ICSEE 2021 JANUARY 2021 | International conference on systems energy and environment
22nd & 23rd January 2021 | GCE Kannur, Kerala, India

# A Review on Instance Segmentation Using Mask-RCNN

Sreya Ramesh C and V Vinod Kumar
Department of Electronics and Communication Engineering
Government College of Engineering Kannur

***Abstract*—Instance segmentation consolidates object detection, where the objective is to classify and localize each objects using a bounding box, and semantic segmentation, where the objective is to characterize every pixel into the given object classes. Mask R-CNN is a deep learning architecture used for instance segmentation. It is an augmentation of the well known Faster R-CNN object detection architecture. Mask R-CNN adds an additional mask branch to the existing Faster R-CNN model. The Faster R-CNN produces two things for each object in the picture. Its class label and the bounding box co-ordinates. Mask R-CNN adds an extra branch to this which yields the object mask too. Mask prediction is done in corresponding with bounding box creation and grouping. This paper contains the idea of how Mask R-CNN performs instance segmentation by using examples of vehicle damage detection and segmentation, Detection and segmentation of oral diseases and segmentation of news paper elements.**

***Index Terms*—Mask R-CNN, Faster R-CNN, Instance segmentation**

## I. Introduction

Instance segmentation is a combination of two sub problems namely object detection and semantic segmentation. Object detection means, to find and classify various objects in an image. In semantic segmentation an object class is assigned for each pixels. Instance segmentation not only gives the location and class of a particular object in an image but also gives the size details of the object by creating an object mask. For example when monitoring a cancer cell, finding the rate of growth of the cell is very important. Here Object detection alone cannot meet the requirement. Object detection only point out the presence of the cancer cell. To get the complete picture of cell growth instance segmentation plays a vital role.

In the area of computer vision, instance segmentation has prominent significance. It has dynamic importance in the field of traffic control system, medical imaging, satellite imaging and robotics. As far as the autonomous vehicles are concerned, detection and segmentation of nearby vehicles can be done through instance segmentation. It has captured world wide attention in medical imaging for the detection and segmentation of brain tumor, cancer cell etc. The images from the satellite could be converted into maps spotting household buildings, commercial buildings and so on for variety of purposes. In surgery robots, instance segmentation has relevant role to play.

One of the most general way to carryout instance segmentation is through Mask R-CNN[5] algorithm. It is a deep learning algorithm which is capable of both object detection and semantic segmentation. R-CNN family has so many object detection algorithms such as R-CNN[1], Fast R-CNN[2], Faster R-CNN[3] etc. Mask R-CNN is the for most one which is capable of both segmentation and object detection. In this paper section II contains architectural details of Mask R-CNN. The section III deals with vehicle damage detection and segmentation using Mask-RCNN[6]. Detection and segmentation of oral diseases[7] are mentioned in section IV. Section V of this paper explains segmentation of news paper elements[8].

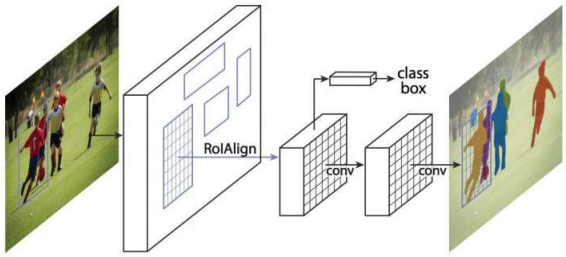


Fig. 1. Mask R-CNN Framework[5]

## II. MASK R-CNN

Mask R-CNN is an extension of Faster R-CNN in which the later is used to detect and classify various objects in an image. Class label and bounding box co-ordinates are assigned to various objects in an image during object detection. Besides these, Mask R-CNN provides object mask for each object through semantic segmentation. Since there are two phases, the model has two parts. An architecture similar to Faster R-CNN is used for object detection, while a Fully Convolutional Network (FCN)[9] is used to carry out semantic segmentation. Figure 1 shows the Mask-RCNN framework and the detailed workflow is shown in Figure 2.

