



# Vertebra Segmentation for Clinical CT Images Using Mask R-CNN

Renjie Wang<sup>(✉)</sup>, Jennifer Hui Yi Voon, Da Ma, Setareh Dabiri,  
KarteeK Popuri, and Mirza Faisal Beg

School of Engineering Science, Simon Fraser University, 8888 University Drive,  
Burnaby, Canada

{da\_ma, faisal-lab}@sfu.ca

**Abstract.** Spine disease is a growing problem in modern society and has been debilitating for every age-group. Researches have shown that more than 266 million people are facing degenerative spine disease and low back pain. CT scanning is a fast, painless, non-invasive diagnostic imaging modality that provides high spatial accuracy in obtaining the 3D structure of the vertebral. However, in real-life scenario, the clinic CT image might not cover the whole spine and the field of view might be hard to determine. Henceforth, this project aims to create and validate an automatic method that can detect, locate, and classify each vertebra from the partial field of view using deep learning. We used Mask R-CNN, a deep neural network aimed to solve the instance segmentation problem in machine learning or computer vision, and produce features such as bounding boxes, classes, and masks to identify each vertebra. This auto-detection method was validated on an open source dataset which has been used on Computational Spine Imaging (CSI 2014). The dataset was chosen by a radiologist with an eight years involvement with thoracic and lumbar spine column scans, and the data of twenty patients were collected using standard CT scanning protocol. The accuracy of the vertebra mask on 210 test images has been increased up to 99.9% DICE Coefficient in Mask R-CNN compare with 69.2% Dice Coefficient in another Deep-learning-based semantic segmentation framework U-Net.

**Keywords:** Spine disease · Vertebra segmentation · Partial clinic CT image · Mask R-CNN

## 1 Introduction

Spine diseases have been debilitating for every age-group. The vertebral spinal column is an important support structure in the human body. The human spinal column consists of 33 vertebrae [1] (7 cervical vertebrae, 12 thoracic vertebrae, 5 lumbar vertebrae, 5 fused sacral vertebrae, and 4 fused coccygeal vertebrae) connected by ligaments, joints and intervertebral discs. The lumbar spine bears a large load at the lower levels and forms the junction of the active and fixed segments which are observed to be the most common site of low back pain [2]. Most of the symptoms of vertebral diseases are neck and shoulder pain, headache, vertigo, and lumbosacral pain. Some complications can lead to lower limb pain where individuals are not able to stand upright. Serious cases

may lead to paralysis. Computed Tomography (CT) is a non-invasive, fast, and accurate 3D imaging modality that has a good spatial resolution to produce images with excellent image contrast between bones and tissues. Therefore, CT is capable to provide a faster and comprehensive display of spinal anatomy and has higher sensitivity in the detection of the bone disorder compared to other imaging modalities. However, clinic CT images might only cover a partial spine field of view center around the site of the pain, which poses difficulty for segmentation algorithms in automatically identifying the varying levels of vertebrae present in the given field of view.

The main obstacles in developing these automatic segmentation methods are the similarities in shapes of the different vertebra and the capability of the system to process images from different imaging scanners. Furthermore, the alignment of the images with different field of views is also one of the main challenges in developing a clinically applicable tool. However, with the continuous development in the field of deep learning, many of these challenges can potentially be overcome. Henceforth, this project aims to create and validate an automated computer-aided diagnosis system that can detect, locate, and classify the thoracic and lumbar vertebral bodies on CT images using Mask R-CNN.

## 2 Background

### 2.1 Regions with Convolutional Neural Network (R-CNN)

In Conference on Computer Vision and Pattern Recognition (CVPR) 2014, the third year of deep learning in full swing, R-CNN was proposed by Ross *et al.* using a convolutional neural network to detect targets. The first step is through a method proposed in 2012, called selective search which extracts 2000 regions from an image. Simply speaking, the image is divided into several blocks by traditional image processing methods, and then several blocks belonging to the same target are taken out by an SVM which is the core of selective search. In the second step of feature extraction, Girshick *et al.* directly relied on the latest achievement of deep learning at that time, Alexnet (2012) which trains a network only for feature extraction by using image classification dataset. In the third step, a support vector machine (SVM) is used to combine the target's label (category) and the size of the bounding box. Therefore, the SVM is also trained separately [3].

