iVS Dataset and ezLabel: A Dataset and a Data Annotation Tool for Deep Learning Based ADAS Applications

Yu-Shu Ni 1, Vinay M. Shivanna 1,*, and Jiun-In Guo 1,2,3

1 Department of Electrical Engineering, Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan; ni0880125.ee08g@nctu.edu.tw (Y.-S.N.); jiguo@nctu.edu.tw (J.-I.G.)
2 Pervasive Artificial Intelligence Research (PAIR) Labs, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan
3 Wistron-NCTU Embedded Artificial Intelligence Research Center, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan
* Correspondence: vinay.ms23.c@nycu.edu.tw

Abstract: To overcome the limitations of standard datasets with data at a wide-variety of scales and captured in the various conditions necessary to train neural networks to yield efficient results in ADAS applications, this paper presents a self-built open-to-free-use ‘iVS dataset’ and a data annotation tool entitled ‘ezLabel’. The iVS dataset is comprised of various objects at different scales as seen in and around real driving environments. The data in the iVS dataset are collected by employing a camcorder in vehicles driving under different conditions, e.g., light, weather and traffic, and driving scenarios ranging from city traffic during peak and normal hours to freeway traffics during busy and normal conditions. Thus, the collected data are wide-ranging and captured all possible objects at various scales appearing in real-time driving situations. The data collected in order to build the dataset has to be annotated before use in training the CNNs and so this paper presents an open-to-free-use data annotation tool, ezLabel, for data annotation purposes as well.

Keywords: ADAS; data; dataset; annotation; object detection; scales and sizes

1. Introduction

Advances in neural network (NN) technologies have given prominence to the process of detecting, classifying, and recognizing numerous objects in real-time to benefit vehicular systems. However, the detection of objects at various scales in real-time and in real traffic environments to aid Advanced Driving Assistance Systems (ADAS) in vehicles is a challenging task.

In recent years, deep learning algorithms are facilitating brand new ways to realize the real world. By learning from a large quantity of well-collected datasets, convolutional neural networks (CNNs) possibly will ascertain hidden criteria within the black box. This accelerates the formation of self-driving cars, because humans can work without thinking about the employment of tedious and complex manual skills. Researchers only have to design a sufficiently efficient architecture and enable it to learn and function exceptionally, without causing accidents and mishaps.

CNNs [1–4] have made sensational progress in various research fields such as speech recognition, natural language processing (NLP), machine translation, bioinformatics, board game programs, agricultural surveys, and particularly computer vision (CV). CNNs offer a technique to select effective and efficient perception models that produce hierarchies of visual features. CNN models that are well-modeled and trained thoroughly, employing end-to-end strategies, have exceeded human abilities in visual perception [5].

Object detection is an important task in various CV applications such as surveillance, autonomous driving, medical diagnosis, smart cities, industrial diagnosis, etc. While object detection has been intensively studied and has advanced tremendously in the recent
past, supported by deep learning technologies, innumerable challenges continue to exist during attempts to harness these technologies to various dissimilar practical and real-time applications. Notably, these challenges are due to the demands of high computational cost, huge data and abundant memory requirements while implementing deep learning models, particularly in embedded systems, usually with restricted computational apparatus required for these expensive and challenging object detection models based deep learning methods. In fact, object detection is one of the key features of self-driving cars and ADAS, and the prerequisite of these systems is that they can detect prominent objects as accurately as possible, implying that the detection and recognition algorithms are obliged to dismiss incorrect inferences, while testing and enhancing the recall rate. The process of detecting and recognizing objects at a further distance from the object is another key to this technology. The detection and recognition efficiency of the CNN models employed in ADAS vehicles is directly proportional to the data used to train them. Therefore, this paper presents an “iVS dataset” built with the primary focus of aiding object detection for autonomous vehicles. The iVS dataset overcomes the shortcomings of already available standard open datasets that are usually built comprising data from western countries, and hence lack the two-wheeler and pedestrian classes of data, to the extent found in eastern countries. Additionally, this paper introduces a semi-automatic object annotation tool entitled, ‘ezLabel’, which can be used for data annotation purpose, and is used for all data annotation purposes in the introduced iVS dataset.

