Hochschule Bonn-Rhein-Sieg

Master Thesis Proposal

Analysis of Object Detection models in CT scans

Ramit Sharma
Matrikel Number : 9030410

October 16, 2020

First Supervisor : TBD
Second Supervisor : Laura Anger
Third Supervisor : Ha Bach

Introduction

According to the iData Research's medical imaging procedures analysis, over 75 million CT scans are performed each year in the United States alone. This number is forecasted to grow to reach 84 million procedures by 2022 [18]. Analysis of CT scans for diagnosis of disease is a tedious task and requires a lot of human effort and working hours, and a small human error in the diagnosis could put the patient's life to risk. So to minimize this risk a lot of research is being done to perform automatic diagnosis of CT scans. In platforms like Kaggle we can find competitions like RSNA pneumonia detectiion challenge [32], COVID 19 CT scans [20] where they provide labeled data to solve the problem of automatic diagnosis of CT scans. The datasets like DeepLesion dataset [3], covid-19-chest-xray-lung-bounding-boxes-dataset [5] have been provided by the medical institutes to openly involve people to develop systems to perform object detection in CT scans. So, in this project we intend to look into the various datasets that are available for object detection in CT scans. We also intend to survey the various object detection models which could be used to perform object detection in CT scans. The object detection models could be broadly classified into two categories. They are:

- **One stage approach** : Unified one stage approach refer to architectures which directly predict class probabilities as well as bounding box offsets from images with single feed forward Convolutional Neural Network(CNN) in monolithic setting which does not involve generation of proposal region or post classification that encapsulates all computation using single network.
- **Two stage approach** : Two stage approaches are region-based frameworks. In case of two stage approach region proposals which are category independent are generated from an image. CNN features are then extracted from these regions. After that category specific classifiers are utilized to determine the label of the categories for the proposals.

We would implement two models.The first implementation would belong to the category of one stage model, and second implementation would be belong to the category of two stage approach. One of the factors that impacts the analysis of CT scans the most is the resolution of images, hence we intend to evaluate the performance of both the models at various resolution and find out which model performs better even at low resolution. This evaluation

would also help us to analyse how the performance of the object detection models get affected by varying resolution. We would also look into the frames per second attribute of the models at various resolution and analyse how the frames per second the model can predict changes when resolution is varied.

Related work

- **3D Object Detection from CT Scans using a Slice-and-fuse Approach**

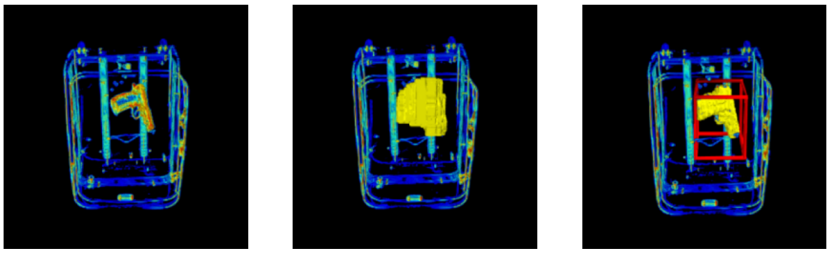
  3D X-ray Computed Tomography screening is used not only in medical imaging [15, 19] but also in baggage screening in case of airport security [2, 9, 7, 10, 8, 31]. CT scans have several favorable properties in comparison to other techniques of 3D scanning. Some of the favorable properties of CT scanning are listed below [30] :

  - It is capable of high resolution even at the sub-millimetre scale
  - It gives full 3D voxel representation which is occlusion free
  - It is non-intrusive

  The performance of object detection and segmentation techniques on 3D baggage CT scans can still be improved [8, 10]. In the case of medical imaging, the intra-class variability is quite less and the objects are mostly the same shape. Hence detecting objects in medical images is quite easy as compared to baggage CT scans as the variability of shapes is quite high. Another problem that is faced while detecting objects in baggage CT scans is that baggages are mostly cluttered with objects of the same shape such as keys, shoes, cell phones, etc.

  The size of baggage CT scan is very large. In each dimension it usually has hundreds of voxels. The data is of high resolution . If full 3D resolution is kept, it would be hard to take advantage of deep neural architectures because of the limited computational resources.

