Proceedings of 2018 IEEE
International Conference on Mechatronics and Automation
August 5 - 8, Changchun, China

Consecutive Feature Network for Object Detection

Jiaming Huang[1,2], Xiaosong Lan[1], Shuxiao Li[1], Chengfei Zhu[1], Hongxing Chang[1]
[1]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China.
{huangjiaming2016, lanxiaosong2012, shuxiao.li, chengfei.zhu, hongxing.chang}@ia.ac.cn

***Abstract* - Feature Pyramid Network (FPN) is one of the best object detection algorithms in the current object detection field, which uses convolutional neural network (CNN) to detect different scaled objects in an image. However, FPN's feature fusion method ignores the influence of the consecutive feature, which hinders the information flow. In this paper, we proposed an end-to-end image detection model called CFN (Consecutive Feature Network) to overcome this problem and speed up the detection process. Under the premise of equal accuracy, the novel feature fusion method we propose can detect faster than other methods. In the feature fusion module, features from consecutive layers with different scales are merged instead of compartmental layers, which will be fed to the classification and regression subnet to predict the final detection results. On the PASCAL VOC 2007 test, without any data augmentation training skills, our proposed network can achieve 77.1 mAP (mean average precision) at the speed of 3.9 FPS (frame per second) on a single Nvidia 1080Ti GPU. Code will be made publicly available.**

Index Terms – Object detection, deep convolutional neural network, feature fusion, multi-level feature.

I. INTRODUCTION

With the arrival of the era of the big data and the greatly improved hardware computing capabilities, object detection based on deep convolutional neural networks (DCNN) has achieved significant advances in recent years. The current DCNN detectors of state-of-art can be divided into two categories: (1) the two-stage approach, including R-CNN [1], SPP-net [2], Fast R-CNN [3] and Faster R-CNN [4], and (2) the one-stage approach, including YOLO [5], YOLOv2 [6] and SSD [7]. In the two-stage approach, a series of candidate object boxes is first generated, and then they are further sent to predict the corresponding category and perform bounding box regression. The two-stage approach has been achieving dominant detection accuracy on most of the challenging benchmarks, including PASCAL VOC [8] and MSCOCO [9]. While one-stage approach uses the idea of regression, that is to say, given the input image, the bounding boxes and the corresponding categories of the targets can be returned at multiple positions in the image without requiring the generation of those candidate object boxes. The main advantage of this method is its high detection speed. However, its detection accuracy is usually not as good as the two-stage detector.

From accuracy perspective, recently, multi-level feature maps are attached to both the two-stage approach and the one-stage approach to improve the detection accuracy. The one-stage Single Shot MultiBox Detector (SSD) [7] is the first network architecture to combine predictions from multi-level feature maps with different resolutions to naturally handle objects of various sizes. Subsequently, Deconvolutional Single Shot Detector (DSSD) [10] merges the low-level feature maps and the high-level feature maps which are generated by a new deconvolutional module to produce fused feature maps for predictions. Compared with SSD, DSSD improves the detection accuracy by almost 4 mAP. Similarly, two-stage Feature Pyramid Network (FPN) [11] merges low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections. But these connections bring a larger amount of calculations, which harms detection speed.

In addition, some two-stage methods try to add another prediction branch to improve the detection accuracy. He et al. [12] add a branch for predicting an object mask in parallel with the existing branch for bounding box recognition, effectively achieving accurate object detection and semantic segmentation. Although the correctness of the region proposals and the effect of object detection have been slightly improved, it will bring high computational complexity and slow down the detection speed.

From speed perspective, the two-stage approach is lack of competitiveness compared with the one-stage approach, which is largely due to two large fully connected layers for per RoI (Region of Interest) classification and regression. In recent years, there are some works on fully convolutional networks and depthwise separable networks, including R-FCN (Region-based Fully Convolutional Networks) [13], Xception [14], MobileNet [15] and MobileNetV2 [16], which greatly enhance the detection speed. However, the detection accuracy is their deficiency.

