JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Deep Learning based Object Detection using Mask RCNN

Hirendra R. Hajare[1], Bhagyashri S. Chandekar[2]
Department of Computer Science Engineering, Ballarpur Institute of Engineering, Ballarpur, Maharashtra, India[1]
Department of Computer Science Engineering, Ballarpur Institute of Engineering, Ballarpur, Maharashtra, India[2]
hirendrahajare@gamil.com[1]
bhagyashrichandekar30@gmail.com[2]

Abstract: Object detection aims to acknowledge all instances of a known class of objects in a picture, like people, vehicles, or faces. In recent times deep learning techniques was applied for detecting objects and people existing methods have some problems with viewpoint changes and occlusion. This paper proposes an answer for automatic object detection by implementing instance segmentation at a pixel level, and a number of other RCNN techniques were also inferred during this paper. The proposed Mask RCNN detects the photographs and marks them with bounding boxes, class labels still as masks. The implemented mask RCNN model was trained and tested with COCO dataset in order that the model is capable to detect objects in an exceedingly congested image. The model was also tested during a custom dataset to test the model performance, and therefore the model obtained 94% mAP for the custom dataset. This paper also mentions the long run scope of this work in order that the robustness and also the reliability of the model may be improved further.

IndexTerms **- Computer vision; CNN; Deep learning; Object detection; Mask RCNN; Fast R-CNN.**

I. INTRODUCTION

Computer vision is a section of AI and it trains the computer to grasp the visual world. The machine can identify and recognize objects with digital images. It reflects the human vision system and helps the pc to grasp the objects in a picture like humans. The prime task of the pc vision is to create the work simple and faster. In computer vision, a number of the applications of CNN are object detection, face recognition, and self-driving cars [7]. artificial intelligence (AI) technique aids the machines to think like humans and replicates the activities of humans infiltrated from everyday lives. The goal is to copy the human brain and function independently. one amongst the benefits of using artificial intelligence is that AI together with higher cognitive process systems can take any critical decision without human intervention [25]. Deep learning is a synthetic intelligence function which is employed in computer vision for better performance and accuracy. In machine learning and computer vision, deep learning techniques like CNN have shown better performance and efficiency [4].

Image classification could be a computer vision task, and therefore the role of image classification is to classify various objects in a picture. Image Classification is required when an image contains two or more objects looking on some attributes of groups or class [8]. the data processing that's carried out during the classification helps to categorize images into different groups. In image classification, input is taken and returns the image containing objects with a category label. Object localization identifies the placement of 1 or more objects in an exceedingly picture which might be enveloped with bounding boxes. Object Localization aims to locate the foremost visible object. Localization of objects is implied in order that the objects are precisely located [9]. In recent times, object detection has gained lot of attention within the field of computer vision [10]. Over a quick period, the vision group has rapidly enhanced beholding and semantic segmentation results [3]. Image classification and localization of objects are incorporated with within the operating phase of object detection. Object detection method localizes and classifies each object in a picture. Thus, in object detection a picture consisting one or more objects would be processed and therefore the output would lean as a picture with one or more bounding boxes together with a category label. Object detection is employed in an exceedingly picture for counting items, determine and track their exact locations by marking them accurately. Object detection techniques for the Mean Average Precision (mAP) metric are typically planned and evaluated [1].

Deep learning approaches use convolutional neural networks (CNNs) to perform end-to-end, unsupervised object detection, which eliminates the requirement for features to be defined and extracted separately. Deep learning-based object detection models are typically divided into two parts. An encoder takes a picture as input and processes it through a series of blocks and layers that learn to extract statistical features which will be used to locate and label objects. The output of the encoder is then passed to a decoder, which predicts bounding boxes and labels for every object. Mask R-CNN, also called Mask RCNN, may be a Convolutional Neural Network (CNN) that's cutting-edge in image and instance segmentation. Faster R-CNN, a Region-Based Convolutional Neural Network was wont to build Mask R- CNN.

