# DETSEG - SEGMENTATION BASED ON OSTEOLYTIC LESIONS LOCALISATION IN MULTIPLE MYELOMA PATIENTS

## HOW DETECTING A REGION OF INTEREST AROUND A LESION IMPROVES SEGMENTATION PERFORMANCE

NGUYEN QUANG KHOI

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2039045

COMMITTEE

Dr. L.L. Sharon Ong
Dr. Eric Postma

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

July 4, 2022

ACKNOWLEDGMENTS

I would like to thank DR. L.L. Sharon Ong for her supervision and support during this thesis. She has offered me a lot of invaluable insight and direction. This thesis would not be in the state today without her involvement. I also want to thank Martijn van Leeuwen, a researcher of the WeCare project. He helped me understand CT data provided by ETZ, which is a modal I have no experience with. He also offered great advice to help me debug programs.

# DETSEG - SEGMENTATION BASED ON OSTEOLYTIC LESIONS LOCALISATION IN MULTIPLE MYELOMA PATIENTS

## HOW DETECTING A REGION OF INTEREST AROUND A LESION IMPROVES SEGMENTATION PERFORMANCE

NGUYEN QUANG KHOI

**Abstract**

This thesis proposes a two-stage approach, DetSeg, towards osteolytic lesion segmentation in whole-body low-dose CT scans (WBLDCT). Osteolytic lesion is one of the definitive symptoms of Multiple Myeloma diagnosis. Applying Deep Learning to segment osteolytic lesions in WBLDCT may help radiologists detect osteolytic lesions. Current standard methods segments directly (e.g. U-Net models) on CT scans. However, the lesions only make up a small section of the images, leading to unbalanced classes. Hence, this thesis aims to explore how localising lesions with object detection to obtain a region of interest prior to segmentation can improve segmentation performance downstream. A one-stage approach, Mask R-CNN, and a two-stage approach, DetSeg, are experimented. The latter comprises Yolov5 + 2D U-Net. It shows a better segmentation quality in comparison with Mask R-CNN. DetSeg also achieved a higher Dice score than standard methods. DetSeg inference speed is also fast enough for a real-time system.

## 1 INTRODUCTION

### 1.1 *Project Definition*

Multiple myeloma (MM) is a hematologic malignancy with a low survival rate if not detected early (*Multiple Myeloma: Statistics*, 2021). Bone lesions are a symptom commonly seen in MM patients; hence, detecting bone lesions throughout the body is a way to detect MM early. One promising approach is to inspect full-body CT scans of patients. Manual inspection is usually time-consuming and error-prone (Maskell, 2019). Deep Learning may help radiologists in detecting osteolytic lesions.

Segmenting osteolytic lesions on WBLDCT has already been explored in previous work with a U-Net model (Hoff, 2021; Xu et al., 2018). Training segmentation models directly on CT scans is problematic for several reasons. Firstly, lesions are overwhelmed by non-lesion information in a particular CT scan, which leads to non-balance classes. Secondly, lesions can grow outside of bone tissue, making their patterns harder to learn. Therefore, this thesis explores the localisation of regions of interest (ROI) around potential lesions with object detection models prior to segmentation. In this thesis, two approaches are implemented, one-stage and two-stage procedures. The one-stage model is a Mask R-CNN. It utilises a Region Proposal Network (RPN) that proposes potential regions and then simultaneously detects/segments them. The two-stage model is DetSeg would detect ROI with an object detection model, cropping around ROI, segment on them afterwards. The object detection model is Yolov5, while the object segmentation model is 2D U-Net. The proposed DetSeg architecture results are compared with the baseline models and Mask R-CNN.

There is no published work up to date that investigates the inference speed of osteolytic lesion detection models. Yet the number of CT examinations in our society keeps increasing year after year. Over 80 million CT examinations are performed annually in the US, in comparison to only 3 million in 1980 (*Radiation risk from medical imaging*, 2021). Increasing trends in CT examinations, coupled with the rise of radiologists' workload, show that accuracy is not the whole story. A fast Deep Learning model is also needed. This thesis investigates the inference speed of each component in the proposed method. The result is also compared with the baseline model.

This thesis is part of the "Implementation of an optimised AI model for the detection and monitoring of osteolytic bone lesions" research project. The research project is a part of We Care, a research collaboration between Tilburg University and Elizabeth-TweeSteden Hospital (ETZ).

## 1.2 *Motivation*

Multiple myeloma is the second most common hematologic malignancy (Kazandjian, 2016). Two characteristic hallmarks for MM are >10% percentage of clonal bone marrow plasma cell plus CRAB symptoms (Filho et al., 2019). CRAB stands for hypercalcemia, renal failure, anaemia, and bone lesions. If the malignancy has advanced to a distant stage, the 5-year survival rate is 53%. However, if we detect it in the early stage, the 5-year survival rate is 75% (*Multiple Myeloma: Statistics*, 2021). A definitive diagnosis criterion is osteolytic lesions with diameters $> 5mm$.

Radiographic skeletal survey is the default choice in detecting MM bone lesions. However, it may reveal lesions only when over 30% of trabecular bone has been lost (Dimopoulos et al., 2009). Alternative imaging techniques are 3D computed tomography (CT), Magnetic resonance imaging (MRI), or PET/CT. CT allows the detection of smaller bone lesions (Horger et al., 2005) while MRI is more sensitive in the detection of bone marrow infiltration (Dutoit & Verstraete, 2016). As PET/CT imaging requires extensive training to operate and is more expensive, an approach of using only whole-body low-dose CT scans is worthwhile studying. Filho et al. (2019) also suggested that whole-body low-dose CT is a good trade-off between cost and quality.

Manual inspection of CT scans is time-consuming, error-prone, and can be subject to intra and inter-observer variability. In one study, the discrepancy rate between radiologists is 26%; moreover, the discrepancy rate of their interpretations on different occasions can be up to 32% (Abujudeh et al., 2010). So, developing a tool for automatic detection and segmentation into the clinical workflow could save radiologists time and aid in decision making.

The advance in Deep Learning (DL) has significantly transformed the field of computer vision. Instead of manually constructing features like edge or curve detectors, we now can let machines extract patterns in data for us (Lecun, Bengio, & Hinton, 2015). Lecun et al. (2015) introduced the foundation of modern deep learning architecture in computer vision, Convolutional Neural Network (CNN). Building upon CNN, models like YOLO for object detection (Redmon, Divvala, Girshick, & Farhadi, 2016) - object detection; and U-Net for image segmentation (Ronneberger, Fischer, & Brox, 2015) have been created.

In order to detect bone lesions, a straightforward approach is to apply a model like U-Net (Ronneberger et al., 2015) to assign a class label to each voxel in CT scans. However, recent works suggest segmentation quality is not sufficient Hoff (2021). One reason is the heterogeneity of lesions. Another reason is that lesions account for a very small part of CT scans. This thesis aims to localise lesions before segmenting with an object detection model.

Inference speed is also an important metric as models need to fit with radiologists' workload. In comparison with 1980, the number of CT scans conducted annually rose from 3 million to 80 million (*Radiation risk from medical imaging*, 2021), a 26 folds increase. The number of annual cross-sectional images interpreted by radiologists represents a 10-fold increase from 1990 to 2010. This trend leads to the number of images interpreted per minute rising from 2.9 to 16.1 in the same period, even after adjusting for staff changes (McDonald et al., 2015). Yet no paper has investigated inference speed on osteolytic lesion detection models. This metric is benchmarked and discussed in this thesis.

## 1.3 *Research Questions*

Osteolytic lesions are small, heterogeneous and can grow outside of bones. Moreau et al. (2020) proposed segmenting bone tissue first to improve lesion segmentation performance. The Dice score improvement is not significant enough. This thesis suggests a more general method, localising lesions with an object detection model followed by segmentation on patches containing lesions. Yolov5 (Jocher, 2020) would be applied to predict coordinates of potential bone lesions. Then the performance of such a model will be gauged. Finally, those predictions are fed into a 2D U-Net to see if it can improve the performance of segmentation tasks. Hence, the main research question of this thesis is:

> *To what extent does localization of potential lesions with object detection models prior to segmentation improve segmentation performance?*

There are two approaches to this idea. The first is one-stage, where segmentation and detection are done simultaneously based on an RPN. The one-stage model to be implemented in this thesis is Mask R-CNN. The second approach is two-stage, where object detection and segmentation are conducted consecutively. The segmentation model only segments on the patches localised by the object detection model. The two-stage approach to be implemented is Yolov5 + 2D U-Net. The first sub-question of this thesis is:

> *How do the two approaches in localising before segmentation, Mask R-CNN and Yolov5 + 2D U-Net, compare against each other?*

As radiologists nowadays have to read an increasingly high number of CT scans per minute. It is required that a practical system needs to be fast enough for usage. Therefore, a sub-question in this thesis is:

> *To what extent can the DetSeg architecture inference speed fit into radiologists' workload?*

# 2 BACKGROUND

## 2.1 *Multiple Myeloma*

Multiple Myeloma (MM) is a malignant disease of plasma cells in the bone marrow. Healthy plasma cells, which account for less than 5% in the bone marrow, produce normal antibodies to help fight against diseases. In the case of MM, cancerous plasma cells accumulate in the bone marrow and produce abnormal antibodies and light chain. These plasma cells adhere to Bone Marrow Stromal cells to grow and survive. They secrete the protein IL3 Lee et al. (2004) that can decrease osteoblast cell activity, which builds bone. Furthermore, they also secrete other substances like DKK1 Heider et al. (2009) to stimulate osteoclast cell activity, which destroys bones. This dynamic would result in osteolytic lesions in MM patients. Hence, detecting osteolytic lesions is a definitive criterion for diagnosing this disease. Filho et al. (2019) proposed a diagnosis plan to detect and evaluate MM (Figure 1). Inspection of whole-body low-dose CT (WBLDCT)is the first step in detecting osteolytic lesions. Osteolytic myeloma lesions, which are $> 5mm$ in diameters, are detected in this step then the patients would be diagnosed with MM. Treatment would then be started without further follow-up investigation (The left branch of Figure 1). Otherwise, those patients would need to undergo whole-body MRI (WBMRI) to detect MM or Smouldering MM (The right branch of Figure 1). Looking at this diagnosis plan, it can be concluded that accurate detection of these lesions well at the WBLDCT step is essential. Patients detected early have a 5-year survival rate of 75% (*Multiple Myeloma: Statistics*, 2021).