In our opinion, considering real-time applications in industry, we intend to improve the feature fusion method of FPN and speed up the process of two-stage detectors under the premise of ensuring certain accuracy. In this paper, we design a novel two-stage object detection framework, called CFN, which utilizes our new consecutive feature fusion method to predict object bounding box and corresponding category. As shown in Fig. 1, the latter feature map performs element-wise summation by convolution and up-sampling operations with the previous feature map's convolution operation. On one hand, this consecutive feature fusion method can increase the richness of the feature expression, and on the other hand, the relationship between the previous and latter feature map can be closer and strengthen the flow of information. The experimental results show that our proposed model is superior to other methods under the premise of equal detection accuracy.

978-1-5386-6072-0/18/$31.00 ©2018 IEEE

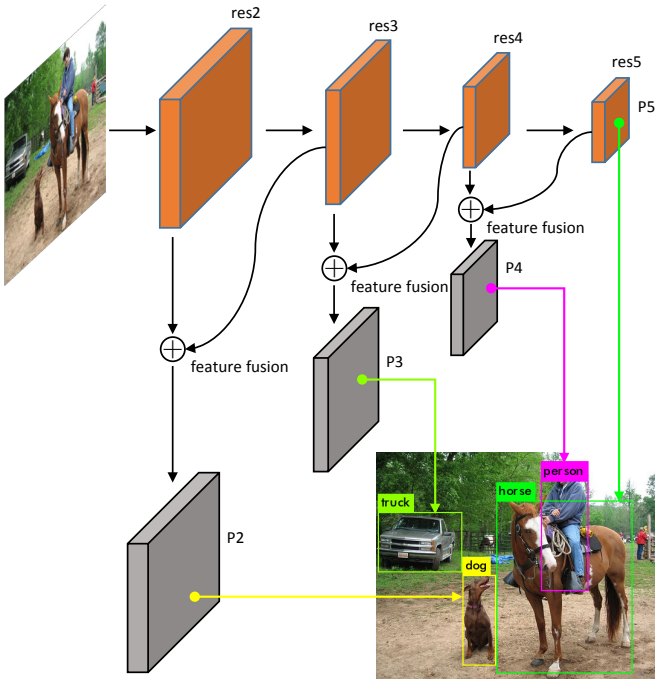

Fig. 1 Our proposed consecutive feature network model

The rest of the paper is organized as follows. Section II introduces related object detection methods and related technologies involved. The proposed consecutive feature network architecture is described in Section III in detail. Section IV gives the experimental results and discussion. The last section concludes the paper.

II. Related Work

A. Traditional Hand-Crafted Approach

Early object detection methods are mainly based on the sliding window and consist of three procedures: 1) using sliding windows of different sizes to convolve with the original image as candidate regions; 2) extracting the feature expression of candidate regions by hand-engineered features; 3) classifying using the trained classifier. As one of the most famous methods, Viola and Jones [17] proposed the face detection method based on haar-like feature and adaboost classifier, which achieves effective detection accuracy with high efficiency. Then, Dalal [18] proposed to use image histograms of oriented gradient (HOG) to extract features for pedestrian detection. After that, Felzenszwalb [19] proposed a multi-scale deformable part module (DPM) based on HOG and support vector machine (SVM) for performance extension. However, with the development of modern deep convolutional neural network, object detectors like R-CNN [1] show dramatic improvements in accuracy, which can be mainly divided into the two-stage approach and one-stage approach.

B. Two-Stage Approach

The dominant model in modern object detection is based on a two-stage approach. The first stage generates a series of candidate object proposals that may contain objects from the whole image pyramid space, such as Selective Search [20], EdgeBoxes [21], and the second stage classifies the candidate object proposals from the first stage into object categories or background and locates the accurate object regions by classification network and regression network. Region Proposal Network (RPN) [4] upgraded the first stage and generated the candidate object proposals by using network features instead of extra feature extraction, which achieved end-to-end training and test. Following RPN, numerous extensions, such as Deformable Convolutional Networks [22] and FPN [11], have been proposed to increase the detection accuracy by enhancing the transformation capacity of CNN model and exploiting inherent multi-scale, pyramidal hierarchy of deep convolutional networks respectively. On the other hand, for the purpose of improving the detection performance of the second stage, Dai et al. [13] proposed to use the position-sensitive score maps generated by a fully convolutional network to replace the fully connected network after the RoI (Region of Interest) pooling layer, which accelerated the real-time detection process to some extent. Subsequently, some research works are conducted on modifying the convolution mode in recent years. Xception [14] presented to use depthwise separable operation (a depthwise convolution followed by a pointwise convolution) to replace the regular convolution in the backbone convolutional neural network architecture, which drastically increase the speed of the deep convolutional neural network.