When R-CNN came out in 2014, it overturned the previous target detection scheme and greatly improved the accuracy. However, the problem of R-CNN is also obvious. The time-consuming selective search usually takes 2 s for a frame of an image. The time-consuming serial CNN forward propagation needs to go through an Alexnet feature extraction for each ROI, which costs about 47 s for all ROI features. The three modules are trained separately, and when training, they consume a lot of storage space [4].

### 2.2 Fast R-CNN

In the face of this situation, Ross *et al.* proposed Fast R-CNN in 2015 to solve the problem of R-CNN [5]. First, the selective search method is still used to extract 2000 candidate

boxes, and then another neural network was used to extract the features of the whole image. Then, an ROI pooling layer is used to extract the corresponding features of each ROI from the whole graph features. The classification and bounding box are corrected by a Fully Connected (FC) layer. Therefore, instead of the serial feature extraction method of R-CNN, a neural network is used to directly extract features from the whole image (which is why ROI pooling is needed). Most parts of the R-CNN can be trained together except the selective search which is still a time-consuming process [4].

### 2.3 Faster R-CNN

Faster R-CNN [6] was later proposed to solve the time-consuming selective search needs completely and speed up the whole process. In faster R-CNN Instead of selective search, the region to be detected is directly generated through a region proposal network (RPN). With this approach, when generating an ROI region, the time is reduced 200 times. Firstly, the shared convolution layer is used to extract features for the whole image, and then the resulting feature maps are sent to RPN. RPN generates the frame to be detected (specifies the position of ROI) and corrects the bounding box of ROI for the first time. The following steps are the same as Fast R-CNN. According to the output of RPN, the ROI pooling layer selects the features corresponding to each ROI on the feature map and sets the dimension as a fixed value. Finally, the FC layer is used to classify the frames, and the target bounding box is modified for the second time. These improvements make the Faster R-CNN become an end-to-end training process.

Nevertheless, ROI pooling is the result of rounding directly. The value taken directly with round is that the output after ROI pooling may not match the ROI on the original image. The rounding operation of the ROI pooling layer causes the offset of the bounding box. Also, quantization has little effect on ROI classification but is harmful to pixel by pixel prediction which causes that the features obtained by each ROI are not aligned with ROI [6].

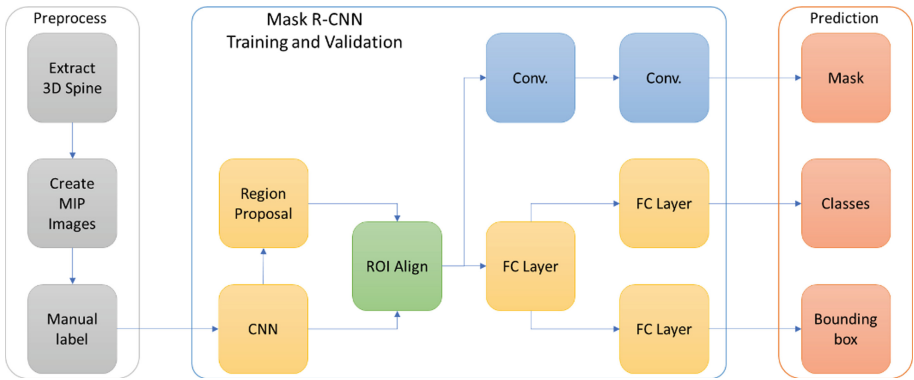
### 2.4 Mask R-CNN

Mask R-CNN directly inherits the 2016 Faster R-CNN, and the main innovation of Mask R-CNN is ROI align instead of ROI pooling in Faster R-CNN. Instead of rounding, ROI align uses bilinear interpolation to find the corresponding features of each bounding box which makes the features obtained for each ROI better align the ROI region on the original image. The output dimension of ROI align can be more accurate in predicting masks. In the training phase of the mask branch,  $K$  masks prediction graphs (one for each class) are output, and average binary cross-entropy loss training is used instead of SoftMax loss. The loss function of a multi-task loss on each sampled ROI is defined as [7]:

