

Received November 4, 2020, accepted November 18, 2020, date of publication November 24, 2020, date of current version December 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040290

# VAID: An Aerial Image Dataset for Vehicle Detection and Classification

HUEI-YUNG LIN<sup>1</sup>, (Senior Member, IEEE), KAI-CHUN TU<sup>2</sup>, AND CHIH-YI LI<sup>2</sup>

<sup>1</sup>Department of Electrical Engineering, Advanced Institute of Manufacturing with High-Tech Innovation, National Chung Cheng University, Minhsiung 621, Taiwan

<sup>2</sup>Department of Electrical Engineering, National Chung Cheng University, Minhsiung 621, Taiwan

Corresponding author: Huei-Yung Lin (lin@ee.ccu.edu.tw)

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 106-2221-E-194-004, and in part by the Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI) through the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.


**ABSTRACT** The availability of commercial UAVs and low-cost imaging devices has made the airborne imagery popular and widely available. The aerial images are now extensively used for many applications, especially in the area of intelligent transportation systems. In this work, we present a new aerial image dataset, VAID (Vehicle Aerial Imaging from Drone), for the development and evaluation of vehicle detection algorithms. It contains about 6000 images captured under different traffic conditions, and annotated with 7 common vehicle categories for network training and testing. We compare the of vehicle detection results using the current state-of-the-art network architectures and various aerial image datasets. The experiments have demonstrated that training the networks using our VAID dataset can provide the best vehicle detection results. Our aerial image dataset is made available publicly at <http://vision.ee.ccu.edu.tw/aerialimage/> and the code is available at [https://github.com/KaiChun-RVL/VAID\\_dataset](https://github.com/KaiChun-RVL/VAID_dataset).

**INDEX TERMS** Aerial image dataset, vehicle detection and classification, convolutional neural network.

## I. INTRODUCTION

Nowadays, the availability of low-cost image acquisition systems and easy-to-use unmanned aerial vehicles (UAVs) has made the aerial imaging more convenient and popular. It is now possible to acquire a large number of high-quality aerial images without elaborate planning and a considerable amount of time. The aerial images have been adopted in many tasks such as cartography, precision agriculture, landscape archaeology and urban studies for many decades. One specific application is to detect and classify the vehicles in aerial images. It is gradually adopted to intelligent transportation for vehicle identification, traffic flow estimation and parking space allocation, etc. Thus, it is the future trend to use aerial images for transportation and vehicle related applications.

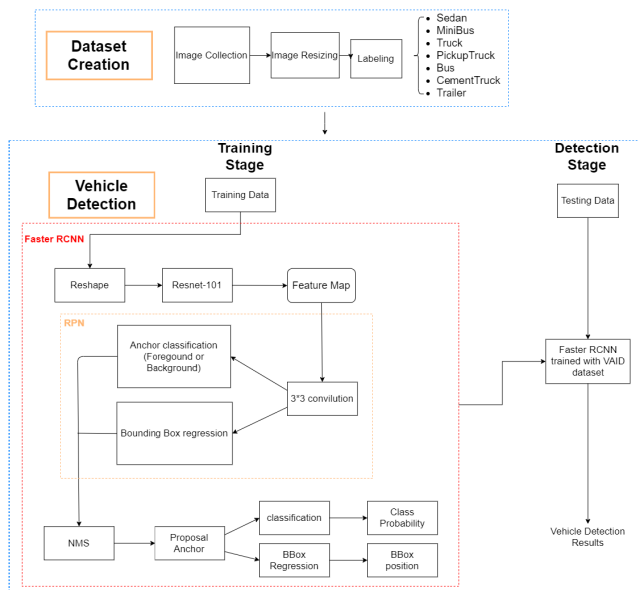
The aerial images are able to cover a variety of scenes from the sky, consisting of forests, rivers, buildings, bridges and roads, etc. In remote sensing applications, various kinds of satellite imagery are used in the fields of geography, land surveying and many earth science disciplines. They are also

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin .

frequently used for the detection of man-made structures, both static constructions and movable targets such as vehicles and vessels. Due to the recent progress on machine learning techniques, we are now able to achieve high object detection rates in cluttered scenes. The detection and classification of vehicles using aerial images have become more feasible with deep neural networks.

The techniques for vehicle detection using aerial images can be classified into two categories, the conventional machine learning methods and the deep learning approaches [1]. For the machine learning methods, low-level image features such as edge, corner, shape, texture and color are extracted for training and classification. Shao *et al.* propose a vehicle detection framework which use local binary patterns combined with histograms of oriented gradient for vehicle detection [2]. The differences in color are used for detection with the blob-like areas extracted from prominent color and grayscale features [3]. There also exist traditional computer vision techniques which use frame difference [4] and optical flow [5] for moving vehicle detection.

For the deep learning approaches, convolutional neural networks (CNNs) have significant improvement on object



**FIGURE 1.** A system flowchart of the proposed method for vehicle detection and classification in aerial images. It consists of creating the aerial image dataset and testing the network architecture for vehicle detection.

detection and classification in the past few years [6]. Liu *et al.* adopt the Single Shot MultiBox Detector (SSD) [7] with the network trained using DLR Vehicle Aerial dataset (DLR-MVDA) and Vehicle Detection in Aerial Imagery dataset (VEDAI) for vehicle detection [8]. In [9], Sommer *et al.* use these two aerial image datasets to evaluate the performance of Fast R-CNN [10] and Faster R-CNN [11] networks. Lu *et al.* [12] evaluate the performance of YOLO, YOLOv2 and YOLOv3 [13]–[15] using COWC, VEDAI and DOTA datasets for training and testing. Similarly, the performance comparison of Faster R-CNN and YOLOv3 for vehicle detection is carried out by Benjdira *et al.* using their aerial image dataset [16].

Compared to the general object detection tasks, there are additional issues for the vehicle detection in aerial images as follows.

- The target size is usually much smaller.
- The targets tend to have monotonic appearance.
- The images are easily affected by illumination changes.
- There might be a large number of vehicles in an image.
- The target aspect ratio could be large.