## 2.2 *Object Detection in Biomedical Imaging*

Object detection is the task of predicting bounding boxes around objects of interest in an image. Object detection comprises two tasks, object localisation and classification at the same time. One of the main issues is the limitation of convolutional layers (CL). CL are very good at learning local image features compared to hand-engineered features. However, CL tends to lose object localisation. This is not a problem with classification because this task only concerns an image class, not localisation. There are two types of models to solve this problem, two-stage and one-stage detectors. Two-stage detectors consist of two components, a Region Proposal Network (RPN) that generates ROI from a feature map. ROI is then fed into a detector that classifies and regresses bounding box coordinates. One-stage detectors instead directly learn the class probabilities and bounding box regression on input images. The former are Faster R-CNN, Mask R-CNN

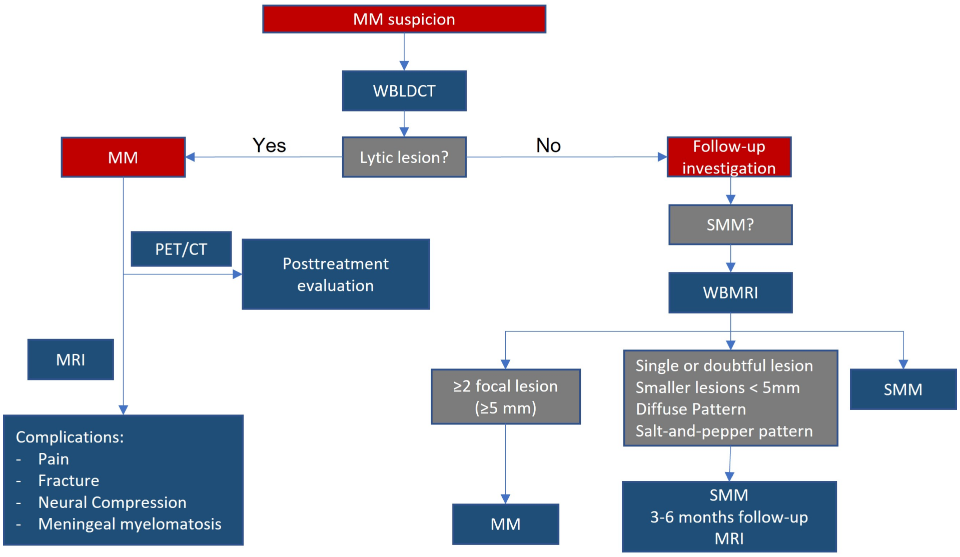


Figure 1: Multiple Myeloma diagnosis and follow-up, adapted from (Filho et al., 2019)

(He, Gkioxari, Dollár, & Girshick, 2017; Ren, He, Girshick, & Sun, 2015) while the latter are Yolov5, SSD (Jocher, 2020; Liu et al., 2016).

Object Detection has since been applied extensively in Medical Imaging. SSD, Faster-RCNN, YOLO, EfficientDet, and RetinaNet are benchmarked for pulmonary nodule detection (Traoré, Ly, & Akhloufi, 2020). VinDr-SpineXR is an ensemble model that combines a classification model with an object detection model to detect abnormalities in spine X-rays (Nguyen et al., 2021).

In the case of MM, radiologists are often interested in osteolytic lesions' size and shape. For example, lesions with diameters $> 5mm$ are a crucial break point for diagnosing MM. That information cannot be extracted from object detection models. Hence, an AI tool for this application would need to incorporate object segmentation and detection in MM diagnosis.

## 2.3 *Object Segmentation in Biomedical Imaging*

In contrast to the object detection problem, segmentation would classify each pixel of an image into a category (e.g., benign, malign). Object segmentation faces a similar but different problem to object detection. CL are good at generating feature maps for input images, but spatial

dimension is lost in the process. Long, Shelhamer, and Darrell (2015) came up with the idea of Fully Convolutional Network (FCNN) to solve this problem. FCNN is an Encoder-Decoder architecture that uses Transposed Convolution to upsample feature maps back to original image sizes. Based on the foundation of FCNN, 2D U-Net(Ronneberger et al., 2015) was proposed as an efficient model for semantic segmentation in Biomedical Imaging. Since then, several other models took inspirations from 2D U-net sprung, V-net, W-net, and nnUnet (Isensee, Jaeger, Kohl, Petersen, & Maier-Hein, 2021; Milletari, Navab, & Ahmadi, 2016; Xia & Kulis, 2017). Besides segmenting directly on images, another method is to segment on ROI generated by a Region Proposal Network (He et al., 2017).
There have been many kinds of research on object segmentation in Biomedical Imaging. Dong, Yang, Liu, Mo, and Guo (2017) studied U-Net to segment brain tumours. Lei et al. (2019) uses V-net to segment prostate in Ultrasound images. Mask R-CNN, a model that segments based on ROI proposed for detection, has also been studied to segment Polyp (Qadir et al., 2019).

Difficulties in Biomedical Imaging segmentation lie in the nature of medical images dataset. The first difficulty is dataset size. Medical dataset is sensitive and is usually protected under strict privacy law in many countries. Hence, applying data augmentation is essential. Secondly, class imbalance can affect segmentation models' performance. The class(es) of interest usually only accounts for a small percentage of pixel/voxel in medical images. Thirdly, local objects like organs, bones, or body fluids can occlude objects of interest. Last but not least, the variability of imaging techniques proves challenging to find a one-size-fits-all solution. For instance, as CT scans spatial dimension is presented in three dimensions, it is possible to train on individual slices or the whole scan. Each may give a different result.

### 2.4 *Osteolytic Lesion Segmentation*

Deep Learning has been applied to segment bone lesions by some researchers. Moreau et al. (2020) applied nnU-Net to segment bone lesions on PET/CT images in the context of metastatic breast cancer. Xu et al. (2018) incorporated V-net to W-net to segment osteolytic lesions in the case of MM. Their methods only show a high Dice coefficient with PET/CT, but not with only WBLDCT.

One way to improve performance is to utilise specialised data transformation techniques. Osteolytic lesions occur in the bone tissue, so training with both segmented bone tissue and lesions can improve Dice coefficient (Moreau et al., 2020). Another way is to combine different imaging techniques to learn more features. Xu et al. (2018) presented this idea with a W-net. CT scans are first fed into a V-net, and then PET/CT scans along with output from the first V-net are fed into a second V-net. The Dice Coefficient is 0.7298, an impressive result compared to 0.2637 of training V-net on CT scans alone. Nevertheless, it is not always possible to conduct both imaging techniques on the same patient at the same time.

## 2.5 *DetSeg Architecture*

An essential obstacle in computer vision is developing a mechanism to filter out irrelevant images' details. Faster R-CNN does this by using an RPN to find regions that contain objects. He et al. (2017) extended this model for segmentation by applying a FCN on each ROI generated by the RPN. Although this model works well on a standard dataset as COCO, there are two drawbacks to this method. First, RPN is intended to find regions that contain objects to improve object detection performance. It is not suited for localising complicated patterns. In addition to that, a single FCN is very limited and may not be enough to segment well.

A different solution is to combine specialised object detection and object segmentation models. The intuition behind the idea is that detection is used to localise and filter out irrelevant pixels; segmentation is then conducted on high-resolution patches that centre around objects of interest. It is shown that by conducting segmentation on detected glomerular using U-Net or DeepLab_v3 in contrast with simply applying Mask R-CNN, Dice coefficient improves (Jha et al., 2021). The architecture also reports state-of-the-art results on flank wear area segmentation (Lin et al., 2021). This architecture has never been tested on osteolytic bone lesion segmentation before. Hence, this thesis aims to explore whether it can improve lesion segmentation performance.

To this date, no studies have benchmarked the inference speed of this architecture with regard to radiologists' workload. Hence, this thesis will explore this new direction to show that not only it can provide better accuracy, but also practical to use.

## 2.6 *Contributions*

Applying deep learning to segment osteolytic lesions in WBLDCT is still a new area of application. This research contributes to the body of literature concerning MM detection and its practical aspect. Xu et al. (2018) shows a decent performance with V-Net/W-Net, but it requires PET/CT while

WBLDCT is more available as an early diagnosis tool. Hoff (2021) tested with different data augmentation techniques, transfer learning, and bone segmentation techniques on the ETZ dataset but did not seem to achieve a good result. This thesis proposed a new approach called DetSeg, which detects lesions as a localisation technique. Segmentation can then be applied to lesion localisation. The result shows that DetSeg outperforms Mask R-CNN and segmentation models that segment directly on the dataset. This result would serve as a baseline for other researchers to improve upon. It is possible to replace some components of Dseg while still keeping the general working intact, making it a very flexible architecture.

Moreover, this thesis also benchmarks a practical aspect of any detection system, inference speed. The inference speed of either component is benchmarked. The benchmarks would provide invaluable insight for two reasons. Firstly, it serves as a baseline for any future system. The trade-off between accuracy/speed can now be analysed in future research. Secondly, it shows that not only does it outperforms the accuracy of other models, but it also does not sacrifice any practical usage.

As far as the author is aware, there has not been any work of a similar approach as DetSeg in Osteolytic lesions in WBLDCT before. Hence, the work in this thesis is the first of its kind. Being able to diagnose MM early can mean a significant increase in survival rate. Every incremental progress can mean the difference between life and death for many patients. The author hopes that DetSeg may lead to a new direction in the battle against MM.

## 3 METHODS

### 3.1 *Dataset*

The CT scans used in this thesis belong to Elizabeth-TweeSteden Hospital (ETZ) in Tilburg, Netherlands. The patients in each scan are 18 years or older. There are 96 full-body CT scans from 79 patients acquired by various scanners, each with a maximum of 20 lesions. The scans are DICOM images with a resolution of either 768 x 768 or 512 x 512 pixels, which are then combined to make 3D axial slices. The slice thickness is either 2.5 *mm* or 3.0 *mm*; pixel spacing varies between 0.42 *mm* to 0.98 *mm* in every direction. These lesions are annotated using 3D Slicer (Cie) by radiology residents. They are saved as uncompressed binary label maps.

Lesions vary in locations, size, and morphology. They can appear in upper, lower limbs, vertebrae, pelvis, ribs and other locations. The size can be between 0.3857 $mm^2$ and 2653.72 $mm^2$. There are 2972 lesions with a mean area of 288.29 $mm^2$ and a standard deviation of 411.6 $mm^2$. Lesion distribution can be seen in Figure 3. The red and black dotted lines are 10%; 90% quantile and mean, respectively. One important break point of diagnosing MM is lesions with diameters $> 5\ mm$. If we approximate a lesion as a circle, then the approximated area is $\frac{\pi}{4} \times D^2 = 19.63\ mm^2$. From Figure 3 we can see the breakpoint for this area is roughly around the 10% quantile.

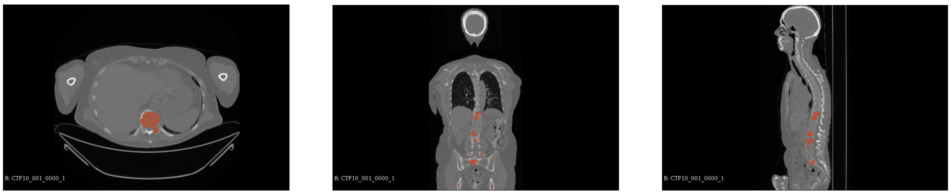


Figure 2: Axial, Coronal, and Sagittal scans from left to right, respectively. Lesions are annotated in red

## 3.2 *Data preparation*

### 3.2.1 *CT scan transformation*

CT image files in DICOM format voxel values are in the Hounsfield scale, which describes radiodensity. A value of -1000 HU is the radio density of air, while HU values of bones can range from 300 to 1900. Since all object detection architectures in this thesis work with images, the values are normalised to be in the range [0;1]. Voxel values $V_i(x, y, z)$ of the scan $i$ is normalized as follows.:

$$V_i(x, y, z) = \frac{V_i(x, y, z) - Min(V_i(x, y, z))}{Max(V_i(x, y, z)) - Min(V_i(x, y, z))} \tag{1}$$

Both CT scans and their masks are then sliced axially to create 2D images as a new dataset. After pre-processing, the dataset contains 43996 axial slices, and 2265 of them contain lesions.