2. Data Description

This section introduces the data descriptors of the intelligent Vision System (iVS) dataset, available at [https://github.com/ivslabnctu/IVS-Dataset](https://github.com/ivslabnctu/IVS-Dataset).

2.1. Data Introduction

The dataset published at the aforementioned GitHub link is entitled “iVS dataset”, where ‘iVS’ stands for ‘Intelligent Vision Systems’, a research group from National Yang Ming Chiao Tung University (NYCU) located in Hsinchu city, Taiwan, which built the iVS dataset. The iVS dataset is an extensive, manually annotated 2-D dataset consisting of over 95K high-definition (1920 × 1080) frames collected using different camcorders equipped on the dashboards of specific cars in real driving environments in Taiwan. The images in the dataset are specific to object detection applications and are captured by four-wheeled automobiles driving along the roads under distinct driving conditions such as peak hour traffic and normal hour traffic on both highways/freeways and urban and rural roads across the island of Taiwan. The data captured is from different traffic conditions such as peak hour traffic, normal and weekend days, and nighttime traffic. It consists of crowded city roads with few to many pedestrians, two-wheelers such as bikes, scooters and motorcycles, and four-wheelers such as cars, buses, mini-vans, buses, trucks, etc. These captured images are comprised of scenes from city roads and streets, universities campuses, rural and urban road and highways, under diverse weather conditions such as clear air-space, sunny, foggy, cloudy, and rainy conditions. Daylight, nighttime and twilight conditions constitute the lighting conditions of these particular images, comprising almost every kind of driving experience across the different hours in a day, weather conditions, and driving scenarios, seen in some of the examples shown in Figure 1.

2.2. Data Annotation

The dataset annotates the four commonly seen object types into four different categories of objects, ‘scooter’, ‘vehicle’, ‘pedestrian’, and ‘bicycle’. It should be noted that given ground truth labels are apparent only on probably hazardous objects on the roads that are essential to be perceived, implying that the scooters and bicycles parked beside the roads, without riders, are not labeled. Such unlabeled images are separated from the labeled ground truth images into four subsets based on the pre-defined annotation rules.
The annotation rules of the proposed iVS dataset are as follows:

1. Vehicle: Four-wheeled machines with an engine, particularly utilized for moving humans or goods by road on land are defined as "Vehicle", namely hatchbacks, vans, sedans, buses and trucks.
2. Pedestrian: The "Pedestrian" class is categorized based as the people on the road excluding those riding the two-wheeled vehicles such as motorbikes, scooters, and bikes.
3. Scooter: The third class "scooter" is considered as a combined set of the compact bounding boxes for scooters and motorbikes.
4. Bikes: The last class "bikes" is defined as objects with two bigger wheels, but no rearview mirrors and license plates.

The data annotation of the presented iVS dataset is carried out using our self-built data annotation tool entitled ‘ezLabel’, which is now available as an open-to-free-use tool for researchers to annotate their specific data in applications.

ezLabel: An Open-to-Free-Use Data Annotation Tool

The ezLabel tool designed and developed by our team is now available as an open-to-free-use tool all kinds of data annotation applications. It can be accessed by signing up at https://www.aicreda.com/. The ezLabel tool supports video data in different formats such as MP4, WMV, AVI and MOV, providing 95K frames of open data. The tool uses filters to select the video data followed by display of the information regarding the annotated objects in the video. The tool provides storage space for both raw videos and annotated videos. The annotated files are supported in XML and JSON formats.

The unique feature of the ezLabel tool is that users can add new categories, for instance, types and attributions not available in the tool. Additionally, the users can define a mission for their respective projects in order to annotate the objects.

Data annotation with ezLabel is semi-automatic using the bounding boxes with at least two frames involved. First, a point from the initial position of the object is found and keeping it as a reference, tracing occurs around the object bodyline until the last position of the object. Once the tracing is done and "confirm" is clicked, the selection of the object to be annotated is completed. Following object tracing, the users can annotate the objects, either from the list of the names in the tool or by defining a new name when one does not exist.
2.3. Training Data

The iVS dataset consists of 89,002 annotated training images and 6400 annotated testing images of resolution \(1920 \times 1080\) presented in “ivslab_train.tar”. It comprises the image and ground truth data, similar to the PASCAL VOC [6] dataset format. The objects present in images of the iVS dataset are broadly classified into four classes: ‘vehicle’, ‘pedestrian’, ‘scooter’, and ‘bikes’, as aforementioned in Section 2.2. The proposed dataset has a total of 733,108 objects that include 74,805 pedestrians, 497,685 vehicles, 153,928 scooters, and 9690 bicycles, as shown in Table 1.