C. One-Stage Approach

Object detection research also strives to increase the speed of detection. Considering the high efficiency, one-stage detector arises spontaneously, which pays more attention to the real-time performance of network model. YOLO (You Only Look Once) [5] simplifies object detection as a regression problem, which directly predicts the object bounding boxes and the corresponding class probabilities without proposals generation. Benefit from this method, YOLO can run object detection at a very high speed but the accuracy is not satisfactory enough. YOLOv2 [6] is the enhanced version of YOLO and it improves the YOLO by removing the fully connected layers and adoptes anchor boxes like the RPN. Similarly, combining the regression idea of YOLO and the anchor mechanism of RPN [4], SSD [7] further improves detection performance by producing predictions of different scales from multiple feature maps. In order to improve SSD's detection accuracy, especially for small objects, Fu et al. [10] propose the DSSD method, which introduces additional context via deconvolution layers into object detection.

In conclusion, the one-stage detectors have made good progress, but their accuracy still trails that of two-stage detectors. The goal of this paper is to improve the detection performance of FPN [11] and design a better two-stage detector called Consecutive Feature Network through better consecutive feature fusion method.

III. Network Architecture

Having achieved significant advances in computer vision challenges, deep CNN model with multiple layers has shown great power in object detection. As shown in Fig. 1, this section

describes our proposed object detection framework. Where res2, res3, res4 and res5 are the blocks of the deep residual network (ResNet) [23]. Besides, P2, P3, P4 feature maps are generated by our feature fusion method respectively, and P5 feature map is the convolutional result of res5 block. We first introduce the backbone CNN for feature extraction in section III.A. Then in section III.B, we elaborate the proposed consecutive feature fusion module in detail. Finally, the predictions are made independently at all levels that combines high-level semantic information and low-level local information, and we supervise losses at these layers in section III.C.

A. Backbone CNN

Recent research reveals that the neural network depth plays a crucial role in computer vision task, such as classification, object detection, segmentation, etc. But deeper stacked neural networks are more difficult to train, which is the result of vanishing/exploding gradients. In order to make full use of the benefit of deep expression, our model is built on the deep residual network (ResNet) [23] framework. As shown in Fig. 2, the main difference between plain network and ResNet is the shortcut connection.

Until recently, the majority of neural networks consisted of linear sequence layers, such as VGG [24]. As depicted in Fig. 2(a), each layer computes a function F and the output x_n of the n-th layer is expressed as follow

$$x_n = F\left(x_{n-1}; W_n\right) \tag{1}$$

Here, W_n is the weight parameters of the layer.

However, ResNet consist of a sequence of residual units. As shown in Fig. 2(b), the output x_n of the n-th residual unit in a ResNet is expressed as follow

$$x_n = x_{n-1} + F\left(x_{n-1}; W_n\right) \tag{2}$$

Here, $F(x_{n-1}; W_n)$ denotes the residual, which is computed by weight parameters W_n.

It has been evidently observed that ResNet has superiority over traditional plain network in training, which can be explained by the further gradient flow within the network. In order to prove this, consider the n-th and m-th residual units where n>m. Through (2) and recursion, we can get following formula

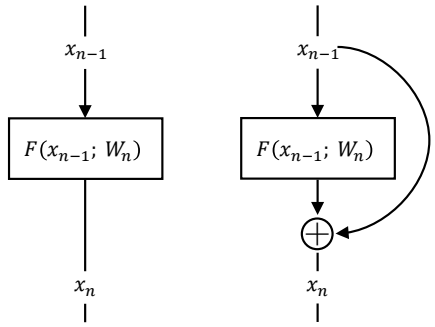

Fig. 2 (a) Layer in a plain network. (b) Residual unit

$$x_n = x_m + \sum_{i=m}^{n-1} F\left(x_i; W_{i+1}\right) \tag{3}$$

Here, i denotes the index of residual units.