### A. Backbone Model

ResNet network [4] is used to extract features from image in Mask R-CNN similar to the ConvNet that is used in Faster R-CNN. The most important function of this segment is to

Sreya Ramesh C is associated with the Department of Electronics and Communication, Government College of Engineering Kannur, Kerala, India. (e-mail:sreyaramesh.c16@gmail.com)
V Vinod Kumar is a faculty of Department of Electronics and Communication, Government College of Engineering Kannur, Kerala, India. (e-mail:vinod.gcek@gmail.com)

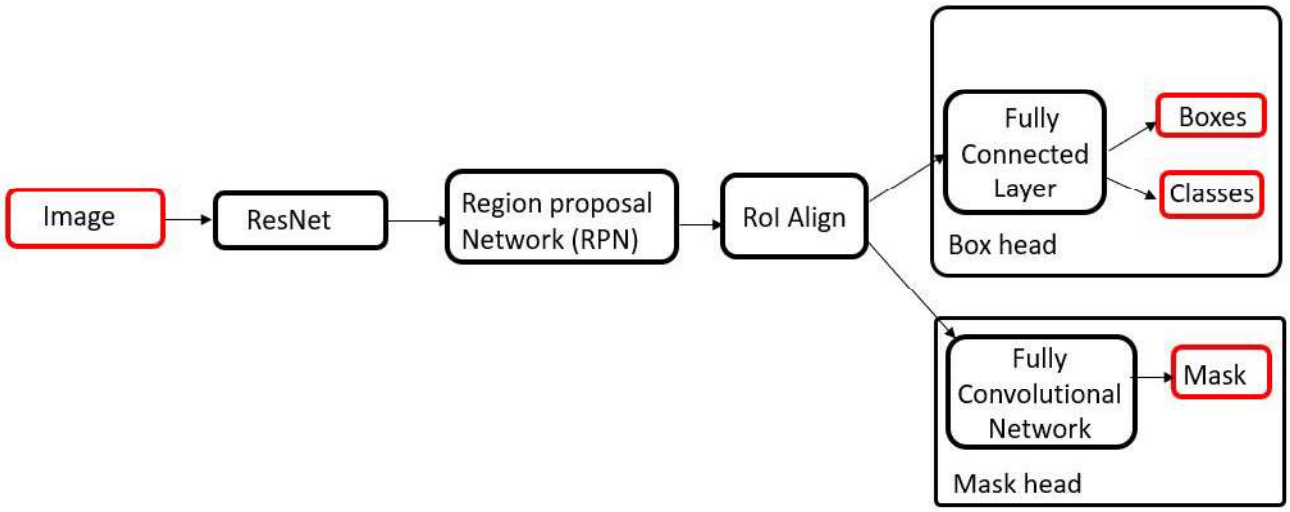


Fig. 2. Mask R-CNN Workflow

extract low level as well as high level features from the given image.

### B. Region Proposal Network (RPN)

A Region Proposal Network is applied to the feature map which is obtained from the previous step. If an object is present in a particular region or not is basically predicted through this network. The Intersection over Union (IoU) with the ground truth box is computed for all predicted regions. The following equation shows how the Intersection over Union can be calculated.

$$IoU = \frac{\text{Area of the intersection}}{\text{Area of the union}} \quad (1)$$

When IoU is greater than or equal to 0.5 it is considered as region of interest. Otherwise it will be neglected.

### C. RoI Align

RoI Align is used to make fixed size regions through max pooling. Usually Faster R-CNN uses RoI pooling for making region of interest in to fixed size. There is loss of data due the quantization of stride value. For example if the stride value is 2.42, it will be rounded off to 2. But in the case of RoI align no quantization of stride value occurs. It is achieved through bilinear interpolation [11].

### D. Box head

Box head uses the fixed size region of interest to predict class label and bounding box for object in that particular region. For that it uses a fully connected layer that consist of softmax layer and linear regression layer[3]. Softmax layer helps to give class label for distinct objects while regression layer is capable of tightening the bounding box.

### E. Mask head

After getting the fixed size region interest , Mask prediction branch can be added to the existing architecture. It consists of a Fully Convolutional Network which is capable of semantic segmentation. A pixel level comparison is carried out to find the layout of each object that is present in an image. After semantic segmentation each region of interest is given with a binary mask.

## III. Vehicle Damage Detection and Segmentation

The detection and segmentation of vehicle damage plays an important role in transportation field. An improved Mask-RCNN model is used to meet this function. In order to make the dataset, the pictures of the damaged part of the vehicle is collected. This dataset is divided into two parts namely testing dataset and training dataset.

Usually in Mask-RCNN, ResNet-101 is used as the back-bone network. There are 101 layers in ResNet-101. But too many layers will certainly reduce the processing speed of the entire network. Considering this problem ,ResNet-50 with Feature Pyramid Netwotk (FPN)[10] is applied instead of Resnet-101. In order to save the computation time, Region Proposal Network set IoU value as 0.8. The improved Mask-RCNN model for vehicle damage detection is shown in Figure 3.