Table 1. Details of the iVS dataset.

| Description                        | Count  |
|------------------------------------|--------|
| Annotated Training Images          | 89,002 |
| Annotated Validation/Testing Images| 6400   |
| Resolution                         | \(1920 \times 1080\) |
| Total object in the iVS dataset    | 733,108|
| Pedestrians                        | 74,805 |
| Vehicles                           | 497,685|
| Scooters                           | 153,928|
| Bikes                              | 9690   |

Unlike standard datasets such as PASCAL VOC [6], ImageNet [7], MS COCO [8], and the KITTI series [9] to name a few, the proposed iVS dataset contains objects averaging above five per frame, occurring in different scales and aspect ratios. This implies that the ability of the CNN detectors to detect distinct road objects at different scales concurrently is crucial for reaching a higher mean average precision (mAP), which proves the performance efficiency of a given CNN model, and the same can be achieved with the iVS dataset.

Although there are plenty of open source datasets, there are no dedicated datasets for ADAS applications. Additionally, almost all the datasets are built from environments in western countries that have uncomplicated road conditions with less dense data for ‘scooter’ and ‘pedestrian’ classes. In order to overcome the aforementioned shortcomings, the iVS dataset is presented along with the belief that it forms a dedicated dataset for ADAS applications comprising complicated road scenes, as shown in Figure 2, from Eastern countries such as Taiwan that consists of denser data from all classes, including ‘scooter’ and ‘pedestrian’ classes.

Figure 2. Examples of complications in traffic environments in Taiwan.
The researchers using this proposed iVS dataset and/or the ezLabel data annotation tool are requested to cite this paper. The applications of the proposed dataset are discussed in Section 3.

3. iVS Dataset Applications

The iVS dataset was evaluated by feeding it into a CSP-Jacinto-SSD CNN model and used in other applications as discussed in Sections 3.2 and 3.3.

3.1. Demonstration of the CSP-Jacinto-SSD CNN Model

The overall architecture of the CSP-Jacinto-SSD CNN model is depicted in Figure 3. The features of the CSPNet [10] are supplemented by JacintoNet [11]. The JacintoNet model is a plain, lightweight model comprised of convolution, group convolution, and max-pooling layers. The Cross Stage Partial (CSP) network attribute has proven to increase the correctness and lower model parameters and complexity at the same time. The common functionality of CSP is to separate the feature maps involving two parts through channels at each input stage. Of these two parts, the first sends the input conventionally into the convolution block, while the second deflects all layers and connects with the output convolution block, which resembles the final output block. As depicted in Figure 3, each of the blue and green squares is to be considered as a convolution block. The aforementioned CSP features are described by the blue arrows and the outputs from each stage by the red arrows. The feature channels are increased by using the 1x1 convolution prior to the convolution block, and the features from the CSP layer are merged using 1x1 convolution after the convolution block. The bounding box outputs are processed using the feature maps employed for dense heads, labeled by the out1 to out5 labels.

![Figure 3. The architecture of the CSP-Jacinto-SSD CNN model.](image)

The dense heads used in the method are quoted to a Single Shot Multibox Detector (SSD), with a few modifications to the anchor boxes. The multi-head SSD method published in [12] is employed. As shown in Figure 4, an additional location exists consisting of anchor boxes with offset 0 at dense head level 2–4, or alternatively solely to the original offset 0.5. This attribute increases the density of the anchor boxes, enhancing the object detection recall rate, and is specifically used for lightweight SSD models that require additional anchor boxes in order to govern the probable appearance locations of the objects.
The anchor box settings are slightly dissimilar compared to the original SSD model. In this paper, the anchors 1:2 are modified to anchors 1:1.5 as this makes the anchor borders denser, and 1:3 anchors are maintained. The base size of the anchor boxes is also modified in this paper in comparison to the base size of the anchor boxes. A juxtaposition of the proposed method to that of the basic SSD is represented in Table 2. These anchor sizes fit better with our model input size of 256 × 256.