In this paper, we introduce a new aerial image dataset, VAID (Vehicle Aerial Imaging from Drone), for vehicle detection and classification. Extended from the previous work using modified Faster R-CNN [17], we compare the advantages, disadvantages and results of vehicle detection in aerial images with several well-known network architectures. Figure 1 shows the system flowchart of the proposed framework for the evaluation of vehicle detection algorithms. It consists of creating our VAID image dataset, and training and testing on the aerial images using various network structures for comparison.

## II. RELATED WORKS

Due to the applications in traffic control, parking management, and security purposes, the detection of vehicles in aerial images has been studied for many decades [18]–[20]. Compared with the vehicle detection from close range or ground viewpoints, the technical requirements are very different since the targets are much smaller and contain less features to distinguish from the environment [21], [22]. The image quality is also degraded in general due to the long range acquisition in the atmosphere. To detect and recognize objects from the air, remote sensing is one of the earliest research fields which adopt the image-based approach [1]. Many techniques have been developed for a variety of applications, and are not restricted to the detection of ground objects. It is then followed by the computer vision community to investigate the object detection or specifically vehicle detection algorithms in airborne images.

Prior to the popularity and success of deep neural networks adopted for object detection and recognition, conventional machine learning methods heavily rely on hand-crafted feature extraction for image classification. When applied to the vehicle detection from aerial images, commonly used features including shape, color, corner, texture, disparity, as well as histogram of oriented gradient (HOG) and scale-invariant feature transform (SIFT). They are combined with various classifiers such as support vector machine (SVM), random forest (RF), AdaBoost, and bag-of-words (BoW) for detection and recognition [23]–[25]. Although more recent works on aerial image analysis have gradually moved to deep learning based approaches, there still exist newly proposed conventional methods because of the low complexity and computational cost. Nevertheless, these techniques are designed for some specific uses rather than general purposes.

Among the few noticeable improvement for traditional methods, Chen *et al.* present a fast classification algorithm using a set of sparse representation dictionaries [26]. A multi-order descriptor is proposed to extract the vehicle feature in aerial images. By introducing the superpixel segmentation and patch orientation, their results on high-resolution images are superior to those obtained from commonly used HOG+SVM, LBP+PLS (Local Binary Patterns and Partial Least Squares), and sparse representation methods. Xu *et al.* proposed an enhanced Viola-Jones detector for vehicle identification from aerial imagery [27]. A road orientation adjustment stage is adopted to improve the original isotropic detection results. The method is further applied to improve the accuracy of vehicle tracking. Liu *et al.* also start the design of a vehicle detector from the orientation issue [28]. They develop a fast oriented region search algorithm to detect the position, size, and orientation of an object. A modified vector of locally aggregated descriptors is used to represent an object and distinguish the proposals from the background. The experiments carried out on public datasets, VEDAI and Munich 3K, have shown some significant results compared to the existing approaches. For training data collecting and labeling, Cao *et al.* propose an efficient

**TABLE 1.** A summary of the aerial image datasets currently available and used for our evaluation and comparison of vehicle detection and classification algorithms. It shows the number of images in the dataset, the image resolution, the actual scale of a pixel, and the typical size of a vehicle in pixel. Some images from the datasets are shown in Figure 2.

Dataset	Number of Images	Image Resolution	Pixel Scale	Vehicle Size
VEDAI	1,250	512 × 512	25cm	10 × 20
		1,024 × 1,024	12.5cm	20 × 40
COWC	53	2,000 × 2,000	15cm	24 × 48
		19,000 × 19,000		
DLR-MVDA	20	5,616 × 3,744	13cm	20 × 40
KIT-AIS	241	300 – 1,800	12.5cm – 18cm	15 × 25
				20 × 40
VAID	5,985	1137 × 640	12.5cm	20 × 40

and labor-light scheme which only works on region-level group annotation [29]. A weakly supervised, multi-instance learning algorithm is developed to learn the weak labels. A multi-instance SVM is then trained to classify from the density map derived from the positive regions. To deal with the scale and orientation variations, shadow, and partial occlusion, Cao *et al.* present an affine-function transformation-based object matching framework [30]. Similar to the previous approach, superpixel segmentation is adopted to generate non-redundant patches, followed by detection and localization with a threshold matching cost. Their results obtained from two UAV image datasets demonstrate that good performance can be derived comparable to Faster R-CNN.

With the recent success of convolutional neural networks for object recognition, they have also been applied to aerial images for vehicle detection. Since the target size is one major issue for aerial imagery, the algorithms often need to emphasize the capability of small object detection. In [31], Zhong *et al.* propose a method which cascades two convolutional neural networks to improve the detection accuracy without decreasing the speed. The first network is used to generate a set of vehicle-like regions, followed by the second network for feature extraction and decision making. They adopt multi-feature maps with different hierarchies and scales, and achieve high recall rates and low computation costs on two public aerial image datasets. Mandal *et al.* propose a single-stage detector, AVDNet, specifically designed for small-size vehicle detection in aerial images [32]. The feature vanishing problem for small objects is mitigated by the use of residual blocks at multiple scales. Their algorithms are evaluated on four datasets, and a better performance compared to the well-known frameworks such as YOLOv3, Faster R-CNN and RetinaNet is reported. For the applications which require *in situ* real-time processing, He *et al.* present a compressed MobileNet capable of 110 fps processing speed [33]. It is built on the light weight network MobileNet and considers the tradeoff between accuracy and computation. Their algorithm is also implemented on a mobile phone with acceptable 15 fps inference speed. With the similar objective to reduce the hardware requirement, Ringwald *et al.* evaluate several popular detection frameworks for best accuracy/speed trade-off [34]. They build upon SSD to construct a network, UAV-Net, for aerial imagery. The impressive 0.4

MB model size makes it suitable for real-time operations on an embedded platform such as Jetson TX2.