  Even if the storage and memory constraints are ignored, the time required to train such a deep model is quite high. If the input volume in spatial resolution of some 3D convolutional layers is largely downsampled, the small targets objects would be removed in subsequent feature extraction and is missed out in detection results.



(A) Input baggage CT (B) One-stage SliceNet (C) Two-stage SliceNet

Figure 1: Classification diagram

When the objects are large for eg. rifles that are long in just one dimension, then it should be trained with a large 3D reception field as they used 3D fully convolutional network. However, this would cause an insufficient memory problem. So designing a detection algorithm that has high detection accuracy, requires less training time, and has real-time speed becomes essential.

In this thesis, they proposed a new slice and fuse strategy which reduces the computational complexity for segmentation and object detection in 3D volumes of high resolution. In the slice and fuse approach first, the 3D volume is sliced into multiple 2D slices. Then segmentation and detection are performed in individual 2D slices and 2D predictions in 3D space are pooled.

As shown in figure 2 in the slicing stage the input volume is divided into 3D slices and each slice is projected into a 2D image. This slicing operation is repeated along XY, YZ, XZ directions and three sets of two-dimensional images are obtained. In the fusion stage, 3D volumetric predictions, one for each direction are reconstructed from 3 sets of 2D predictions. The final 3D prediction is obtained from the fusion of the selected two most confident predictions. This 3D prediction helps subsequent region proposals as well as classification functions.

Their strategy relies on two main observations. The first observation is that if the whole baggage CT scan is projected onto a single two-dimensional plane which causes severe occlusion among cluttered as well as targeted objects. A simple method to get rid of heavy occlusion.

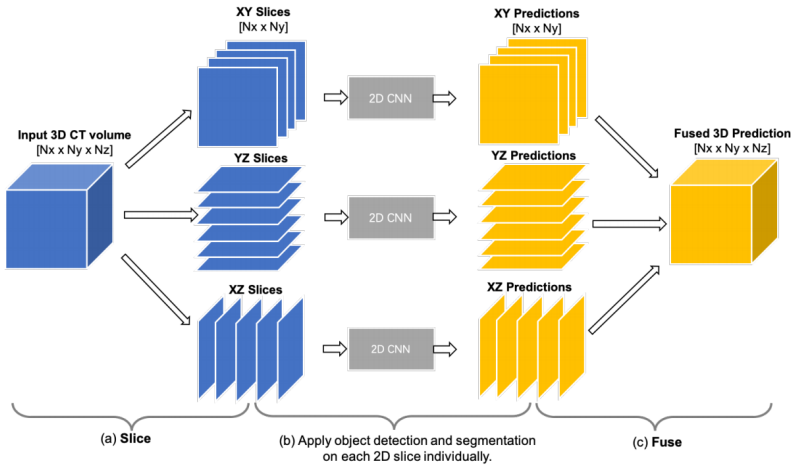


Figure 2: Slice and fuse diagram

The slicing method is quite suitable for the task of object detection since no matter how huge the input volume is, one receptive field is focused only at a time. This provides them the flexibility to divide the whole input into several slices and perform the task of detecting objects on each object. There exists an optimal, sub-optimal, and noninformative viewpoints given specific object categories [21].

A pistol could easily be recognized if the grip panel as well as the barrel can be seen in the projection. They fuse the 2 most confident predictions to get the voxel prediction. This helps to suppress the false prediction that is generated from a confusing viewpoint and guarantees the consistency of prediction among different viewpoints.

The effectiveness of the given method is verified using two 3D object detection methods referred to as SliceNets. The proposed strategy is incorporated into two state-of-the-art detection frameworks i.e. one stage strategy [27, 34, 25] and two-stage strategy [13, 12, 37, 24] . The result of these algorithms is shown in figure 1. In the case of both the algorithms input, volume is sliced and projected into 2D images.

In a single-stage object detector which is referred to as Retinal-SliceNet, each slice is utilized to predict the location of the bounding box as well as corresponding confidence scores. Linear fusion is used to obtain the

3D bounding boxes, two most confident predictions are used to get the final predictions.

U-Slicenet the two-stage object detector which is based on slice and fuse strategy is used only for the region proposal stage. 2D-UNet [20] is fed with input slice to get the pixel-level labeling, fusion operation is used to obtain the voxel level labeling.

