Mask R-CNN algorithm losses.

3 System Design and Components

3.1 Materials

In this study, vehicle detection from the UAV is realized by applying to Mask R-CNN as deeper. The vehicles in the image can be detected with instance segmentation at the end of the training process.

The DJI Mavic PRO model drone, which has a 12.3 MP resolution camera supported CMOS sensor, is used as the UAV [20]. The drone can also take 1080p, 4K and HD camera shooting. The drone used in this study is shown in Figure 3. In this study, 282 images taken from

Figure 3. UAV used in this study (DJI Mavic PRO) [20]

the UAV trained with the Mask R-CNN algorithm (1000 iterations at each step and 50 iterations in total) and the vehicles in the UAV image

were detected as a bird's-eye. The images are pictures that mostly contain vehicles and are taken from different angles and locations. The images are initially divided into two parts as test and train datasets. The vehicles in all images are polygonally drawn with the VGG image annotator, which is a basic and useful drawing tool [21].

Additionally, the study is realized with hardware which is Nvidia Geforce GTX 1070 Ti (8GB – GDDR5), 16 GB RAM, GPU, Intel (R) Core (TM) 4 cores CPU and 240 GB SSD + 1 TB Harddisk [22]. The experiment is also realized by Tensorflow and Keras libraries based on Python language. Moreover, Tensorboard graphic interface at Tensorflow library is used for determining error rates at the end of the training. By this way, the observation of the error rates due to the iterations can be made at Ubuntu platform. At the same time, the different images, which were taken from different locations as the eye-bird view on the web and not taken from the UAV, were also tested in the test process.

3.2 Methods

Nowadays, deep learning tools based on advanced graphic cards have become very popular for the object-detection in computer vision. The deep learning libraries such as Tensorflow, Keras are very useful platforms for applying to deep learning model [23]. The image is taken by the deep learning platforms and it is converted to numpy arrays which are defined in Tensorflow library.

The Mask R-CNN algorithm loss should be decreased in Mask R-CNN for a high-quality segmentation. When the Mask R-CNN algorithm is applied for the vehicle segmentation with the default backbones such as MS Coco, Resnet-50, vgg16, a segmentation weight file is obtained at the end of the last iteration of 1^{st} training process [24]. The segmentation losses are observed with the obtained weights and the 2^{nd} training process is started with these weights for decreasing the losses. The Deep Mask R-CNN is a useful method for decreasing the algorithm loss by using the trained weights. The segmentation is observed at the end of each training process until the segmentation edges can be improved. The working principle of the proposed method is shown in Figure 4. The proposed method can also be repeated more

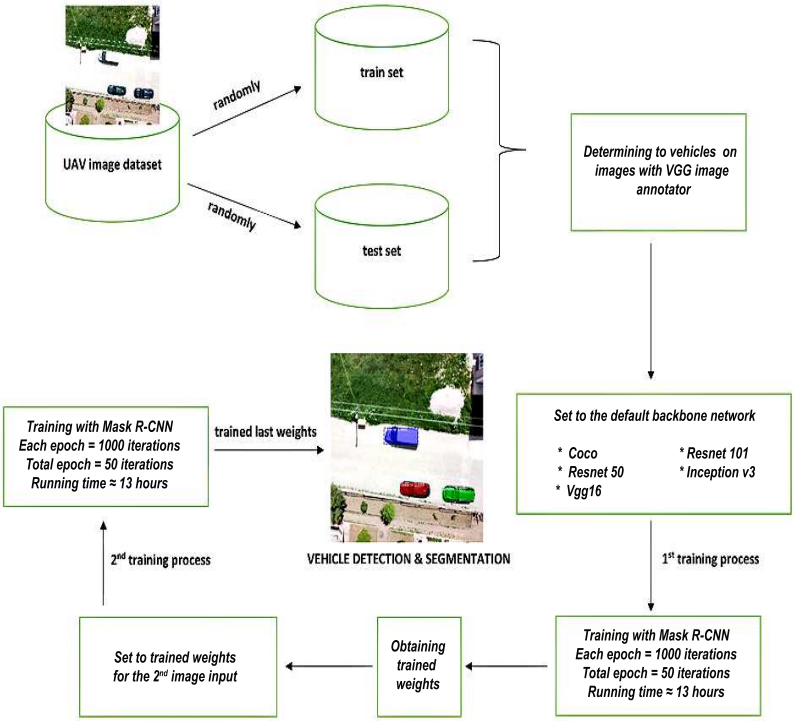

Figure 4. Working principle of the proposed method

than two times until the accuracy of the edges reaches the least error. When the Mask R-CNN loss reaches a constant value at the end of the training process, the experiment is ended.