Furthermore, if the loss is l, we can use the chain rule of calculus and express the partial derivative of l with respect to the output x_m of the m-th layer.

$$\frac{\partial l}{\partial x_m} = \frac{\partial l}{\partial x_n}\frac{\partial x_n}{\partial x_m} = \frac{\partial l}{\partial x_n} + \frac{\partial l}{\partial x_n}\sum_{i=m}^{n-1}\frac{\partial F\left(x_i; W_{i+1}\right)}{\partial x_m} \tag{4}$$

Next, we will find

$$\frac{\partial l}{\partial W_m} = \frac{\partial l}{\partial x_m}\frac{\partial x_m}{\partial W_m} = \frac{\partial x_m}{\partial W_m}\left(\frac{\partial l}{\partial x_n} + \frac{\partial l}{\partial x_n}\sum_{i=m}^{n-1}\frac{\partial F\left(x_i; W_{i+1}\right)}{\partial x_m}\right) \tag{5}$$

Then, we can see that the latter is relevant to the depth n. And the former is independent of the depth n, which won't hinder the gradient flow from deep unit to shallow unit. Therefore, we exploit ResNet-50 framework as our backbone CNN.

B. Feature Fusion Module

In general, the deep neural network structure is composed of a stack of convolutional layers, which reduces the memory usage in the spatial dimension by down-sampling pooling. At the same time, it drastically increases the dimensions of feature channels to ensure classification accuracy. After all, the deep convolutional neural network can increase the diversity of high-level semantic expression. Therefore, during the design of the feature fusion module, we should take into account the

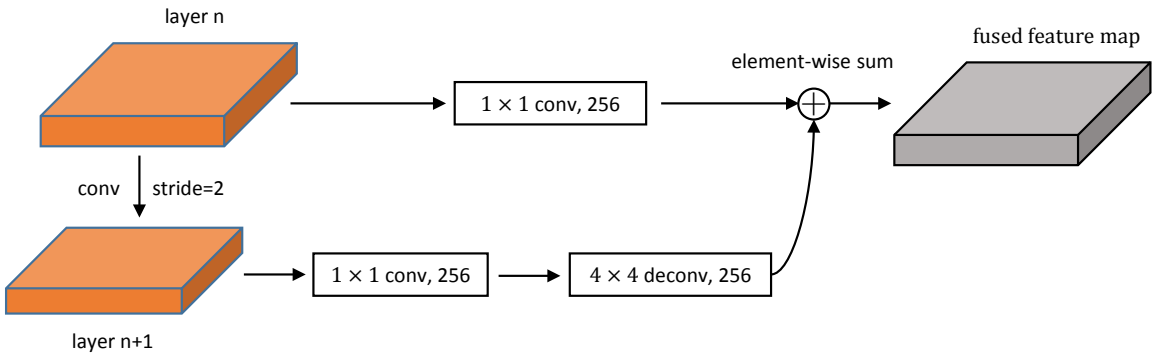

Fig. 3 Our proposed feature fusion module

matching problem of the spatial dimensions of feature maps and the dimensions of feature map channels.

As depicted in Fig. 3, the latter $n+1$-th layer is half of the spatial dimension of the n-th layer, on the contrary, the dimension of feature channels is twice that of it. Thus, when the number of feature channel does not match, we can change the channel dimension by 1×1 convolutional layer. On one hand, 1×1 convolutional layer can reduce the channel dimension, and on the other hand, it can reduce the connection parameters and computational complexity. Then, we use deconvolution layer with a stride of 2 to upsamle the $n+1$-th layer's spatial resolution after 1×1 convolution. Finally, the upsampled feature map is merged with the corresponding former feature map that goes through a 1×1 convolutional layer to reduce channel dimensions by element-wise summation. In table I, the fused feature maps are shown in detail. Compared with FPN's top-down and compartmental connection mode, our consecutive connection mode enhances the information flow and makes the detection speed faster. Our experiment shows that our proposed consecutive feature fusion method is better than that of FPN.