II. RESEARCH METHODOLOGY

Object detection is one among the pc vision tasks, it helps to locate the objects and also labelling them. It draws a bounding box around the objects in a picture. Deep learning is that the most advanced preferable approach for detecting objects. Object detection is helpful in vehicle detection, pose estimation, and autonomous vehicles. In earlier days, objects are often detected by using Convolutional Neural Networks. CNN-based representation tries to capture a good range of discriminative appearance variables and also it'll show a sensitivity of localization which is significant for accurate object localization [12]. In CNN, objects are often detected with the assistance of bounding boxes but at the identical time, it can detect only one object. to recognize multiple objects in a picture and also for drawing bounding boxes round the image RCNN was used. RCNN aims to require the input image and it identifies main objects well with a help of bounding boxes. For object detection, RCNN uses the selective look for generating region proposals. Regions are often picked out by different colors, textures, scales, and enclosures. to beat the matter of huge number of region selection, RCNN was introduced with better results.

For detecting the little objects an Augmented R-CNN algorithm was proposed. In detecting the tiny objects, the proposed deep learning-based object detection algorithm outperforms the traditional approach [13]. In object detection, multiple object detection could be a difficult task. By comparing RCNN and Deformable part-based (DPM) models, Multiple objects were detected [23]. RCNN works with 2000 region proposals to classify enormous number of regions but it takes an unlimited amount of time for training the network. to overcome that, Fast RCNN was proposed, it just gives the input image to a Convolutional Neural Networks than feeding the region proposals to the CNN. instead of the SVM employed in R-CNN, softmax was utilized by Fast RCNN for classification tasks. In 2017, Zhong-Qiu Zhao et.al proposed Fast RCNN for Pedestrian Detection. A batch normalization layer has been added amongst the convolutional layer and therefore the activation function layer to reduce the training time and also to enhance generalization efficiency. With the assistance of the edge Boxes algorithm, it'll delete redundant windows with calibre and it helps to extend pedestrian detection speed and efficiency [17].

Fast RCNN with Secure Margin in RoI approach increases model performances by adopting the simplest bounding box for RoI. This method creates extended pooling layer with extended output layer for predicting multiple region proposals [5]. Faster RCNN allows the network to review the proposals. In real-world applications including automated driving, pedestrian detection draws lots of focus from isolated object detection. Region Proposal Network in Faster RCNN outperforms well for Pedestrian detection [18].Mask RCNN is an ultra-modern technique for instance segmentation, it works well and faster than the opposite models, for pixel-level segmentation, it absolutely was accepted. Kittinun Aukkapinyo et.al applied Mask RCNN in localization and classification for Rice grain Images. The proposed Mask RCNN will classify and also localize the rice grain automatically. Meanwhile, it performed skillfully when put next to other techniques [11].

2.1 Object Detection Using Mask RCNN

The motive of using Mask RCNN for the detection of objects is, it works well in automatic detection and it detects accurately. The proposed work includes the method of pre-processing, region proposals, ROI alignment, bounding box, fully connected layer, CNN classification, mask generation. VGG Net, ResNet are often used for several object detection tasks. ResNet-101 could be a CNN and it's 101 layers deep, during this paper ResNet-101 has been used as a backbone for Mask RCNN. For ResNet layer as shown in Figure. 1, consider initial mapping as M(x),

$$M(x) = F(x) + x$$

Thus, we would end up fitting with using $F(x) = M(x) - x$, the significance of this model would be bypassing the layers which could hinder the model performance.

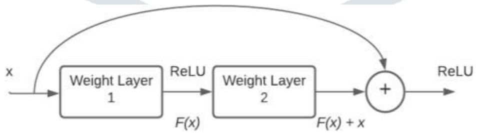


Figure. 1 ResNet Function

A pre-processing step can resize the photographs at a pixel level, and it changes videos or data to the understandable desired format, even some noisy information and missing particulars are cleansed in data cleaning. The features will be extracted with the assistance of region proposals. From region proposals, RoI align is employed to make fixed-size regions of interest. the little rectangle with vertical or horizontal sides that fully surrounds an object is termed a bounding box. The bounding box includes the whole object and these boxes are the lines that are drawn around the object found for visualization. it's one of the image annotations approaches. All the previous outputs possibly fed into a completely connected layer. By using the CNN, images may be classified. Masks are generated with the employment of instance segmentation. Masks are the shaded regions on the objects in a picture.