Table 2. Anchor box base size.

|                      | Base Size               |
|----------------------|-------------------------|
| Original SSD         | 16, 32, 64, 100, 300    |
| Proposed SSD         | 16, 32, 64, 128, 256    |

The detailed specifications of the CSP-Jacinto-SSD model are shown in Table 3. With input size of 256 × 256 pixels, the model complexity is 1.08 G Flops, and there are about 2.78 million parameters. The model is tested on two different types of hardware platforms, namely a powerful GPU, nVidia 1080Ti, and an embedded platform with DSP-based architecture, TI TDA2X. Both run at an extremely high speed, as high as 138 fps on GPU, and achieve a real-time speed of 30 fps on TI TDA2X. The backbone of the CSP-Jacinto-SSD CNN model is trained with the ImageNet classification dataset as a pre-trained model.

Table 3. CSP-Jacinto-SSD CNN model specifications.

|                      | 256 × 256               |
|----------------------|-------------------------|
| Number of parameters | 2.78 M                  |
| Model complexity (Flops) | 1.08 G                |
| Speed on GPU (1080Ti) | 138 fps                 |
| Speed on embedded platform (TI TDA2X) | 30 fps                |

By employing the iVS dataset, the annotation performance is evaluated to ensure that the dataset qualifies for standard ADAS applications. The bounding box is very close to the objects. The iVS dataset is used to train the CNN model to detect the four objects: ‘pedestrian’, ‘vehicle’, ‘bike’, and ‘scooter’. Along with training the CSP-Jacinto-SSD CNN model shown in Figure 3, other models such as You Only Look Once version 5

Figure 4. Dense head levels and anchor boxes in the proposed multi-head SSD.
(YOLOv5) were also trained with the iVS dataset at the beginning, and resulted in a reliable model of front driving recorder for object detection. Table 4 shows the comparison of these two models, of which the CSP-Jacinto-SSD model is of much lower complexity in terms of implementation and lower computing time during the processes of detection and recognition. In the four given categories/classes of the iVS dataset, it achieves excellent accuracy as listed in Table 5, trained and evaluated on the CSP-Jacinto-SSD and YOLOv5 models. An attempt to detect more classes is also made using the MS COCO dataset [8] combined with the proposed iVS dataset. The accuracy of the bicycle and motorbike is improved by 3% mAP and 2.5% mAP, respectively.

Table 4. Comparison of the two models.

|                  | YOLOv5     | CSP-Jacinto-SSD |
|------------------|------------|-----------------|
| Number of parameters | 46.5 M     | 2.78 M          |
| Model Complexity (Flops) | 68.57 G    | 1.08 G          |
| Speed on GPU (1080Ti) | 50 fps     | 138 fps         |
| mAP(iVS dataset Test set) | 32.4 %     | 23.8 %          |

Table 5. Tested results of the four classes of the iVS dataset.

| Model          | Vehicle AP(%) | Pedestrian AP(%) | Bikes AP(%) | Scooter AP(%) | Total mAP(%) |
|----------------|---------------|------------------|------------|--------------|--------------|
| YOLOv5         | 51.3          | 37.4             | 13.4       | 27.5         | 32.4         |
| CSP-Jacinto-SSD| 40.7          | 27.9             | 10.8       | 15.8         | 23.8         |

The estimation of the performance of a CNN model is made using the precision, average precision and mean average precision (mAP). The correctness of the methods used for object detection implemented using numerous CNNs and image-processing algorithms is estimated using the general metric Average precision (AP). The AP calculates the average of the precision recall values. Precision estimates the accuracy of the predictions performed by a method, which is the ratio of accurate predictions. On the other hand, Recall estimates the gradation of the predicted positives. In this paper, AP is predicted using Equation (1) where \( r \) represents the recall rate and \( \hat{r} \) the recall precision value. The average of the interpolated precision [13] is utilized to assess both the detection and classification processes. The purpose of interpolating the precision-recall graph this way reduces the influence of the wiggles in the precision-recall values, generated by the small-scale ranking variations. Similarly, the average of AP gives the mean average precision (mAP).