TABLE I
FEATURE FUSION RULES

fused feature map	component
P2	res2 + res3_up
P3	res3 + res4_up
P4	res4 + res5_up
P5	res5

C. Multi-Level Supervised Learning

Generally, the deep layers have stronger semantic information and the shallow layers have higher resolution local information. In this section, we combine deep layers and shallow layers for object detection. For the feature maps merged in section 3.2, they will be sent to R-CNN subnet for predictions simultaneously. In the R-CNN series, the R-CNN subnet refers to the class-specific classification network and the bounding box regression network, thus, the whole network losses are composed of four parts, including RPN's classification loss and bounding box regression loss, R-CNN subnet's classification loss and bounding box regression loss respectively. Finally, the whole network multi-task losses are supervised as follow

$$L = \sum_p l_p = \sum_p rpn\_cls_p + rpn\_bbox_p + rcnn\_cls_p + rcnn\_bbox_p \quad (6)$$

Where p denotes the merged feature map.

In next section, the implementation details are presented by contrasting experiments.

IV. Experiment and Analysis

Our model is trained on the collection of PASCAL VOC2007 trainval and PASCAL VOC2012 trainval that has 20 object categories, and then evaluates the results on the PASCAL VOC2007 test which has 4952 images. We implement our network by well-known open-source deep learning library Caffe. The backbone ResNet-50 model is pretrained on the ILSVRC classification dataset. And then we finetune our proposed model by stochastic gradient descent (SGD) with learning rate of 10^{-4}, and the learning rate decays by a factor of 0.1 every 50000 steps. We train 160000 iteration on a single Nvidia 1080Ti GPU.

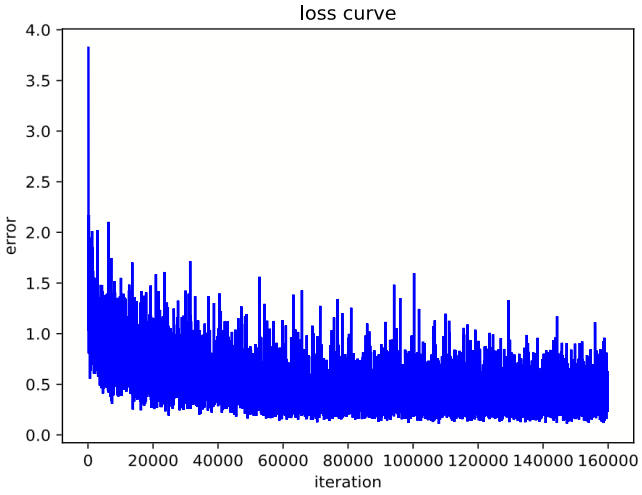

Fig. 4 The training loss curve

A. The Training and Test Results

The training loss curve is shown in Fig. 4. From training loss curve, it can be seen that the training loss converges at 60000 iterations, only a slightly little jitter. But the test performance of 60000 iterations' caffe model is far inferior to that of 160000 iterations' caffe model. The main reason is that it's not enough to fit other data when the loss decreases to the convergence point exactly. So it requires further convergence and approximation, and the final test experiment is based on 160000 iterations' caffe model. Besides, we evaluate the object detection accuracy by mean average precision (mAP) and the detection details on PASCAL VOC2007 test dataset are shown in the following table II. In fact, the training dataset plays a

TABLE II
THE DETECTION RESULTS ON PASCAL VOC 2007 TEST DATASET

mAP	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
77.1	85.9	83.2	77.0	67.4	59.9	85.1	86.1	89.1	58.0	81.5
table	dog	horse	motorbike	person	plant	sheep	sofa	train	tv	
71.0	87.9	84.6	82.8	78.5	49.3	75.7	77.5	87.1	73.9	

significant role in detection performance. When we trained only on VOC2007 trainval dataset, the detection accuracy is 6 mAP lower than our model. Thus, we can see that without data augmentation training skills, the detection accuracy of our model is predominant.

B. Inference Time

In this section, we compare several two-stage object detection methods with ours. As shown in table III, except that the detection result of Faster R-CNN model based on VGG [24] comes from the original paper, the other two experimental results are conducted on a single Nvidia 1080Ti with Intel(R) Core(TM) i7-6850k CPU @ 3.60GHz. Besides, the mAP and FPS results are averaged after many tests. Obviously, our CFN model outperforms more accurate than Faster R-CNN model by nearly 4 mAP, but the speed is slightly slower, which is the result of complicated multi-level supervised learning. In the case where our CFN model is almost as accurate as the FPN model, our model detects 15 percent faster than FPN model, which benefits from the lower amount of computation that our proposed consecutive feature fusion method brings.