$$L = L_{cls} + L_{box} + L_{mask} \quad (1)$$

### 3 Method

The proposed pipeline consists of three main steps as shown in Fig. 1. First, in the preprocessing step, the vertebral column in the provided 3D data was extracted to have a clear view of the vertebral column and converted into a maximum intensity projection (MIP) image. The vertebrae on the MIP image is manually identified and labelled. Second, in the training and validation step, Mask R-CNN is used to train with the preprocessed dataset to identify and locate the vertebra of the MIP image. Finally, in the prediction step, based on the result of training, the Mask R-CNN predicts each vertebra in the test image with a bounding box, a class label, a mask, and a score of intersection-over-union (IOU).



**Fig. 1.** The flowchart of the vertebra segmentation

#### 3.1 Preprocessing

This step aims to preprocess and generate ground-truth data for training the neural network. All the images in the provided dataset on SpineWeb [8] were resampled to an isotropic resolution of  $1\text{ mm} \times 1\text{ mm} \times 1\text{ mm}$  using linear interpolation. Then, intensity outside the bone intensity range of 100HU (Hounsfield unit) and 1500HU is set to 0 to reduce the noise, imaging artifacts, and the influence from the tissues around the vertebral column. A spinal canal that allows the spinal cord to pass through the vertebra body can be detected by circle detection on every axial slice of the 3D data showing in Fig. 2. Circle detection was set to detect circles with a radius within a range according to the anatomical knowledge. The circle in the spinal canal with a radius  $R_{min}$  keeps expanding while moving away from the bones until the circle hits the bone and cannot expand further or reaches  $R_{max}$ .

The moving and expanding process is iterated on every axial slice of the image, and the location of the center of the circles on every slice is recorded and k-mean clustering is applied to divide the detected circles into 3 clusters as shown in Fig. 3. The 3 main areas that contribute to the detected circles are the vertebral column, and the sides of the

hip bone locating on both sides of the vertebral column. Therefore, a cuboid is extracted around the middle cluster by cropping the hipbone. This step prevents the hip bone to block the L5 in the MIP image. Then, the extracted 3D data is converted into a 2D sagittal MIP image by projecting a line from the sagittal view of the 3D data and retain the maximum intensity over all the voxels along that line.



