# An adaptive dead fish detection approach using SSD-MobileNet

1st Guoyan Yu
*School of Mechanical and Power Engineering*
*Guangdong Ocean University;*
*Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang)*
Zhanjiang, China
yugy@gdou.edu.cn

2nd Lin Wang
*School of Mechanical and Power Engineering*
*Guangdong Ocean University*
*Guangdong Provincial Ocean Equipment and Manufacturing Engineer Technology Research Center*
Zhanjiang, China
912589683@qq.com

3rd Mingxin Hou*
*School of Mechanical and Power Engineering*
*Guangdong Ocean University;*
*Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang)*
Zhanjiang, China
Corresponding Author:
houmx@gdou.edu.cn

4th Yicha Liang
*School of Mechanical and Power Engineering*
*Guangdong Ocean University;*
*Guangdong Provincial Ocean Equipment and Manufacturing*
*Engineer Technology Research Center*
Zhanjiang, China
1193876227@qq.com

5th Taihua He
*School of Mechanical and Power Engineering*
*Guangdong Ocean University;*
*Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang)*
Zhanjiang, China
taihuahe@163.com

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9326648

**Abstract: Aiming at the difficulty in the recognition of dead fish in a large cage, a method of detecting dead fish on water surface based on SSD-MobileNet is proposed. In this paper, the models of SSD-MobileNet V1, V2, and V3 are compared, which are suitable for an embedded mobile terminal. Through theoretical and experimental analysis, it has been proved that SSD-MobileNet V3 is not only real-time object detection algorithm, but also a new architecture combining hardware network architecture search (NAS) and NetAdapt. SSD-MobileNet V3 realizes the cooperation between automatic search algorithm and network design, and the complementary method improves the overall technical level.**

***Keywords: SSD-MobileNet, real-time object detection, NetAdapt, dead fish***

## I. Introduction

In recent years, computer vision tracking is of great significance for the applications of intelligent video surveillance, human-computer interaction, vehicle navigation, etc. Under the background of the rapid development in the field of science and technology, the total amount of marine economy is increasing, which is serving as an important engine to promote the development of the national economy. More and more countries are beginning to increase the utilization and development of the oceans. China's sea area is huge and rich in resources, but its utilization rate is not high. In order to promote the development of marine economy, the government has put forward the idea of marine strategic development. The application of computer vision in fisheries industry is still at a preliminary stage. Globally, the statistics recently published on the official website of the Food and Agriculture Organization of the United Nations (FAO) have shown that whole fish production enters directly into the market without careful testing or selection.

The original traditional algorithms based on filtering and tracking have achieved good performance in speed and accuracy, but when there are interference factors such as deformation, occlusion, scale change, illumination change

This work is supported by the Fund of Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang)(No. ZJW-2019-01), key scientific research projects of universities in Guangdong Province(No. 2018KZDXM038) and program for scientific research start-up funds of Guangdong Ocean University.

Authorized licensed use limited to: Rutgers University. Downloaded on May 15,2021 at 02:10:36 UTC from IEEE Xplore. Restrictions apply.

and rotation during tracking, tracking performance will be much worse. LeNet[1] network structure was proposed by Lecun in 1998, which was created to solve the problem of handwritten numeral recognition. This is a solution for ten kinds of tasks. In 2012, the Alexnet[2] proposed by Hinton proposed a novel structure and dropout method in the Imagenet image recognition competition, which reduced the error from more than 25% to 15%, having a great impact in the image recognition field. Ross Girshick proposed R-CNN[3-5] in 2014. R-CNN is the first algorithm that successfully applies deep learning to object detection. After R-CNN, YOLO[6] is another framework proposed by Ross Girshick to solve the problem of object detection speed. VGGNet[7] was designed by the computer vision group of Oxford University and Google, of which the main purpose is to study the influence of network depth on model accuracy, and to build the whole network by small convolution superposition. Its parameters are up to 138 Million, and the size of the whole model is more than 500 MB. ResNet[8] was proposed by Kaiming's team in 2015. A jump connection structure is introduced to prevent the gradient from disappearing and f increase the depth of the object detection . SSD[9] is an one stage object detection algorithm proposed by Wei Liu on ECCV. Compared with Faster R-CNN, SSD has a significant advantage in speed; while compared with YOLO, it has better performance in mAP.