### III. VAID AND AERIAL IMAGES DATASETS

Currently, there are not many public datasets available for vehicle detection in aerial images. Some datasets, such as VIRAT video dataset, are designed for video surveillance and action recognition [35]. For the existing aerial image datasets, there are also some problems such as containing only a very limited number of categories, imprecise bounding boxes, small image sizes, etc. Several popular datasets for vehicle detection in aerial images include VEDAI, COWC, DLR-MVDA, DOTA and KIT-AIS. The description of these datasets are shown in Table 1. VEDAI (Vehicle Detection in Aerial Imagery) dataset is made available by Razakarivony and Jurie [36], and originated from the public Utah AGRC database.<sup>1</sup> It contains a total of 1,250 RGB and NIR images with the resolution of 512 × 512 and 1024 × 1024 captured at about the same height. The dataset is manually annotated with 9 classes of objects ('plane', 'boat', 'camping car', 'car', 'pick-up truck', 'tractor', 'truck', 'van', and others) and a total of 2,950 samples. Each image consists of 5 vehicles in average, and the vehicle size is about 0.7% of an image. The annotation of each sample includes the sample class, the center point coordinates, direction and the four corner point coordinates of the ground-truth. 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.

COWC (Cars Overhead With Context) dataset created at LLNL contains the overhead imagery collected from six major cities [37]. All images are standardized to 15 cm per pixel at ground level, so the vehicles span about 24 to 48 pixels. The objective of this dataset is mainly for vehicle counting, so the annotation is different from the datasets for vehicle detection and classification. The labeled images in COWC dataset only mark the center point of a vehicle with a red dot. It does not provide the category or bounding box information. There are totally 32,716 annotated vehicles in the dataset, with additional 58,247 negative samples.

<sup>1</sup><https://gis.utah.gov/data/>



**FIGURE 2.** Some images from our VAID dataset and four other datasets for comparison and evaluation: VEDAI, DLR-MVDA, COWC and KIT-AIS. The vehicles in aerial images usually appear much smaller than most objects in general image recognition and classification datasets.

In DOTA (Dataset for Object deTecton in Aerial images) dataset, 2,806 aerial images from different sensors and platforms are collected at the resolution of  $4000 \times 4000$  [38]. It contains more than 188k instances with different scales, orientations, shapes, and labeled by quadrilaterals instead of commonly used bounding boxes. Although the dataset is large in terms of the number of images and instances per image, it aims to provide for general purpose use with

only two vehicle classes (large and small) out of the total 15 categories. This makes it unsuitable for object detection on vehicle specific applications.

DLR-MVDA dataset contains 20 large scale-aerial images [39]. The images are captured with more realistic road scenes and the vehicle detection is more challenging. KIT-AIS is a dataset with the images taken from an airplane at about 330 m above the ground [40]. It has 228 high resolution



**FIGURE 3.** In our VAID dataset, 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. The sample images are shown in the figure from the left to the right accordingly.

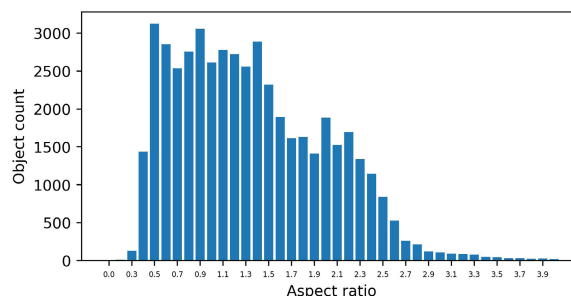


**FIGURE 4.** Some aerial images in our VAID dataset captured using a drone. It consists of different road types and traffic scenes.

images ( $5161 \times 3744$ ), but there is only one annotated vehicle category for network training.

This paper introduces a new vehicle detection dataset, VAID (Vehicle Aerial Imaging from Drone), with the aerial images captured by a drone.<sup>2</sup> We collect about 6,000 aerial images under different illumination conditions and viewing angles from different places in Taiwan. The images are taken with the resolution of  $1137 \times 640$  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 shows some example images from our VAID dataset as well as four other datasets, VEDAI, DLR-MVDA, KIT-AIS and COWC. It can be seen that the vehicles are much smaller compared to the objects in general recognition and classification datasets.

Although the vehicles are divided into the 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



**FIGURE 5.** The distribution of the object’s aspect ratio.

the names of the classes and the bounding box coordinates. Figure 3 shows several cropped vehicle images from different categories.

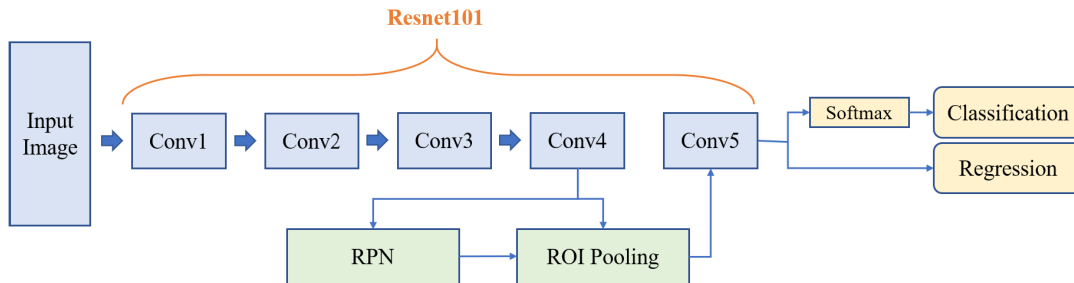
The images in the dataset are taken by a drone (DJI’s Mavic Pro). To keep the sizes of the vehicles consistent in all images, the altitude of the drone is maintained at about 90 – 95 meters from the ground during video recording. The output resolution is  $2720 \times 1530$  at 2.7K and the frame rate 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  $110 \times 45$  pixels. In the VAID dataset, the images are scaled to the resolution of  $1137 \times 640$ , and a sedan in the images is about the size of  $40 \times 20$  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 4 shows some of the dataset images with various road and traffic scenes. There are totally 7 categories for vehicle classification in our VAID dataset. The images

<sup>2</sup>VAID Dataset: <http://vision.ee.ccu.edu.tw/aerialimage/>

**TABLE 2.** Some statistics of our VAID dataset. It shows the number of images and the number of vehicles in each category for three different types of image acquisition locations: university campus, urban area, suburb.

	Image	Sedan	Minibus	Truck	Pickup truck	Bus	Cement truck	Trailer
University Campus	3,257	11,385	406	605	611	292	17	36
Urban Area	1,118	18,349	95	1,014	822	225	6	10
Suburb	1,610	10,596	0	1,568	1,578	63	168	758
Total	5,985	40,330	501	3,187	3,011	580	191	804



**FIGURE 6.** The modified Faster R-CNN architecture for vehicle detection and classification proposed in this paper. It also serves as the baseline for the comparison with other network models.

