Computers, Materials & Continua
DOI:10.32604/cmc.2021.017073
Article

Tech Science Press

Real-Time Recognition and Location of Indoor Objects

Jinxing Niu[1,*], Qingsheng Hu[1], Yi Niu[1], Tao Zhang[1] and Sunil Kumar Jha[2]

[1]Institute of Mechanics, North China University of Water Resources and Electric Power, Zhengzhou, 450011, China
[2]IT Fundamentals and Education Technologies Applications, University of Information Technology and Management in Rzeszow, Rzeszow, 100031, Poland
[*]Corresponding Author: Jinxing Niu. Email: njx.mail@163.com

Received: 18 January 2021; Accepted: 20 February 2021

Abstract: Object recognition and location has always been one of the research hotspots in machine vision. It is of great value and significance to the development and application of current service robots, industrial automation, unmanned driving and other fields. In order to realize the real-time recognition and location of indoor scene objects, this article proposes an improved YOLOv3 neural network model, which combines densely connected networks and residual networks to construct a new YOLOv3 backbone network, which is applied to the detection and recognition of objects in indoor scenes. In this article, RealSense D415 RGB-D camera is used to obtain the RGB map and depth map, the actual distance value is calculated after each pixel in the scene image is mapped to the real scene. Experiment results proved that the detection and recognition accuracy and real-time performance by the new network are obviously improved compared with the previous YOLOV3 neural network model in the same scene. More objects can be detected after the improvement of network which cannot be detected with the YOLOv3 network before the improvement. The running time of objects detection and recognition is reduced to less than half of the original. This improved network has a certain reference value for practical engineering application.

Keywords: Object recognition; improved YOLOv3 network; RGB-D camera; object location

1 Introduction

Object recognition and location have important application value in the service robot, industrial automation, self-driving, and other fields. In the traditional methods, the objects in the scene are mostly segmented by the features of the objects to achieve the purpose of recognition [1–4]. According to the nature of feature extraction, the object recognition method can be classified into global feature and local feature. Global features refer to the overall attributes of the image, including features such as color, texture, and shape, which are characterized by good invariability, simple calculation, and intuitive representation [5]. Local features refer to the features extracted from the local area of the image, including features such as edges, corners, lines, etc., which have a small degree of correlation between features and will not affect the detection and matching of other features due to the disappearance of some features under occlusion [6]. Traditional scene object recognition methods can only segment and recognize simple objects with low recognition accuracy and efficiency. In the case of complex indoor scenes, object recognition effect is even worse. For the location of scene objects, the common method is to use binocular stereo vision to calculate the image parallax map and obtain the location information of each pixel in the whole scene image, which may have defects such as low calculation accuracy and slow speed. In recent years, with the development and application of deep learning technology in the direction of machine vision, it provides a new research direction for scene object recognition [7]. For example, Yang et al. [8] proposed an improved fast-YOLO model, combined with semi-global matching SGM algorithm to process binocular image data, and realized the detection and location of scene pedestrians. Peng et al. [9] used the Mask R-CNN neural network model combined with the nearest point search algorithm (ICP) to realize the recognition and location of scene objects in binocular vision.

In view of the problems such as low accuracy of scene object recognition and location, this article proposes an improved YOLOv3 neural network combining the with RGB-D camera to realize real-time identification and location of indoor objects.

2 Method

2.1 Object Recognition

Compared with other neural network of object detection and recognition, (such as Fast R-CNN [10], Faster R-CNN [11], Mask R-CNN [12], etc.), The YOLOv3 neural network model uses regression to extract the features of objects in the image, without generating a large number of candidate windows, it directly uses a single neural network model to predict and classify objects in the input image, realizing end-to-end objects detection and recognition [13]. The YOLOv3 neural network model can achieve relatively fast object detection and recognition while ensuring high accuracy, so it is more suitable for object detection and recognition scenes with high real-time requirements.

In our experiments, it is found that when YOLOv3 network is applied to object recognition in indoor scenes, there are problems such as slow recognition speed and object missing detection in dark places. Aiming to solve the above problems, this article proposes an improved YOLOv3 network to increase the real-time performance of the network and the robustness of object detection and recognition. The DarkNet53 is used by YOLOv3 as the backbone network, the network is characterized by the use of a deep residual network (ResNet) to alleviate the problem of gradient disappearance caused by increasing depth. When the network is deeper, the effect of feature transmission is gradually weakened, which will reduce the robustness of the network to object detection and recognition. This article proposes a method to build the backbone network of YOLOv3 by combining the deep residual network (ResNet) with the densely connected convolutional networks (DenseNet).