The above methods are effective in different fields, but they are not friendly to the embedded system. In many practical applications, such as no driver and virtual reality, the object detection task needs to be completed in time on the embedded platform with limited computing. In this paper, we present an energy-efficient implementation of object detection systems on an embedded platform. To make better use of this architecture, we adopt the SSD-MobileNet algorithm, which is suitable for embedded systems. It can not only ensure accuracy, but also improve the detection speed.

## II. Detection Algorithm

In this section, the algorithm model is introduced to determine the most suitable model, with regard to the trade-off between speed and accuracy. Generally speaking, there are two ways for objects detection. The first one is a two-stage algorithm based on suggested areas, such as R-CNN[3-5]; the other is a one-stage algorithm based on regression, such as YOLO and SSD. In particular, SSD is a common single level object detection algorithm.

### A. *Single Shot MultiBox Detector (SSD)*

SSD uses the regression idea of YOLO to transform the object detection into a simple regression problem for processing. At the same time, SSD adopts the detection method based on the pyramid feature level, and carries out the softmax classification and position regression on the feature images of Multiscale.

As shown in Fig. 1, the network architecture of SSD is mainly composed of the basic network part at the front end and the additional feature extraction part at the back. In the basic network, the convolution network of VGG - 16 is used to extract the basic features. The additional feature extraction part is a series of convolutional networks, which are used to extract the advanced features of the object. In VGG-16 model, convolution layer, conv 6 and conv 7, are modified by the down sampling methods of fc 6 and fc 7, respectively, which reduce a large amount of computation. The rest is the specially created convolution layer. Each group uses 1 convolution core descent channel at first, and then adopts 3 convolution core to increase the number of channels.

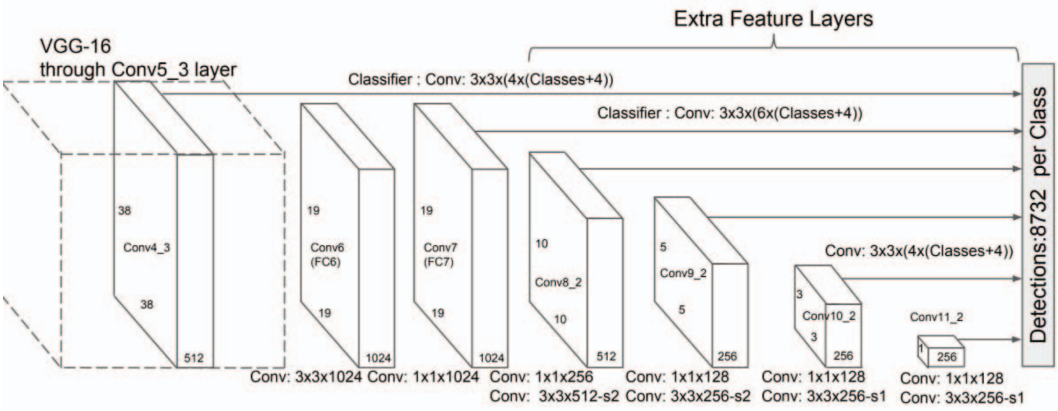


Fig. 1. Network architecture of SSD

Authorized licensed use limited to: Rutgers University. Downloaded on May 15,2021 at 02:10:36 UTC from IEEE Xplore. Restrictions apply.

## B. *MobileNet*

MobileNet is a new generation of convolutional neural networks proposed by Google. The model balances performance and fluency, which is very suitable for deployment on mobile platforms, such as embedded systems with low hardware configuration and relatively poor computing power.