2.2 Mask RCNN:

Mask RCNN is an extension of Faster RCNN to pixel-level segmentation and it's a deep neural network. Faster R-CNN model is that the basis for the Mask R-CNN model. Faster RCNN returns a category label and bounding box for each object. a category label, bounding box, and additionally, a mask is returned still in Mask RCNN. Mask RCNN targets to resolve instance segmentation, the most intention of image instance segmentation is to acknowledge the objects present in a picture at a pixel level. Instance segmentation detects and segments each distinct object of interest existing in a picture. Mask RCNN accomplishes the instance segmentation tasks with a help of adding an object mask prediction branch to the present branch for bounding box identification. On top of the CNN features of Faster R-CNN, a totally Convolutional Network (FCN) was attached. The mask section will take the Region of Interest

(ROI) and can predict masks by using FCN. Mask RCNN predicts binary masks for each class. Earlier time objects were detected using YOLO, Faster RCNN then on but it doesn't give assurance about the form of the objects. Mask RCNN is extremely fast and it's easy to generalize tasks. Advantages of preferring mask RCNN over other models are Mask RCNN provides semantic segmentation also as bounding box and also it runs at the speed of 5 frames per second (fps). Mask RCNN may be applied in medical level segmentation, multi-organ segmentation, human pose estimation, drone object detection, and self-driving cars.

Loss function of Mask RCNN is calculated using,

$Loss\ (RPN) = RPN\_Class\ Loss + RPN\_BBox\ Loss$

Where, $Loss\ (RPN)$ is Loss of Region Proposal Network.

Loss (Mask RCNN) = Loss (class labels prediction) + Loss (Bounding box prediction) + Loss (Mask prediction)

Where, Loss (Mask RCNN) is Loss of Mask RCNN.

$Total\ Loss = Loss\ (RPN) + Loss\ (Mask\ RCNN)$

2.3 Overall Architecture Of The Proposed System

The mask RCNN for detecting the objects is pre-trained with COCO dataset, and for detecting objects.The proposed architecture diagram is shown in Figure 2.

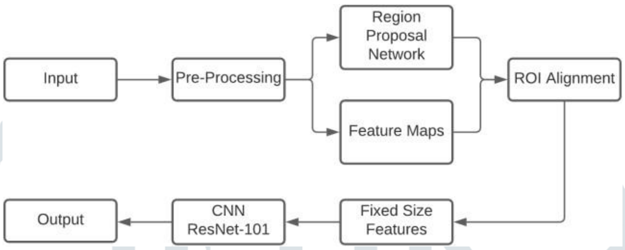


Figure 2. Architecture Diagram

2.3.1 Input & pre-processing

The input image which has n number of pixels is taken and it will pre-process the given input. A utility function is used to get information about the object from NumPy array about the bounding boxes details including shapes, minimum, and maximum values. The step is followed by Batch Normalization can affect during training process.

2.3.2 Region Proposal Network

RPN generates region proposals for the objects in the image. Region proposals draws multiple number of bounding boxes in the input image. It scans an image in a small window fashion and also discovers areas that has objects. Anchors are the regions which is scanned by RPN.

2.3.3 Feature Maps – FPN

To represent objects at multiple scales, a Feature Pyramid Network (FPN) was applied. The high-level features in the first pyramid are taken by second pyramid for upgrading the feature extraction pyramid, also passing those layers to lower layers.

2.3.4 ROI Alignment

For data pooling from RoI pooling, RoI Align removes hard quantization. The size of the proposed regions can be determined by using RoI Alignment or RoI Pooling.