### 3.2.2 *Data Augmentation*

Data Augmentation is a crucial process whenever dataset size is short. As in this dataset, the percentage of axial slices that contain lesions is only around 5%. Hence, translation, flipping, and rotating are applied to lesion slices. Custom functions are written to handle data augmentation

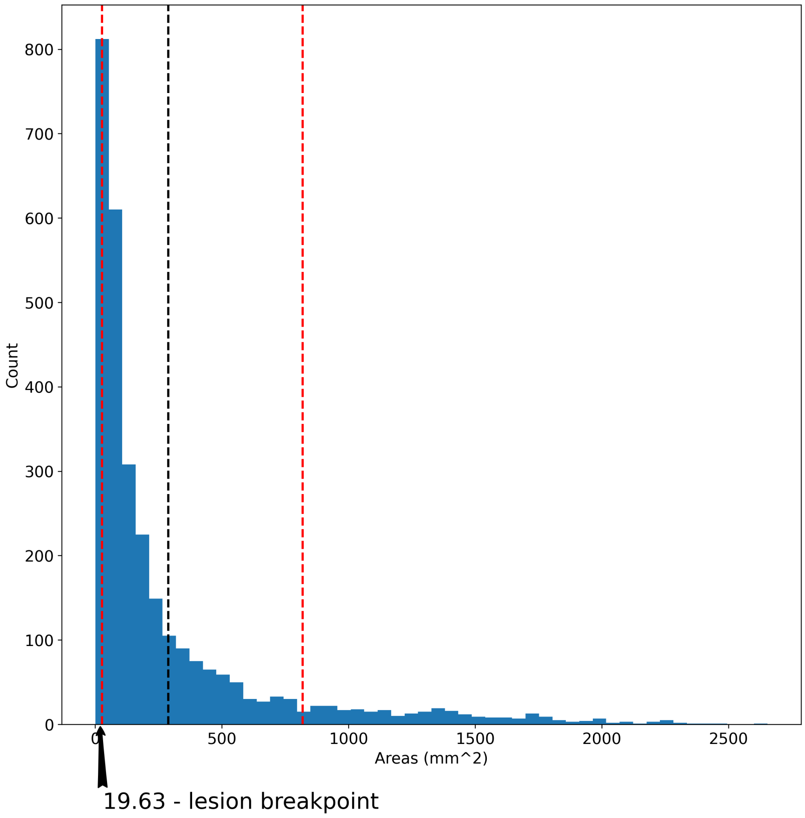


Figure 3: Lesion distribution

for this project. For each lesion with slices, the image is first translated randomly between −200 to 200 pixels in either width and height directions. Then, there is a 50% chance that it is flipped left-right or up-down. Finally, it is rotated randomly between −30 and 30 degrees with a 50% chance. This augmentation process is applied repeatedly 10 times for each lesion slice. After augmenting, the dataset has in total 66646 images with 22650 of them containing lesions. The whole pipeline for transformation and augmentation can be seen in Figure 4. Besides data augmentation techniques applied by the author, Yolov5 and Mask R-CNN each automatically apply their own augmentation techniques. The former applies mosaic augmentation, while the latter applies Scale Jittering augmentation.

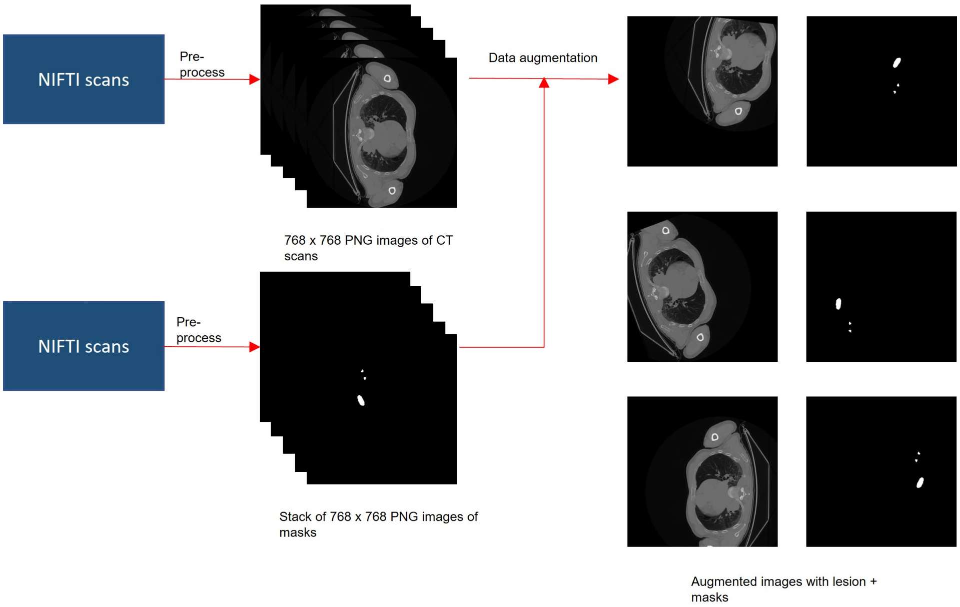

NIFTI scans
Pre-process
768 x 768 PNG images of CT scans
Data augmentation
NIFTI scans
Pre-process
Stack of 768 x 768 PNG images of masks
Augmented images with lesion + masks

Figure 4: Data Transformation and Augmentation pipeline

Table 1: Yolov5 bounding box label file for a lesion image with 3 lesions (row names is only for illustration)

| Class number | center_x | center_y | width | height |
|---|---|---|---|---|
| 0 | 0.368 | 0.244 | 0.016 | 0.018 |
| 0 | 0.38541 | 0.2929 | 0.0156 | 0.0182 |
| 0 | 0.3593 | 0.4414 | 0.0481 | 0.0703 |

## 3.3 *Annotated Label Transformation*

### 3.3.1 *Yolov5 label transformation*

The original dataset does not have the bounding box target information required to train and evaluate Yolov5. Hence, custom processing functions need to be written to process the masks and create bounding box labels. For each lesion, the centre, bounding box height and width are extracted. They are then normalised by dividing by the corresponding image height and width. Afterwards, they are written into a text file. Each lesion image would have a text file detailing bounding box labels on each line. An example is illustrated in table 1.

### 3.3.2 *Mask R-CNN instance segmentation mask transformation*

Mask R-CNN is an instance detection/segmentation architecture and was originally evaluated on the COCO dataset. Hence, it is necessary to create instance masks and prepare the data conforming to COCO format. First, instance masks are extracted from semantic segmentation masks per Figure 5. Each instance mask is then transformed into a bitmask with values of 0 for the background class and 1 for the lesion class. Finally, they are encoded into run-length encoding (RLE) format. RLE is a compression format used for the COCO dataset because it saves space by encoding only the size of masks, pairs of offset and counts that mark how many labelled pixels are since the offset.

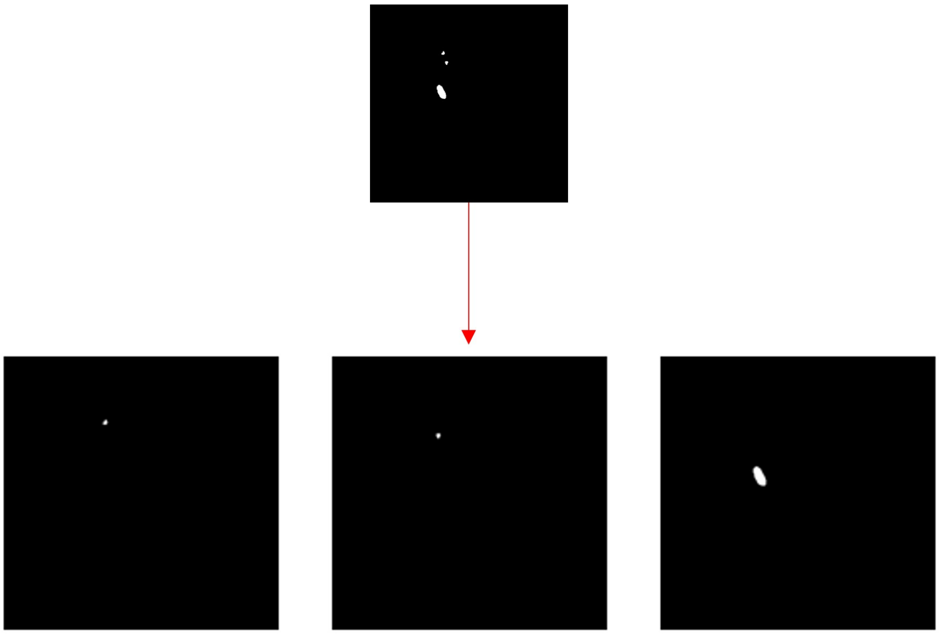


Figure 5: Instance masks creation from a semantic segmentation mask

### 3.3.3 *2D U-Net data transformation*

A 192 × 192 patch is cropped around the centre of each detected lesion by Yolov5. These cropped images are then transformed into numpy arrays and rescaled to $[0, 1]$ range before training. The rescale equation can be seen in Equation 2:

$$I(x, y) = \frac{I(x, y)}{255} \tag{2}$$

## 3.4 *Models*

### 3.4.1 *Yolov5*

Traditionally, object detection models like R-CNN, Fast-RCNN, or Faster-RCNN would try to propose potential regions and then predict their classes. Yolov5 belongs to the family of Yolo models, which take a different approach by predicting directly on images using grids (Redmon et al., 2016). Instead of training a convolutional model to find potential regions, Yolo would look at the image once and make predictions on each cell. Yolov2 added anchor boxes that let them detect multiple objects in the same grid cell (Redmon & Farhadi, n.d.). Multi-scale features for the object detection model and some adjustments to the network architecture were added to Yolov3 (Redmon & Farhadi, 2018). The most recent significant changes came with Yolov4, which proposed SPP and YAN for feature aggregation. Moreover, Mosaic augmentation and Self-Adversarial Training (SAT) were introduced that would improve training performance (Bochkovskiy, Wang, & Liao, 2020). In this project, Yolov5 would be used for two reasons. Firstly, it is developed on the Pytorch framework instead of Darknet. This framework choice makes it easier to train and adjust. Secondly, Yolov5 offers more flexibility in model sizes and data enhancement configurations while not sacrificing performance. Yolov5 architecture can be seen in Figure 6.