TABLE III
COMPARATIVE EXPERIMENTS WITH OTHER METHODS

Method	mAP	FPS
Faster R-CNN (VGG 16) [4]	73.2	5
FPN [11]	77.3	3.4
CFN (ours)	77.1	3.9

C. The Influence of Multi-Level Supervision

Generally, combining the deep and shallow layers can bring improvement of detection accuracy. In this section, we do ablation study on whether reducing level influences the detection or not and what effect it brings. As shown in table IV, we can see that the detection performance based only on single level is far inferior to that of our all levels'. Furthermore, compared with multi-level without res6 which denotes the pooling result of the res5, our model based on all levels also outperforms more accurate than it. These comparisons indicate that the multi-level supervision plays a significant role in detection performance and our multi-level feature is superior to single-level feature for object detection.

TABLE IV
ABLATION EXPERIMENTS

Level	mAP	FPS
baseline on P5	73.8	5.4
multi-level, without res6	75.9	4.1
Our all levels	77.1	3.9

V. CONCLUSION AND FUTURE WORK

We propose a novel two-stage network architecture for object detection, which exploits our new consecutive feature fusion method to predict object bounding box and corresponding category. And the proposed network utilizes the rich feature expression capability of deep convolutional neural network and combines semantic information with multi-level feature maps. The final experimental results show that our network model is superior to Faster R-CNN model and FPN model in detection accuracy aspect and detection speed aspect respectively. In conclusion, our proposed model achieves competitive detection accuracy and speed in object detection. In future work, we are going to further improve our detection speed with the help of depthwise separable convolution, which is widely used in mobile and embedded applications. And we hope that our proposed feature fusion method can be extended to other computer vision tasks, such as semantic segmentation, etc.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China under Grants 61573350.

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, 2014, pp. 580-587.

[2] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.

[3] R. Girshick, "Fast R-CNN," *2015 IEEE International Conference on Computer Vision (ICCV)*, Washington, DC, USA, 2015, pp. 1440-1448.

[4] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[5] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788.

[6] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," arXiv: 1612.08242, 2016.

[7] W. Liu, et al, "SSD: Single Shot Multibox Detector," 2016 *European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016, pp. 21-37

[8] M. Everingham, L. V. Gool, C. K. Williams, J. Winn and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," in *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, 2010.

[9] T. Lin, et al, "Microsoft COCO: Common Objects in Context," 2014 *European Conference on Computer Vision*, Zurich, Switzerland, 2014, pp. 740-755.

[10] C. Fu, W. Liu, A. Ranga, A. Tyagi and A. C. Berg, "DSSD: Deconvolutional Single Shot Detector," arXiv: 1701.06659, 2017.

[11] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, 2017, pp. 936-944.

[12] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2018, pp. 2980-2988.

[13] J. Dai, Y. Li, K. He and J. Sun, "R-FCN: Object Detection via Region-Based Fully Convolutional Networks," arXiv: 1605.06409, 2016.

[14] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, 2017, pp. 1800-1807.

[15] A. G. Howard, et al, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv: 1704.04861, 2017.

[16] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov and L. Chen, "Inverted Residuals and Liner Bottlenecks: Mobile Networks for Classification, Detection and Segmentation," arXiv: 1801.04381, 2018.

[17] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *2001 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition (CVPR), Kauai, Hawaii, USA, 2001, pp. 511-518.

[18]B. Triggs and N. Dalal, "Histograms of Oriented Gradients for Human Detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, 2005, pp. 886-893.

[19]P. F. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," in *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, 2010.

[20]J. R. Uijlings, K. E. Sande, T. Gevers and A. W. Smeulders, "Selective Search for Object Recognition," in *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154-171, 2013.

[21]C. L. Zitnick and P. Dollar, "Edge Boxes: Locating Object Proposals from Edges," *2014 European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 2014, pp. 391-405.

[22]J. Dai, et al, "Deformable Convolutional Networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2018, pp. 764-773.

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

[24]K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv: 1409.1556, 2014.