2.1.1 Deep Residual Network

In a neural network, more abundant image features are obtained by stacking the number of layers of the network, so the increase in the number of network layers means that the extracted features of different layers are richer. However, this is not the case. As the network deepens, the performance of the network quickly reaches saturation, and then even performance degradation begins. In order to solve the problem of network performance degradation, He et al. [14] proposed a deep residual network model, which greatly solved the problem of performance degradation caused by network deepening. Depth of the residual network is the core of the block as a unit, as shown in Fig. 1. Each cell block is composed of a series of layers and a shortcut. The shortcut connects the input and output of the module together, and then add at the element level crossing the middle tier. It does not produce additional parameters and increase the complexity of the calculation, and ensure that the performance of the network after deepening will not be worse than before.

In a deep residual network, because adding shortcut identity mapping, it can express the input and output relationships at level l by:

$$X_l = H_l\left(X_{l-1}\right) + X_{l-1}. \tag{1}$$

In this formula, X_{l-1} is the input image feature, l is the layer number of neural network, $H_l(\cdot)$ stands for nonlinear operational combinations including BN (Batch Normalization), ReLu, 3×3 convolution, and X_l is the output image feature.

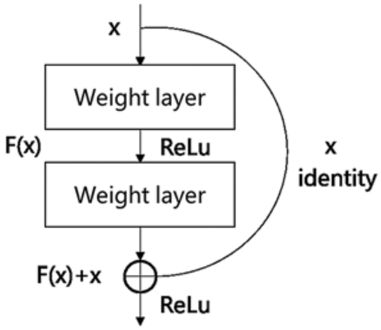


Figure 1: Deep residual network unit

2.1.2 Densely Connected Convolutional Networks

In the deep residual network, with the deepening of the network, there are more parameters of the network, and the utilization rate of the network structure is not high. In 2017, Huang et al. [15] proposed a densely connected convolutional networks (DenseNet). The output of each layer of the network model is imported into all subsequent layers. Unlike the deep residual network, the densely connected network structure uses a concatenate structure. This structure can reduce the network model parameters. It is possible to retain all levels of low-level features in high-level features as much as possible, and further realize the multiplexing and fusion of network multi-layer features. In this paper, the densely connected convolutional network is applied to the backbone network of the YOLOv3 network, combining with the deep residual network, to improve the robustness of indoor object detection and recognition. The densely connected convolutional networks are composed of two parts: the Dense Block and the Transition Layer, as shown in Fig. 2.

In DenseNet, the input and output relationship of layer l can be represented by the following formula since the current layer is closely connected to all the subsequent layers:

$$X_l = H_l\left(\left[X_0, X_1, \ldots, X_{l-1}\right]\right) \tag{2}$$

where $[X_0, X_1, \ldots, X_{l-1}]$ represents image feature stitching from the image input layer to the $l-1$ layer. $H_l(\cdot)$ contains six continuous nonlinear transformations, such as BN, ReLu, 1×1 convolution, BN, ReLu, and 3×3 convolution.

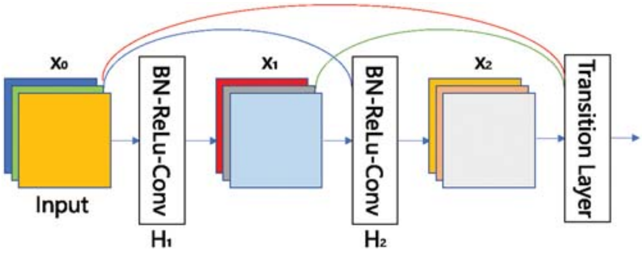


Figure 2: Densely connected convolutional networks

2.1.3 Improved YOLOv3 Networks

Deep residual network and densely connected convolutional networks are combined as the backbone network of YOLOv3. After many times of improvement and experimental verification, an improved network model is finally obtained, and the new backbone network structure is shown in Tab. 1.