3. METHODS:

3.1 Masking Algorithm

The proposed work Mask RCNN has undergone the following process has shown in Figure 3 below.

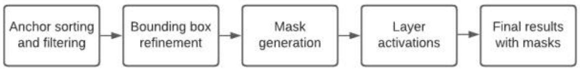


Figure. 3 Object Detection Steps

The Region Proposal Network (RPN) will show all the anchors with anchor box refinement. The RPN could be a fully convolutional network which will predict the item bounds and at every position, it predicts object scores. it's trained to come up with excellent region proposals. Once the anchor sorting and filtering process are done it then goes under the method of bounding dox refinement. In object detection, bounding boxes are the rectangle boxes drawn round the objects in a picture or a video. Bounding box refinement shows the dotted lines and within the second stage refinement approach is applied to the solid lines. Both the dotted lines and therefore the solid lines are illustrated within the positive regions. Masks are generated within the next step. Masks are the shaded regions at a pixel level. Masks are placed on the image within the correct location. Then it's moved to the layer activations. this is often very helpful to look at the activations of the various layers to determine whether it's all zeros and noises randomly.

3.2 Convolutional Neural Network

A Convolutional Neural Network (CNN) could be a kind of artificial neural network that's optimized to process pixel data and is employed in image recognition and processing. As a result, Convolutional Neural Networks are the fundamental and basic building blocks for the image segmentation computer vision task (CNN segmentation).

The Convolutional Neural specification is split into three layers.

1.Convolutional layer: Using filters and kernels, this layer helps to abstract the input image as a feature map.

2.Pooling layer: This layer aids within the down sampling of feature maps by summarising the presence of features in feature map patches.

3.Fully connected layer: a totally connected layer connects every neuron in one layer to each neuron within the next layer.

By combining the layers of a CNN, the neural network can find out how to spot and recognise the item of interest in a picture. Simple Convolutional Neural Networks are accustomed classify images and detect objects in images with one object.

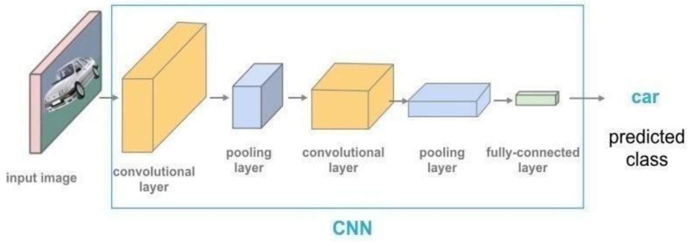


Figure 4. The CNN architecture

A simple CNN architecture isn't optimal in a more complex situation with multiple objects in an image. Mask R-CNN is a cutting-edge architecture based on R-CNN that is ideal for these situations (also referred to as RCNN).

3.3 Region-based Convolutional Neural Networks

Region-based Convolutional Neural Networks(R-CNNs)[1] are a sort of machine learning model employed in computer vision and image processing. the first goal of any R-CNN, which is specially designed for object detection, is to detect objects in any input image and define boundaries around them. The R- CNN (7,1)model uses a mechanism referred to as selective search to extract information about the region of interest from an input image. The rectangle boundaries can represent a section of interest. depending on the scenario, there can be over 2000 regions of interest. This area of interest is processed by CNN to generate output features. These output features are then fed into the SVM (support vector machine) classifier, which classifies the objects presented within a locality of interest.

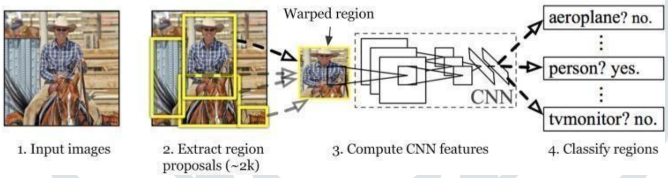


Figure 5. R-CNN

The image above depicts the procedures of an R-CNN while detecting an object with it. We use the region extraction algorithm to extract regions of interest within an image using the R-CNN. The total number of regions can be increased to 2000. The model manages the size to be fitted for the CNN foreach region of interest, where CNN computes the features of the region and SVM classifiers classify what objects are presented in the region.

