Hochschule Bonn-Rhein-Sieg

Master Thesis Proposal

Object Detection in CT scans - Analysis of the Interaction of Data Resolution and Detection Results

Ramit Sharma
Matrikel Number : 9030410

October 16, 2020

First Supervisor : To be foundl
Second Supervisor :To be found
Third Supervisor :Laura Anger

Abstract

To be written

Introduction

Related work

- **VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection** :

  3D object detection based on point cloud can be used in many applications like household robots, autonomous navigation, augmented virtual reality. This project uses LiDAR based depth information to localize objects as well as characterize their shapes.The problem with LiDAR point clouds is that it is sparse and has quite variable point density owing to various factors like non-uniform sampling of three dimensional space, occlusion, relative pose etc.

  To encounter these challenges several approaches crafted feature representations manually for point clouds tuned for three dimensional object detection. Some approaches project point clouds in a perspective view and feature based extraction techniques are applied. Quite a few approaches like [18, 58] point clouds are rasterized into 3D voxel grid and each voxel is encoded with handcrafted features. A major drawback of these approaches is that it introduces an information bottleneck which hinders the ability of these models from exploiting the three dimensional shape information as w ell as invariances required for the detection task. In [40] Qi et al. introduced PointNet. PointNet is a deep neural net that uses point clouds to learn the point-wise features. In [39] Qi et al. introduced PointNet++, an enhanced version of PonintNet. The major advantage of PointNet++ was that it could learn from local structures at various scales.The LiDARs have around 100k points and training on such huge number of points results in high computational complexity and heavy memory requirements.

  To overcome this drwaback they present Voxelnet, a 3D detection framework which utilizes point clouds to learn a discriminative feature representation and 3D bounding boxes are predicted in end-to-end fashion as shown in figure 1 .

  . They came up with a new Voxel Feature Encoding (VFE) layer, that combines the point-wise features using a locally aggregated feature and helps to enable inter-point interaction in a voxel. The local three dimensional shape inforamtion are characterized complex features learnt by stacked multiple VFE layers. The voxelnet transforms the point

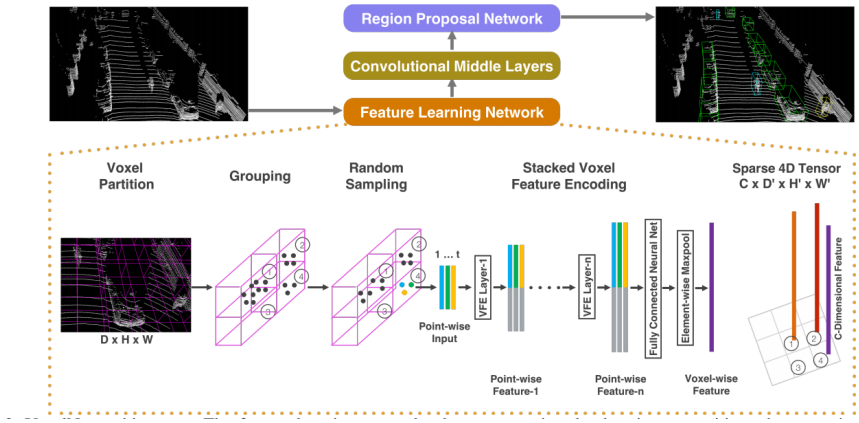


Figure 1: Voxelnet architecture [65]

cloud to a higher dimensional volumetric representation. The transformation of point clouds to higher dimension is achieved by dividing the point cloud in equally spaced three dimensional voxels , encoding each voxel by stacked VFE layer and using 3D convolution to aggregate the local voxel features, these voxel features are nothing but point cloud transformed to higher dimension. The volumetric representation is used by RPN(Region Proposal Network) which gives the detection result. The major advantage of voxelnet is that it takes advantage from both sparse points structure as well as parallel processing on voxel grid.

- **Deep Hough Voting for 3D Object Detection in Point Clouds**:

  In [41] , Qi et al. proposed Votenet a 3D detection framework which is based on point cloud, it processes the raw point cloud directly and does not rely on any 2D detectors , this model is based on generalized Hough voting process used for object detection.