Let the output of Yolov5 be a tensor of shape $S \times S \times A \times (C+5)$, where S is the shape of the grid, A is the number of anchors, C is the number of classes and 5 are the predictions. The network predicts 4 coordinates $t_x, t_y, t_w, t_h$ for each bounding box. If the cell offset from the image top left is $(c_x, c_y)$ and $p_w, p_h$ are the bounding box width and height prior, respectively. The final prediction is simply confidence score. The predictions can be formalized as follows:

$$
\begin{aligned}
b_x &= \sigma(t_x) + c_x \\
b_y &= \sigma(t_y) + c_y \\
b_w &= p_w e^{t_w} \\
b_h &= p_h e^{t_h}
\end{aligned}
$$

Yolov5 can be trained by optimizing the loss function defined in Equation 3. $\mathcal{L}_{CIoU}$ is the Complete Intersection over Union loss that penalises the 4 coordinates errors (Zheng et al., 2019). $\mathcal{L}_{confidence}$ tries to maximizes IoU between predicted boxes and ground-truth. $\mathcal{L}_{class}$ penalises class misclassification. Besides $\mathcal{L}_{CIoU}$, the other losses are Binary Cross-Entropy.

$$\mathcal{L}_{Yolov5} = \mathcal{L}_{CIoU} + \mathcal{L}_{confidence} + \mathcal{L}_{class} \tag{3}$$

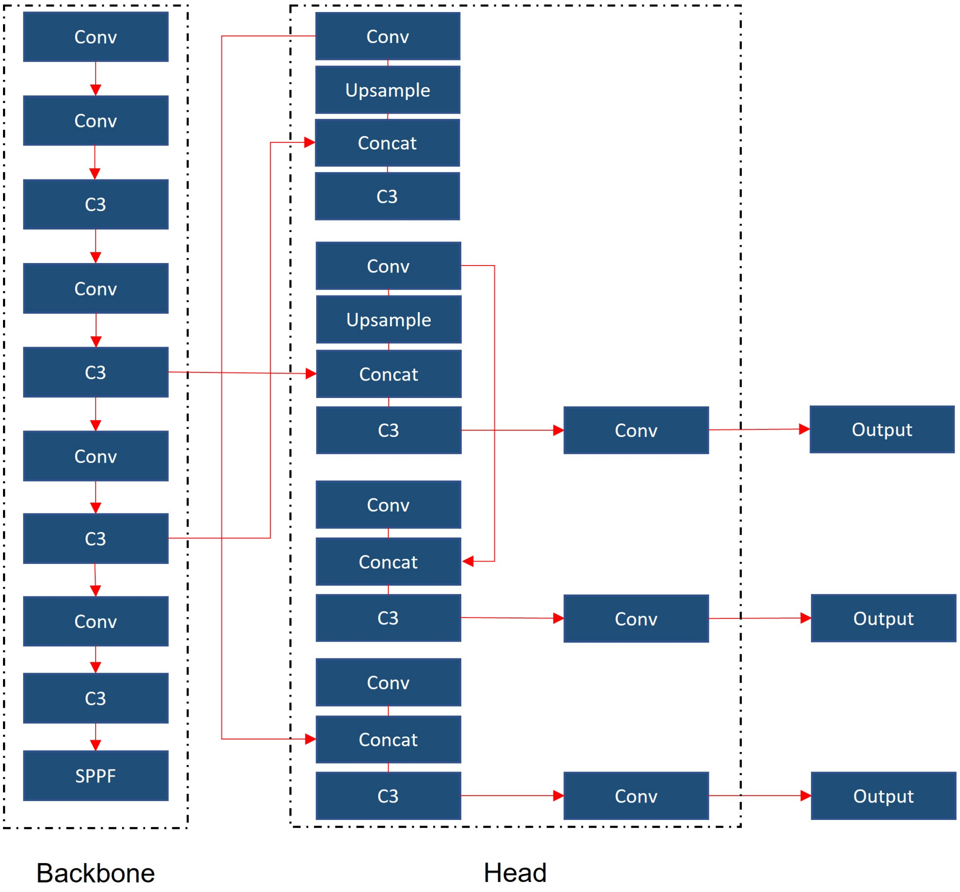


Figure 6: Yolov5 architecture

### 3.4.2 *2D U-Net*

2D U-Net is a lightweight, efficient semantic segmentation model Ronneberger et al. (2015). Although convolutional layers are effective at learning local properties in images, they would lead to the loss of original image size. Another problem is that deep convolutional layers would "forget" properties in earlier layers. 2D U-Net tried to solve these problems by utilising an encoder-decoder approach. There are two main components in 2D U-Net, a contracting and detracting path. The Contracting path stacks consecutive convolutional layers to learn feature maps of the image.

Convolutional layers with max pooling would downsample images. Feature maps are then connected to a detracting path that would apply transposed convolution to expand the feature map size back to the original. Skip connections are used to let the model learn from earlier layers. In this project Dropout layers are also added between convolutional layers to alleviate overfitting (Srivastava, Hinton, Krizhevsky, & Salakhutdinov, 2014). The model architecture can be seen in Figure 7.

Let the input of 2D U-Net be a tensor of shape $W \times H$, then the output is a probability map of shape $W \times H$ as a result of the sigmoid function. The loss function is the Binary Cross-Entropy function defined as in Equation 4. $\hat{p}_{ij}$ is the $[0;1]$ predicted value of the pixel $ij$ while $p_{ij}$ is the ground-truth label.

$$\mathcal{L}_{U-Net} = -\sum_{i=1}^{W}\sum_{j=1}^{H} \hat{p}_{ij} \log p_{ij} + (1 - \hat{p}_{ij}) \log (1 - p_{ij}) \tag{4}$$

### 3.4.3 *Mask RCNN*

Mask R-CNN expands upon Faster-RCNN by adding a branch to predict instance segmentation masks (Figure 8). The RPN layer would propose ROI, which is then fed separately into detection and segmentation heads. Furthermore, there is a RoIAlign layer to improve segmentation performance inbetween. An advantage of Mask RCNN is that it segments on the same ROI proposed for detection as Faster RCNN. Hence, leading to using one network (RPN) for two tasks. Along with the fact that it is the *de facto* method of doing instance detection/segmentation in computer vision, it is chosen to compare for its similarity to the thesis idea. The implementation chosen for this project is the ResNet-50+FPN+LSJ based on Detectron2 framework (Facebook AI Research (FAIR), 2022). This implementation automatically applies Large Scale Jittering (LSJ), which can improve Average Precision (AP) performance (Ghiasi et al., n.d.).

The loss function of RPN is defined as follows:

$$\mathcal{L}_{RPN} = \mathcal{L}_{rpn\_cls} + \mathcal{L}_{rpn\_reg}$$

$\mathcal{L}_{RPN}$ let RPN learns to propose ROIs that contain objects. $\mathcal{L}_{rpn\_cls}$ learns whether RPN can separate objects from background while $\mathcal{L}_{rpn\_reg}$ learns the localisation of objects.

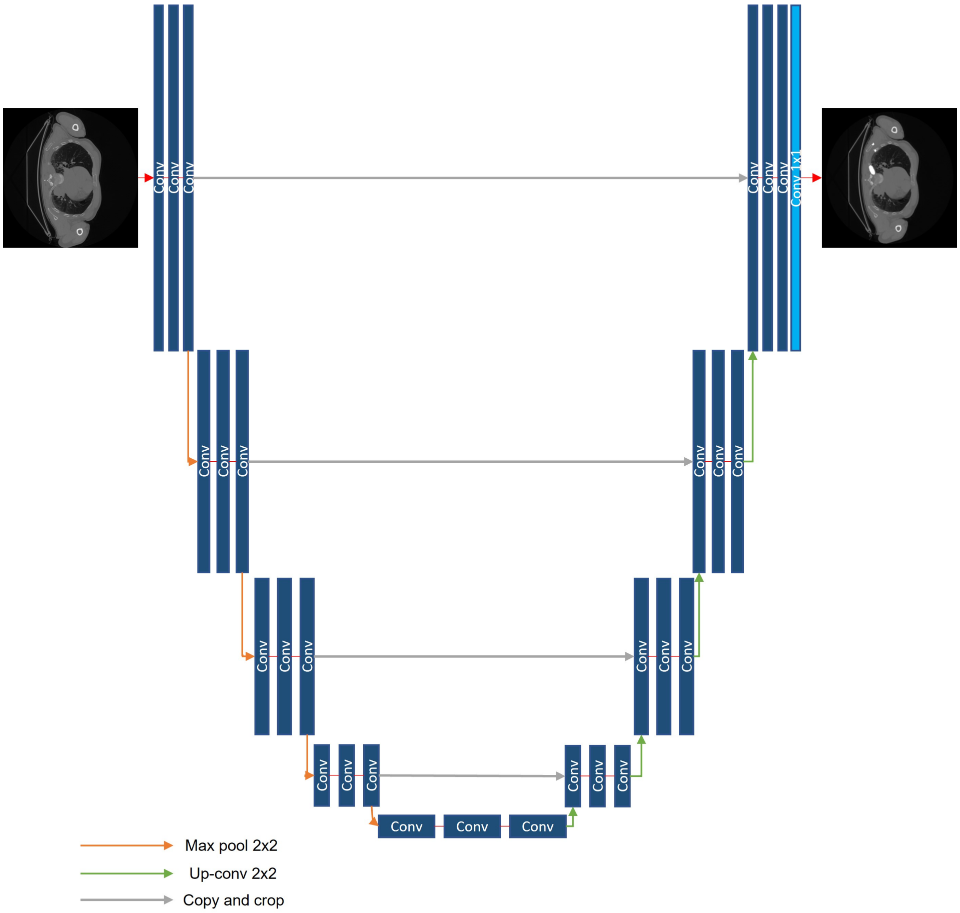


Figure 7: 2D U-Net architecture, adapted from (Ronneberger et al., 2015)

The loss function of Mask R-CNN is then formalised in Equation 5. $\mathcal{L}_{mrcnn\_bbox}$ and $\mathcal{L}_{mrcnn\_class}$ are losses of the detection head. They penalise incorrect localisation and class misclassification, respectively. The final loss, $\mathcal{L}_{mrcnn\_class}$ is the loss for the segmentation head.

$$\mathcal{L}_{Mask\ R-CNN} = \mathcal{L}_{RPN} + \mathcal{L}_{mrcnn\_bbox} + \mathcal{L}_{mrcnn\_class} + \mathcal{L}_{mrcnn\_mask} \tag{5}$$

## 3.5 *Evaluation Metrics*

### 3.5.1 *Mean Average Precision(mAP)*

The evaluation metric for object detection is the Average Precision (mAP). This metric is the area under the precision-recall curve (Equation 6). Specifically, the $AP@[.50 : .05 : .95]$ and $AP@0.5$ are used. The threshold

