Mobilenet-SSDv2: An Improved Object Detection Model for Embedded Systems

Yu-Chen Chiu, Chi-Yi Tsai*, Mind-Da Ruan, Guan-Yu Shen and Tsu-Tian Lee
Department of Electrical and Computer Engineering
Tamkang University
151 Ying-chuan Road, Tamshui District, New Taipei City 251, Taiwan R.O.C.
*Email: chiyi_tsai@gms.tku.edu.tw

Abstract*—Object detection plays an important role in the field of computer vision. Many superior object detection algorithms have been proposed in literature; however, most of them are designed to improve the detection accuracy. As a result, the requirement of reducing computational complexity is usually ignored. To achieve real-time performance, these superior object detectors need to operate with a high-end GPU. In this paper, we introduce a lightweight object detection model, which is developed based on Mobilenet-v2. The proposed real-time object detector can be applied in embedded systems with limited computational resources. This is one of the key features in the design of modern autonomous driving assistance systems (ADAS). Besides, we also integrate a feature pyramid network (FPN) with the proposed object detection model to effectively improve detection accuracy and detection stability. Experimental results show that the proposed lightweight object detection model achieves up to 75.9% mAP in the VOC dataset. Compared with the existing Mobilenet-SSD detector, the detection accuracy of the proposed detector is improved about 3.5%. In addition, when implemented on the Nvidia Jetson AGX Xavier platform, the proposed detector achieves an average of 19 frames per second (FPS) in processing 720p video streams. Therefore, the proposed lightweight object detector has great application prospects.

Keywords*—single-shot multibox detector (SSD), mobilenet-v2, mobilenet-ssd, feature pyramid network, embedded systems

I. INTRODUCTION

Object detection is an essential function in the development of ADAS and self-driving cars. In order to obtain a robust and accurate object detector, many advanced detection algorithms have been proposed in the literature based on machine learning techniques. In recent years, deep learning-based convolutional neural network (CNN) methods become an emergency research topic due to their remarkable improvements on the applications object classification [1], object recognition [2] and object detection [3-6]. Although the deep CNN model provides really strong and robust feature representation for object detection and recognition, it usually is computationally expensive and requires a high-end hardware platform to perform the model inference computation. However, in the application of self-driving cars, object detection processing usually needs to be performed on an embedded platform with limited computing resources. Thus, only a neural network model with a small network size and a small amount of calculations can satisfy this situation. How to transplant the object detection network model into the embedded platform to operate while maintaining a considerable accuracy and stability is an urgent problem in the development of self-driving systems.

The major object detection algorithms can be divided into one-stage and two-stage algorithms. Among two of them, only the one-stage algorithms can be implemented in the embedded platform with real-time performance. In recent years, SSD [5] and YOLO [6] are two of the most popular one-stage object detectors.

After that, RetinaNet [7] with an improved loss function and M2det [8] based on a FPN model [9] are also excellent object detection methods. However, in order to improve accuracy rate, these new object detectors use extremely high computational backbone networks such as VGG16 [2] or ResNet [10] in the design of feature extraction networks. These backbone networks are too large for embedded platforms and cannot achieve real-time processing speed. Thus, when these advanced object detectors are implemented on embedded platforms, the results in practical applications are not as good as expected. Finally, most embedded platforms still choose the traditional SSD or YOLO detectors for object detection applications.

On the other hand, the authors in [11] mentioned that the YOLO detector is not effective in detecting dense multiple targets and large objects. Instead, the SSD detector is more suitable for detecting a large number of objects with different scales and different types because it considers multiple feature maps of different scales. Therefore, in this work we decided to employ the SSD detector and to improve it so that the modified detector can perform in real-time on the embedded platform while increasing the accuracy rate of object detection. Because the proposed object detector needs to run in an embedded system, the computational complexity of the neural network model is also an important design condition. In a self-driving system, the calculation of the object detector must satisfy fast and low-latency features. Hence, a small, fast, and lightweight feature extraction network is a necessary choice. Currently, Mobilenet-v2 [12] and Shufflenet-v2 [13] network models have achieved a good balance between computing speed and object recognition accuracy. Therefore, considering the limitations of the embedded platform, we chose Mobilenet-v2 network, which is supported by many embedded platforms, as the backbone network for the design of the proposed lightweight object detection model.