Fig. 2. Spinal canal detection

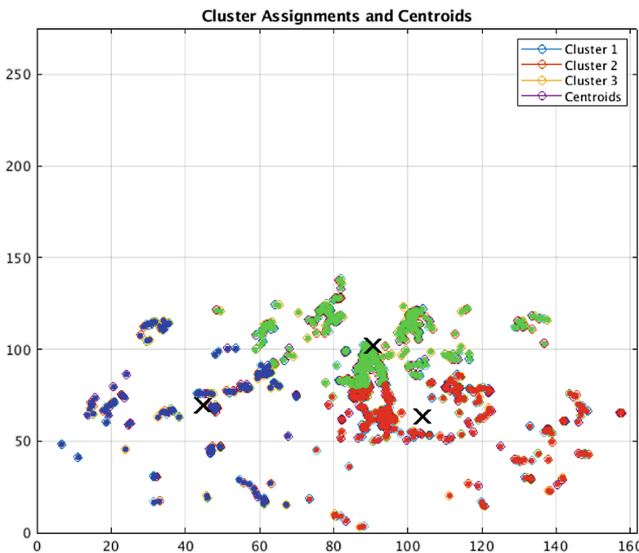
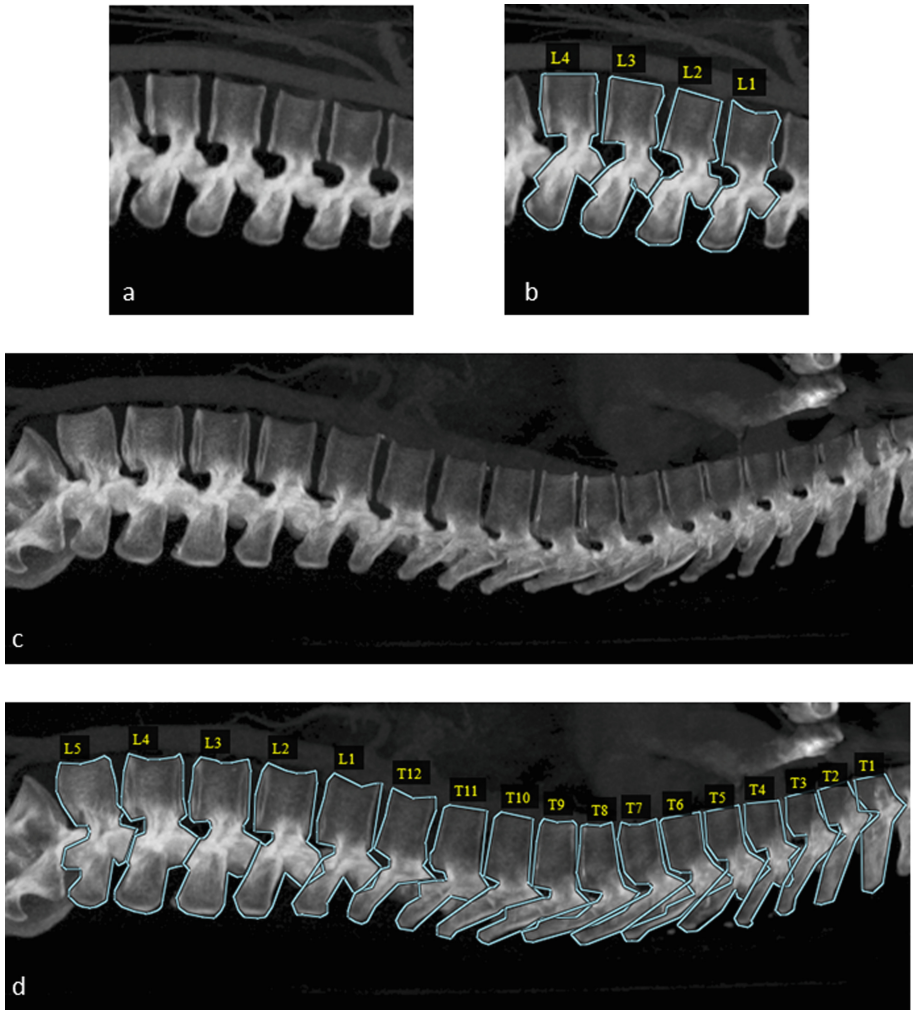


Fig. 3. Cluster Assignments and Centroids

To generate more training data, data augmentation was introduced. The augmented MIP images was generated from the full vertebral column by cropping the raw images into patches of images with small field of view (FOV) containing 4 to 17 vertebrae. The MIP image was then manually labelled on the VGG Image Annotation website [9] to produce an annotation file identifying each vertebra as shown in Fig. 4.



**Fig. 4.** a: Sagittal partial MIP image; b: Manual labelling of the sagittal partial MIP image c: Sagittal full MIP image d: Manual labelling of the sagittal full MIP image

### 3.2 Mask R-CNN-Based Vertebra Segmentation

Vertebrae was localized and classified by a trained Mask R-CNN model. Mask R-CNN generates bounding boxes and segmentation masks for each instance of every object detected in the image using the Feature Pyramid Network and a ResNet 101 backbone to extract features of the image [7]. The training of this model has been implemented on Python, Keras 2.0.8 and TensorFlow 1.15 and on the Compute Canada GPU to reduce the training time significantly.