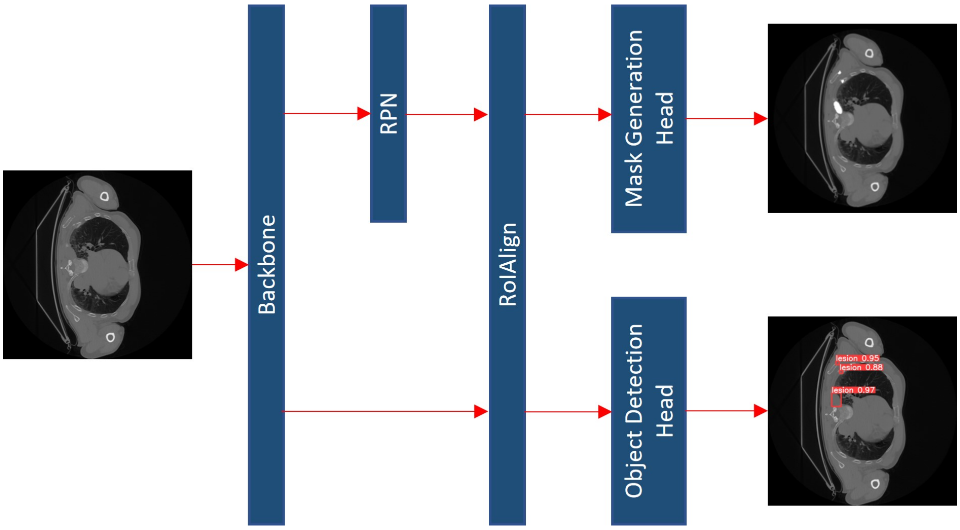


Figure 8: Mask R-CNN architecture

value for this metric is IoU.

$$AP = \int_0^1 precision(r)dr \tag{6}$$

### 3.5.2 *Dice coefficient*

A standard metric for either object detection or segmentation tasks is Intersection-over-Union (IoU). The equation for this metric can be seen in Equation 7. X and Y can be bounding boxes or segmentation masks depending on the task. A similar metric, the Dice coefficient, can be derived as in Equation 8. IoU score tends to punish individual mistakes harsher than Dice score. A special property of them is that they are always positively correlated.

$$IoU(X,Y) = \frac{|X \cap Y|}{X \cup Y} = \frac{TP}{TP + FP + FN} \tag{7}$$

$$IoU(X,Y) = \frac{Dice(X,Y)}{2 - Dice(X,Y)} \tag{8}$$

Dice coefficient/F1-score is a standard metric to evaluate the quality of predicted segmentation masks. It is used in this thesis instead of IoU. The Dice coefficient is calculated as follows in the thesis:

$$Dice(X,Y) = \frac{2|X \cap Y|}{X \cup Y} = \frac{2 * TP}{2 * TP + FP + FN} \tag{9}$$

## 3.6 *One stage approach: Mask R-CNN*

The standard Mask R-CNN is trained and evaluated on the whole-body CT scans provided by ETZ hospital. Mask-RCNN makes a prediction for each instance in each image. Predictions of instances in the same image are combined to create a single binary mask. This mask is then compared with the ground truth to evaluate the Dice score, precision, and recall. Object detection is evaluated by comparing detected bounding boxes with ground truth bounding boxes. Inference speed is benchmarked by running predictions on the test set.

## 3.7 *Two stage approach: DetSeg*

### 3.7.1 *Lesion Detection: Yolov5*

Yolov5 is chosen as the object detection as it is an efficient and fast model. It can predict objects based on three different scales (Figure 6). Hence, it is fitting as lesion sizes are homogeneous. Yolov5 is also capable of high inference speed as it does not propose ROI. Yolov5 is trained on the same dataset as Mask R-CNN. Afterwards, a 192 × 192 around the centre of each predicted lesion is cropped. Those cropped images serve as training data for the segmentation model. 192 × 192 patches are also cropped in segmentation masks at the exact predicted coordinates as ground truth. Object detection is evaluated by comparing detected bounding boxes with ground truth bounding boxes. The inference speed is benchmarked by running predictions on the test set.

### 3.7.2 *Lesion Segmentation: 2D U-Net*

2D U-Net is chosen as it is a standard segmentation model in biomedical imaging. It is lightweight in terms of parameters. It is also suitable as a baseline for future development. If it can perform well, then it is reasonable to conclude that better segmentation models can improve it even further. 2D U-net is trained on cropped images and binary masks produced by lesion localisation predicted by Yolov5. The final layer with sigmoid produces a probability map for each input image. The map is binarised at the threshold of 0.5 to create the final segmentation mask. The probability map is used for minimising loss function, but the other one is used for evaluation. Inference speed is benchmarked by running predictions on the test set.

### 3.7.3 *Optimizing DetSeg*

DetSeg optimizes a loss function formalised in Equation 10. More generally, DetSeg can be formalised as in Equation 11. On the one hand, Mask R-CNN two heads rely on how well RPN optimises $\mathcal{L}_{RPN}$ to perform well. On the other hand, DetSeg decouples this co-dependence to let two components focus on their specialities. This modification has two advantages. First, optimising two parts independently without creating side effects is easier. Secondly, the segmentation part of DetSeg is trained on the result of the detection part. Hence, it can correct mistakes made by the detection component.

$$\mathcal{L}_{DetSeg} = \mathcal{L}_{Yolov5} + \mathcal{L}_{U-Net} \tag{10}$$

$$\mathcal{L}_{DetSeg\_general} = \mathcal{L}_{obj\_detection} + \mathcal{L}_{obj\_segmentation} \tag{11}$$

### 3.7.4 *Inference speed*

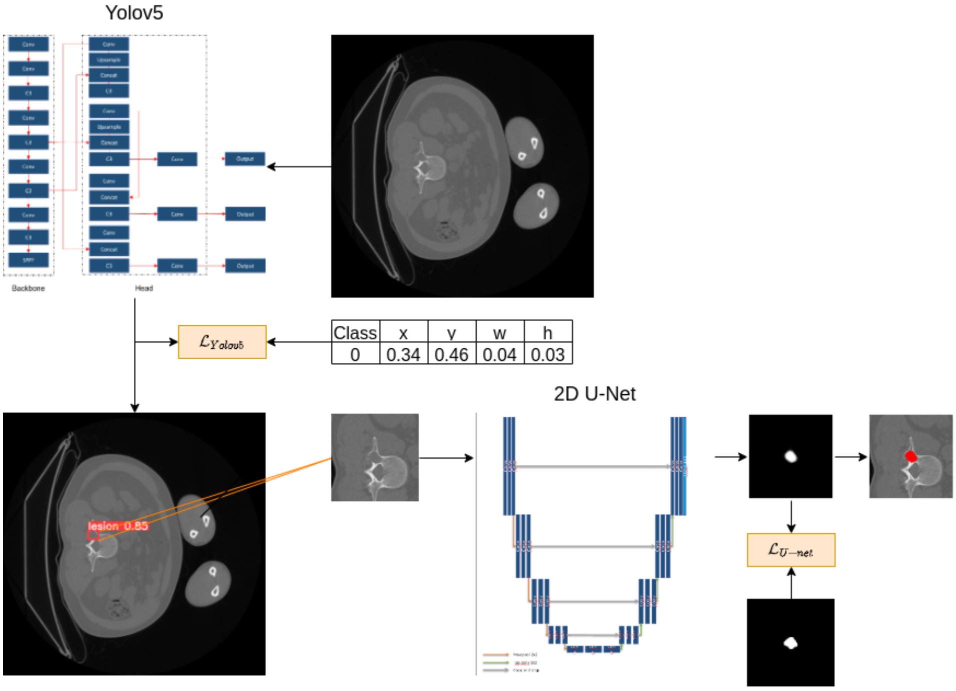
Although the architecture utilises two models, it is still efficient in terms of training time and speed. Since the segmentation model is trained only on the detected patches, training size decreases in a significant order. Thus, the training time of the whole architecture is not increased significantly. Furthermore, by careful selection of models, the inference speed would still be practical. Inference speed is benchmarked by calculating the average of each image prediction time.

## 3.8 *Implementation Details*

The dataset is divided into 5 folds. Each training iteration would be conducted by training on 4 folds with a train/validation split of 90%/10%. The remaining fold would then be used as a test set. This process would then be repeated in a cross-validation manner. Averages and standard deviations across 5 test sets are then reported.

### 3.8.1 *Mask R-CNN*

The dataset is trained directly using Mask R-CNN then predicted masks are made on hold-out sets. Instance masks predicted by the model are combined to create semantic masks. Segmentation evaluation is then made on them afterwards. Each iteration is trained for 100 epochs with a mini-batch size of 16. The optimizer is SGD (momentum = 0.9), initialized at



| Class | x | y | w | h |
|---|---|---|---|---|
| 0 | 0.34 | 0.46 | 0.04 | 0.03 |

Figure 9: DetSeg architecture

0.01, weight decay rate of $4 * 10^{-5}$. The learning decreases to 0.001 and 0.0001 after 88 and 96 epochs, respectively.

### 3.8.2 *DetSeg*

The dataset is first trained directly using Yolov5. Each iteration is trained for 300 epochs with a mini-batch size of 64. The model employs an early stopping mechanism of 20. The optimiser is SGD (momentum:0.937) with an initial learning rate of 0.001. The learning rate schedule is Cyclical Learning Rate (CLR) (Smith, 2015) with a lower bound learning rate and upper bound of 0.01 and 0.1, respectively. After training, Yolov5 would detect lesions across the whole dataset and save predictions in the same format as Table 1. For each predicted lesion, a $192x192$ patch is cropped around the centroid. This process is applied to both CT scans and segmentation masks to create training images and targets.

$2D$ U-Net is then trained on those patches in 150 epochs each iteration. The learning rate schedule is also CLR with a lower and upper bound of 0.0001 and 0.01, respectively. Early stopping with the patience of 20 is employed. The loss function is Binary Cross Entropy.

### 3.9 *Software, Hardware and Libraries*

All training was made possible on an NVIDIA Quadro RTX 6000 24GB GPU provided by the hospital. The programming language for this thesis is Python 3.9.7. NiBabel, Pydicom, pickle, and pyplot are the libraries used for Data transformation. The Data Augmentation pipeline utilises Pytorch framework. The Mask R-CNN implementation comes from detectron2 framework (Facebook AI Research (FAIR), 2022). Yolov5 implementation comes from (Jocher, 2020). 2D U-Net implementation is programmed in Keras TensorFlow. All training, inference, or evaluation code uses either TensorFlow or Pytorch. Visualisation is made by 3D Slicer, Matplotlib.

## 4 RESULTS

### 4.1 *Lesion Detection Results*

Lesion detection evaluation metrics can be seen in Table 2. The average of a 5-fold cross-validation setting (80%-20% split) and standard deviation are reported. The Yolov5 model achieved a higher $mAP@0.5$ of 0.3808 score compared to the Mask R-CNN model, which has a score of 0.1338. Though Mask R-CNN $mAP@[.50 : .05 : .95]$ is better than Yolov5. It means that Mask R-CNN is better at lesion detection when averaging across different IoU thresholds while Yolov5 performs better at 0.5. Figure 14 illustrates detection results in three categories, "Good detection", False Positive (FP) cases, and False Negative (FN). Good detection is determined where predicted boxes coincide to a high degree with ground truth boxes. FP means there are lesions detected but are not there. FN means Yolov5 missed one or more lesions. FP case is not an issue because 2D U-Net would learn to correct the mistake afterwards. Since a 192 × 192 patch is cropped around the predicted centre, the case where predicted boxes and ground-truth boxes do not collide completely is also not an issue. The training time for Yolov5 is around 2.5 days, while Mask R-CNN takes around 3 days for each fold iteration.