**TABLE 3.** The number of vehicles in different class used for training, validation and testing for each class in the VAID dataset.

Class	Sedan	Minibus	Truck	Pickup Truck	Bus	Cement Truck	Trailer
Training	9,976	117	806	705	133	40	195
Validation	10,024	116	759	769	158	54	233
Testing	20,330	268	1622	1,537	289	97	376

are divided into 3 regions, namely, urban area, suburb and university campus. Some statistics are shown in Table 2. Another important statistic regarding the distribution of the object's aspect ratio is shown in Figure 5.

#### IV. EXPERIMENTS AND EVALUATION

To evaluate the effectiveness of the proposed VAID dataset, two experiments are carried out with different object detection techniques and several aerial image datasets. First, our VAID dataset is used to train five popular object detection architectures, including Faster R-CNN, YOLOv4, MobileNetv3, RefineDet and U-Net, for performance comparison. The network architecture with the best performance for vehicle detection and classification in this experiment is considered for further evaluation. Second, the selected network structure is trained separately using different aerial image datasets, including VEDAI, DLR-MVDA, COWC, KIT-AIS and VAID. The trained neural network models are then tested on a new dataset for performance evaluation. It provides the comparison on the effectiveness of the training sets. The hardware used for the evaluation is a PC with an Intel i7-8700k CPU, 16GB RAM and Nvidia GTX1080Ti GPU. The software tools for the development include Ubuntu 16.04, cuda 10.0, cudnn 7.4.2, tensorflow-gpu 1.4, pytorch 1.4.0, Keras 2.2.4 and opencv-python 4.2.0.

In our VAID dataset, there are totally 5,985 aerial images with the vehicles classified into seven categories. It is split into three parts, with 1,512 images for training, 1,534 images for validation, and 2,939 images for testing. Table 3 shows the detailed information for each class in training, validation, and

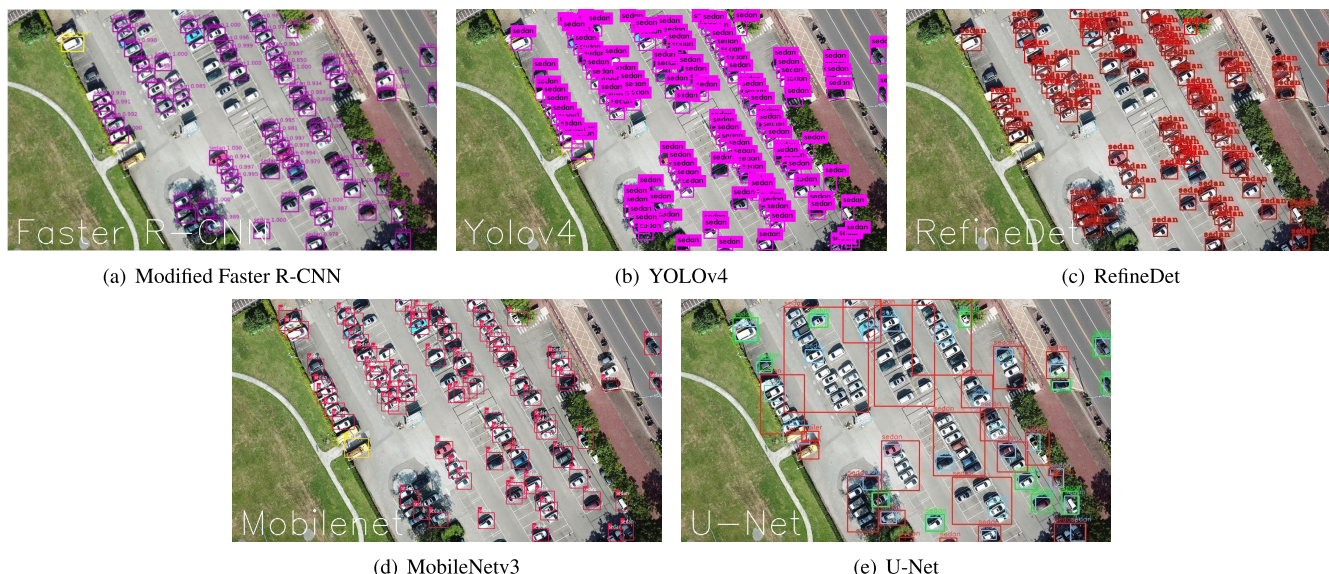
testing sets. It can be seen that the number of training samples is unbalanced among the classes. Thus, training the network with fewer samples is an important issue to achieve better classification results. We use our modified Faster R-CNN model as the baseline for benchmarking. First, the ReLU (Rectified Linear Unit) activation function is used on the RPN (Region Proposal Network) layer. As shown in Table 4, this provides slightly better results compared to the original network and the modifications with other activation functions. Second, we replace the feature extraction model with ResNet-101. Finally, the aspect ratio is changed from  $[0.5, 1, 2]$  to  $[0.2, 0.5, 1, 1.2, 2]$ . Our modified Faster R-CNN architecture is illustrated in Figure 6. For the evaluation of other network models (YOLOv4, MobileNetv3, RefineDet, U-Net), we use the default settings without further changes.

The network model evaluation on the VAID dataset is tabulated in Table 5. It shows the mAP (mean average precision), precision, recall and F-1 score for Modified Faster R-CNN, YOLOv4, MobileNetv3, RefineDet and U-Net.<sup>3</sup> Figure 7 shows the vehicle detection results of a parking lot image using different network models. There are several important observations from the network outputs and evaluation results. First, Modified Faster R-CNN has 90.12% mAP but with very low precision. This is due to a large number of incorrect predictions of the 300 anchor boxes in the network model. Second, U-Net reports very high precision but only with a relatively low mAP (at 85.38%). It is caused by the use of pixel-level segmentation to define the bounding box

<sup>3</sup>The code is available at [https://github.com/KaiChun-RVL/VAID\\_dataset](https://github.com/KaiChun-RVL/VAID_dataset)