Based on a streamlined architecture, MobileNet V1[10] applies depthwise separable convolutions. MobileNet V1 build a lightweight deep network and introduces two hyper-parameters, which efficiently make a better balance between latency and accuracy. It uses depth separable convolution simplifies the traditional network structure. As shown in Fig. 2 , the depth separable convolution is used in MobileNet V1 to replace ordinary convolution network, which greatly reduces computation. When the sizes of the input and output image of channels are the same, the depth separable convolution is 1/9 –1/8 of the traditional one.

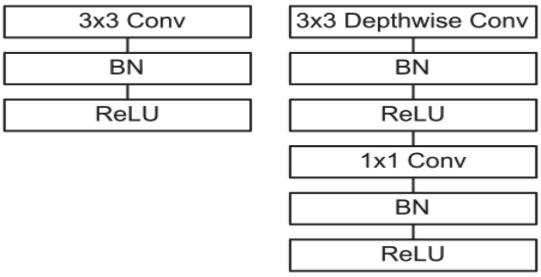


Fig. 2. Left: Standard convolutional, Right: Depthwise Separable convolutions

MobileNet V2 [11] is an improvement of MobileNet V1, which is also a lightweight convolutional neural network. MobileNet V2 takes full advantage of inverted residual structure. To avoid the damage of ReLU to the feature, the linear layer is used in MobileNet V2, to replace the nonlinear activation of Rely after the channel number becomes less. As shown inFig. 3, Relu can save the complete information of the input manifold, but only if the input manifold is located in the low dimensional subspace of the input space.

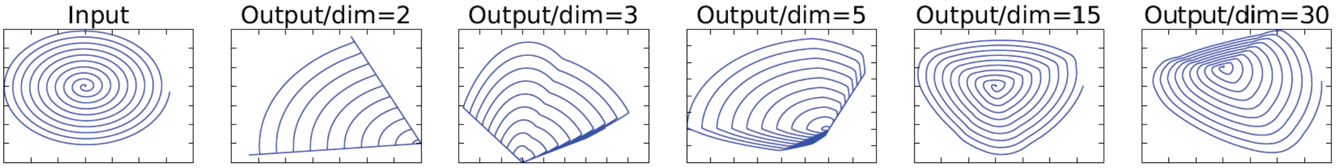


Fig. 3. Examples of ReLU transformations

MobileNet V3[12] introduces: (1) complementary search techniques, (2) new efficient versions of nonlinearities that are practical for the mobile setting, (3) different network design with high efficiency, and (4) a new segmentation decoder. MobileNet V3 is applied for developing the most satisfactory mobile computer vision architectures that can optimize the accuracy-latency balance on embedded device.

MobileNet V3 first uses MnasNet[13] for rough structure search, and adopts reinforcement learning to select the optimal configuration from a set of discrete choices. After that, MobileNet V3 uses NetAdapt to fine-tune the architecture, which embodies the supplementary function of NetAdapt[14]. It can adjust the underutilized activation channels. As shown in Fig. 4, MobileNet V3 introduces Squeeze-and-Excitation[15]. The core idea of this neural network is to improve the quality of the representation produced by the network by explicitly modeling the interdependencies between the characteristic channels of network convolution. Specifically, learning automatically captures the importance of each feature channel, improves useful features according to the result, and inhibits features that are not of much use to the current task. To this end, Mobilenet V3 proposes a mechanism that allows the network to perform feature recalibration. Through this mechanism, the network can learn to use global information to selectively emphasize informative features and suppress less useful features.

Authorized licensed use limited to: Rutgers University. Downloaded on May 15,2021 at 02:10:36 UTC from IEEE Xplore. Restrictions apply.