Table 2: Lesion Detection metrics

| Model | $mAP@0.5$ | $mAP@[.50 : .05 : .95]$ |
|---|---|---|
| Yolo v5 | **0.3808** ± 0.0415 | 0.1639 ± 0.0262 |
| Mask R-CNN | 0.1338 ± 0.0271 | **0.31092** ± 0.0516 |

## 4.2 *Lesion Segmentation Results*

The results can be seen in Table 3. The average of a 5-fold cross-validation setting (80%-20% split) and standard deviation are reported. Not only DetSeg perform better than Mask R-CNN, but it is also better than segmenting directly on the CT scans. We can see that a regular 2D U-Net is enough to get a better result than 3D U-Net or 3D nnUnet. Transforming the data with bone segmentation then segment also does not perform as well as DetSeg. Moreover, DetSeg trained the segmentation head very fast due to a smaller dataset and less elaborate than other models. 2D U-Net head of DetSeg trained for only 3 hours for each fold iteration. Overall, the training time for DetSeg took roughly the same time as Mask R-CNN while performing better.

A more in-depth analysis of the result can be seen in Table 4. An analysis of precision and recall would go into detail about what kind of errors DetSeg makes. High precision means DetSeg does not classify lesions where they are not there while classifies lesions shape correctly. A high recall means DetSeg does not miss many lesions or classify lesions smaller than they are. Around 60% segmentation is of "high quality", which means having a precision of $> 0.7$ and recall of $> 0.5$. However, around 25% of them are of the opposite case. The illustration for this analysis is in Figure 10. As we can see from this figure, DetSeg makes a mistake when lesions are in a non-rounded shape like others. Big lesions are also an issue since most lesion area is below 1000 $mm^2$ (Figure 3). More illustration of the result is in Figure 15.

Table 3: Lesion Segmentation metrics

| Model | Dice | Precision | Recall |
|---|---|---|---|
| DetSeg | **0.49** ± 0.0079 | **0.5652** ± 0.036 | **0.4596** ± 0.0415 |
| Mask R-CNN | 0.2496 ± 0.0121 | 0.1235 ± 0.0371 | 0.2092 ± 0.0375 |
| 2D U-Net + Bone Pre-seg (Hoff, 2021) | 0.3807 ± 0.0217 | 0.4754 ± 0.0336 | 0.3746 ± 0.0219 |
| 3D U-Net + Genesis TL + Bone Post-seg (Hoff, 2021) | 0.3192 ± 0.0372 | 0.3032 ± 0.0226 | 0.4307 ± 0.0782 |
| 3D nnU-Net (Hoff, 2021) | 0.3433 | 0.4189 | 0.3408 |

## 4.3 *Inference Speed Benchmarks*

The inference speed of both the detector - Yolov5 and the segmentor - 2D U-Net is benchmarked. As observed in Figure 11, Yolov5 offers flexibility

Table 4: Lesion Segmentation analysis

| Precision/Recall | $< 0.5$ | $> 0.5$ | $> 0.7$ |
|---|---|---|---|
| $< 0.5$ | **0.25** $\pm 0.0613$ | $0.006 \pm 0.0079$ | $0.0225 \pm 0.0230$ |
| $> 0.5$ | $0.0073 \pm 0.0044$ | $0.0096 \pm 0.0040$ | $0.0369 \pm 0.0201$ |
| $> 0.7$ | $0.064 \pm 0.0378$ | **0.1001** $\pm 0.0509$ | **0.4976** $\pm 0.1530$ |

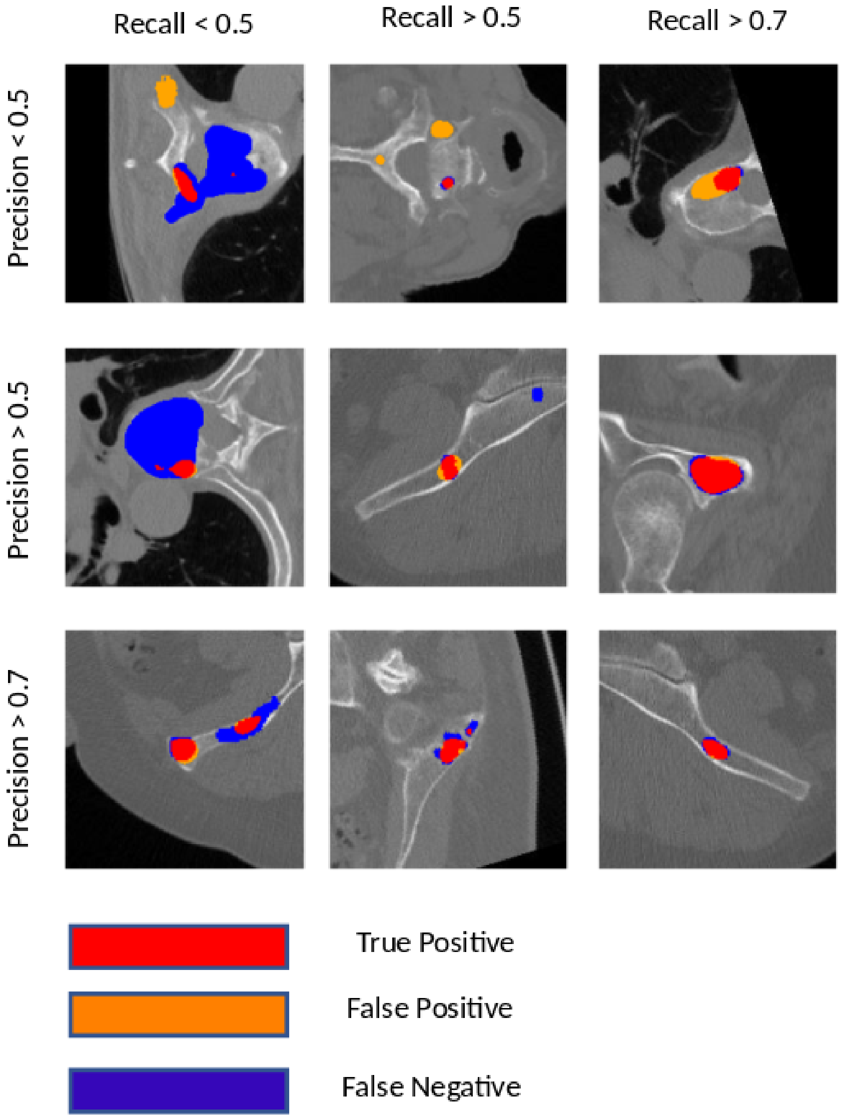


Figure 10: DetSeg Segmentation analysis illustration

in terms of inference speed. It is possible to run at 1.5$ms/img$ without sacrificing too much $mAP$@0.5, precision, and recall. The inference speed of 2D U-Net is also very fast, $1ms/img$. In total, DetSeg can be run at $2.5ms/img$ without sacrificing too much detection quality compared with Mask R-CNN at $4ms/img$. According to (McDonald et al., 2015), radiologists spent around 16.1 images per minute. Hence, the inference speed of DetSeg fits well into radiologists' workload. The only bottleneck that affects the actual running time of DetSeg is data movement, which is discussed in the Discussion section.

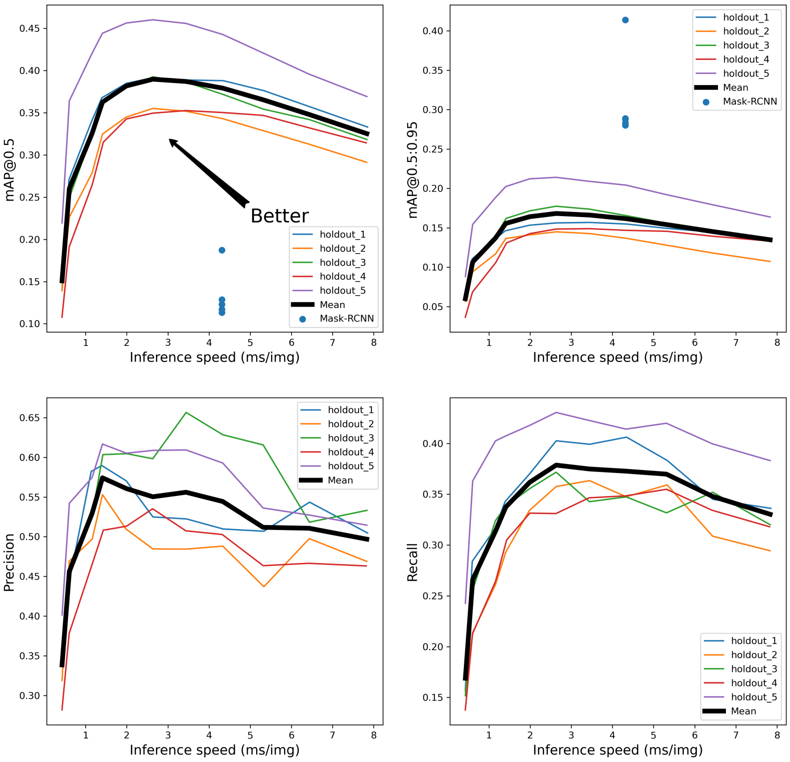


Figure 11: Yolov5 inference speed benchmark

Table 5: 2D U-Net head inference speed benchmark

| Model | inference speed(ms/img) |
|---|---|
| Mask R-CNN | $4ms$ |
| 2D U-Net | $1ms$ |

# 5 DISCUSSION

The goal of this thesis is to develop a new method to improve osteolytic lesion segmentation in WBLDCT. The idea is to localise lesions first before segmenting. The novel method proposed is DetSeg, which comprises an object detection model followed by an object segmentation model. The former would localise lesions while the latter segment on images cropped from predicted localisation. In particular, Yolov5 + 2D U-Net is utilised. DetSeg was compared with Mask R-CNN, a model that does both at the same time. The result showed that DetSeg performs segmentation better than Mask R-CNN. The performance is also better than many segmentation models that train directly on the dataset. Inference speed was also benchmarked to show that it also infers fast enough for practical use.

## 5.1 *DetSeg*

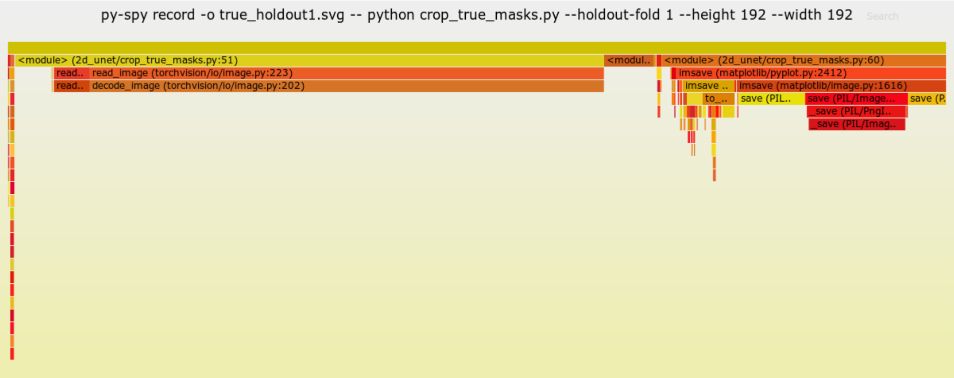
The first sub-question for this thesis is "How do the two approaches in localising before segmentation, Mask R-CNN and Yolov5 + 2D U-Net, compare against each other?". DetSeg performs better at segmentation than Mask R-CNN, although Mask R-CNN does have a similar localisation technique, an RPN. It seems to indicate that even though RPN helps improve detection performance, it has very little effect on improving segmentation performance. Another advantage of DetSeg is that even though the training of two models is required, it is much faster than training two independently. The downstream 2D U-Net did not train on the whole dataset but only a tiny portion. The training of Yolov5 for this thesis took 2.5 days, while 2D U-Net only took 3 hours. Thus, it is straightforward to fine-tune hyper-parameters to make them achieve even better results.