**TABLE 4.** Several different activation functions used in the modified Faster R-CNN for comparison (mAP).

	Original Faster R-CNN	Modified Faster R-CNN	Modified Faster R-CNN (softplus)	Modified Faster R-CNN (ELU)	Modified Faster R-CNN (ReLU)
Sedan	90.0%	90.2%	90.2%	90.2%	<b>90.2%</b>
Minibus	95.0%	<b>97.9%</b>	92.2%	96.6%	95.6%
Truck	83.4%	84.9%	<b>87.5%</b>	87.1%	86.8%
Pickup Truck	76.8%	78.7%	79.3%	78.6%	<b>79.6%</b>
Bus	88.9%	89.4%	89.6%	89.8%	<b>90.3%</b>
Cement Truck	94.2%	97.6%	93.8%	97.6%	<b>98.1%</b>
Trailer	<b>85.6%</b>	83.8%	81.5%	84.7%	84.5%
Average	87.7%	88.9%	87.7%	89.2%	<b>89.3%</b>



**FIGURE 7.** The results and comparison of vehicle detection in a parking lot view aerial image using the different network models, modified Faster R-CNN, YOLOv4, RefineDet, MobileNetv3 and U-Net.

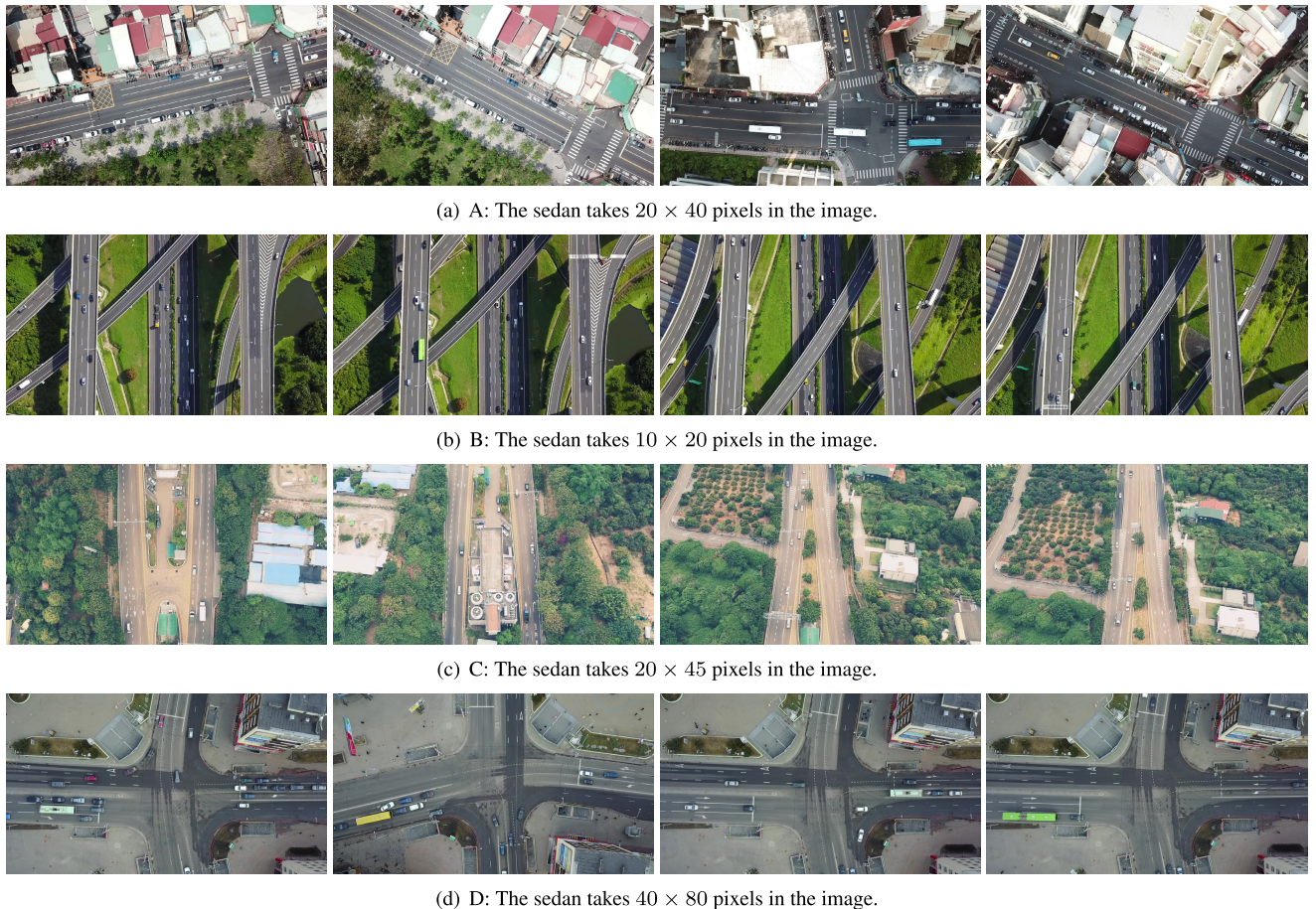
**TABLE 5.** The network model evaluation on the VAID dataset. It shows the mAP, precision, recall and F-1 score for Modified Faster R-CNN, YOLOv4, MobileNetv3, RefineDet and U-Net.

	Modified Faster R-CNN	YOLOv4	MobileNetv3	RefineDet	U-Net
Sedan	90.22%	<b>98.49%</b>	70.46%	89.08%	67.20%
Minibu	90.80%	<b>96.04%</b>	89.02%	90.14%	94.36%
Truck	89.34%	<b>96.44%</b>	64.92%	82.21%	83.46%
Pickup Truck	<b>88.93%</b>	57.25%	75.73%	84.59%	82.80%
Bus	90.87%	<b>97.03%</b>	87.67%	90.46%	97.84%
Cement Truck	90.96%	69.94%	90.30%	80.68%	<b>91.24%</b>
Trailer	89.75%	<b>95.45%</b>	78.14%	86.64%	80.74%
mAP	90.12%	<b>96.91%</b>	73.9%	86.26%	85.38%
Precision	0.041	<b>0.94</b>	0.2818	0.2420	0.9099
Recall	<b>0.9755</b>	0.97	0.8807	0.9640	0.9016
F1 score	0.0731	<b>0.96</b>	0.4178	0.3739	0.9057

for U-Net, which reduces the number of false detection. However, if the objects are very close to each other, they tend to be considered as a single large target as shown in Figure 7(e). Third, MobileNetv3 has the lowest mAP among all network models. As indicated in Figure 7(d), it cannot deal with the nearby objects very well. The main problem is the feature map extraction. For other models, including YOLOv4, RefineDet and U-Net, the next higher dimension feature map is used to regenerate the feature map. However, the use of the raw feature map makes MobileNet hard to distinguish the object features, and have the bounding box

regression perform well. Finally, YOLOv4 provides the best performance in terms of mAP, precision, F1 score (and with the recall slightly worse than Modified Faster R-CNN), and is selected for the experiments on the dataset evaluation. In general, all network models perform fairly well for the vehicle detection. However, if the viewing angle of the camera with respect to the ground is too large, all models cannot provide good results.