\[
AP = \frac{\sum (r_{n+1} - r_n) \hat{p}(r_{n+1})}{\hat{p}(r_{n+1}) \max_{r \geq r_{n+1}} p(r)}
\]  

(1)

From the results tabulated in Table 5 which show the mAPs of the YOLOv5 model and CSP-Jacinto-SSD tested on the iVS dataset, it can be inferred that the detection of the objects and scene in the iVS dataset is much more difficult and thus it can be concluded that the iVS dataset contributes to enhancing the training and testing capabilities of the CNNs model used in ADAS applications.

3.2. Task Specific Bounding Box Regressors (TSBRRs)

The design considerations of the Task Specific Bounding Box Regressors (TSBRRs) [14] aim to make the implementation in low-power systems feasible, in which the TSBRRs model uses the iVS dataset for its training and testing. The TSBRRs system disjoins the object detection CNNs of dissimilar sizes and implements the corresponding algorithms. For large objects, a NN with a large visually receptive field is employed; small objects
detection is carried out using the network with a smaller receptive field, and features such as fine grain are used for precise predictions. The mechanism of Conditional Back Propagation makes distinct types of NNs execute data-driven learning for preset criteria and perceives the presentation of objects of dissimilar sizes without diminishing each other. The architecture of multiple paths objects bounding box regressors can concurrently detect objects at distinct scales and aspect ratios. The framework and algorithm designed in this method make the models able to extract robust features under the same training data, which can be applied to different weather conditions (sunny, night, cloudy and rainy days) and different countries. Only a single inference of NN is needed for every frame to aid the detection and recognition of multiple kinds of object, namely motorbikes, buses, bicycles, cars, trucks, and pedestrians, and find the respective exact positions.

The TSBBRs network is trained for about 120 epochs on the training part of the proposed iVS dataset. During the training process, a batch size of 100 is employed with 0.9 momentum and 0.0005 s decay. The learning procedure schedule is as described below.

A learning rate of $10^{-3}$ is used in the first epoch. If triggered with large learning rates, unstable gradients force the presented model to diverge. We continue to train with $10^{-2}$ for 70 epochs, then at $10^{-3}$ for 30 epochs and finally at $10^{-4}$ for 20 epochs. To increase the robustness of the presented model, an execution-time data augmentation technique is used during the training. For data augmentation, we implement a tailored layer to randomly dropout and transform the color style of the original images, as seen in Figure 5 for the iVS dataset. All the training and inference for the proposed model is built upon the Caffe [15] framework.

![Figure 5](image_url) Specimens of augmented data, (a) original images, (b) dropout, (c) color transformation.

3.3. Competitions Based on the iVS Dataset

There were three competitions held so far that required the contestants to use the proposed iVS dataset in their methods to prove its efficiency. The details of these competitions are tabulated in Table 6.
### Table 6. Competitions using iVS dataset.

| Name of the Competition                  | Number of Registered Contestants |
|------------------------------------------|----------------------------------|
| IEEE MMSP-2019 PAIR Competition          | 87                               |
| IEEE ICME-2020 GC PAIR Competition       | 128                              |
| ACM ICMR-2021 GC PAIR Competition        | 308                              |

#### 3.3.1. IEEE MMSP-2019 PAIR Competition

The embedded deep learning object-detection model competition organized in collaboration with IEEE is entitled the IEEE 21st International Workshop on Multimedia Signal Processing (MMSP-2019) PAIR competition. The competition was devoted to object detection for sensing technology in autonomous vehicles, aiming at detecting small objects in extreme situations using embedded systems. We, the organizers, provided the iVS dataset with 89,002 annotated images for training and 1500 annotated images for validation. The competition was divided into a qualification round and a final round. There were 87 teams registered in the competition, of which 14 submitted the compositions of their teams. In the qualification round, the participants’ models were tested using 3000 testing images and qualified for the final round depending on the mAP of the models. There were five teams that submitted their respective final models which could be realized on the nVidia Jetson TX-2 in the final round. In the finals, their models were again tested with a different set of 3000 images and the winners were chosen based on achieving the target accuracy requirement. Team R.JD was declared winner of the IEEE MMSP-2019 PAIR competition [16].

