5th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Automated vehicle detection and instance segmentation from high-resolution UAV imagery using YOLOv7 model

Esra Yildirim*[1], Umut Gunes Sefercik [1], Taskin Kavzoglu [1]

[1]*Gebze Technical University, Faculty of Engineering, Department of Geomatics Engineering, Kocaeli, Türkiye*

Keywords
Deep learning
UAV imagery
Vehicle detection
Instance segmentation
YOLOv7

Abstract

Automatic vehicle detection from unmanned aerial vehicles (UAVs) is an important task in the remote sensing domain and plays a pivotal role in many applications such as traffic monitoring, parking lot management, search and rescue tasks. Inspired by the success of the deep learning paradigm in image processing applications, many object detection, and tracking approaches have been developed and successfully employed in UAV-based object detection studies. In this study, automatic vehicle detection and instance segmentation was conducted using YOLOv7, which is the latest version of the You Only Look Once (YOLO) model from high-resolution UAV data obtained from Gebze Technical University campus in Turkey. For this purpose, vehicle images were collected from the UAV data of the study area, and the vehicles in the images were manually annotated with the LabelMe annotation tool. With the created dataset, the YOLOv7 algorithm was trained and tested with a transfer learning approach on Google Colab's virtual machine. Experimental results revealed that the YOLOv7 model achieved the Precision, Recall, and mAP@0.50 values for the bounding boxes and masks of vehicles as 99.79%, 97.54%, and 99.46%, respectively.

1. Introduction

Nowadays, there is a huge increase in the utilization of unmanned aerial vehicles (UAVs) for a wide range of applications, including disaster management, smart agriculture, transportation, and surveillance. Compared with conventional satellite systems, UAVs afford several capabilities such as high spatial resolution, a large field of view, low cost, flexible and effective data acquisition [Ammar et al. 2021]. Considering these unique capabilities, UAVs have become an indispensable technology in various image processing applications including automatic object detection, tracking, and image classification. In this context, the detection of numerous objects such as trees [Yildirim et al. 2022], vehicles [Tang et al. 2017a], buildings [Boonpook et al. 2018], and pedestrians [Shao et al. 2021] from UAV data has recently attracted increasing attention from researchers. Specifically, the recognition of vehicles from UAV data is a significant research topic as it has many useful applications including traffic management, surveillance, search and rescue tasks.

Traditional vehicle detection algorithms in aerial imagery mostly adopt sliding window search and hand-crafted features (e.g., histogram of oriented gradients, local binary patterns). However, the sliding window search produces a large number of candidate windows, leading to high computational complexity. Moreover, manually extracted features have restricted representation power for the target object. Due to these drawbacks, it is arduous for traditional vehicle detection methods to achieve real-time performance and high detection accuracy [Tang et al. 2017b].

With the advent of the deep learning paradigm, many convolutional neural network-based object recognition architectures have been recently developed, and they have been increasingly employed in object detection applications from aerial images [Yildirim and Kavzoglu 2022]. Among these, YOLO models based on the "You Only Look Once" approach are the most well-known object detectors as they can achieve real-time performance and high detection accuracy [Redmon et al. 2016]. Furthermore, the YOLOv7, the latest version of YOLO, performs instance segmentation tasks as well as object detection [Wang et al. 2022].

The main objective of this study is to automatically detect vehicles from high-resolution UAV data and to

* **Corresponding Author**

*(esrayildirim@gtu.edu.tr) ORCID ID 0000-0002-4951-0488
(sefercik@gtu.edu.tr) ORCID ID 0000-0003-2403-5956
(kavzoglu@gtu.edu.tr) ORCID ID 0000-0002-9779-3443

Cite this study

Yıldırım, E, Sefercik, U. G., & Kavzoglu, T. (2022). Automated vehicle detection and instance segmentation from high-resolution UAV imagery using YOLOv7 model. 5th Intercontinental Geoinformation Days (IGD), 116-119, Netra, India

obtain pixel-wise masks for each detected vehicle. To achieve this goal, the YOLOv7 model was employed in the vehicle dataset generated from the UAV data obtained from the Gebze Technical University (GTU) campus in Turkey. The performance of the model was investigated using Precision, Recall, and mAP evaluation metrics.