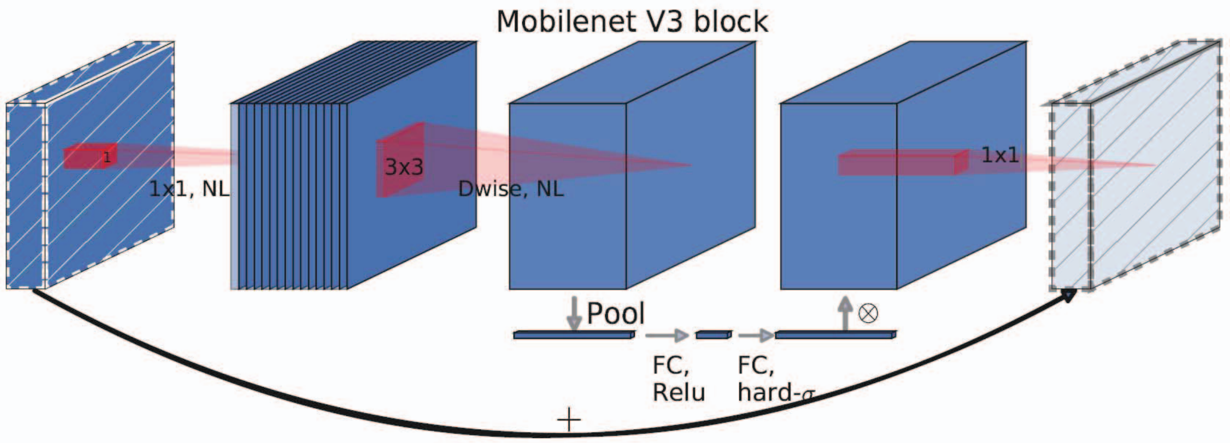


Fig. 4. MobileNetV2 + Squeeze-and-Excite.

### C. *SSD-MobileNet*

As a basic network, MobileNet can extract image features, which can reduce the amount of parameters meanwhile ensuring the performance as much as possible. Therefore, the network is very suitable for mobile devices. By balance the time and accuracy requirements, MobileNet with an appropriate accuracy and speed are constructed based on the width and resolution factor. The basic idea of network structure is to completely separate the correlation between channels, thus greatly reducing amount of calculation and parameters.

SSD-MobileNet algorithm includes two parts. One is the MobileNet network located in the front end, by the initial characteristics of the target can be extracted (based on VGG-16); the second one is the multi-scale feature detection network in back end, and it is to obtain the typical features of the front-end network under different conditions.

As shown in Fig. 5, It is obvious that the information of six scales points to the last detection module, which can help to estimate the target location, classification as well as confidence. At last, the repeated predicted targets are filtered out by non-maximum suppression (NMS) module. The model can effectively extract the information of feature map, and accurately identify the probability and position of the dead fish; it is characterized by the advantages in keeping invariance of displacement and high detection speed, and moreover, it holds good robustness to the changing target.