Table 1: The new backbone network structure

Network layer	Number of convolution kernels	Convolution kernel size	Output size
Conv1	32	$3 \times 3/1$	208×208
Conv2	64	$3 \times 3/1$	208×208
Conv3	128	$3 \times 3/2$	208×208
Max pooling		$2 \times 2/2$	104×104
Dense block		$\begin{bmatrix} 64 & 1 \times 1 \\ 128 & 3 \times 3 \end{bmatrix} \times 2$	104×104
Res block (1)	256	$3 \times 3/2$	52×52
		$\begin{bmatrix} 128 & 1 \times 1 \\ 256 & 3 \times 3 \end{bmatrix} \times 6$	52×52
Res block (2)	512	$3 \times 3/2$	26×26
		$\begin{bmatrix} 256 & 1 \times 1 \\ 512 & 3 \times 3 \end{bmatrix} \times 6$	26×26
Res block (3)	1024	$3 \times 3/2$	13×13
		$\begin{bmatrix} 512 & 1 \times 1 \\ 1024 & 3 \times 3 \end{bmatrix} \times 6$	13×13

In the improved combined network, the input image is subjected to 1 densely connected convolution networks and 3 deep residual networks to extract features. Compared with the original DarkNet53 backbone network, the new combined network uses densely connected convolution networks and shallow images features can be communicated to deep convolutions better and faster, realizing multi-feature multiplexing and fusion. It can effectively increase the information transmission efficiency and gradient transmission efficiency of the entire network, which is conducive to the combination of up sampling and shallow features. The new combined network still uses the residual network as feature extraction network. The feature extraction network structure has three residual blocks, which improves the ability to select and extract features. Fig. 3 shows the new backbone network.

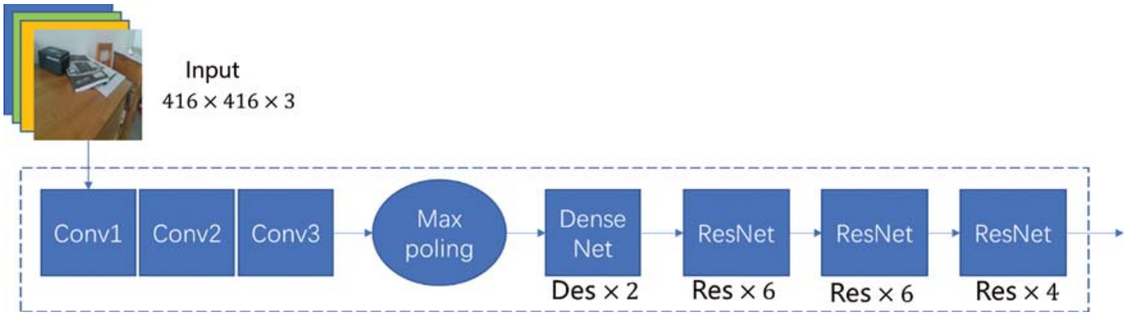


Figure 3: The new backbone network diagram

2.2 Object Location

The RGB-D camera used in this article is a RealSense D415 depth camera produced by Intel. This camera can obtain both RGB and depth maps, and can easily build point cloud maps based on the obtained RGB and depth maps. It has binocular infrared camera combined with infrared structured light coding technology, and it can obtain the depth information of the scene quickly and conveniently.

In order to ensure that the collected RGB image and depth map are at the same time and in the same perspective, it is necessary to align the collected RGB image and depth map. The alignment operation can convert the depth map coordinate system to the RGB image coordinate system, so that the pixels in the RGB image are correspond to the depth values expressed in the depth map. The depth value in depth map is first converted to the space point in the world coordinate system, and then the space point is projected into the point in the RGB image. Each pixel in the RGB image will have a one-to-one correspondence with the pixel in the depth map.

In order to locate the object in the scene, the improved YOLOv3 network is used to detect and identify the object in the indoor scene. The pixel coordinates of the center point of the object frame are calculated. The depth value of the center point is read by applying the one-to-one correspond relationship between RGB image and depth map, and the actual space distance is calculated according to the depth value.

3 Experiments and Results

3.1 Experiments

In order to verify the effect of the networks, we do some experiments with a RealSense D415 camera. RealSense SDK 2.0 is used to provide API interface driver for RealSense D415 RGB-D sensors, and it can read the RGB image and depth map with the frame rate at 25 fps.

In this experiment, 3584 indoor scene pictures are collected, 2560 of which are used as training set data and 1024 as verification set data. LabelMe annotation tool is used to label the objects in the indoor scene pictures. The objects are labeled mainly include cups, people, books, tables, chairs, mobile phones and computers. The annotation information is exported as a Json file. The marking process is shown in Fig. 4.