2. Method

2.1. Study area and UAV data acquisition

GTU Campus is located in Kocaeli province in the Northwest side of Turkey (Figure 1). The area is nearly 2.5 km^2 and covered by different land cover classes such as buildings, roads, and varied vegetation. The topography is mostly flat and orthometric elevation is between 2 m and 50 m.

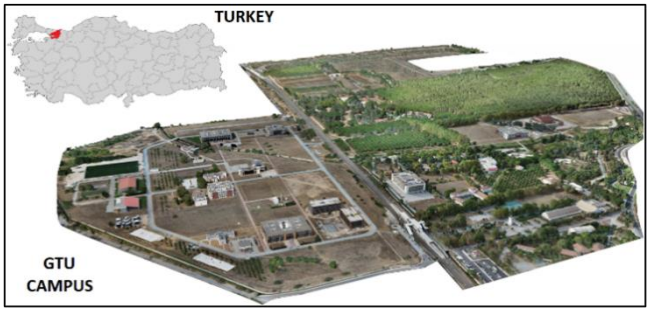


Figure 1. The UAV orthomosaic of GTU Campus

For the acquisition of high-resolution UAV orthomosaic, DJI Phantom 4 Pro V2 UAV was used. The geometry of the captured aerial photos was corrected by using 86 mobile ground control points (GCP) which were measured by CHC-i80 GNSS receiver. Table 1 shows the specifications of used materials for UAV data acquisition.

Table 1. Specifications of used materials

DJI Phantom IV Pro V2.0 UAV	
Specification	**Value**
Camera	4K, HD, 1080p, 1", effective pixel resolution 20 MP
Gimbal	3-axis (pitch, roll, yaw)
Hover accuracy range	± 0.1 m V, ± 0.5 m H (Vision) ± 0.3 m V, ± 1.5 m H (GPS)
Flight duration	Max. 30 minutes
Weight and speed	1375 g, Max. 20 m/s in S-mode
Operating temperature	0° to 40°C
CHC-i80 GNSS Receiver	
GNSS technology	GPS, GLONASS, GALILEO, BeiDou, SBAS, NavIC
Positioning accuracy RTK	± 0.8 cm H, ± 1.5 cm V with initialization reliability >99.9%
Network-RTK	Available

According to the land use and land cover in the Campus, the UAV flights were organized as polygonal, bundle-grid, and circular. While the nadir camera view is applied in polygonal flights, 70° was preferred for bundle-grid and circular flights. In all flights, minimum front and side overlap ratios were applied as 80% and 60%, respectively. The flying altitude was chosen as 80 m for polygonal and bundle grid flights and 30 m for circular flights. Totally, 8333 RGB aerial photos with ≤2.2 cm GSD were captured and used for orthomosaic generation [Sefercik et al. 2022].

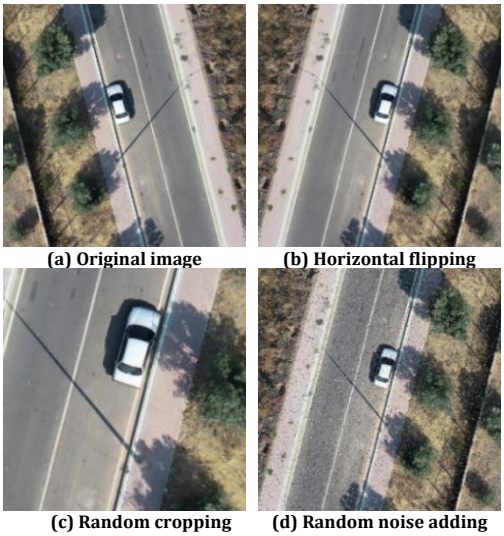
2.2. Data preparation

To build a vehicle dataset, a total of 200 images with 512x512 pixel-sized were collected from the UAV data. The vehicles in each image were manually labeled with the polygon shape using the open-source image labeling tool LabelMe [Wada, 2016] and ground-truth masks were obtained. Figure 2 depicts the annotation process of a sample image in the dataset and the corresponding ground-truth instance segmentation mask. Unlike semantic segmentation, instance segmentation identifies each target as a different instance, regardless of its class. Thus, each vehicle in the image was masked in distinct colors as can be seen from the generated ground-truth mask. Afterward, the annotated dataset was divided into training, validation, and testing datasets at a ratio of 70:20:10, respectively.