Firstly the input image is given to the ResNet50 + FPN backbone model. Then the feature map obtained from model is

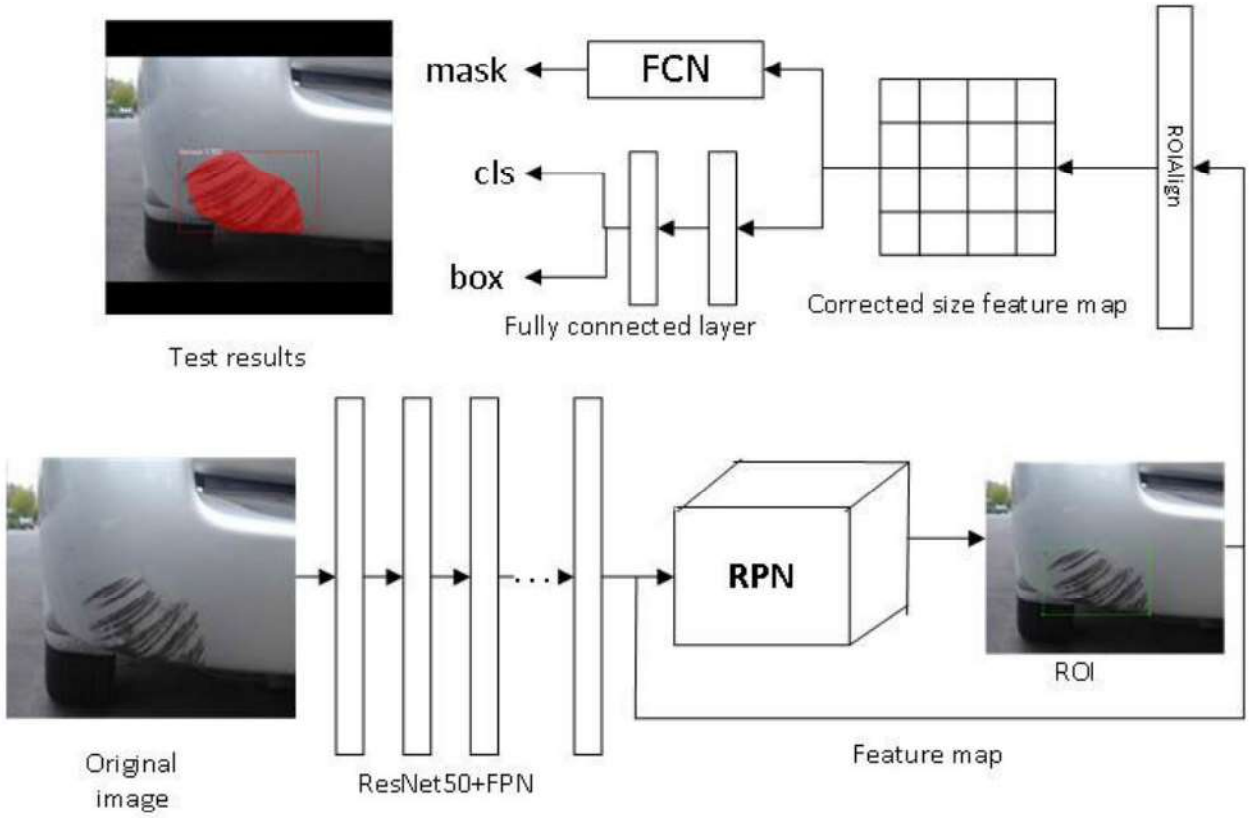


Fig. 3. Mask R-CNN model for car damage detection and segmentation[6]

passed through the Region Proposal Network. It gives Region of Interest (ROI) as output. After that RoI Align is applied to get fixed size proposals. Eventually the workflow divert into two sections. Out of the two, first section involves object classification and regression through a Fully Connected layer and the second section involves segmentation through a Fully Convolutional Network.

## IV. Detection and Segmentation of Oral Diseases

Image segmentation using deep learning has major impact in analysing medical images. Mask R-CNN can be applied in detecting oral diseases. It is used to segment cold sores as well as canker sores.

Here the backbone network combines ResNet-101 + FPN and ResNet-50 + FPN in order to get low level and high level features. Figure 4 shows the workflow of Mask-RCNN in detecting and segmenting oral diseases.

The Convolutional Neural Network (CNN) extract features from input image and produces feature maps. Region Proposal Network (RPN) produce region proposals. RoI Align convert all the proposals into same size. Finally class prediction and mask prediction are done independently. Figure 5 shows the overall output after detection and segmentation.