To improve the detection accuracy of the simplified object detection network model, we also employed the technology of the inverted residual module in Mobilenet-v2 and the FPN architecture to improve the detection performance of the proposed object detector, termed as Mobilenet-SSDv2. Experimental results show that the proposed Mobilenet-SSDv2 detector achieves an accuracy rate of 75.9% mAP in the Pascal VOC [14] test set and a processing speed of 19 FPS running on the Nvidia Jetson AGX Xavier platform. Moreover, the memory usage of the proposed detector is only 32MB, which is helpful in the development of modern embedded ADAS.

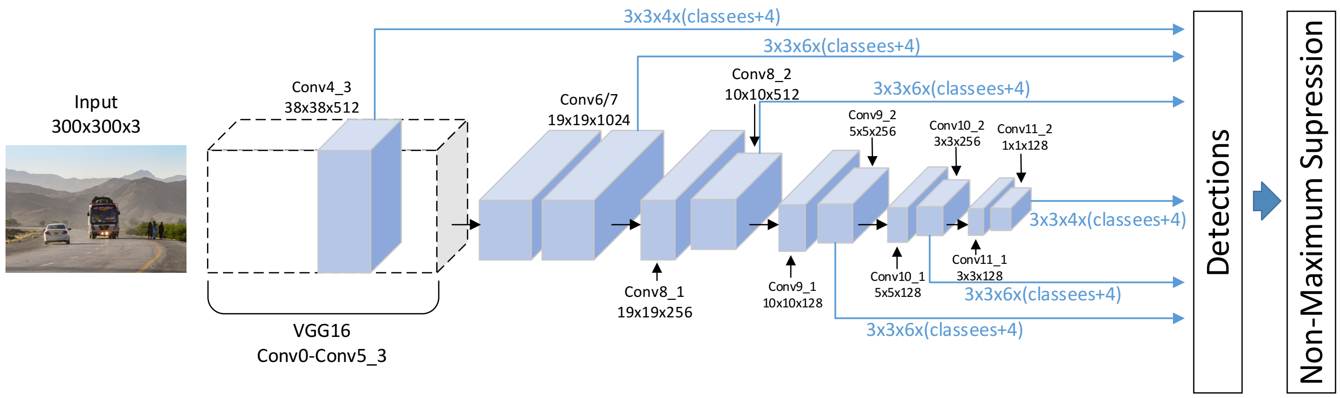


Fig. 1. Network architecture of VGG16-SSD.