In the second experiment, we evaluate the aerial image datasets DLR-MVDA, VEDAI, COWC, KIT-AIS and VAID using YOLOv4 for vehicle detection. The network is trained



**FIGURE 8.** 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.

using the individual image datasets separately and tested on a new dataset (with the aerial images acquired from different places) for performance evaluation. Because COWC and KIT-AIS provide only one category ('vehicle'), we modify the labels of all datasets to a single vehicle class as a basis for comparison. If an image is larger than  $1137 \times 640$ , it is cut to several  $1137 \times 640$  sub-images for processing. Some classes which are not vehicle related such as 'boat', 'plane' and 'other' in VEDAI are removed from the dataset. The annotation in COWC only provides a dot on the center of a target, so we set a  $20 \times 20$  bounding box on each object for IoU (Intersection over Union) computation.

The new testing data for the evaluation of different network models are selected from four other image acquisition scenarios. Figure 8 shows some example images in the testing dataset. Scene A consists of the aerial images acquired from two different locations in a city (see Figure 8(a)). Scenes B, C, D are the airborne traffic scene videos obtained from YouTube, which are recorded above two highways and one expressway in Taiwan, and a crossroad in Belarus. As shown in Figure 8(b), the highway images in Scene B contain several roads in different altitudes, and the objects may have different scales even belong to the same category. In Scene C,

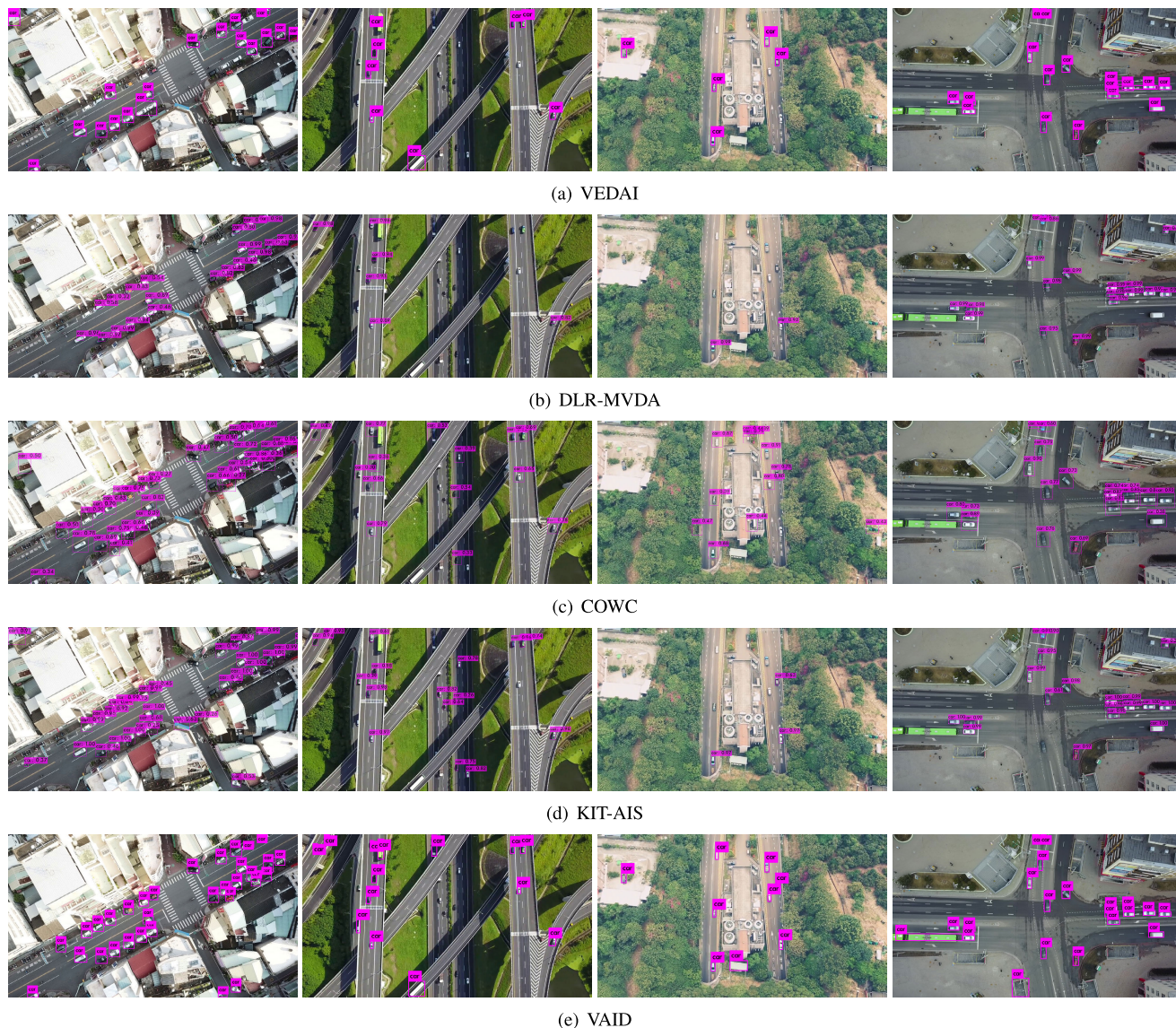
**TABLE 6.** The evaluation results (mAPs) of different scenes (A, B, C and D) using VEDAI, DLR-MVDA, COWC, KIT-AIS and our VAID datasets for network training. VEDAI, DLR-MVDA, COWC, KIT-AIS and VAID in the training set contain 967, 3,046, 86, 912, 113 and 967 images, respectively. The image sizes used in the training data are  $1,024 \times 1,024$ ,  $1,137 \times 640$ ,  $1,137 \times 649$ ,  $767 \times 669$  and  $1,024 \times 1,024$ , respectively. Scenes A, B, C and D in the testing set contain 99, 17, 31 and 35 images, respectively. The sizes of the images in Scene A, B, C and D are  $1,137 \times 640$ ,  $1,280 \times 720$ ,  $1,280 \times 720$  and  $1,920 \times 1,080$ , respectively.