In this study, transfer learning was used by initializing the network with a pre-trained weight trained on the MS COCO dataset [10]. The input image size of the

R-CNN network was set to be  $1024 \times 1024$ , with 18 classes including vertebra from Thoracic 1 to Lumbar 5 and background. 1260 training and 210 testing images were generated from 12 and 2 full vertebral columns, respectively. The model was trained with a learning rate of 0.001 and a batch size of 5 and the detection of the minimum confidence score is set at 0.9.

The two main parts of Mask R-CNN are: a) the RPN which generates the bounding box of the detected vertebra and b) the binary mask classifier which generates a mask on each vertebra. The image passing through the CNN generates a feature map. Then, the RPN makes use of CNN to identify the ROI using a lightweight binary classifier that displays positive and negative anchors. The ROI align network outputs multiple bounding boxes and wrap them into a unified dimension. The connected layers then perform image classification using the SoftMax and boundary box detection using the regression model. Finally, the mask classifier allows the network to generate the mask for every class without competition among classes.

## 4 Results

If the trained model detects a vertebra, the results provide 4 valid information including a bounding box, a class label, a score, and a mask for the vertebra in the bounding box as showing in Fig. 5. The bounding box has accurately identified 13 vertebrae with 13 bounding boxes. The class label demonstrates 13 vertebrae from L5 to T5, and no vertebra has been left out. The detected vertebra has presented with a different colour mask to differentiate each vertebra in pixel-level. Only 1 vertebra in thoracic miss a small portion of the mask, but the IOU score of every vertebra with IOU are all over 99% (Fig. 5), demonstrating promising level of accurate. In general, the mask of lumbar is more accurate than the mask of thoracic, and the results might miss-predict or over-predict due to vertebra similarity. Since all test images are different, the results of remaining samples might not show as same as Fig. 5. Besides, even though two or more vertebrae overlap with each other, the algorithm could still detect each vertebra with a different colour because the binary classification with the mask is only processed

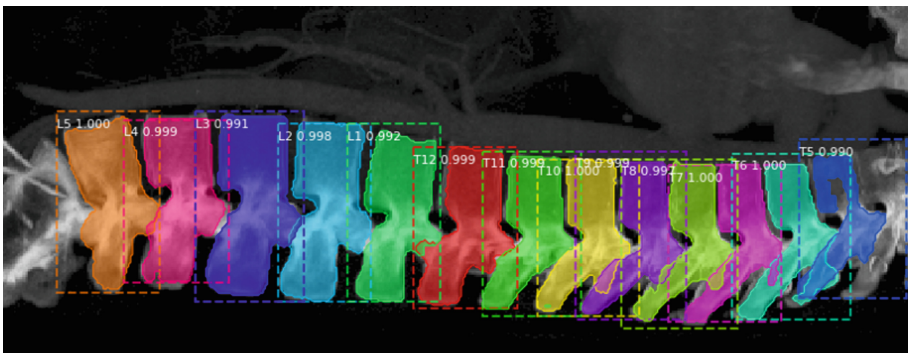
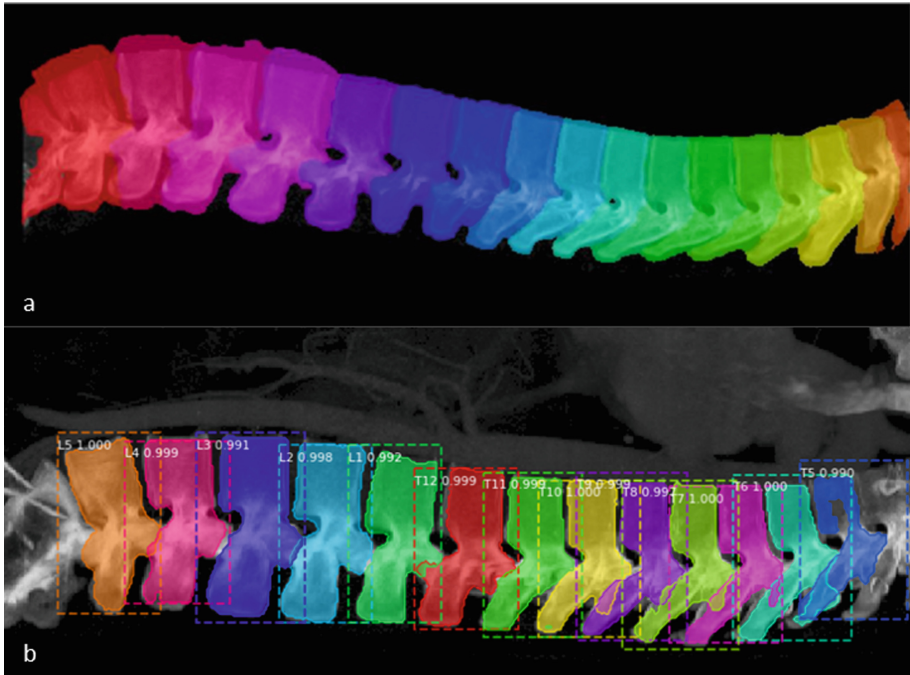