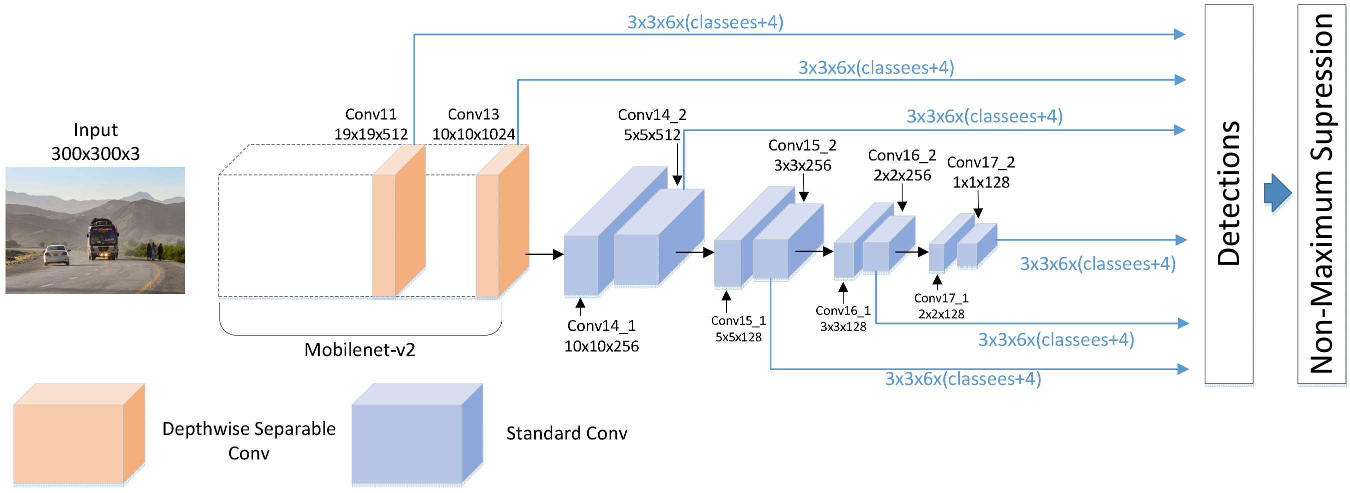


Fig. 2. Network architecture of Mobilenet-SSD.

TABLE I. DIMENSIONS OF FUTURE MAPS USED IN VGG16-SSD AND MOBILENET-SSD DETECTOR MODELS

Detector Model	**VGG16-SSD**	**Mobilenet-SSD**
Layer 1	38x38x512	19x19x96
Layer 2	19x19x1024	10x10x1280
Layer 3	10x10x512	5x5x512
Layer 4	5x5x256	3x3x256
Layer 5	3x3x256	2x2x256
Layer 6	1x1x128	1x1x128

II. ARCHITECTURE OF SINGLE SHOT MULTIBOX DETECTOR

A. *VGG16-SSD*

SSD is one of the most popular one-stage object detectors used in several detection applications. Although its detection accuracy is not as good as existing two-stage target detectors, its main advantage is that it has a fast calculation speed. Figure 1 shows the existing VGG16-SSD network architecture, which uses the VGG16 network as the backbone network model of the object detector. The VGG16 network provides six feature maps with different dimensions for the back-end network to detect multi-scale objects. Then, a non-maximum suppression (NMS) process is applied on the detection outputs of the network model. For each group with multiple overlapping detection outputs, the detection output with the highest confidence score is selected as the final detection result for that group.

Although the VGG16 network model has good feature extraction capabilities, its network architecture is too large for embedded platforms. As a result, VGG16 may exceed the maximum system memory and is difficult to achieve real-time performance when running on embedded systems.

B. *Mobilenet-SSD*

To reduce the computational complexity of the VGG16-SSD detector, Google applied the Mobilenet network model [15] to replace the VGG16 network, improving the real-time performance of the SSD detector. Figure 2 shows the existing Mobilent-SSD network architecture, which uses the second-generation Mobilenet network, called Mobilenet-v2, as the backbone network model of the SSD detector. The Mobilent-SSD detector inherits the design of VGG16-SSD that the front-end Mobilenet-v2 network provides six feature maps with different dimensions for the back-end detection network to perform multi-scale object detection. Since the backbone network model is changed from VGG-16 to Mobilenet-v2, the Mobilent-SSD detector can achieve real-time performance and is faster than other existing object detection networks.

Table 1 shows the dimensional comparison of the multi-layer feature maps extracted from the VGG16-SSD and Mobilenet-SSD backbone networks. The table shows that the feature map dimension extracted by the Mobilenet-v2 network at the first layer is twice smaller than the feature map dimension of the VGG16 network. Thus, on the corresponding feature map, the object detection range of the