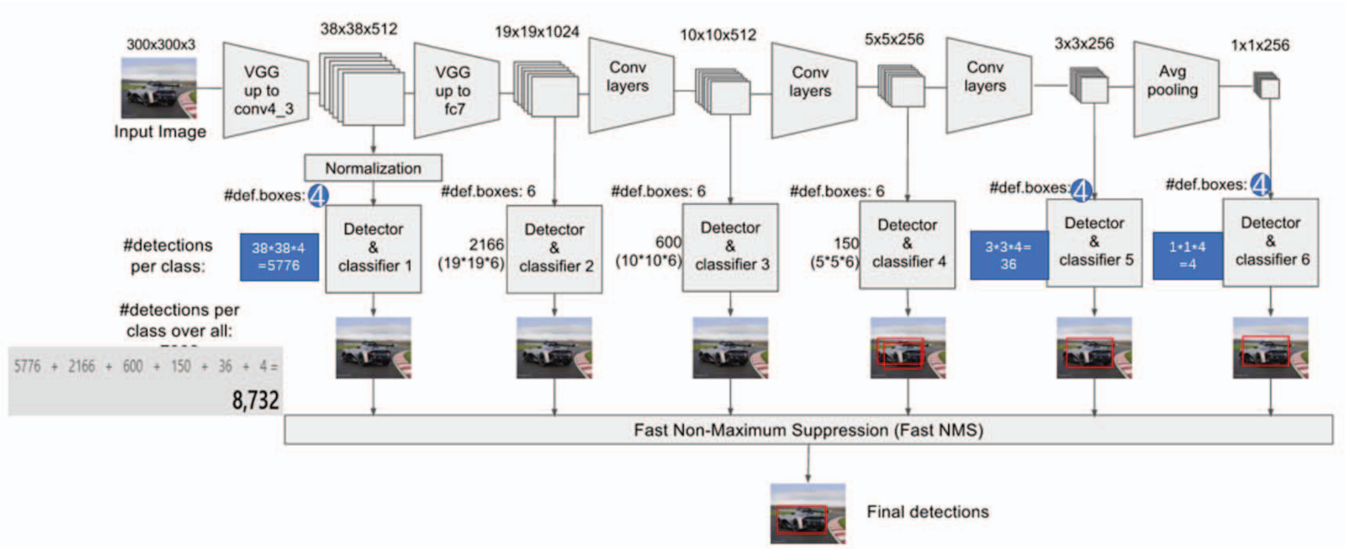


Fig. 5. Network architecture of SSD-MobileNet

## III. Design Of Experiment

In this paper, the TensorFlow platform is used to build a multi-target recognition deep learning model. The main steps of model training based on TensorFlow are: preparing data set; Labelimage calibration data set; training data set; generating recognition model. In engineering applications, we can choose to put the model into the cloud or embedded

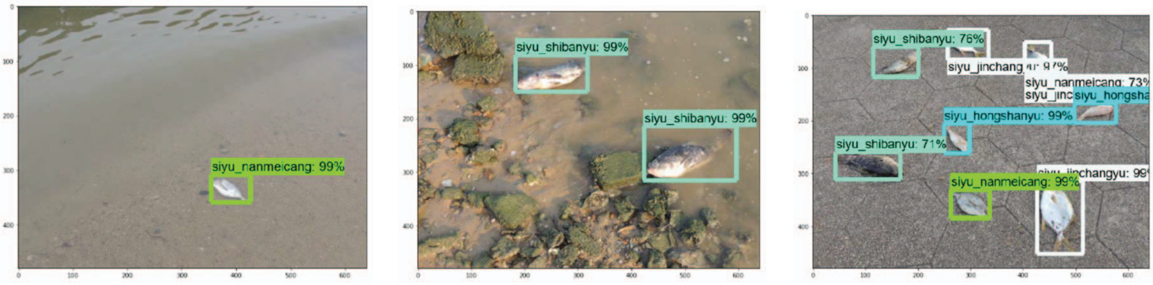
Authorized licensed use limited to: Rutgers University. Downloaded on May 15,2021 at 02:10:36 UTC from IEEE Xplore. Restrictions apply.

devices for edge computing.

The experimental data set is a real picture taken at the seaside. There are 2343 pictures in total. Each picture contains one or more dead fish. The image annotation tool, Labelimage is used to annotate 4 kinds of common fish in the data set, including South American warehouse, Redwood, Grouper and Golden pomfret, to generate the corresponding. xml file, which contains the label information and location information of all the target objects in corresponding image. 1789 of 2343 pieces were picked up randomly as training groups, and the other 554 pieces were taken as test ones. Before the training, the brightness, contrast and saturation of the image are changed randomly by the data enhancement method. At the same time, the image is randomly cut and mirrored, which improves the generalization ability and detection ability of the network training model.