	VEDAI	DLR-MVDA	COWC	KIT-AIS	VAID
Scene A	78.17%	62.22%	81.32%	77.01%	<b>93.17%</b>
Scene B	61.44%	69.47%	31.62%	<b>77.39%</b>	60.63%
Scene C	62.51%	62.88%	64.28%	54.06%	<b>80.75%</b>
Scene D	83.00%	81.17%	83.65%	91.05%	<b>97.91%</b>
Average	71.28%	68.94%	65.22%	74.88%	<b>83.12%</b>

as illustrated in Figure 8(c), there are some vehicles parking on the roads with different orientations. The images in Scene D consist of the road scenes acquired in Belarus, with the vehicle size larger than those in the training dataset (see Figure 8(d)).

Table 6 shows the evaluation results of different scenes (A, B, C and D) using VAID, VEDAI, DLR-MVDA, COWC and KIT-AIS as training datasets. The details and specifications of the training and testing data are also provided.





**FIGURE 9.** The vehicle detection results using YOLOv4 trained on different aerial image datasets, (a) VEDAI, (b) DLR-MVDA, (c) COWC, (d) KIT-AIS, and (e) VAID. The images from left to right correspond to Scene A, B, C and D, respectively. Scenes A, B and C contain the road images acquired in Taiwan, and the vehicles such as trucks and trailers are rare in other datasets.

Although the IoU threshold for VAID, VEDAI and KIT-AIS is 0.5, it is set as 0.25 for DLR-MVDA and COWC. This is due to the imprecise ground-truth bounding boxes for DLR-MVDA (too small) and COWC (too large), and the mAPs will be close 0 if the IoU of 0.5 is used. Figure 9 shows some example images of the detection results using different training datasets. Scenes A, B and C contain the road images acquired in Taiwan, and the vehicles such as trucks and trailers are rare in other datasets. This causes the classification problem for certain types of vehicles, and results in low mAP for VEDAI, DLR-MVDA and COWC. Using our VAID dataset for network training, high accuracy results are obtained for Scenes A, C and D. Our low mAP result of Scene B is mainly due to the much smaller vehicle size (about  $20 \times 10$ ) compared to those in VAID (about

$40 \times 20$ ) for training. In Scene B, the vehicles in the images are at different elevations (on the viaducts). Our dataset images are collected at approximately the same height, while other datasets including KIT-AIS, MVDA, and VEDAI contain images taken at different heights. Moreover, KIT-AIS has the images not only acquired from multiple heights, but also similar to Scene B, as illustrated in Figure 2(d). Consequently, the networks trained using our dataset perform not as good as using VEDAI, DLR-MVDA and KIT-AIS in Scene B. Nevertheless, the overall accuracy for the network trained on our dataset provides much better performance.

### V. CONCLUSION

In this paper, we present a new aerial image dataset for the development and evaluation of vehicle detection algorithms.

The dataset contains 6,000 images captured under different illumination conditions, and are available for public access. To illustrate the effectiveness of our dataset, the performance evaluation of vehicle detection techniques is carried out on widely used network architectures and training datasets. The experimental results have demonstrated that training the deep neural networks using our VAID dataset can provide the best vehicle detection rate on an independent testing dataset. In the future, the aerial image dataset will be extended with diverse imaging conditions and maintained for public access and benchmarking.