3.4 Fast R-CNN

Instead of performing maximum pooling, we use ROI pooling in fast R-CNN to use a single feature map for all regions [2]. This warps ROIs into a single layer, and the ROI pooling layerconverts the features using maximum pooling. Because max pooling is also active here, we can consider fast R-CNN to be an upgrade to the SPPNet. Instead of generating layers in the shape of a pyramid, it only generates one layer.

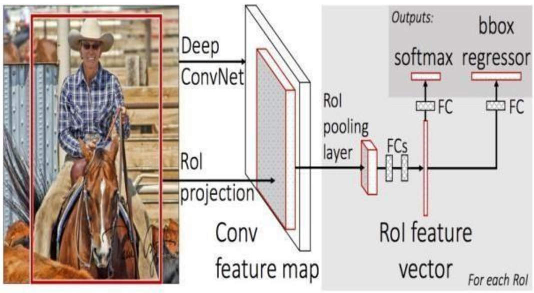


Figure 6. Faster R-CNN

3.5 Mask R-CNN

Mask R-CNN, also known as Mask RCNN, is a Convolutional Neural Network (CNN) that is cutting-edge in image and instance segmentation. Faster R-CNN, a Region-Based Convolutional Neural Network, was used to build Mask R- CNN[4].The first step in understanding how Mask R-CNN works is to grasp the concept of Image Segmentation. The task of computer vision the process of dividing a digital image into multiple segments is known as image segmentation (sets of pixels, also known as image objects). This segmentation is employed in order to locate objects and boundaries (lines, curves, etc.)

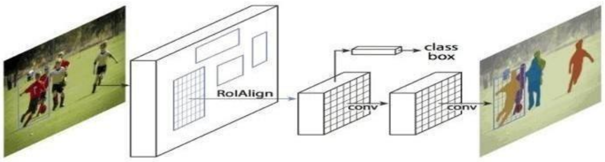
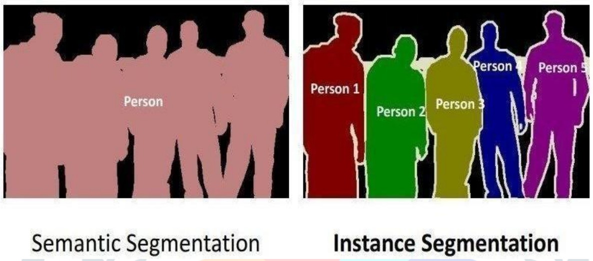


Figure 7: Mask R-CNN – The Mask R-CNN Framework for Instance Segmentation

Mask R-CNN encompasses two major types of image segmentation [6]:

3.5.1. Semantic Segmentation: Semantic segmentation assigns each pixel to one of a fixed set of categories without distinguishing between object instances. Semantic segmentation, in other words, is concerned with the identification/classification of similar objects as a single class at the pixel level. All objects were classified as a single entity, as shown in the image above (person). Background segmentation is another name for semantic segmentation, which separates the image's subjects from the background.



Semantic Segmentation **Instance Segmentation**

Figure 8: Semantic Segmentation and Segmentation of Instances

3.5.2 Segmentation of instances: Instance Segmentation,[7] also referred to as Instance Recognition, is worried with accurately detecting all objects in a picture while also precisely segmenting each instance. As a result, it combines object detection, object localization, and object classification. In other words, this kind of segmentation goes above and beyond to differentiate each object classified as an analogous instance. for example Segmentation, all objects are persons, as shown within the example image above, but this segmentation process separates every person as one entity. Semantic segmentation is additionally called foreground segmentation because it emphasises the image's subjects instead of the background. Mask R-CNN was created with Faster R-CNN. While Faster R-CNN produces two outputs for every candidate object, a category label and a bounding-box offset. Mask R-CNN adds a 3rd branch that produces the article mask. the extra mask output differs from the category and box outputs therein it requires a far finer spatial layout of an object to be extracted. Mask R-CNN may be a Faster R-CNN extension that works by adding a branch for predicting an object mask (Region of Interest) alongside the present branch for bounding box recognition.