The proposed method relies on the assumption that a 3D object could be classified using a 2D viewpoint. In case such as a triangular pyramid, this assumption does not hold true and hence it becomes impossible to detect if it is a square pyramid or a triangular prism, and hence the model fails.

They evaluated U-SliceNet on the IDSS 3D baggage CT dataset and for 3D semantic segmentation. In the case of the Real scan dataset, Retinal-SliceNet gives an accuracy of 95.26 percent and U-SliceNet gives an accuracy of 98.18 percent.

- **Mask R-CNN** In [16] they propose Mask R-CNN which is a framework for instance segmentation. The vision domain has seen significant improvement in semantic segmentation and object detection owing to the development of powerful baseline systems like Fast/faster RCNN [12, 37] and Fully Convolutional Network (FCN) [28] . These methods are not only conceptually intuitive but also offer robustness, flexibility along with fast training. The authors assert that their main motive behind this paper is to develop a framework for instance segmentation.

  Instance segmentation is a complex task as it requires the correct detection of the object along with the segmentation of each instance. Hence it combines the task of object detection in which is the aim is to classify individual objects as well as localize them using a bounding box, with semantic segmentation where the aim is to classify every pixel to a set of categories.

  This method is an enhanced version of Faster R-CNN [37], it adds a branch for the prediction of segmentation masks on the RoI (Region of Interest) in parallel with the branch of bounding box regression and classification. Mask R-CNN is easy to implement as well as train as it utilizes the Faster R-CNN framework that facilitates a multifarious range of architectural designs.

Even though Mask R-CNN is an extension of Faster R-CNN, but the construction of a proper mask branch is important for obtaining good results. Faster R-CNN is not implemented to handle pixel-to-pixel alignment that occurs between inputs as well as outputs. This can be seen in RoIPool [17, 12], the core operation used for attending to instances. In this operation coarse spatial quantization is performed for feature extraction. To fix this misalignment ROIAlign is proposed. ROIAlign is a quantization free layer that preserves the exact spatial locations.

ROIAlign improves the mask accuracy relatively by 10 to 50 percent and has bigger gains in case of stricter localization metrics.

The authors claim that Mask R-CNN is better than previous state-of-the-art techniques in COCO instance segmentation as well as object detection task.

- **Focal Loss for Dense Object Detection**

  Lin et al. [25] proposed Focal Loss for Dense Object Detection.

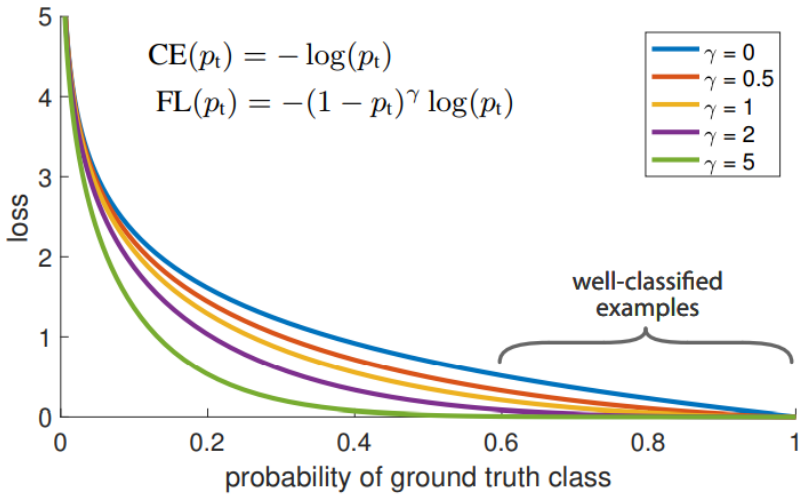


Figure 3: focal loss

Present state-of-the-art techniques for object detection rely on a two-stage proposal driven mechanism. The popular R-CNN framework [14] has two stages, the first stage creates a sparse set of object locations.

In the second stage, each location is classified as a foreground class or background using CNN (Convolutional Neural Network).

Even though the two-stage detectors have been very successful, a question that arises is that, can a simple one-stage object detector achieve similar accuracy? One stage detectors like YOLO [34, 36] and SSD [27, 11] are faster and have an accuracy 10 to 40 percent less than the state of the art techniques for two-stage object detection.