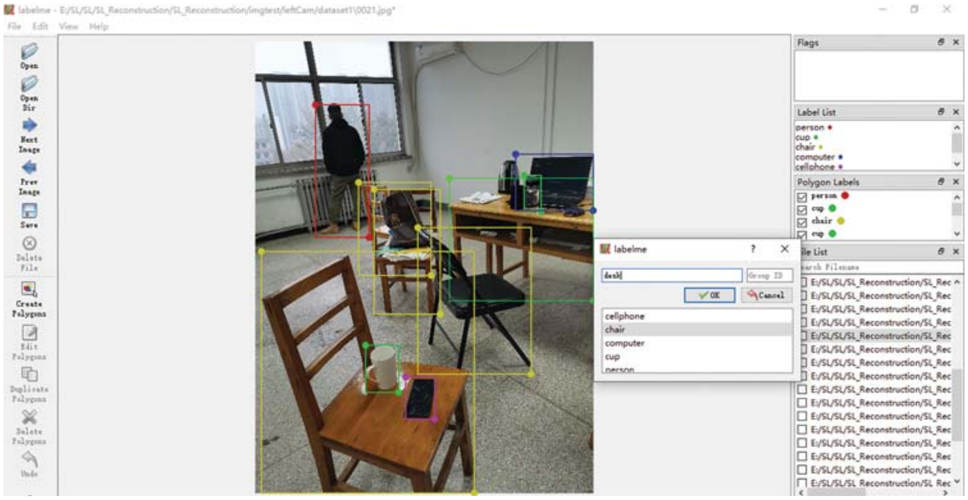


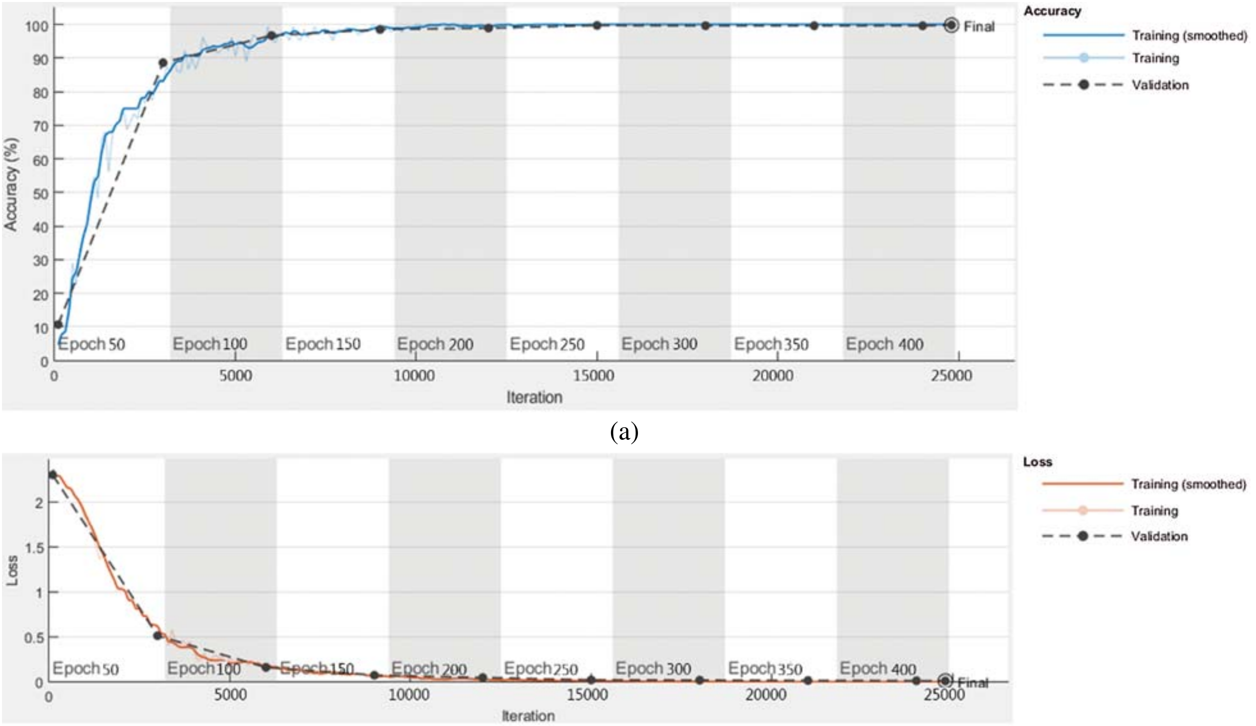
Figure 4: Labeling scene objects in LabelMe