## IV. Experimental Results

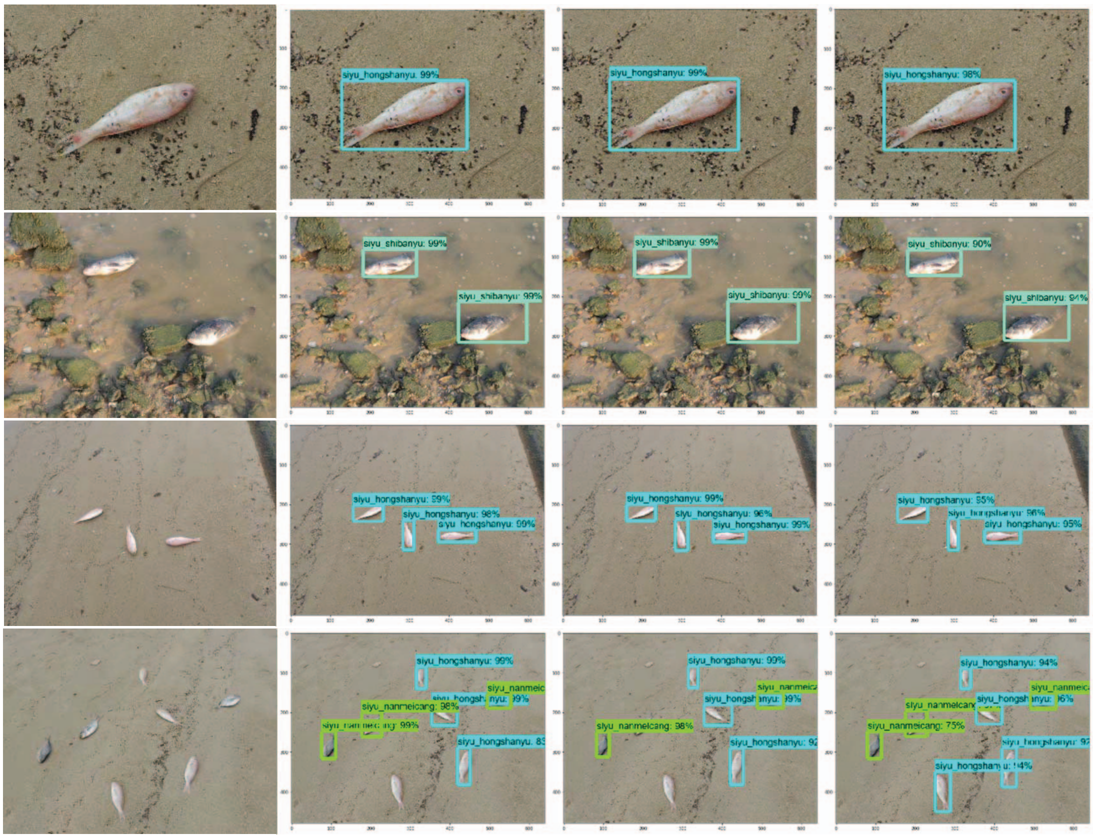
The experiment platform is on GPU. The running time of each step is about 0.4s, and 50K steps are trained in total. Three control experiments were done.Except the model is different, all other conditions are the same.



(a) MobileNet V1 (b) MobileNet V2 (c) MobileNet V3

Fig. 6. Experimental Results

As shown in Fig. 6, the first is MobileNet V1, the second is MobileNet V2, and the third is MobileNet V3. All the network structures are the same except for MobileNet. After the training, the trained model is saved and then we start to do some tests. There are 20 pictures in the images folder in test image. The average detection accuracy of these 20 images is 95%. And the highest accuracy is 99%.



(a) Original ( b ) MobileNet V1 (c) MobileNet V2 (d) MobileNet V3

Fig. 7. Original, MobileNet V1, V2, and V3

Authorized licensed use limited to: Rutgers University. Downloaded on May 15,2021 at 02:10:36 UTC from IEEE Xplore. Restrictions apply.