Fig. 5. Sample result of Mask R-CNN segmentation.

in the bounding box. The mask and score could potentially be improved by training more images and patients since the dataset only includes less than 20 full spines of patients.

The method of Semantic segmentation which predicts the object in pixel-level has been attempted before Mask R-CNN as shown in Fig. 6 (a). Since there is no bounding box restricting the boundary for the vertebra, the prediction has an extra layer outside of the edge of each vertebra. Compare both methods, Mask R-CNN presents better results in pixel-level for 210 test images. The following figure shows the DICE coefficient for each vertebra applying both Semantic segmentations using U-Net and Mask R-CNN. Most of the vertebra have a higher DICE coefficient in Mask R-CNN than Semantic segmentation which proved that the combination of object detection and instance segmentation is more accurate than instance segmentation alone. Also, vertebra has either over 0.9 or 1.0 DICE coefficient which means that the confidence level of Mask R-CNN is very accurate while training (Fig. 7).



**Fig. 6.** a: Result of Semantic segmentation b: Result of Mask R-CNN



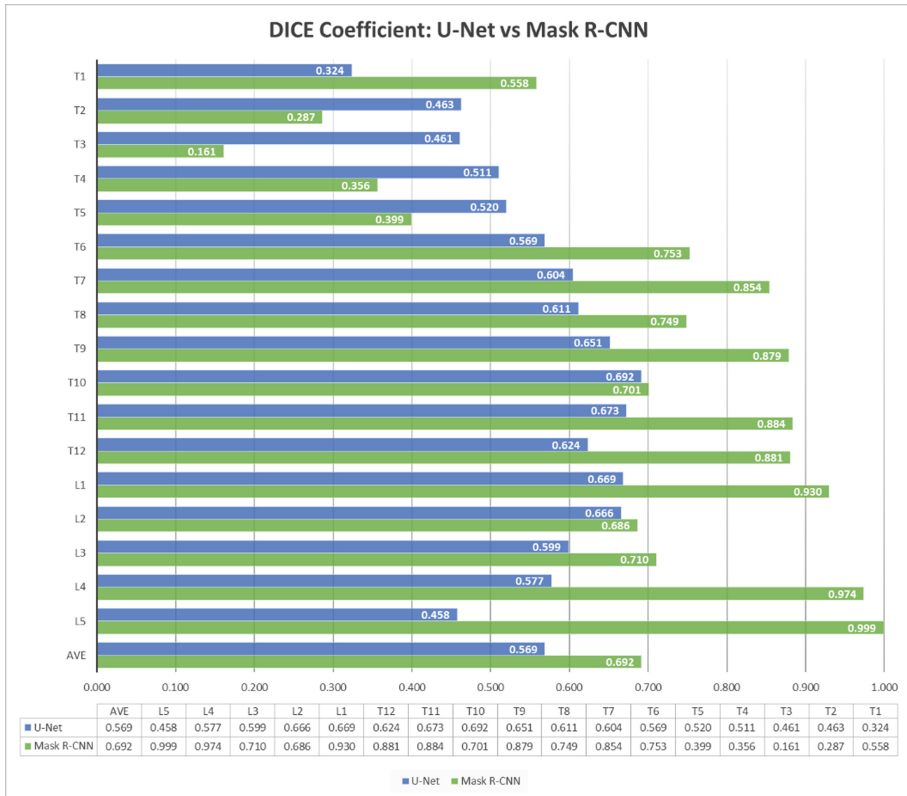


Fig. 7. DICE Coefficient Comparison

## 5 Conclusion and Discussion

The results of this study demonstrated successful localization and classification of each vertebra with high accuracy proved by the validation results of bounding box and the class label. Overlapping among two or more vertebrae are detectable, and test accuracy could be potentially increased by training more datasets. Also, the combination of object detection and instance segmentation performs better compared with instance segmentation alone because the binary classification for instance segmentation is performed in the bounding box.