#### 3.3.2. IEEE ICME-2020 GC PAIR Competition

The embedded deep learning object-detection model compression competition for traffic in Asian countries entitled the IEEE International Conference on Multimedia and Expo (ICME-2020) Grand Challenges (GC) PAIR competition focused on the object detection and recognition of objects using sensing technology in autonomous vehicles. The target of the competition was to successfully implement the designed model on an embedded system with low complexity and low model sizes [17]. The dataset used in the competition was an iVS dataset comprising data mostly from Asian countries that contain some of the harshest driving environments, such as crowded streets with scooters, bicycles, pedestrians and vehicles. The contestants were provided with 89,002 annotated training images and 1000 validation images. In the evaluation process, 5400 testing images were used, 3000 of them in the qualification stage and the remainder in the final stage. A total of 133 teams joined the competition. The top 10 teams with the highest accuracies were chosen for the final round of the competition. The overall best models that could be implemented on nVidia Jetson TX2 were chosen as winners based on the accuracy (mAPs), model size, computational complexity, and speed on nVidia Jetson TX2. The winners of the IEEE ICME-2020 GC PAIR competition were team USTC-NELSLIP, followed by the team “BUPT_MCPRL” and the team “DD_VISION”.

#### 3.3.3. ACM ICMR-2021 GC PAIR Competition

The 2021 embedded deep learning object detection model compression competition for traffic in Asian countries was entitled the ACM International Conference on Multimedia Retrieval (ICMR-2021) Grand Challenges (GC) PAIR competition. The Grand Challenges of the competition were devoted to technologies to detect and recognize objects in autonomous driving situations [18]. The main objective of the competition was to detect objects in traffic using low complex and CNNs with small model size in Asian countries such as Taiwan that contain various harsh driving conditions. The target detected objects comprised pedestrians, bicycles, vehicles, and scooters in crowded and non-crowded traffic situations. Again, the iVS dataset was used in the competition consisting of 89,002 annotated images provided for model training and 1000 images for validation. Additionally, 5400 testing
images were employed in the validation of the contest’s submissions, of which 2700 were utilized in the qualification stage of the competition, and the rest in the final stage. There was a total of 308 teams registered in the competition, and the top 15 teams with the leading detection accuracy qualified for the last stage of the competition. Of these 15 teams, only 9 submitted the final results within the set deadlines. The submitted models from all the finalists were evaluated on the MediaTek Dimensity 1000+ mobile computation platform [19]. The overall best model was awarded to team “as798792”, followed by the team “Deep Learner” and the team “UCBH”. Two special awards, “Best accuracy award” and “Best bicycle detections”, were awarded to the same team “as798792", and another special award of “Best scooter detection” was awarded to the team “abcda”.

4. Conclusions
This paper has presented an open dataset entitled ‘iVS dataset’ that is available at https://github.com/ivslabnctu/IVS-Dataset, specifically for object detection purposes for the ADAS applications. In addition, this paper has also presented an open-to-free-use data annotation tool entitled ‘ezLabel’ that is used to annotate the iVS dataset. The iVS dataset comprises of data captured from real driving environments varying from peak hour to non-peak hour traffic and from city roads to freeways across Taiwan, R.O.C. The data is captured in different lighting and weather conditions to best suit the purpose of training the CNNs to be used in the ADAS applications. In the future, more emphasis will be given to enhancing the iVS dataset with real-time cases that are not available in Taiwan’s Cities such as snow (day and night), and focus on detecting other objects such as road signs and characters on road signs. Thus, extra object types could be included in the proposed iVS dataset to support additional object detection information for autonomous vehicles.