(a) Annotation of a sample image (b) Generated ground-truth mask

Figure 2. Dataset labeling process with LabelMe

In the implementation of deep learning models in object detection studies, a high-quality dataset containing a large number of images immensely enhances the training performance and prediction accuracy of the model. Therefore, different variations of the existing images were obtained in this study by utilizing several data augmentation techniques to boost the generalization capability of the model and prevent overfitting. For this aim, the training dataset was extended by horizontal flipping, randomly cropping 0 and 50 percent of the image, and adding salt and pepper noise to 5 percent of the pixels in the image. The samples of horizontally flipped, cropped and noise added images are illustrated in Figure 3.



(a) Original image (b) Horizontal flipping

(c) Random cropping (d) Random noise adding

Figure 3. Data augmentation methods

2.3. YOLOv7 algorithm

In YOLOv7, the latest detector in the YOLO series, several changes were made in architecture to improve the detection accuracy and speed of the model. An extended efficient layer aggregation network (E-ELAN) based on the use of expanding, shuffling, and merging cardinality was proposed on its backbone. This network continuously improves the learning capability without losing the original gradient path. In addition, it was introduced the utilization of a compound model scaling approach to retain the features that the model had in the initial design and thereby maintain an optimal structure. Moreover, the planned re-parametrized convolution was proposed in YOLOv7. To improve training, label designers and soft labels were introduced. This process generates two types of soft labels, namely course labels and fine labels. This mechanism is significant because it enables fine and coarse labels to be dynamically adjusted during the training process [Wang et al. 2022].

In addition to the foregoing, there are brand-new features that were not available in the previous versions of YOLO. The first one is instance segmentation that enables YOLO to segment objects pixel-wise. The second is pose estimation, which is beneficial for understanding body movement.

2.4. Design and implementation

For the experiment, Python programming language and the PyTorch framework were utilized. The training and testing processes of the model were carried out on the cloud-based Google Colab environment, which provides access to the NVIDIA Tesla T4 GPU. In the study, the transfer learning approach was adopted because a relatively small data set was generated. That is, instead of training the model end-to-end from scratch, the training process of the YOLOv7 model was initialized from the pre-trained weights obtained on the COCO dataset. The hyperparameters utilized in the training phase of the model are given in Table 2.

Table 2. Experimental hyperparameter configuration

Hyperparameters	Values
Image size	640 x 640
Epochs	100
Batch size	16
Learning rate	0.01
Momentum	0.937
Weight decay	0.0005

3. Results

As a result of the training and validation process of YOLOv7, three types of losses were generated, namely bounding box loss, segmentation loss, and objectiveness loss. As can be seen from Figure 4, all loss values showed a decreasing trend during the training and no overfitting was observed in the model. The training loss converged in the early stages of the training while the validation loss converged at the end of the training. After 100 epochs of training, the minimum value was reached in the training and validation loss curves.