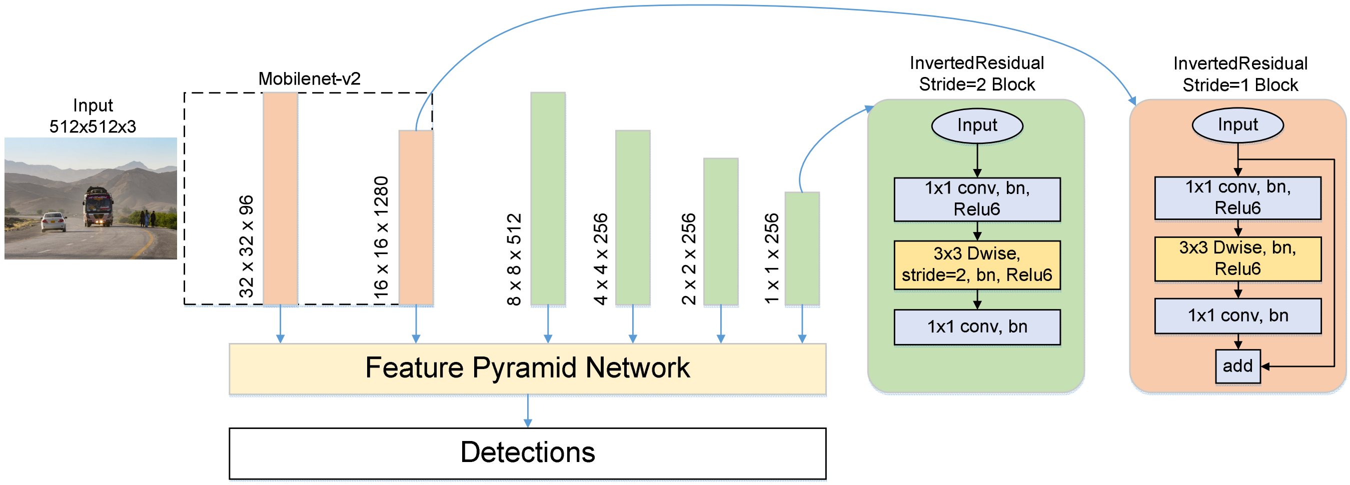


Fig. 3. Network architecture of the porposed Mobilenet-SSDv2.

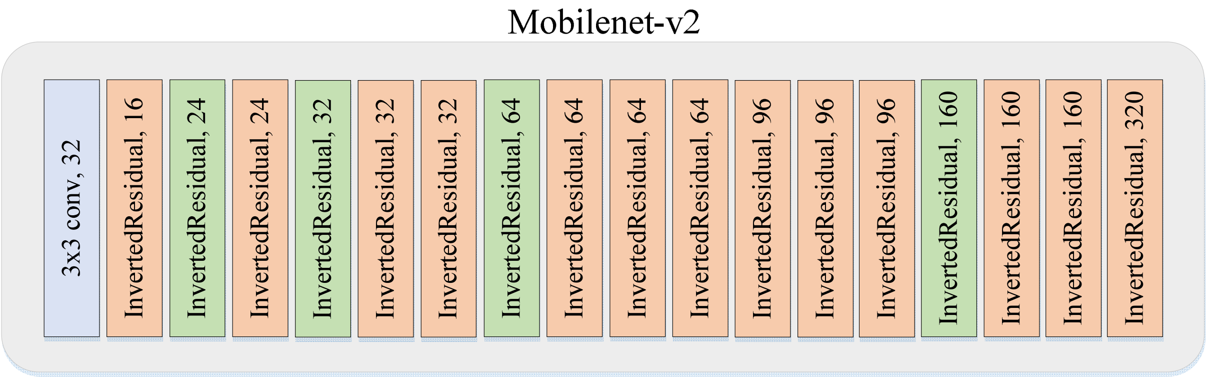


Fig. 4. Network architecture of Mobilenet-v2.

Mobilenet-v2 network is only half of the VGG16 network. This disadvantage leads to poor detection accuracy of the Mobilenet-v2 network in practical applications. On the other hand, because the 38x38 feature map provided by the Mobilenet-v2 network is a shallow feature, it is difficult to extract an effective image feature map. Therefore, in this study we try to use the FPN module to fuse the output feature map of the Mobilenet-v2 network to effectively improve the detection accuracy without increasing too many parameters.

III. The Proposed Mobilenet-SSDv2

In this section, we introduce the design of the proposed object detector. In order to meet the requirement of running on an embedded platform, the backbone network we chose is Mobilenet-v2. Figure 3 shows the network architecture of the proposed Mobilenet-SSDv2 detector, which improves the SSD detector based on Mobilenet-v2 and FPN technology and maintains the memory usage of the network model.