As shown in Fig. 7, in the experiment, MobileNet V1, V2, and V3 network architectures are used to train the same data set respectively, and the model detection results are shown. By the analysis of Fig . 7, several conclusions can be derived. Among all the global sets of images detected, MobileNet V3 is the best detector. It is obvious that from the figure that the SSD-MobileNet V3 network has realized a good balance in detection performance and accuracy. For example, in the last line of Figure 7, the dead fish in the picture is too small, SSD-MobileNet V1 and SSD-MobileNet V2 are missed target. SSD-MobileNet V3 can get a desirable effect. This is because Mobilenet V3 integrates three models: the deep wise separable convolutions of MobileNet V1, the inverted residual of MobileNet V2 and the lightweight attention model of MnasNet based on the squeeze and exclusion structure. MobileNet V3 is the next generation of MobileNet, which is designed based on complementary search technology and new architecture. After the combination of the hardware network architecture search (NAS) and NetAdapt algorithm, a new architecture is formed. MobileNet V3 realizes how the automatic search algorithm and network design work together improve the overall technical level by using complementary methods. Despite this, the three models behave very similarly on this set of images. By the comparison among the three methods of SSD-MobileNet V1, V2, and V3, it can be concluded that MobileNet V3 is usually more challenging.

## V. Conclusion

SSD-MobileNet is a new type of network model for target recognition, which combines the advantages of light-weight MobileNet network save storage space and low energy consumption, At the same time, it has the characteristics of high efficiency and high precision of SSD network. MobileNet V1, V2, and V3 are compared in the self-made dataset for training and eval. By eval and testing the training model, it can be seen that MobileNet V3 can guarantee the performance in both speed and accuracy. In this experiment, one of the goals of MobileNet V1 and V2 is missed. In the future, we will further explore and make every effort to get an ideal network through experiments.

## Acknowledgment

The authors would like to thank the reviewers and editors for their time to help improve this paper.

## References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol.60 , pp. 1097-1105, May 2017.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014.

[4] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan, "Scale-aware fast R-CNN for pedestrian detection," IEEE transactions on Multimedia, vol. 20, no. 4, pp. 985-996, 2017.

[5] Young-Jin Cha, Wooram Choi, Gahyun Suh, and Sadegh Mahmoudkhani, "Autonomous structural visual inspection using region□based deep learning for detecting multiple damage types," Computer□Aided Civil and Infrastructure Engineering, vol. 33, no. 9, pp. 731-747, 2018.

[6] Mohammed A. Al-masni , Mugahed A. Al-antari , Jeong-min Park , Geon Gi , Tae-Yeon Kim , Patricio Rivera , Edwin Valarezo , Mun-Taek Choi , Seung-Moo Han, and Tae-Seong Kim, "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system," Computer methods and programs in biomedicine, vol. 157, pp. 85-94, 2018.

[7] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," CoRR, abs/1512.03385, 2015.

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "Ssd: Single shot multibox detector," In European conference on computer vision pp. 21-37, October 2016.

[10] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[11] Mark Sandler, Andrew G. Howard, Menglong Zhu AndreyZhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," mobile networks for classification, detection and segmentation. CoRR, abs/1801.04381, 2018.

[12] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo

Authorized licensed use limited to: Rutgers University. Downloaded on May 15,2021 at 02:10:36 UTC from IEEE Xplore. Restrictions apply.

Chen, Mingxing Tan, WeijunWang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam, "Searching for mobilenetv3," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1314-1324, 2019.

[13] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, and Quoc V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," CoRR, abs/1807.11626, 2018.

[14] Tien-Ju Yang, Andrew Howard, Bo Chen,Xiao Zhang, Alec Go, Mark Sandler, Vivienne Sze, and Hartwig Adam, "Netadapt: Platform-aware neural network adaptation for mobile applications," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 285-300, 2018.

[15] Zhao Lin , Kefeng Ji, Xiangguang Leng, and Gangyao Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images." IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 5, 751-755, 2018.

Authorized licensed use limited to: Rutgers University. Downloaded on May 15,2021 at 02:10:36 UTC from IEEE Xplore. Restrictions apply.