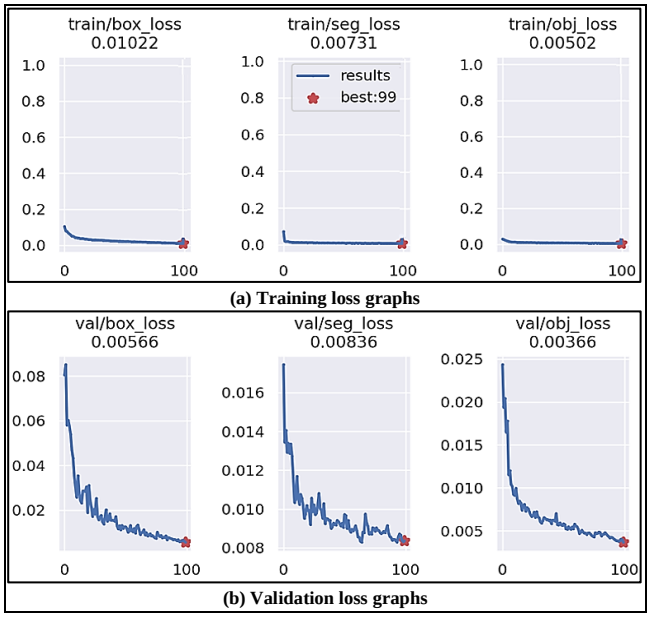


Figure 4. Convergence of the training and validation loss curves

To assess the performance of the trained model, Precision, Recall, and mean average precision (mAP), calculated at the intersection over union (IoU) threshold value of 0.50, accuracy metrics were utilized (Table 3). The model achieved the Precision, Recall, and mAP@0.50 values for both the bounding boxes and masks of vehicles as 99.79%, 97.54%, and 99.46%, respectively. Considering the computational burden, the model exhibited superior performance in detection and segmentation tasks with only training for about 49 minutes. Additionally, the prediction time of the model per test image is about 62.3 milliseconds.

Table 3. Performance evaluation of the YOLOv7 model

Metrics	Values
Precision (%)	99.79
Recall (%)	97.54
mAP@0.50 (%)	99.46
Training time	48 min 43 sec
Average detection speed (ms)	62.3

Apart from the accuracy assessment and computation time evaluation, the visual detection and segmentation results of the model on the testing dataset are given in Figure 5 to better investigate the performance of the model. It was observed that the YOLOv7 model successfully predicted the bounding boxes and segmentation masks of the vehicles, and they fitted the object boundaries correctly. Moreover, the model performed well in detecting vehicle objects with diverse backgrounds such as shadows, parking lots, and roads. It was also observed that the model gave remarkable results in identifying different vehicle types, orientations, and sizes. Even though many vehicle objects were located very adjacent to each other in the image, the model was able to accurately draw the boxes and masks of these objects. When the confidence scores in the upper left of the bounding box drawn for each vehicle were examined, it was obvious that the model detected each vehicle with a confidence score of more than 90%.