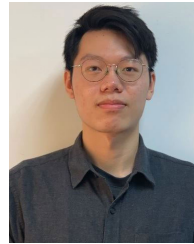
## REFERENCES

- [1] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.
- [2] W. Shao, W. Yang, G. Liu, and J. Liu, "Car detection from high-resolution aerial imagery using multiple features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 4379–4382.
- [3] D. Lenhart, S. Hinz, J. Leitloff, and U. Stilla, "Automatic traffic monitoring based on aerial image sequences," *Pattern Recognit. Image Anal.*, vol. 18, no. 3, pp. 400–405, Sep. 2008.
- [4] A. C. Shastry and R. A. Schowengerdt, "Airborne video registration and traffic-flow parameter estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 4, pp. 391–405, Dec. 2005.
- [5] H. Yalcin, M. Hebert, R. Collins, and M. J. Black, "A flow-based approach to vehicle detection and background mosaicking in airborne video," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, p. 1202.
- [6] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.
- [7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and C. Alexander Berg, "SSD: Single shot multibox detector," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. The Netherlands, 2016.
- [8] T. Tang, S. Zhou, Z. Deng, L. Lei, and H. Zou, "Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks," *Remote Sens.*, vol. 9, no. 11, p. 1170, Nov. 2017.
- [9] L. W. Sommer, T. Schuchert, and J. Beyerer, "Fast deep vehicle detection in aerial images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 311–319.
- [10] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
- [11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [12] J. Lu, C. Ma, L. Li, X. Xing, Y. Zhang, Z. Wang, and J. Xu, "A vehicle detection method for aerial image based on YOLO," *J. Comput. Commun.*, vol. 06, no. 11, pp. 98–107, 2018.
- [13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [14] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
- [15] J. Redmon and A. Farhadi, "YOLO v.3," Univ. Washington, Seattle, WA, USA, Tech. Rep., 2018.
- [16] B. Benjdira, T. Khurshed, A. Koubaa, A. Ammar, and K. Ouni, "Car detection using unmanned aerial vehicles: Comparison between faster R-CNN and YOLOv3," in *Proc. 1st Int. Conf. Unmanned Vehicle Syst.-Oman (UVS)*, Feb. 2019.
- [17] C.-Y. Li and H.-Y. Lin, "Vehicle detection and classification in aerial images using convolutional neural networks," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2020, pp. 775–782.
- [18] A. Regester and V. Paruchuri, "Using computer vision techniques for parking space detection in aerial imagery," in *Advances in Computer Vision*, K. Arai and S. Kapoor, Eds. Cham, Switzerland: Springer, 2020, pp. 190–204.
- [19] G. Guido, V. Gallelli, D. Rogano, and A. Vitale, "Evaluating the accuracy of vehicle tracking data obtained from unmanned aerial vehicles," *Int. J. Transp. Sci. Technol.*, vol. 5, no. 3, pp. 136–151, Oct. 2016.
- [20] K. Sakai, T. Seo, and T. Fuse, "Traffic density estimation method from small satellite imagery: Towards frequent remote sensing of car traffic," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 1776–1781.
- [21] Y. Zhou, L. Liu, L. Shao, and M. Mellor, "Fast automatic vehicle annotation for urban traffic surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 6, pp. 1973–1984, Jun. 2018.
- [22] C. Lai, H. Lin, and W. Tai, "Vision based ADAS for forward vehicle detection using convolutional neural networks and motion tracking," in *Proc. 5th Int. Conf. Vehicle Technol. Intell. Transp. Syst. (VEHITS)*, O. Gusikhin and M. Helfert, Eds. Heraklion, Crete, Greece: SciTePress, May 2019, pp. 297–304.
- [23] B. Ma, Z. Liu, F. Jiang, Y. Yan, J. Yuan, and S. Bu, "Vehicle detection in aerial images using rotation-invariant cascaded forest," *IEEE Access*, vol. 7, pp. 59613–59623, 2019.
- [24] S. U. Raj, M. Veera Manikanta, P. S. Sai Harsitha, and M. Judith Leo, "Vacant parking lot detection system using random forest classification," in *Proc. 3rd Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Mar. 2019, pp. 454–458.
- [25] H. Zhou, L. Wei, C. P. Lim, D. Creighton, and S. Nahavandi, "Robust vehicle detection in aerial images using Bag-of-Words and orientation aware scanning," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7074–7085, Dec. 2018.
- [26] Z. Chen, C. Wang, H. Luo, H. Wang, Y. Chen, C. Wen, Y. Yu, L. Cao, and J. Li, "Vehicle detection in high-resolution aerial images based on fast sparse representation classification and multiorder feature," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2296–2309, Aug. 2016.
- [27] Y. Xu, G. Yu, X. Wu, Y. Wang, and Y. Ma, "An enhanced violajones vehicle detection method from unmanned aerial vehicles imagery," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 7, pp. 1845–1856, Jul. 2017.
- [28] Liu, Ding, Zhu, Xiu, Li, and Li, "Vehicle detection in aerial images using a fast oriented region search and the vector of locally aggregated descriptors," *Sensors*, vol. 19, no. 15, p. 3294, Jul. 2019.
- [29] L. Cao, F. Luo, L. Chen, Y. Sheng, H. Wang, C. Wang, and R. Ji, "Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning," *Pattern Recognit.*, vol. 64, pp. 417–424, Apr. 2017.
- [30] S. Cao, Y. Yu, H. Guan, D. Peng, and W. Yan, "Affine-function transformation-based object matching for vehicle detection from unmanned aerial vehicle imagery," *Remote Sens.*, vol. 11, no. 14, p. 1708, Jul. 2019.
- [31] J. Zhong, T. Lei, and G. Yao, "Robust vehicle detection in aerial images based on cascaded convolutional neural networks," *Sensors*, vol. 17, no. 12, p. 2720, Nov. 2017.
- [32] M. Mandal, M. Shah, P. Meena, S. Devi, and S. K. Vipparthi, "AVDNet: A small-sized vehicle detection network for aerial visual data," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 494–498, Mar. 2020.
- [33] Y. He, Z. Pan, L. Li, Y. Shan, D. Cao, and L. Chen, "Real-time vehicle detection from short-range aerial image with compressed MobileNet," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8339–8345.
- [34] T. Ringwald, L. Sommer, A. Schumann, J. Beyerer, and R. Stiefelhagen, "UAV-net: A fast aerial vehicle detector for mobile platforms," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 544–552.
- [35] S. Oh, A. Perera, N. Cuntoor, and C. C. Chen, "A large-scale benchmark dataset for event recognition in surveillance video," in *Proc. CVPR*, Jun. 2011, pp. 3153–3160.
- [36] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, Jan. 2016.
- [37] T. Nathan Mundhenk, G. Konjevod, A. Wesam Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *Computer Vision–ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 785–800.
- [38] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

- [39] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938–1942, Sep. 2015.
- [40] M. Y. Yang, W. Liao, X. Li, Y. Cao, and B. Rosenhahn, "Vehicle detection in aerial images," *Photogramm. Eng. Remote Sens.*, vol. 85, no. 4, pp. 297–304, 4 2019.



**HUEI-YUNG LIN** (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from The State University of New York at Stony Brook, USA. He joined the Department of Electrical Engineering, National Chung Cheng University, Taiwan, as an Assistant Professor, in 2002, where he is currently a Full Professor. He was the Director of the Research Liaison Division, Office of Research and Development, from 2009 to 2013, where he was the Director of the Academic Development Division, from 2012 to 2014. He is also a Professor and the Director of the Robot Vision Laboratory, National Chung Cheng University. He is the author of more than 150 journal and conference papers. He has written two book chapters. He also holds 11 U.S. patents and eight Taiwan patents. His research interests include computer vision, robotics, machine learning, and image processing. He is a Senior Member of OSA. He serves as an organizing committee member and a program committee member of more than 50 international conferences.



**KAI-CHUN TU** received the B.S. degree in electrical engineering from National Chung Cheng University, Taiwan, in 2020, where he is currently pursuing the master's degree in electrical engineering. His research interests include machine learning and computer vision.



**CHIH-YI LI** received the B.S. degree in electrical engineering from Chung Yuan Christian University, and the M.S. degree in electrical engineering from National Chung Cheng University, Taiwan. Her research interests include computer vision and image processing.

...