IV.RESULTS AND DISCUSSION

The proposed model has been trained with different objects and shapes to detect objects for a given input image, however, the training has been constrained to the requirements of detection needs. The model performance is good for most of the detection tasks, and performance depends primely on model training. There are pre-trained weights available for the mask RCNN model and those weights can be used to detect objects, and these weights can be updated during further training of the model. Mask RCNN proposed in this paper uses ResNet-101 as the backbone to extract features from input images.

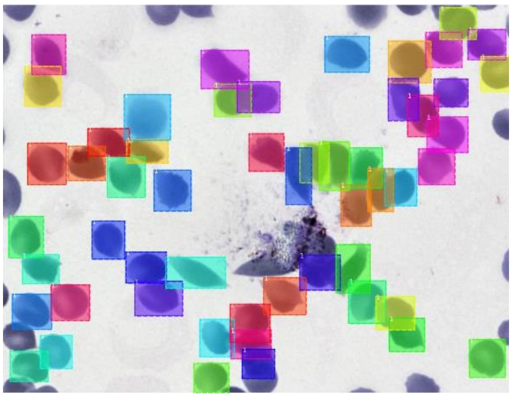


Figure 9. Applying Mask on Images

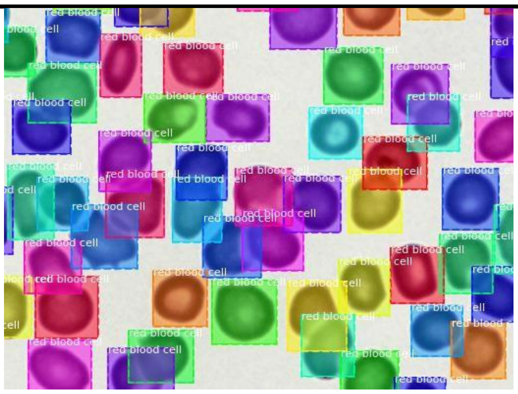


Figure 10. Object Detection of Masked Images

Thus, Mask RCNN may be used to detect and apply mask objects with more precision and accuracy. The model was again trained with a custom dataset to detect objects, and therefore the custom dataset was generated with shapes, in order that the model can identify shapes and label them. The model was trained and tested with validation set, and also the model obtained mAP of 94% for the custom-made dataset. These characteristics shows that the model will be trained for tiny objects which can be clustered during a single image and therefore the model may detect large and medium sized objects in a picture. The comparison of model performance other models has been shown Table.1. In Figure 10, the model was compared with existing models used for object detection, however the models were trained with different datasets. The work may be further extended by testing the model performance with the dataset employed by other existing models. Then the model may be further fine-tuned to extend the reliability of detections and to extend the robustness of the model.

Table 1. Various object detection models with mean average precision (map)

Object detection models	mAP
[12] CNN	78.2%
[21] RCNN	53.3%
[24] SGFr-RCNN	74.6055%
[26] AF-RCNN	56.4%
[19] PV-RCNN	83.90%
[28] Fast RCNN	70.0%
[27] Faster RCNN	78.8%
Proposed Model	94%

Mask RCNN solves instance segmentation problem at a pixel level which is a major significance of this model. To reduce human efforts, automatic object detection algorithm was developed, and it helps to detect objects automatically and precisely. This paper proposed Mask RCNN for object detection using COCO dataset, and the model was also tested with custom dataset and 94% of mAP was achieved. The efficiency of our approach is proved by the considerable experimental outcomes of taking the object detection. This paper also shown Mask RCNN using instance segmentation outperforms RCNN, Fast RCNN, Faster RCNN, and other models for object detection. In future the work can be further improved by increasing the accuracy and performance of the proposed model specific for medical research application.