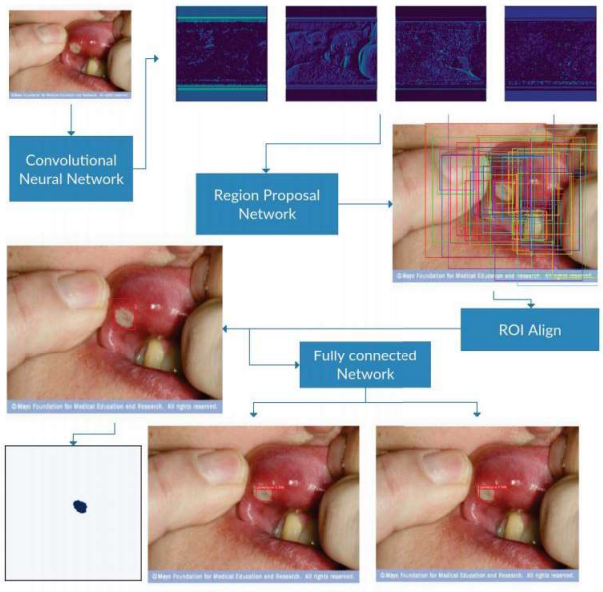


Fig. 4. Mask R-CNN Workflow of oral disease detection and segmentation[7]

Fig. 5. Final output of oral disease detection and segmentation[7]

## V. SEGMENTATION OF NEWSPAPER ELEMENTS

Segmentation of news paper contents into articles and other elements has acquired importance in the modern world. Mask R-CNN can help in segmenting and classifying various contents in a news paper.

To get the accurate model for the segmentation, huge number of data is needed for training. By using technique like image augmentation method and transfer learning method[12] the problem can be reduced to a great extent.

The model takes news paper images as input and provide an output with separate mask and label for each content. Figure 6 shows the final output after the segmentation of news paper elements.

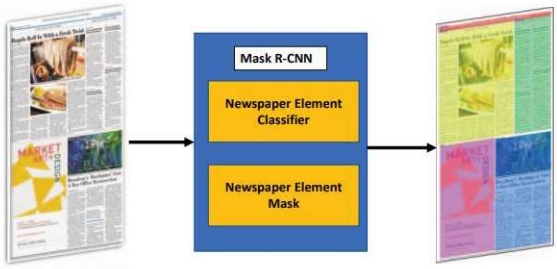


Fig. 6. Mask R-CNN model for classification and segmentation of news paper elements[8]


Fig. 7. Output of classification and segmentation of news paper elements[8]

## VI. CONCLUSION

This paper tries to demonstrate the working of Mask R-CNN and how it is capable of performing instance segmentation. Instance segmentation has wide applications in the area of agriculture, robotics, traffic control system and satellite imaging. Three different situations through which instance segmentation using Mask R-CNN is performed are mentioned in this paper. The small variations occurred according to the situation has also been mentioned.

## REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, vol. 13, no. 1, pp. 580–587.
[2] R. Girshick, "Fast R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015. pp. 1440–1448.
[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[5] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask RCNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.
[6] Q. Zhang, X. Chang and S. B. Bian, "Vehicle-Damage-Detection Segmentation Algorithm Based on Improved Mask RCNN," in IEEE Access, vol. 8, pp. 6997-7004, 2020, doi: 10.1109/ACCESS.2020.2964055.
[7] R. Anantharaman, M. Velazquez and Y. Lee, "Utilizing Mask R-CNN for Detection and Segmentation of Oral Diseases," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 2197-2204, doi: 10.1109/BIBM.2018.8621112.
[8] J. A. Almutairi and M. Almashan, "Instance Segmentation of Newspaper Elements Using Mask R-CNN," 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 2019, pp. 1371-1375, doi: 10.1109/ICMLA.2019.00223.
[9] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.
[10] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and ´ S. Belongie. Feature pyramid networks for object detection. In CVPR, 2017.
[11] S. Wang and K. Yang, "An image scaling algorithm based on bilinear interpolation with VC++," in Proc. Techn. Autom. Appl., 2008, pp. 168–176.
[12] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" CoRR, vol. abs/1411.1792, 2014. [Online]. Available: http://arxiv.org/abs/1411.1792
[13] N Dinesh Reddy, Minh Vo, Srinivasa G.Narasimhan,"Occlusion-Net: 2D/3D Keypoint Localization using Graph Networks",IEEE Conference on CVPR,2019.
[14] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z.Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In CVPR, 2017.
[15] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. IJCV, 2013.
[16] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Doll´ar. Learning to refine object segments. In ECCV, 2016.
[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV, 2014.
[18] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In CVPR, 2017.
[19] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In CVPR, 2014
[20] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In ECCV. 2014.