The loss values are variants for different segmentation experiments. In our study, the single-vehicle class is defined for the segmentation and the Mask R-CNN scans vehicle fields of the image in the algorithm. Thanks to instance segmentation, more than one vehicle can be detected with different colorization in the deep Mask R-CNN. In this way, vehicles are distinguished from each other in the image.

The method is not only applied to static images but it can also be applied to dynamic images such as video and gif extension images. YOLO algorithm based on CNN is faster for the dynamic images but it uses only bounding boxes for the vehicles segmentation [25]. In our study, because the masking and instance segmentation are used in Mask R-CNN, the Mask R-CNN is preferred for a high-quality vehicle segmentation. The vehicles can be segmented with different masking in the deep Mask R-CNN.

3.3 Experimental Results and Performance

When the proposed method is applied to different pre-trained backbone networks, the performances are observed for each backbone weight. Each experiment is trained with different backbones and the trained weights, which are obtained at the end of the 1st training, are sent to the network as the initial weights again. The aim of the deep Mask R-CNN is to decrease Mask R-CNN losses which consist of the masking loss, classification loss, and bounding box loss for a high-quality segmentation.

The classification models that are used in the study have different performances. Ms Coco model is the default (baseline) standard pre-trained model for object detection. The more higher accuracy has been obtained in the preliminary Ms Coco model due to the large-scale dataset for the various computer vision detection. The experiment has been initially executed for the default Ms Coco model. After the default model, the experiment is executed for the other pre-trained models for the performance comparison. For example, while the Vgg16 backbone

network model has fixed-size kernels for the experiment of the study, the Inception v3 model has wider – parallel kernels for the experiment of the study. Skipping connection between the perceptions is a technic that is used in the Resnet models. Resnet models have different parameter numbers and the Resnet-50 has fewer parameters. Due to the skipping technique and fewer parameters, the Resnet-50 network model has provided the most successful results for the experiment.

The four sample images that are taken from two different UAVs and locations are shown in Figure 5. The segmentation results are also shown in Figure 6, Figure 7, Figure 8, Figure 9 and Figure 10. In the experiment figures, the bounding boxes, masking, and prediction accuracy are shown for each vehicle. The accuracy values of each vehicle detection are also shown in top-right of the each detected vehicle as numerical according to related backbone accuracy of the experiments.

Figure 5. Sample Images (Location: Pazaryeri-Bilecik and Turgutlu-Manisa Turkey [26])

The Deep Mask R-CNN experimental results are shown in Table 1. In Table 1, five different backbone network performances are shown

Figure 6. Deep Mask R-CNN (Ms Coco) segmentation (accuracy 92%)

with accuracy and loss. Each pre-trained model is initially sent to CNN at the 1st iteration and the observed weights are also sent to CNN at the 2nd iteration. This recurrent process is applied until the loss function arrives at a fixed value. In this study, the loss values have arrived at a fixed value in the 2nd iteration. In Table 1, the fixed values of a total loss, classification, bounding box, and masking loss are shown in the 2nd iteration. The obtained accuracy is also shown in the 2nd iteration for each backbone network.

3.4 Conclusion

Deep learning techniques are widely used nowadays for a lot of areas such as image processing, natural language processing, network security. In our study, the vehicle detection is made with a deeper Mask R-CNN based on CNN deep learning architecture by using the images, which are obtained from a UAV. A single-class segmentation based on the vehicle is made by using instance segmentation feature of the Mask R-CNN. Thanks to this feature, multi vehicles of an image are seg-

Figure 7. Deep Mask R-CNN (Resnet-50) segmentation (Accuracy: 93%)

Figure 8. Deep Mask R-CNN (Vgg16) segmentation (Accuracy: 88%)

Figure 9. Deep Mask R-CNN (Resnet-101) segmentation (Accuracy: 71%)

Figure 10. Deep Mask R-CNN (Inception-v3) segmentation (Accuracy: 88%)

Table 1. The experimental results for pre-trained backbones.