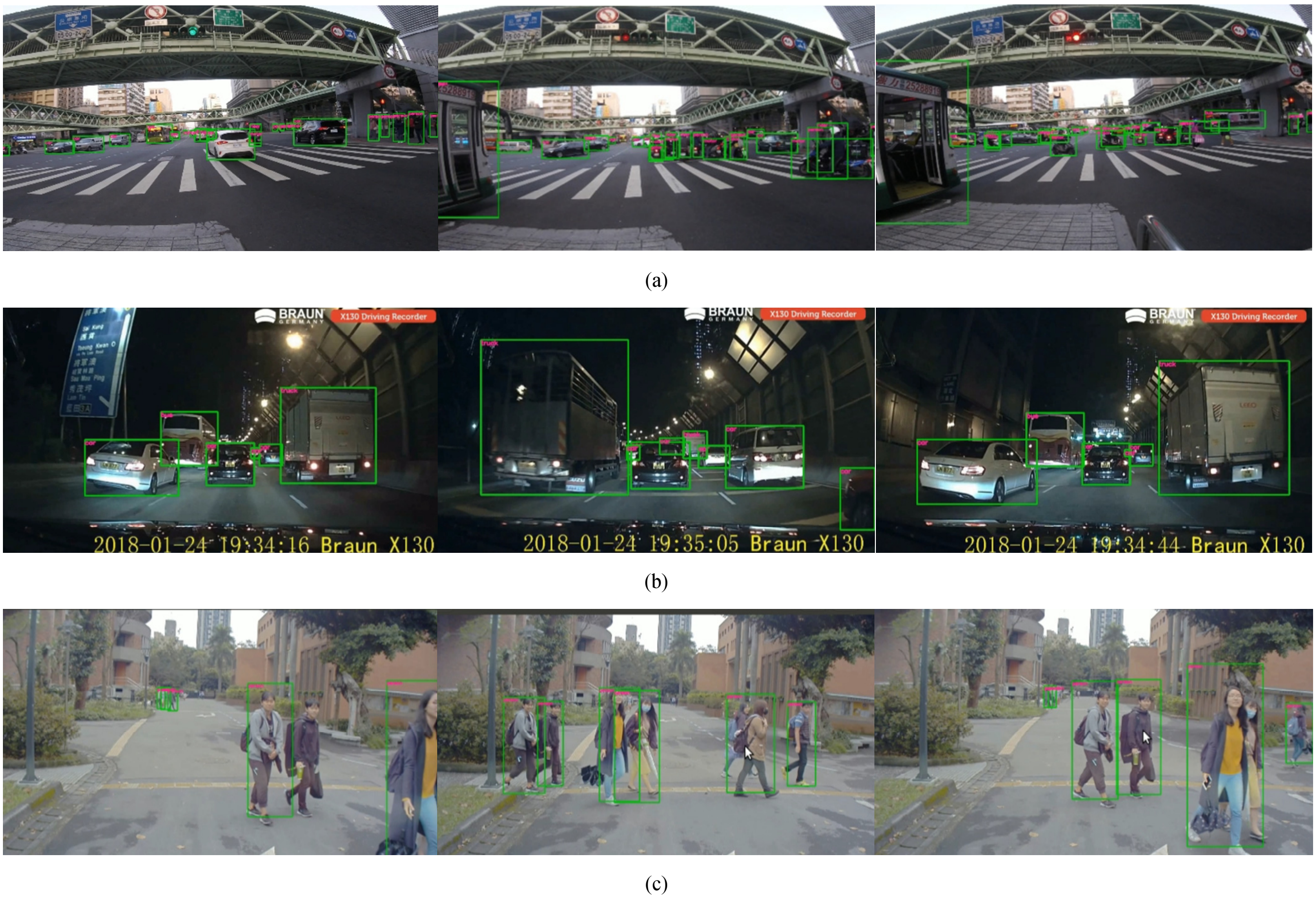
A. Mobilenet-v2

Most lightweight network models use Mobilenet-v2 as the backbone network. Figure 4 shows the network architecture of Mobilenet-v2, which includes a standard convolutional layer and 17 inverse residual modules. As shown in Figure 3, each inverse residual module contains a 1x1 convolutional layer, a 3x3 depth-wise (Dwise) separable convolutional layer, batch normalize (bn) and Relu6 excitation functions. The output feature map is also added with the input feature map without changing its size. The advantage of using the inverse residual module is that it can effectively prevent the gradient vanishing issue in order to correctly transfer the gradient information to the deeper network layer to form an effective training during backpropagation process.

In the traditional Mobilene-SSD, the feature map is down-sampled by a 1/16-scale convolution layer. This setting will cause poor and unstable detection of small objects in the image. To overcome this problem, in our network architecture, we extracted feature maps at 1/16 and 1/32 scales instead. In addition, four inverse residual modules are added after the backbone network to extract feature maps at scales of 1/64, 1/128, 1/256, and 1/512. Finally, these feature maps with different scales are enhanced by the FPN module.

B. Feature Pyramid Network (FPN)

In our detection network architecture, the multi-scale feature map extracted by the backbone network can be directly used as the input information for the back-end detection network. However, experiments show that this approach limits the detection accuracy of the backend detection network. To overcome this problem, we added the FPN at the output of the backbone network to improve the performance of the back-end detection network. Shallow feature maps are suitable for object localization, but the semantic information is insufficient. Deep feature maps are not suitable for localization, but the semantic information is sufficient. This contradiction is a common problem for one-stage target detectors. Therefore, in the proposed detection network architecture, we fused the multi-scale feature maps obtained from the backbone network to enhance detection


BRAUN
X130 Driving Recorder
2018-01-24 19:34:16 Braun X130
2018-01-24 19:35:05 Braun X130
2018-01-24 19:34:44 Braun X130

(a)

(b)

(c)