Despite performing well on segmentation, DetSeg has a drawback: the FN rate of Yolov5. U-Net can learn to correct Yolov5 mistakes, though it cannot do anything if they are not detected in the first place. The author proposes to solve this problem by integrating more data processing techniques. For example, bone segmentation can be proved to be useful (Moreau et al., 2020). Another proposal is to experiment with different architectures.

Even though the result is not satisfactory yet, DetSeg is very flexible. It is possible to experiment with various approaches and get better results. The customisation capability DetSeg offers is one of the most significant advantages. For example, it is possible to detect lesions on a 3D scale and then segment a cube around them with a 3D U-Net. Furthermore, either model in this thesis was not fully optimised. This thesis only aims to show that a baseline level DetSeg can outperform a one-stage approach with elaborate models and data transformation techniques. Unfortunately, due to time constraints, the author could not experiment with other object detection or segmentation architectures.

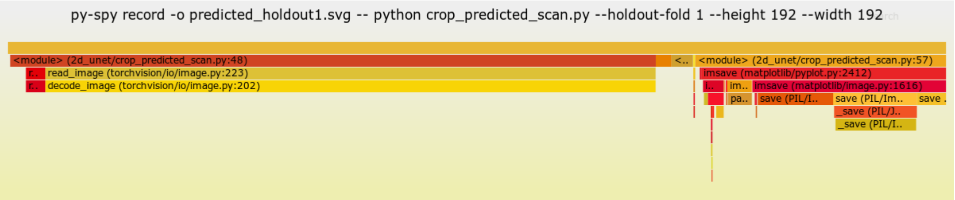
## 5.2 *Inference speed and Data Movement*

The second sub-question for this thesis is "To what extent can the DetSeg architecture inference speed fit into radiologists' workload?". Even though the inference speed of either model is fast, there remain bottlenecks in data movement. This is not an easy problem, so due to time limitations, no thorough analysis has been done yet. The first bottleneck is a transformation from Nifti scans into a format suitable for training, either arrays or images. The second bottleneck is cropping lesions predicted by Yolov5. Figure 12 and 13 illustrate this bottleneck. Whenever Yolov5 makes a prediction, the result needs to be moved from GPU memory to hard disk. Then it needs to be reread into memory for cropping. This pipeline involves two sets of system calls and data movement, which are very expensive. As Figure 13 shows, reading an image into memory alone accounts for a large part of CPU utilisation. A practical implementation needs to build a data pipeline that moves data inside GPUs while training/inference without moving back and forth. In conclusion, data bottlenecks can be resolved by careful engineering without affecting the general performance and working of DetSeg.



```
py-spy record -o true_holdout1.svg -- python crop_true_masks.py --holdout-fold 1 --height 192 --width 192
```

Figure 12: Flame Graph of cropping ground truth masks



```
py-spy record -o predicted_holdout1.svg -- python crop_predicted_scan.py --holdout-fold 1 --height 192 --width 192
```

Figure 13: Flame Graph of cropping predicted lesions

## 5.3 *Limitations*

### 5.3.1 *Dataset size*

The dataset size of the project is not significant, with only 96 CT scans. The number of lesions totals 2972 (673 if counted in 3D). Voxels which are lesions, are highly overwhelmed by non-lesion. There is also a high variance in three feature dimensions: lesion shapes, sizes, and slice distribution. Sometimes lesions can appear in only two or three axial slices of a scan. The homogeneous nature of lesions also makes it very hard to segment edge cases correctly. For example, we can see a "non-rounded" lesion in top left of Figure 10. Another abnormally large lesion can also be seen in the same figure. The current data augmentation setting treats all lesions the same. In future research, it is advised to divide lesions into different categories and augment under-represented categories.

### 5.3.2 *Annotation quality*

Radiologists annotated lesions at ETZ. It is possible that mistakes were made in some parts of the dataset. The author lacks the rigorous training to detect them. Hence, it is not possible at the moment to argue whether they affect performance or how many mistakes were there. More data cleaning needs to be done to correct mistakes. In future research, the author proposes to create a feedback loop where radiologists would work together with the system results and incrementally improve the dataset.

### 5.3.3 *Dataset Modality*

Training on other modalities may improve performance. At the moment, DetSeg trains on 2D axial slices. In real life, radiologists look at Coronal and Sagittal planes to get a different perspective of the scan. So, training on Coronal or Sagittal slices can offer us a different data distribution. Training directly on 3D scans is also a promising approach.

### 5.4 *Future Research*

More experimentation, a data pipeline, and better data are recommended for future research. Firstly, better data pre-processing techniques can be applied. For instance, data augmentation can be applied to sizeable and abnormal lesions. Different architectures should be tested to see which pairs perform better. An important metric is the object detector FN rate. Secondly, for DetSeg to fit into a radiologist workflow, a better data pipeline will be needed. The most important thing is data movement. In training mode, data needs to be moved back and forth because training data is usually large and does not fit into memory. However, data in inference mode should be discouraged from moving between hard disk and memory, preferably staying in memory as long as possible. Last but not least, better data will be required to achieve better performance. The first thing needed is a larger dataset. A primary source is natural growth as ETZ provided more data to the project. Collaboration between hospitals is also advised if possible. The second aspect of a better dataset is better annotation. Including radiologists in a feedback loop with the system is advised to improve the data gradually.

## 6 CONCLUSION

The primary question of this thesis is, "To what extent does localisation of potential lesions with object detection models prior to segmentation improve segmentation performance?". One obstacle in osteolytic lesion segmentation is the homogeneity nature of lesions. Lesions also account for a modest part of CT scans, leading to the unbalanced class problem. An intuition to solve this problem is to localise lesions before segmenting. The author proposes to do this by using Yolov5 to detect lesions and then segment them with 2D U-Net. The architecture is named DetSeg. The architecture was trained in a 5-fold cross-validation setting. Another model with a similar idea that was experimented with is Mask R-CNN, which conducts segmentation at the same time as detection. The result was also compared with previous work by (Hoff, 2021) on the same dataset but with different methodologies. The results show that DetSeg outperforms every segmentation metric. Yolov5 did not offer significantly better detection results than Mask R-CNN, though it can be offset with its superior inference speed. The inference speed of either model was also benchmarked to show that DetSeg fits into radiologists' workflow. Further research is required to benchmark data movement time, a key bottleneck.

This is the first study investigating the idea of localising potential lesions before segmentation on WBLDCT. WBLDCT is affordable while demands fewer technical facilities. Hence, it is worthwhile to explore research directions on this imaging technique. DetSeg is promising as no complicated data transformation techniques, or elaborate models have been applied yet. By experimenting with more techniques or models, the author believes the performance can be improved even further. This work is the first step in developing a real-time system that can accurately segment osteolytic lesions in MM diagnosis.

## REFERENCES

Abujudeh, H. H., Boland, G. W., Kaewlai, R., Rabiner, P., Halpern, E. F., Gazelle, G. S., & Thrall, J. H. (2010, 8). Abdominal and pelvic computed tomography (CT) interpretation: Discrepancy rates among experienced radiologists. *European Radiology*, *20*(8), 1952–1957. doi: 10.1007/s00330-010-1763-1

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020, 4). YOLOv4: Optimal Speed and Accuracy of Object Detection. Retrieved from http://arxiv.org/abs/2004.10934

Dimopoulos, M., Terpos, E., Comenzo, R. L., Tosi, P., Beksac, M., Sezer, O., ... Durie, B. G. (2009). *International myeloma working group consensus statement and guidelines regarding the current role of imaging techniques in the diagnosis and monitoring of multiple Myeloma* (Vol. 23) (No. 9). Nature Publishing Group. doi: 10.1038/leu.2009.89

Dong, H., Yang, G., Liu, F., Mo, Y., & Guo, Y. (2017, 5). Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks. Retrieved from http://arxiv.org/abs/1705.03820

Dutoit, J. C., & Verstraete, K. L. (2016). MRI in multiple myeloma: a pictorial review of diagnostic and post-treatment findings. *Insights into imaging*, *7*(4), 553–569.

Facebook AI Research (FAIR). (2022). *Detectron2.*

Filho, A. G., Carneiro, B. C., Pastore, D., Silva, I. P., Yamashita, S. R., Consolo, F. D., ... Nico, M. A. (2019, 7). Whole-body imaging of multiple myeloma: Diagnostic criteria. *Radiographics*, *39*(4), 1077–1097. doi: 10.1148/rg.2019180096

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., ... Zoph, B. (n.d.). Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. Retrieved from https://cocodataset.org/

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the ieee international conference on computer vision (iccv).*

Heider, U., Kaiser, M., Mieth, M., Lamottke, B., Rademacher, J., Jakob, C., . . . Sezer, O. (2009, 1). Serum concentrations of DKK-1 decrease in patients with multiple myeloma responding to anti-myeloma treatment. *European Journal of Haematology*, *82*(1), 31–38. doi: 10.1111/j.1600-0609.2008.01164.x

Hoff, W. (2021). *Automated segmentation of osteolytic lesions in whole-body CT imaging of multiple myeloma patients using deep learning models* (Tech. Rep.).

Horger, M., Claussen, C. D., Bross-Bach, U., Vonthein, R., Trabold, T., Heuschmid, M., & Pfannenberg, C. (2005). Whole-body low-dose multidetector row-CT in the diagnosis of multiple myeloma: An alternative to conventional radiography. *European Journal of Radiology*, *54*(2), 289–297. doi: 10.1016/j.ejrad.2004.04.015

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, *18*(2), 203–211.

Jha, A., Yang, H., Deng, R., Kapp, M. E., Fogo, A. B., & Huo, Y. (2021, 1). Instance segmentation for whole slide imaging: end-to-end or detect-then-segment. *Journal of Medical Imaging*, *8*(01). doi: 10.1117/1.jmi.8.1.014001

Jocher, G. (2020). *YOLOv5.* Retrieved from https://github.com/ultralytics/yolov5

Kazandjian, D. (2016, 12). *Multiple myeloma epidemiology and survival: A unique malignancy* (Vol. 43) (No. 6). W.B. Saunders. doi: 10.1053/j.seminoncol.2016.11.004

Lecun, Y., Bengio, Y., & Hinton, G. (2015, 5). *Deep learning* (Vol. 521) (No. 7553). Nature Publishing Group. doi: 10.1038/nature14539

Lee, J. W., Chung, H. Y., Ehrlich, L. A., Jelinek, D. F., Callander, N. S., Roodman, G. D., & Choi, S. J. (2004, 3). IL-3 expression by myeloma cells increases both osteoclast formation and growth of myeloma cells. *Blood*, *103*(6), 2308–2315. doi: 10.1182/blood-2003-06-1992

Lei, Y., Tian, S., He, X., Wang, T., Wang, B., Patel, P., . . . Yang, X. (2019, 7). Ultrasound prostate segmentation based on multidirectional deeply supervised V-Net. *Medical Physics*, *46*(7), 3194–3206. doi: 10.1002/mp.13577

Lin, W. J., Chen, J. W., Jhuang, J. P., Tsai, M. S., Hung, C. L., & Li, K. M. (2021, 12). Integrating object detection and image segmentation for detecting the tool wear area on stitched image. *Scientific Reports*, *11*(1). doi: 10.1038/s41598-021-97610-y

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg,

A. C. (2016). SSD: Single shot multibox detector. In *European conference on computer vision* (pp. 21–37).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 3431–3440).