Mask R-CNN (1st iteration)					
Backbone	***loss***	***mrcnn class loss***	***mrcnn bbox loss***	***mrcnn mask loss***	***Accuracy***
Vgg16	0.1948	0.02296	0.02176	0.101	80%
Inception v3	0.1875	0.02375	0.02061	0.0987	81%
Resnet 101	0.4635	0.05099	0.06606	0.1771	53%
Resnet 50	0.08741	0.01473	5.0135	0.05257	91%
Ms Coco	0.08652	0.01478	4.9783	0.05138	91%
Mask R-CNN (2nd iteration)					
Backbone	***loss***	***mrcnn class loss***	***mrcnn bbox loss***	***mrcnn mask loss***	***Accuracy***
Vgg16	0.1118	0.01413	0.01053	0.0668	88%
Inception v3	0.1151	0.01525	0.01026	0.06763	88%
Resnet 101	0.2899	0.03569	0.03697	0.1225	71%
Resnet 50	0.06656	0.01172	3.8966	0.04082	93%
MS Coco	0.08036	0.01371	4.9783	0.04938	92%

mented with different segmentation and all vehicles in the image are separated from each other. It is aimed that the segmentation loss is decreased by using Mask R-CNN as deeper. In this study, a prototype experiment is made for vehicle detection. If more images are trained and the dataset image diversity becomes more than the current diversity, the accuracy of the detection is greater. Moreover, the dataset images have been taken on sunny days and the models are trained by these sunny images. Vehicle detection can be affected by other weather conditions. The study can be expanded for other weather conditions by being trained in different images such as cloudy, rainy, or foggy weather images. This study also provides ease of use for vehicle detection in many areas. It can also be extended with multi-class segmentation such as buildings, vehicles, and roads. Thanks to multi-class segmentation, different moving objects can be detected with the proposed method.

4 Acknowledgement

The study called "Vehicle Detection from Unmanned Aerial Vehicle Images with Deep Mask R-CNN" has been made in the scope of Bilecik Şeyh Edebali University, Graduate Education Institute, master thesis study.

References

[1] J. Brownlee, "What is Deep Learning," Machine Learning Mastery, Aug. 14, 2019. [Online]. Available: https://machinelearningmastery.com/what-is-deep-learning/

[2] H.Li, C. Li, G. Li, and L. Chen, "A real-time table grape detection method based on improved YOLOv4-tiny network in complex background," *Biosystems Engineering*, vol. 212, pp. 347–359, 2021.

[3] L. Szu-Yin and L. Hao-Yu "Integrated Circuit Board Object Detection and Image Augmentation Fusion Model Based on YOLO," *Frontiers in Neurorobotics* , vol. 15, pp. 155, 2021.

[4] S. Nie, Z. Jiang, H. Zhang, B. Cai, and Y. Yao, "Inshore Ship Detection Based on Mask R-CNN," in *Proc. 2018 IEEE International Geoscience and Remote Sensing Symposium*, (Valencia), 2018, pp. 693–696.

[5] X. Li and S. Cheng, "Pedestrian Gender Detection Based on Mask R-CNN," in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, (Chengdu), 2019, pp. 2082–2086.

[6] S. Vemula and M. Frye, "Mask R-CNN Powerline Detector: A Deep Learning approach with applications to a UAV," in *Proc. 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)*, (San Antonio), 2020, pp. 1–6.

[7] G. Cao, W. Song, and Z. Zhao, "Gastric Cancer Diagnosis with Mask R-CNN," in *Proc. 2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, (Hangzhou), 2019, pp. 60–63.

[8] A. Deshpande, "A Beginner's Guide To Understanding Convolutional Neural Networks," UCLA CS'19, Jul. 29, 2016. [Online]. Available: https://adeshpande3.github.io/adeshpande3.github.io/

[9] R. Yayla and B. Şen, "Research on Region-Based Convolutional Neural Network for Semantic Segmentation," in *Proc. 8th International Conference on Advanced Technologies Conference (ICAT'19)*, (Sarajevo), 2019, pp. 244–249.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition*, (Columbus, OH, USA), 2014, pp. 580–587.