In order to increase the size of the training set, we use transfer learning to train the collected indoor scene data based on COCO dataset. COCO data set is a huge deep learning data set released by Microsoft. It includes about 200000 tagged images, more than 1.5 million object instances, and a total of 80 categories [16]. The experimental platform used in this paper is: Intel Core i7-8300k CPU, GeForce GTX 1080 GPU, 16G RAM, Windows 10 system. The training rate is set to 0.001. Multi-scale strategy is used to generate more training samples by rotating angle, adjusting saturation or exposure, which improves the robustness of the network model for object recognition in different indoor environments.

Fig. 5 is the curve of accuracy and loss with the number of iterations when the improved YOLOv3 network is used to train the indoor scene data set. It can be seen from the figure that when the training network is iterated to about 20,000 times, the accuracy has basically tended to 100%, and the loss value has basically stabilized. From the perspective of parameter convergence, the network training results are ideal.

3.2 Results

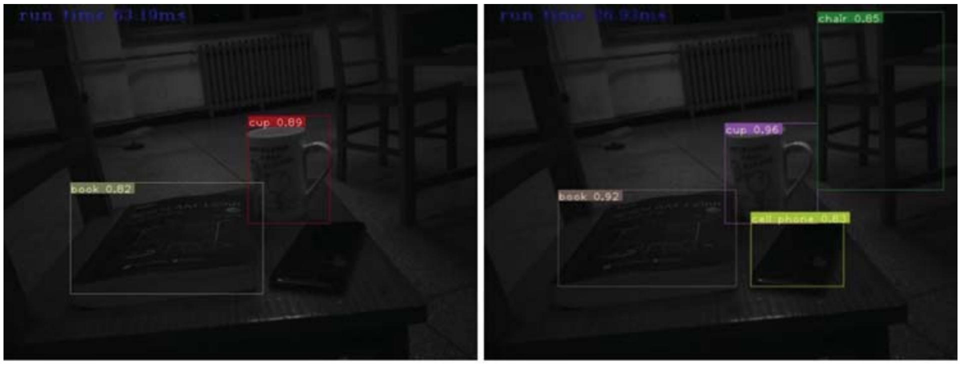
It is found that the improved YOLOv3 network not only effectively improves the robustness of indoor object detection and recognition, but also the speed of detection and recognition is greatly improved, which can achieve the effect of real-time processing. Fig. 6 shows the comparison diagram of object detection and recognition in the same darkroom scene diagram by YOLOv3 network before and after the improvement. It can be clearly seen from the figure that the chair and cellphone can be detected after the improvement which cannot be detected with the YOLOv3 network before the improvement. The running time of the network before and after the improvement is 63.19 ms and 26.93 ms respectively. The improved network has improved the accuracy and real-time performance of detection and recognition.



(a)

(b)

Figure 5: The curve of accuracy and loss with the number of iterations. (a) Accuracy (b) loss value



(a)

(b)

Figure 6: Comparison diagram of YOLOv3 network detection and recognition before and after improvement. (a) Before improvement (b) after improvement

Fig. 7 is the result of real-time detection, recognition, and location of indoor scenes. From the Fig. 7, we can see that the person is 1.57 m away from RGB-D camera, the book is 0.9 m away from RGB-D camera, the cup is 0.844 m away from RGB-D camera, and the chair is 1.092 m away from RGB-D camera. From the experimental results, the improved YOLOv3 neural network model combined with RGB-D camera can effectively identify objects in the scene and determine the position of objects relative to the RGB-D camera.