ACKNOWLEDGMENT

This paper and therefore the research behind it'd not are possible without the exceptional support of my guide, Hirendra R. Hajare. His enthusiasm, knowledge and exacting attention to detail are an idea and kept my work heading in the right direction from my first encounter with data analysis to the ultimate draft of this paper. I'm also grateful for the insightful comments offered by the anonymous peer reviewers at Books & Texts. The generosity and expertise of 1 and everyone have improved this study in innumerable ways and saved me from many errors; people who inevitably remain are entirely my very own responsibility.

REFERENCES

[1] Ammirato, Phil, and Alexander C. Berg. "A mask-rcnn baseline for probabilistic object detection." *arXiv preprint arXiv:1908.03621* (2019).

[2] Songhui, Ma, Shi Mingming, and Hu Chufeng. "Objects detection and location based on mask RCNN and stereo vision." *2019 14th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*.IEEE, 2019.

[3] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[4] Azam, Shoaib, AasimRafique, and MoonguJeon. "Vehicle pose detection using region based convolutional neural network." *2016 International Conference on Control, Automation and Information Sciences (ICCAIS)*.IEEE, 2016.

[5] Saeidi, Mahmoud, and Ali Ahmadi. "Pedestrian detection using an extended fast RCNN based on a secure margin in RoI feature maps." 2018 9th International Symposium on Telecommunications (IST). IEEE, 2018..

[6] Fan, Quanfu, Lisa Brown, and John Smith. "A closer look at Faster R- CNN for vehicle detection." *2016 IEEE intelligent vehicles symposium (IV)*.IEEE, 2016.

[7] Voulodimos, Athanasios, et al. "Deep learning for computer vision: A brief review." *Computational intelligence and neuroscience* 2018 (2018).

[8] Nath, Siddhartha Sankar, et al. "A survey of image classification methodsand techniques." 2014 International conference on control, instrumentation, communication and computational technologies (ICCICCT).IEEE, 2014.

[9] Long, Yang, et al. "Accurate object localization in remote sensing imagesbased on convolutional neural networks." *IEEE Transactions on Geoscience and Remote Sensing* 55.5 (2017): 2486-2498.

[10] Zou, Zhengxia, et al. "Object detection in 20 years: A survey." *arXiv preprint arXiv:1905.05055* (2019).

[11] Aukkapinyo, Kittinun, et al. "Localization and classification of rice-grain images using region proposals-based convolutional neural network." *International Journal of Automation and Computing* (2019): 1-14.

[12] Gidaris, Spyros, and Nikos Komodakis. "Object detection via a multi- region and semantic segmentation-aware cnn model." *Proceedings of theIEEE international conference on computer vision*. 2015.

[13] Chen, Chenyi, et al. "R-CNN for small object detection." *Asian conference on computer vision*. Springer, Cham, 2016.

[14] Wang, Tan, et al. "Visual commonsense r-cnn." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[15] Zhang, Jianming, et al. "A cascaded R-CNN with multiscale attention andimbalanced samples for traffic sign detection." *IEEE Access* 8 (2020): 29742-29754.

[16] Braun, Markus, et al. "Pose-rcnn: Joint object detection and pose estimation using 3d object proposals." *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016.

[17] Zhao, Zhong-Qiu, et al. "Pedestrian detection based on fast R-CNN and batch normalization." *International Conference on Intelligent Computing*. Springer, Cham, 2017.

[18] Zhang, Liliang, et al. "Is faster R-CNN doing well for pedestrian detection?." *European conference on computer vision*. Springer, Cham, 2016.

[19] Shi, Shaoshuai, et al. "Pv-rcnn: Point-voxel feature set abstraction for 3dobject detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[20] Xu, Hang, et al. "Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[21] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[22] Lu, Xin, et al. "Grid r-cnn." *Proceedings of the IEEE/CVF Conference onComputer Vision and Pattern Recognition*. 2019.

[23] Li, Junliang, et al. "Multiple object detection by a deformable part-basedmodel and an R-CNN." IEEE Signal Processing Letters 25.2 (2018): 288-292.