The authors claim that in this paper they propose a one stage detector that gives an equivalent performance on COCO AP compared to the complex two-stage detectors like Feature Pyramid Network(FPN) [24], Mask R-CNN [16] or other variants of Faster R-CNN [37].

In R-CNN detectors class imbalance is addressed using two-stage cascade and sample heuristics. The proposal stage (e.g. Selective search [41], RPN [37] narrows down the candidate object locations to quite a small number (e.g. 1-2k) by filtering out most of the background samples. In the second stage which is the classification stage, the sampling heuristics like foreground to background ratio (1:3) is performed to maintain a balance between foreground and background.

In the case of one-stage detectors, a large set of object locations (e.g. 100k) need to be sampled across the image. A similar sampling heuristic could also be applied in this case, but they do not prove to be sufficient as the training process is dominated the background examples which are easily classified. This problem can be seen in detection techniques like hard example mining [42, 6, 39] or bootstrapping [40, 38] .

Hence in this paper, they try to address this problem of class imbalance by introducing a new loss function. This loss function is a cross-entropy loss that is dynamically scaled, the scaling factor in this loss function decays to zero as confidence increases in the correct class as shown in figure 3 . A major advantage of this scaling factor is that it can automatically down-weight the impact of a large number of easy examples and help focus on the harder training examples. The authors claim that the proposed approach outperforms the single-stage detectors which rely on hard example mining or sampling heuristics. They also show that other instantiations of focal loss also achieve similar results.

- **KIDNEY RECOGNITION IN CT USING YOLOV3**

  In [23] Lemay et al propose to assess the performance of YOLOv3 [35] on Kidney localization in 3D and 2D from CT scans. They also assess the performance of SSD [27] to detect kidneys on various CT scans.

  Organ detection is used in various medical applications like planning surgeries or finding pathologies. It is important to add bounding boxes around scans of organs before performing other image processing techniques like segmentation [1, 22]. Whether it is laparoscopic surgeries [4] or adaptive radiotherapy [29], real-time organ tracking is quite useful. Recognizing kidneys is quite challenging if they consider the type of forms, positioning, textures, and contrasts that are found in CT scans figure 4.

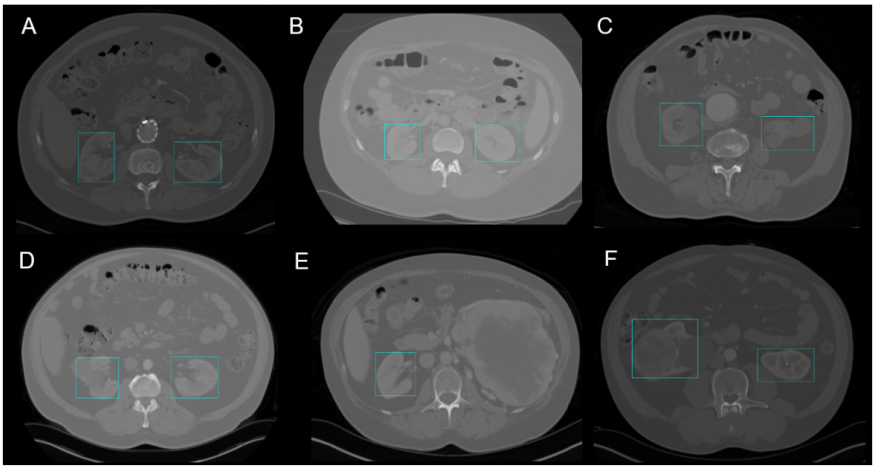


Figure 4: 2D kidney detection by YOLOv3. A-B: Normal kidneys with different CT scan mean intensities. C-D: Cystic kidneys. E: Failed detection of hypertrophied kidney. F: Tumoral kidney.

The authors assert that alongside with Faster R-CNN, SSD (Single Shot Detector), YOLO (You Only Look Once) has been proven to be state-of-the-art for robust object detection systems [22] . YOLO and SSD are real-time models in comparison to Faster R-CNN [43] . The use of models based on YOLO [35] was explored for the localization of organs in three dimensional PET scans [1], for lung cancer prevention [33], for lung nodules detection, and CT scans of nasal cavities [22]. If

the robustness, speed, and accuracy are considered, YOLO was selected for kidney detection. They also tested SSD for this task.