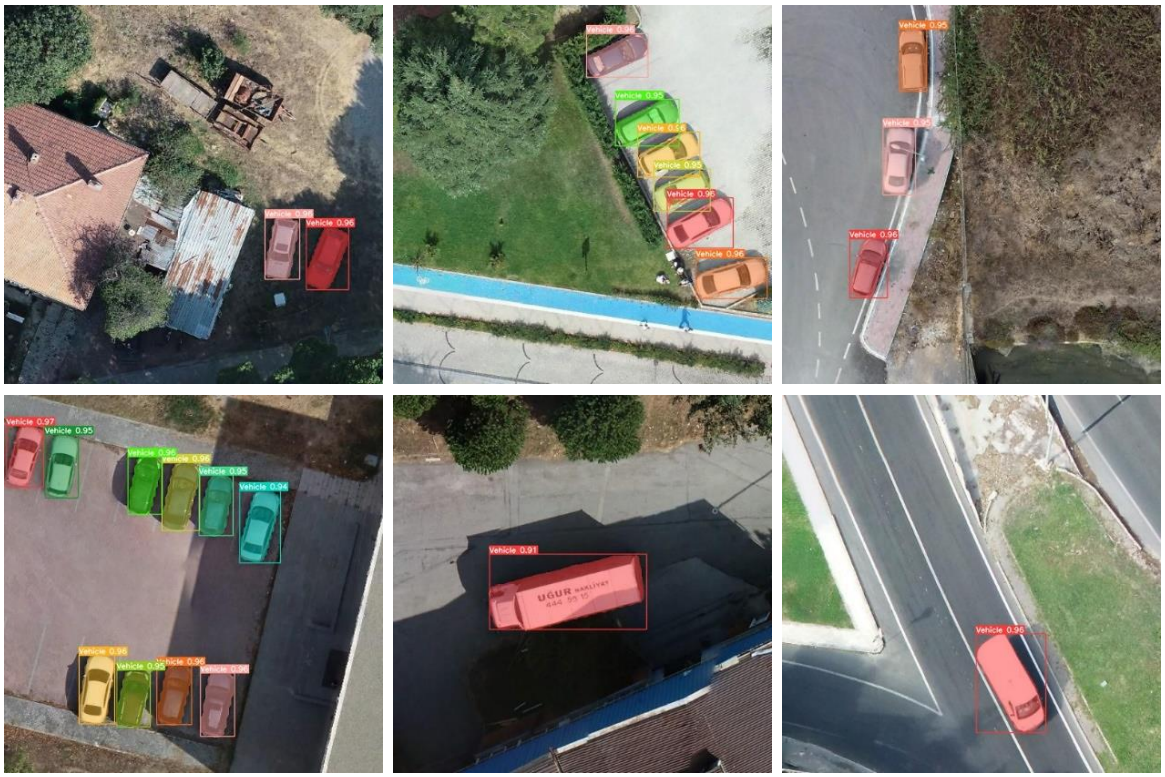


Figure 5. Examples of vehicle detection and instance segmentation results of the YOLOv7 model

4. Conclusion

In this study, automatic vehicle detection from high-resolution UAV imagery was carried out using the state-of-the-art deep learning-based YOLOv7 object detector. Besides, pixel-wise masks of each detected vehicle were obtained by utilizing the instance segmentation feature of YOLOv7. Thanks to this feature, vehicle objects belonging to a single class in an image were segmented as distinct instances, thus each vehicle could be distinguished from the other. According to the experimental results, it was observed that the model could correctly identify vehicles from UAV data with mAP of about 99%. As a result of the study, it was concluded that YOLOv7 achieved satisfactory results in terms of accuracy and speed in detecting vehicles from UAV imagery.

References

Ammar, A., Koubaa, A., Ahmed, M., Saad, A., & Benjdira, B. (2021). Vehicle detection from aerial images using deep learning: A comparative study. Electronics, 10(7), 820.

Boonpook, W., Tan, Y., Ye, Y., Torteeka, P., Torsri, K., & Dong, S. (2018). A deep learning approach on building detection from unmanned aerial vehicle-based images in riverbank monitoring. Sensors, 18(11), 3921.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition, 779-788, Las Vegas, USA.

Sefercik, U. G., Kavzoglu, T., Nazar, M., Atalay, C., & Madak, M. (2022). Creation of a virtual tour .exe utilizing very high-resolution RGB UAV data. International Journal of Environment and Geoinformatics, 9(4), 151-160.

Shao, Z., Cheng, G., Ma, J., Wang, Z., Wang, J., & Li, D. (2021). Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic. IEEE Transactions on Multimedia, 24, 2069-2083.

Tang, T., Deng, Z., Zhou, S., Lei, L., & Zou, H. (2017a). Fast vehicle detection in UAV images. International Workshop on Remote Sensing with Intelligent Processing, 1-5, Shanghai, China.

Tang, T., Zhou, S., Deng, Z., Zou, H., & Lei, L. (2017b). Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. Sensors, 17(2), 336.

Wada, K. (2016). Labelme: Image polygonal annotation with python.

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696.

Yildirim, E., & Kavzoglu, T. (2022). Ship Detection in Optical Remote Sensing Images Using YOLOv4 and Tiny YOLOv4. Innovations in Smart Cities Applications, 913-924, Springer, Cham.

Yildirim, E., Nazar, M., Sefercik, U. G., & Kavzoglu, T. (2022). Stone Pine (Pinus Pinea L.) Detection from High-Resolution UAV Imagery Using Deep Learning Model. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 441-444, Kuala Lumpur, Malaysia.