More dataset could increase the anatomical variability of each vertebra which improve mask accuracy in pixel-level. Some remaining tissue could increase the difficulty to distinguish each vertebra in a noisy black and white image. However, after attempting to remove most of the tissues in the MIP image, the predicted accuracy is lower than the accuracy with noisy tissue. We hypothesize that contextual information from surrounding tissues help the network to distinguish the difference between each vertebra. With additional training data, the accuracy of spine extraction could be further

improved, and the final image would help clinical practitioners to visualize the details of each vertebra.

**Acknowledgment.** This work was supported by National Science Engineering Research Council (NSERC), Canadian Institutes of Health Research (CIHR), and the Michael Smith Foundation for Health Research (MSFHR).

## References

1. Drake, R.L., Vogl, W., Tibbitts, A.W.M., Richard, I.B., Richardson, P.: *Gray's Anatomy for Students*, p. 17. Elsevier/Churchill Livingstone, Philadelphia (2005)
2. Gray, H.: *Gray's Anatomy*, p. 34. Crown Publishers Inc, New York (1977)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation Tech Report (V5)*. University of California, Berkeley, Berkeley (2014)
4. Gandhi, R.: R-CNN, Fast R-CNN, Faster R-CNN, YOLO — Object Detection Algorithms. *Towards Data Science*, 9 July 2018. [Online]. Available: <https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e>. [Accessed 15 Dec 2019]
5. Girshick, R.: *Fast R-CNN*. Microsoft Research, (2015)
6. Everitt: ROI pooling, align, warping. Github, 7 February 2019. [Online]. Available: <https://everitt257.github.io/blog/2019/02/07/ROI-Explained.html>. [Accessed 15 Dec 2019]
7. Kaiming, H., Gkioxari, G., Dollar, P., Girshick, R.: *Mask R-CNN*. Facebook AI Research (FAIR) (2018)
8. SpineWeb: Collaborative Platform for Research on Spine Imaging and Image Analysis. SpineWeb, [Online]. Available: <http://spineweb.digitalimaginggroup.ca/spineweb/index.php?n=Main.Datasets>. [Accessed 31 July 2019]
9. VGG Image Annotator: VGG Image Annotator. (2019). [Online]. Available: <http://www.robots.ox.ac.uk/~vgg/software/via/via-1.0.6.html>. [Accessed 1 Nov 2019]
10. Common Object in Context: ms coco dataset. Common Object in Context, [Online]. Available: <http://cocodataset.org/>. [Accessed 23 Aug 2019]