YOLO takes 2D images as input which is a challenge, as this model needs to be adapted for 3D medical images. So, they counter this challenge by taking every slice of CT scan as a single image as input. They use non-maximum suppression for 3D generalizations after the 2D bounding boxes are generated for every 2D image. This 3D bounding box is generated by grouping 2D boxes using threshold criteria which corresponds to intersection over union (IOU). They trained the model on 14 CT scans that generated 2911 2D images of dimension 512* 512 which had 1200 kidneys. The model was tested on 41 CT scans that generated 7451 2D images. They used histogram equalization to increase the contrast. The results obtained were compared with SSD which used MobileNet architecture to extract features. YOLOv3, as well as SSD both, are a one-step framework [8]. The two models tend to have the same accuracy. They also found that YOLOv3 is 3 times faster than SSD.

Problem Statement

There are many one stage object detection models available like SSD [26], YOLO [35] , focal loss for dense object detection [25]. Similarly many two stage detectors like Fast RCNN [12], Faster RCNN [37], Mask RCNN [16]. But the question which object detection model would perform better in low resolution CT scans is not answered yet. A research on how the resolution of image would affect the overall performance of the models and frames per second the model can predict is not done in the field of object detection. In our research project we would like to address these issues.

The steps we would follow to carry out the research work is shown below:

- To carry out the survey of the various object detection models
- To carry out the survey of various CT scan datasets for object detection
- To select one "one stage object detector" and one "two stage object detector" for object detection and select the corresponding dataset.
- To prepare various datasets at different resolutions
- To implement both the selected one stage as well as two stage object detector models
- To compare the performance of the models at different resolutions
- To analyse the impact of resolution in accuracy of the model and frames per second the model can predict
- If the time permits, we would also try to publish a paper in a journal

Expected Goals

- Minimum
  - To survey the various CT scan datasets available for object detection
  - To survey the various object detection models for CT scans
  - To select two approaches, one that belongs to single stage object detector category and the other that belongs to two stage object detector category
  - To implement the selected two stage object detector model
- Expected
  - To implement the selected one stage object detector model
  - To compare the performance of both models at different resolution and frames per second the model can predict at different resolution
  - To analyse the impact of resolution on the performance of models and frames per second the model can predict
  - To select the model which performs the best even at low resolution
- Maximum
  - To publish a paper in one of the journals

References

[1] Saeedeh Afshari, Aïcha BenTaieb, and Ghassan Hamarneh. Automatic localization of normal active organs in 3d pet scans. In *Computerized Medical Imaging and Graphics, 70:111–118*, 2018.

[2] Rushil Anirudh, Hyojin Kim, Jayaraman J Thiagarajan, K Aditya Mohan, Kyle Champley, and Timo Bremer. Lose the views: Limited angle ct reconstruction via implicit sinogram completion. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6343–6352*, 2018.

[3] NIH Clinical Center. `https://nihcc.app.box.com/v/DeepLesion/folder/50715173939`. Accessed on 22 October 2020.

[4] Toby Collins, Adrien Bartoli, Nicolas Bourdel, and Michel Canis. Robust, real-time, dense and deformable 3d organ tracking in laparoscopic videos. In *In Lecture Notes in Computer Science, pages 404–412. Springer*, 2016.

[5] covid-19-chest-xray-lung-bounding-boxes dataset. `https://github.com/GeneralBlockchain/covid-19-chest-xray-lung-bounding-boxes-dataset`. Accessed on 22 October 2020.

[6] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *In CVPR*, 2010.

[7] Greg Flitton, Toby P Breckon, and Najla Megherbi. A 3d extension to cortex like mechanisms for 3d object class recognition. In *In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 3634–3641. IEEE*, 2012.

[8] Greg Flitton, Toby P Breckon, and Najla Megherbi. A comparison of 3d interest point descriptors with application to airport baggage object detection in complex ct imagery. In *Pattern Recognition, 46(9):2420–2436*, 2013.

[9] Greg Flitton, Andre Mouton, and Toby P Breckon. Object classification in 3d baggage security computed tomography imagery using visual codebooks. In *Pattern Recognition, 48(8):2489–2499*, 2015.

[10] Gregory T Flitton, Toby P Breckon, and Najla Megherbi Bouallagu. Object recognition using 3d sift in complex ct volumes. In *In BMVC, number 1, pages 1–12*, 2010.