Maskell, G. (2019). *Commentary error in radiology-where are we now?* (Tech. Rep.).

McDonald, R. J., Schwartz, K. M., Eckel, L. J., Diehn, F. E., Hunt, C. H., Bartholmai, B. J., ... Kallmes, D. F. (2015, 9). The Effects of Changes in Utilization and Technological Advancements ofCross-Sectional Imaging onRadiologist Workload. *Academic Radiology*, *22*(9), 1191–1198. doi: 10.1016/j.acra.2015.05.007

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016, 6). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. Retrieved from http://arxiv.org/abs/1606.04797

Moreau, N., Rousseau, C., Fourcade, C., Santini, G., Ferrer, L., Lacombe, M., ... Normand, N. (2020). *Deep learning approaches for bone and bone lesion segmentation on 18 FDG PET/CT imaging in the context of metastatic breast cancer*.* doi: 10.0/Linux-x86{_}64

*Multiple Myeloma: Statistics.* (2021). American Society of Clinical Oncology. Retrieved from https://www.cancer.net/cancer-types/multiple-myeloma/statistics

Nguyen, H. T., Pham, H. H., Nguyen, N. T., Nguyen, H. Q., Huynh, T. Q., Dao, M., & Vu, V. (2021). VinDr-SpineXR: A deep learning framework for spinal lesions detection and classification from radiographs. In *International conference on medical image computing and computer-assisted intervention* (pp. 291–301).

Qadir, H. A., Shin, Y., Solhusvik, J., Bergsland, J., Aabakken, L., & Balasingham, I. (2019). Polyp Detection and Segmentation using Mask R-CNN: Does a Deeper Feature Extractor CNN Always Perform Better? In *2019 13th international symposium on medical information and communication technology (ismict)* (pp. 1–6). doi: 10.1109/ISMICT.2019.8743694

*Radiation risk from medical imaging.* (2021, 9).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. Retrieved from http://pjreddie.com/yolo/

Redmon, J., & Farhadi, A. (n.d.). *YOLO9000: Better, Faster, Stronger* (Tech. Rep.). Retrieved from http://pjreddie.com/yolo9000/

Redmon, J., & Farhadi, A. (2018, 4). YOLOv3: An Incremental Improvement. Retrieved from http://arxiv.org/abs/1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 9351, pp. 234–241). Springer Verlag. doi: 10.1007/978-3-319-24574-4{_}28

Smith, L. N. (2015, 6). Cyclical Learning Rates for Training Neural Networks. Retrieved from http://arxiv.org/abs/1506.01186

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Traoré, A., Ly, A. O., & Akhloufi, M. A. (2020). Evaluating Deep Learning Algorithms in Pulmonary Nodule Detection*. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. doi: 10.0/Linux-x86{_}64

Xia, X., & Kulis, B. (2017, 11). W-Net: A Deep Model for Fully Unsupervised Image Segmentation. Retrieved from http://arxiv.org/abs/1711.08506

Xu, L., Tetteh, G., Lipkova, J., Zhao, Y., Li, H., Christ, P., ... Menze, B. H. (2018). Automated Whole-Body Bone Lesion Detection for Multiple Myeloma on 68 Ga-Pentixafor PET/CT Imaging Using Deep Learning Methods. *Contrast Media and Molecular Imaging*, *2018*. doi: 10.1155/2018/2391925

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2019, 11). Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. Retrieved from http://arxiv.org/abs/1911.08287

# APPENDIX A

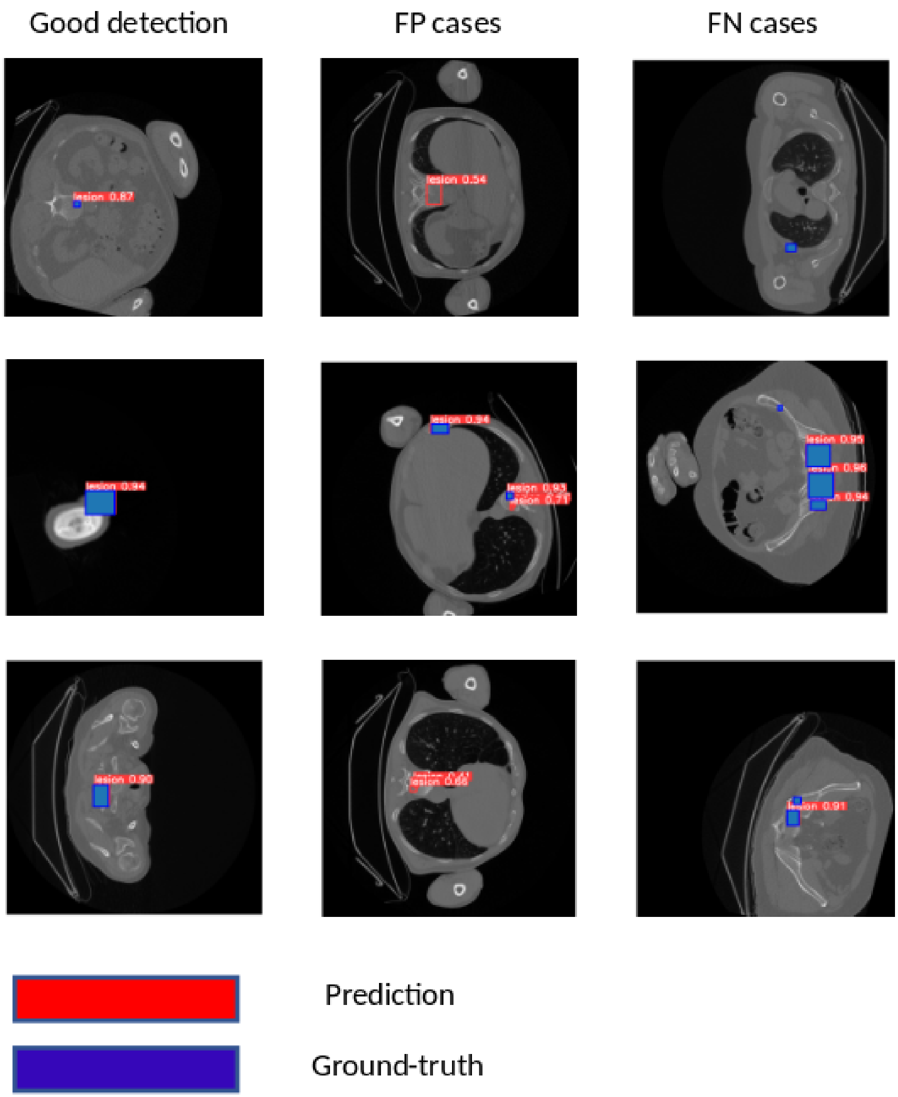


Figure 14: Visualization of lesion detection with Yolov5. The blue rectangles show the bounding-box of the ground truth. The red bounding boxes show the predicted results with the detection probability.

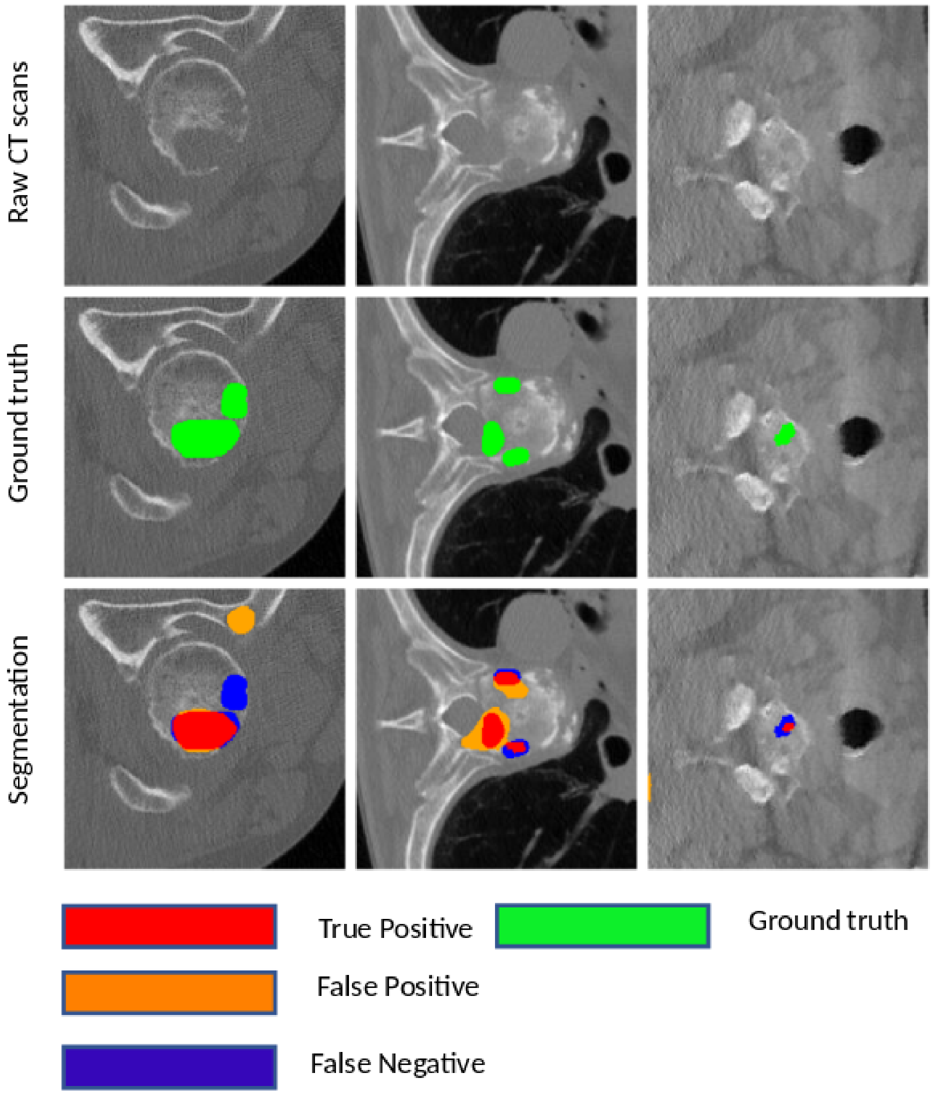


Figure 15: More example of DetSeg lesion segmentation