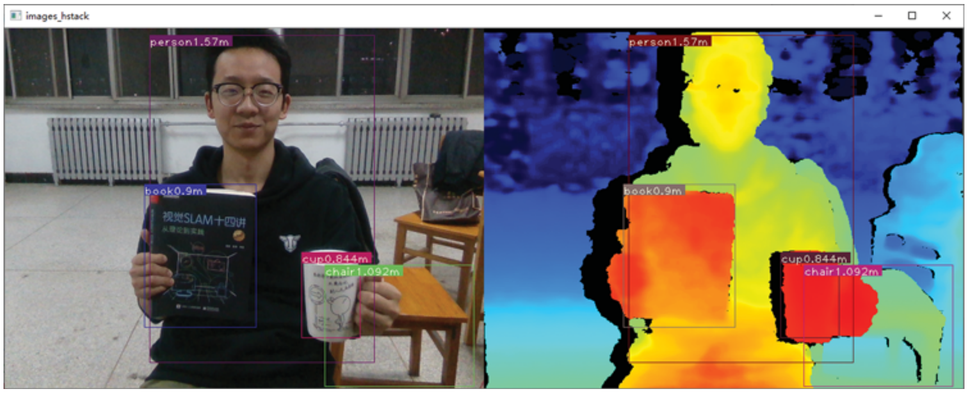


Figure 7: Real-time detection, recognition and location map of indoor scene

4 Conclusion

An improved network model of YOLOv3 is proposed, which backbone network is constructed by combining the dense connection network with the deep residual network. The improved network is utilized to realize real-time recognition and location of indoor scene objects combined with RGB-D camera. Experiment results proved that the detection and recognition accuracy and real-time performance by the new network are obviously improved compared with the previous YOLOV3 neural network model in the same scene. More objects can be detected which cannot be detected with the YOLOv3 network before the improvement. The running time of objects detection and recognition is reduced to less than half of the original. This improved network has a certain reference value for practical engineering application.

Acknowledgement: The authors would like to thank the anonymous reviewers and the editor for the very instructive suggestions that led to the much-improved quality of this paper.

Funding Statement: This work was supported by Henan Province Science and Technology Project under Grant No. 182102210065.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. Sun, X. Zhang, S. Peeta, X. He and Y. Li, "A real-time fatigue driving recognition method incorporating contextual features and two fusion levels," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3408–3420, 2017.

[2] X. Zhang, W. Zhang, W. Sun, T. Xu and S. K. Jha, “A robust watermarking scheme based on ROI and IWT for remote consultation of COVID-19,” *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1435–1452, 2020.

[3] J. Niu, Y. Jiang, Y. Fu, T. Zhang and N. Masini, “Image deblurring of video surveillance system in rainy environment,” *Computers, Materials & Continua*, vol. 65, no. 1, pp. 807–816, 2020.

[4] W. Sun, X. Zhang, X. He, Y. Jin and X. Zhang, “Two-stage vehicle type recognition method combining the most effective Gabor features,” *Computers, Materials & Continua*, vol. 65, no. 3, pp. 2489–2510, 2020.

[5] K. Goto, A. J. Wills Kazuhiro and S. E. G. Lea, “Global-feature classification can be acquired more rapidly than local-feature classification in both humans and pigeons,” *Animal Cognition*, vol. 7, no. 7, pp. 109–113, 2004.

[6] D. G. Lowe, “Local feature view clustering for 3D object recognition,” in *Proc. IEEE CVPR*, Maui, Hi, USA, pp. 682–688, 2001.

[7] H. Wu, Q. Liu and X. Liu, “A review on deep learning approaches to image classification and object segmentation,” *Computers, Materials & Continua*, vol. 60, no. 2, pp. 575–597, 2019.

[8] R. Yang, F. Wang and H. Qin, “Research of pedestrian detection and location system based on stereo images,” *Application Research of Computers*, vol. 35, no. 5, pp. 1591–1595, 1600, 2018.

[9] Q. Peng and Y. Song, “Object recognition and localization based on mask R-CNN,” *Journal of Tsinghua University*, vol. 59, no. 2, pp. 135–141, 2019.

[10] R. Girshick, “Fast R-CNN,” in *Proc. IEEE ICCV*, Santiago, pp. 1440–1448, 2015.

[11] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[12] K. He, G. Georgia, D. Pirotr and G. Ross, “Mask R-CNN,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[13] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018. [Online]. Available: https://arxiv.org/abs/1804.02767.

[14] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, Las Vegas, NV, USA, pp. 770–778, 2016.

[15] G. Huang, Z. Liu, K. Q. Weinberger and L. Maaten, “Densely connected convolutional networks,” in *Proc. IEEE CVPR*, Honolulu, Hi, USA, pp. 4700–4708, 2017.

[16] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona *et al.,* “Microsoft COCO: Common objects in context,” in *Proc. ECCV*, Zurich, Switzerland, pp. 740–755, 2014.