[11] S. Hsu, C. Huang, and C. Chuang, "Vehicle detection using simplified fast R-CNN," in *Proc. 2018 International Workshop on Advanced Image Technology (IWAIT)*, (Chiang Mai), 2018, pp. 1–3.

[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(ICCV)*, (Santiago), 2015, pp. 1440–1448.

[13] R. Shaoqing, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Proc. IEEE,* vol. 39, no. 6, pp. 1137–1149, Jun 2017.

[14] B. Liu, W. Zhao and Q. Sun, "Study of object detection based on Faster R-CNN," in *Proc. 2017 Chinese Automation Congress (CAC)*, (Jinan), 2017, pp. 6233–6236.

[15] J. Lin, C. -. T. Chiu, and Y. Cheng, "Object Detection with Color and Depth Images with Multi-Reduced Region Proposal Network and Multi-Pooling," in *Proc. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Barcelona), 2020, pp. 1618–1622.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. 2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice), 2017, pp. 2980–2988.

[17] R.Yayla and B. Şen, "Region-based Segmentation of Terrain Fields in SAR Images," in *Proc. 28th Signal Processing and Communications Applications Conference Conference (SIU2020)*, Gaziantep, 2019, pp. 1–4.

[18] R. Yayla and B. Şen "A New Classification Approach with Deep Mask R-CNN for Synthetic Aperture Radar Image Segmentation," *Elektronika Ir Elektrotechnika*, vol. 26, no. 6, pp. 52–57, 2020.

[19] C. Ozcan, B. Sen, and F. Nar, "Sparsity-Driven Despeckling for SAR Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 115–119, 2016.

[20] *DJI Mavic Pro User Manual*, DJI technology., Shenzhen, China, 2017.

[21] A. Dutta and A. Zisserman, "The VIA annotation software for images,audio and video," in *Proc. 27th ACM International Conference on Multimedia*, (Nice), 2019, pp. 2276-–2279.

[22] R. Yayla, "Hybrid Intelligent Classification Technique For High-Resolution Sar (Synthetic Aperture Radar) Images," Ph.D. dissertation, Dept. Comp. Eng., Ankara Yıldırım Beyazıt Univ., Ankara, Turkey, 2020.

[23] Z. Zeng, Q. Gong, and J. Zhang, "CNN Model Design of Gesture Recognition Based on Tensorflow Framework," in *Proc. IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, (Chengdu), 2019, pp. 1062–1067.

[24] W. Abdulla, "Mask R CNN for object detection and instance segmentation on Keras and TensorFlow,"

GitHub repository, Mar. 20, 2017. [Online]. Available: https://www.github.com/matterport/Mask_RCNN

[25] M. Liu, X. Wang, A. Zhou, et al., "UAV-YOLO: Small Object Detection on Unmanned Aerial Vehicle Perspective," *Sensors Journal*, vol. 20, no.81, pp. 1–12, 2020.

[26] M. Almislar, "Manisa Turgutlu (Town) 4k aerial view," Youtube, Sep. 5, 2019. [Online]. Available: https://www.youtube.com/watch?v=AFwPmWeUwpM (in Turkish)

Rıdvan Yayla, Emir Albayrak,
Uğur Yüzgeç

Received February 22, 2021
Revised version December 29, 2021
Accepted for publication February 1, 2022

Rıdvan YAYLA
Computer Engineering Department
Bilecik Şeyh Edebali University
Pelitözü Dist. Fatih Sultan Mehmet Boulevard
Gülümbe Campus No:27 11230 BİLECİK / TURKEY
E–mail: ridvan.yayla@bilecik.edu.tr

Emir Albayrak
Computer Engineering Department
Bilecik Şeyh Edebali University
Pelitözü Dist. Fatih Sultan Mehmet Boulevard
Gülümbe Campus No:27 11230 BİLECİK / TURKEY
E–mail: emiralbayrak@gmail.com

Uğur Yüzgeç
Computer Engineering Department
Bilecik Şeyh Edebali University
Pelitözü Dist. Fatih Sultan Mehmet Boulevard
Gülümbe Campus No:27 11230 BİLECİK / TURKEY
E–mail: ugur.yuzgec@bilecik.edu.tr