Fig. 5. Experimental results in different test scenarios: (a) the crossroad at day time, (b) the road at night time, and (c) the road in a campus at day time.

TABLE II. PERFORMANCE OF THE PROPOSED DETECTOR MODEL TESTED ON THE PASCAL VOC DATASET

Detector Model	**Mobilenet-SSDv2 (Ours)**		**Mobilenet-SSD**	
Input Image Resolution	512x512		320x320	512x512
+Mobilenet-v2	√	√		
+FPN		√		
mAP	73.0%	75.6%	68.8%	72.4%
FPS	23	21	20	17
Model Size	32MB	32MB	25MB	25MB

performance. First, we used a 1x1 convolutional layer to unify the number of channels in each feature map. Next, we resized feature maps of each different scale and added them together. Empirically, we found that the fusion process of multi-scale feature maps can effectively improve the object detection effect of the back-end detection network. In addition, we also added an inverse residual module to the FPN network to improve the information transfer between the feature maps during the fusion process. This method helps to effectively improve the information transfer between different channels without increasing the number of model parameters.

IV. EXPERIMENTAL RESULTS

In order to verify the detection performance of the proposed Mobilenet-SSDv2 detector, we used the Pascal VOC [14] dataset in the experiments for network model training and testing. In the meantime, we also compared the object detection performance of the proposed detector with the existing Mobilenet-SSD detector. The input image format of our network model is a 512x512 RGB color image. The optimizer we use to train the proposed network model is the SGD method, and we trained the model for a total of 200 epochs. The initial learning rate is set to 0.001 and is reduced by 10 times at 140 and 180 epochs, respectively.

Table 2. shows the comparison results between the proposed object detector and the existing Mobilenet-SSD detector on the Pascal VOC dataset. First of all, we tested the effect of changing the backbone network model to Mobilenet-v2. The results show that the modified method improves the detection accuracy by about 0.6% mAP compared with the original method. Next, we added the FPN module to fuse multi-scale feature maps, which helps increasing the detection accuracy by 2.6% mAP. Therefore, the proposed Mobilenet-SSDv2 detector can improve the detection accuracy to 75.6% mAP on the Pascal VOC dataset. This result is about 3.2% mAP higher than the existing Mobilenet-SSD detector.