Author Contributions: Conceptualization, Y.-S.N., V.M.S. and J.-I.G.; methodology, J.-I.G.; software, Y.-S.N. and V.M.S.; validation, Y.-S.N. and J.-I.G.; formal analysis, Y.-S.N., V.M.S. and J.-I.G.; investigation, Y.-S.N. and J.-I.G.; resources, J.-I.G.; data curation, Y.-S.N., V.M.S. and J.-I.G.; writing—original draft preparation, V.M.S.; writing—review and editing, V.M.S.; visualization, J.-I.G.; supervision, J.-I.G.; project administration, J.-I.G.; funding acquisition, J.-I.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is fractionally promoted by the “Featured Areas Research Center Program” subordinate to “Center for mmWave Smart Radar Systems and Technologies” under the framework of the Higher Education Sprout Project by the Ministry of Education (M.O.E), Taiwan R.O.C. It is also partially promoted under the Ministry of Science and Technology (M.O.S.T.), Taiwan R.O.C. projects with grants MOST 110-2634-F-009-020, MOST 110-2634-F-009-028, MOST 110-2221-E-A49-145-MY3, and MOST 110-2634-F-A49-004 through Pervasive Artificial Intelligence Research Labs (PAIR Labs) in Taiwan, R.O.C. as well as the partial support from the Qualcomm Technologies under the research collaboration agreement 408929.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS) 25, Lake Tahoe, NV, USA, 3–6 December 2012.
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
3. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
4. Yang, M.-D.; Tseng, H.-H.; Hsu, Y.-C.; Tsai, H.P. Semantic Segmentation Using Deep Learning with Vegetation Indices for Rice Lodging Identification in Multi-date UAV Visible Images. Remote Sens. 2020, 12, 633. [CrossRef]
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015.
6. Everingham, K.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. Int. J. Comput. Vis. 2010, 88, 303–338. [CrossRef]
7. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Computer Vis.* 2015, 115, 211–252. [CrossRef]

8. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, D.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv* 2014, arXiv:1405.0312.

9. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361. [CrossRef]

10. Wang, C.; Liao, H.M.; Wu, Y.; Chen, P.; Hsieh, J.; Yeh, I. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 1571–1580.

11. Mathew, M.; Desappan, K.; Swami, P.K.; Nagori, S. Sparse, Quantized, Full Frame CNN for Low Power Embedded Devices. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 328–336.

12. Lai, C.Y.; Wu, B.X.; Lee, T.H.; Shivanna, V.M.; Guo, J.I. A Light Weight Multi-Head SSD Model for ADAS Applications. In Proceedings of the 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), Taipei, Taiwan, 3–5 December 2020; pp. 1–6.

13. Salton, G.; McGill, M.J. *Introduction to Modern Information Retrieval*; McGraw-Hill: New York, NY, USA, 1986.

14. Lin, G.-T.; Malligere Shivanna, V.; Guo, J.-I. A Deep-Learning Model with Task-Specific Bounding Box Regressors and Conditional Back-Propagation for Moving Object Detection in ADAS Applications. *Sensors* 2020, 20, 5269. [CrossRef] [PubMed]

15. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv* 2014, arXiv:1408.5093.

16. Guo, J.-I.; Tsai, C.-C.; Yang, Y.-H.; Lin, H.-W.; Wu, B.-X.; Kuo, T.-T.; Wang, L.-J. Summary Embedded Deep Learning Object Detection Model Competition. In Proceedings of the IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), Kuala Lumpur, Malaysia, 27–29 September 2019; pp. 1–5. [CrossRef]

17. Tsai, C.-C.; Yang, Y.-H.; Lin, H.-W.; Wu, B.-X.; Chang, E.-C.; Liu, H.-Y.; Lai, J.-S.; Chen, P.-Y.; Lin, J.-J.; Chang, J.-S.; et al. The 2020 Embedded Deep Learning Object Detection Model Compression Competition for Traffic in Asian Countries. In Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 6–10 July 2020; pp. 1–6. [CrossRef]

18. Ni, Y.-S.; Tsai, C.-C.; Guo, J.-I.; Hwang, J.-N.; Wu, B.-X.; Hu, P.-C.; Kuo, T.-T.; Chen, P.-Y.; Kuo, H.-K. Summary on the 2021 Embedded Deep Learning Object Detection Model Compression Competition for Traffic in Asian Countries. In Proceedings of the 2021 ACM International Conference on Multimedia Retrieval (ICMR2021), Taipei, Taiwan, 21–24 August 2021.

19. MediaTek Dimensity 1000+, Flagship 5 G Experiences, Incredible Performance, Supreme AI-Cameras. Available online: https://i.mediatek.com/dimensity-1000-plus (accessed on 5 January 2021).