[11] C.-Y. Fu, A. Ranga W. Liu, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. In *arXiv:1701.06659*, 2016.

[12] Ross Girshick. Fast r-cnn. In *In Proceedings of the IEEE international conference on computer vision, pages 1440–1448*, 2015.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587*, 2014.

[14] ´R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *In CVPR*, 2014.

[15] Sardar Hamidian, Berkman Sahiner, Nicholas Petrick, and Aria Pezeshk. 3d convolutional neural network for automatic detection of lung nodules in chest ct. In *Computer-Aided Diagnosis, volume 10134, page 1013409*, 2017.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *arxiv.org*, 2018.

[17] ´K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *In ECCV*, 2014.

[18] iData Research. `https://idataresearch.com/over-75-million-ct-scans-are-performed-each-year-and-growing-despite-radiatic` Accessed on 22 October 2020.

[19] Kamal Jnawali, Mohammad R Arbabshirani, Navalgund Rao, and Alpen A Patel. 3d convolutional neural network for automatic detection of lung nodules in chest ct. In *Computer-Aided Diagnosis, volume 10575, page 105751C*, 2018.

[20] Kaggle. `https://www.kaggle.com/andrewmvd/covid19-ct-scans`. Accessed on 22 October 2020.

[21] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5010–5019*, 2018.

[22] Cristina Oyarzun Laura, Patrick Hofmann, Klaus Drechselr, and Stefan Wesarg. Automatic detection of the nasal cavities and paranasal sinuses using deep neural networks. In *In IEEE 16th International Symposium on Biomedical Imaging, pages 1154–1157. IEEE*, 2019.

[23] Andréanne Lemay. Kidney recognition in ct using yolov3. In *In Advances in neural information processing systems, pages 91–99*, 2019.

[24] Tsung-Yi Lin, Piotr Doll´ar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *In Advances in neural information processing systems, pages 91–99*, 2015.

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Doll´ar. Focal loss for dense object detection. In *In Proceedings of the IEEE international conference on computer vision, pages 2980–2988*, 2017.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. 2015.

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision, pages 21–37. Springer*, 2016.

[28] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *In CVPR*, 2015.

[29] Martin J Menten, Martin F Fast, Andreas Wetscherek, Christopher M Rank, Marc Kachelrieß, David J Collins, Simeon Nill, and Uwe Oelfke. The impact of 2d cine mr imaging parameters on automated tumor and organ localization for mr-guided real-time adaptive radiotherapy. In *Physics in Medicine Biology, 63(23):235005*, 2018.

[30] Andre Mouton. On artefact reduction, segmentation and classification of 3d computed tomography imagery in baggage security screening. 2014.

[31] Andre Mouton, Toby P Breckon, Greg T Flitton, and Najla Megherbi. 3d object classification in baggage computed tomography imagery using randomised clustering forests. In *In 2014 IEEE International Conference on Image Processing (ICIP), pages 5202–5206. IEEE*, 2014.

[32] Radiological Society of North America. `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge`. Accessed on 22 October 2020.

[33] Sindhu Ramachandran, Jose George, Shibon Skaria, and Varun V.V. Using yolo based deep learning network for real time detection and localization of lung nodules from low dose ct scans. In *Medical Imaging 2018: Computer-Aided Diagnosis, volume 10575, page 53*, 2018.

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788*, 2016.

[35] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. In *Technical report, University of Washington*, 2018.

[36] ´J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *In CVPR*, 2017.

[37] Shaoqing Ren, Ross Girshick Kaiming He, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *In Advances in neural information processing systems, pages 91–99*, 2015.

[38] H. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. In *Technical Report CMU-CS-95-158R, Carnegie Mellon University*, 1995.

[39] A. Shrivastava, A. Gupta, and R. Girshick. Training regionbased object detectors with online hard example mining. In *In CVPR*, 2016.

[40] K.-K. Sung and T. Poggio. Learning and example selection for object and pattern detection. In *In MIT A.I. Memo No. 1521*, 1994.

[41] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. In *IJCV*, 2013.

[42] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *In CVPR*, 2001.

[43] Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu. Object detection with deep learning: A review. In *IEEE Transactions on Neural Networks and Learning Systems, pages 1–21*, 2019.