In comparison of computing speed, when the proposed detection network model processes a 720p video stream on Nvidia Jetson AGX Xavier, the average frame rate is about 23 FPS without using the FPN module. After adding the FPN module, the frame rate can still reach an average of about 21 FPS. Compared with the existing Mobilenet-SSD detector, which provides a processing speed of about 17 FPS on average, our detector has a significant increase in processing speed. Finally, the memory capacity of our network model is 32MB, which is 7MB higher than the existing Mobilenet-SSD. However, the proposed detector can significantly improve the processing speed and detection accuracy.

Figure 5 shows the object detection results of our Mobilenet-SSDv2 detector in different test scenarios. Figure 5(a) shows the test results at a crossroad environment during the daytime. We can clearly see that our detection network model can accurately detect a large number of different types of objects, including buses, cars, motorbikes, and pedestrians. Figure 5(b) shows the test results in a lane environment at nighttime, which contains a large number of different types of cars and trucks. Test results show that our detection network model can also accurately detect various types of cars and trucks in the low-light environment. Figure 5(c) shows the test results in a campus road environment during the daytime, which contains a large number of pedestrians moving on the road. It can be seen from the test results that the proposed detection network model can accurately detect the movement of each pedestrian in the image. Therefore, the above experimental test results can validate the detection accuracy and robustness of the proposed Mobilenet-SSDv2 detector in various scenarios.

V. Conclusions and Future Work

In this paper, we propose a lightweight network architecture with improved feature extraction based on the Mobilenet-v2 backbone network. We combine Mobilenet-v2 and FPN models to enhance the feature map of the input image and effectively improve the detection accuracy of the back-end detection network. In the Pascal VOC dataset, the proposed detection network obtained a 75.6% mAP accuracy and a processing speed of 21 FPS. In addition, the memory capacity of the entire network model is approximately 32MB. This is a great advantage for embedded platforms with limited resources. Experimental results show that the proposed Mobilenet-SSDv2 detector not only retains the advantage of fast processing of the original Mobilenet-SSD detector, but also greatly improves the detection accuracy. These advantages indicate that the Mobilenet-SSDv2 detection model proposed in this paper is more suitable for embedded platforms.

In future work, we will continue to optimize our detection network model, including reducing memory usage and increasing network computing speed.

Acknowledgment

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 108-2218-E-032-004 and Grant MOST 108-2221-E-032-046.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *International Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, pp. 1097-1105, 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, San Diego, USA, 2015.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference Computer Vision and Pattern Recognition*, Columbus, USA, pp. 580-587, 2014.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," IEEE International Conference on Computer Vision, Venice, Italy, pp. 2980-2988, 2017.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," *European Conference on Computer Vision*, Amsterdam, Netherlands, pp. 21-37, 2016.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 779-788, 2016.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal Loss for Dense Object Detection," *IEEE International Conference on Computer Vision*, Venice, Italy, pp. 2980-2988, 2017.

[8] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling. "M2det: A single-shot object detector based on multi-level feature pyramid network" *Thirty-Third AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, USA, 2019.

[9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 936-944, 2017.

[10] K. He, X. Zhang, S. Ren, J. Sun "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770-778, 2016.

[11] Q. Zhao, T. Sheng, Y. Wang, F. Ni, and L. Cai, "CFENet: An accurate and efficient single-shot object detector for autonomous driving," CoRR, arXiv:1806.09790, 2018.

[12] M. Sandler, A. Howard, M. *Zhu*, A. Zhmoginov, and L.-C. Chen "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 4510-4520, 2018.

[13] N. Ma, X. Zhang, H.-T. Zheng, J. Sun "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," *European Conference on Computer Vision*, Munich, Germany, pp. 116-131, 2018.

[14] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98-136, 2015.

[15] MobileNetV2: The Next Generation of On-Device Computer Vision Networks, available online: https://ai.googleblog.com/2018/04/mobilenetv2-next-generation-of-on.html