  The idea of 3D object detection is to recognize and localize objects in three dimensional scene. In this work they aim to estimate semantic classes as well as 3D bounding boxes in the objects of point clouds. 3D point clouds have better robustness and accurate geometry to illumination changes as compared to 2d images, but point clouds are irregular and typical CNNs cannot process them directly. The current 3D detection techniques thus rely on 2D based detectors in various aspects, so

that they don't have to process the irregular point clouds. For example [54, 21] improve 2D detection frameworks like Mask/Faster R-CNN [19, 45] to 3D, the irregular point clouds are voxelized to regular 3D girds and feed it to 3D CNN detectors. The 3D CNN detectors cannot learn due to the sparsity in the 3D point clouds and also has high computational cost because of the 3D convolutions.

In [5, 65] they localize objects using 2D detectors in 2D bird's eye view images. The geometric details which are very important in cluttered indoor environments get sacrificed. In [26, 42] cascaded two-step pipeline has been proposed by first detecting objects which are in front view of the image and then objects in the frustum point clouds are localized by 2D boxes, the architecture of these models is entirely dependent on 2D.

So, the central idea in this project is to develop a 3D detection framework that directly processes the 3D point cloud and does not rely on 2D detectors in any way.

They use Pointnet++ [39] for learning from the point clouds. The point clouds get processed directly which avoids information loss through quantization process and also takes advantage of sparsity of point clouds by limiting the computation to sensed points.

PointNet++ has been very successful in object classification and semantic segmentation, but only few research study to detect 3D object in point clouds encompassing such architectures.

They propose point cloud deep networks and has voting mechanism similar to traditional Hough voting. Voting helps to generate new points that are nearer to the object centers and are grouped and aggregated to get the box proposals. As opposed to the classical Hough voting that has multiple separate modules which are hard to optimize together, Votenet is optimizable. Vote clusters Vote clusters are generated near object centers and are aggregated using a learned module to create box proposals.

They evaluate their approach on two 3D object detection datasets ScanNet and SUN RGB-D. Votenet performs better than the prior architectrures that use RGB as well as geometry in both the datasets . Votenet performs better when object centers are far away from the object surface.

- **Frustum PointNets for 3D Object Detection from RGB-D Data**:

  In [42] they propose a method that leverages advanced 3D deep learning techniques for object localization as well as mature 2D object detectors and achieves efficiencies and high recall for small objects as well.

  As 3D sensors are becoming popular whether it is in mobile devices or in autonomous vehicles, we have more 3D data that needs to be captured and processed. In this work they focus on 3D object detection that classifies the category of the object and estimates 3D bounding box around the object by processing the 3D sensor data.

  As 3D sensor data is mostly in the form of 3D point clouds, how the 3D point cloud can be represented and what sort of deep net architecture could be used for 3D detection is still an open problem. Most of the existing architectures, 3D point clouds are converted to images either by projection [55, 38] or are converted to volumetric grids through quantization [60, 35, 38] and convolutional networks are applied.

  This transformation in representation of data obscures natural 3D patterns as well as invariances of the data. Many papers recently have proposed to process 3D poindt clouds directly escaping the part of converting the data to any other formats. For example [40, 39] propose novel type of deep net architectures which are referred to as PointNets that have shown superior performance in several 3D understanding tasks like semantic segmentation and object classification.

  Even though PointNets classify 3D point cloud or predict semantic class of each point in 3D point cloud, it is not known how this architecture could be used for 3D object detection. One of the challenges addressed by this paper is to efficiently propose location of 3D objects present in a 3D space. Most of the works imitate the practice of image detection, i.e. enumerate 3D bounding boxes using sliding boxes [7] or 3D region proposal networks [53] . The major drawback of these approaches is that computational complexity in case of 3D search increases cubically with respect to resolution as it becomes very expensive for large scenes as well as for real time applications like autonomous driving.

  So in this work they try to reduce the search space using the dimension reduction principle by taking the advantage of 2D object detectors. In order to do so they extract the 3D bounding frustum for an object by 2D

bounding boxes of image detectors. They perform 3D object inastance segmentation as well as 3D bounding box regression consecutively using the variants of PointNet. The 3D mask for the object of interest is predicted by the segmentation network and amodal 3D bounding box is estimated by the regression network.

In bird's eye view detection [1] and KITTI 3D object detection [3], this approach achieves leading positions. In comparison to previous state of the art [4] , this method is 8.04 percent better on 3D car AP. This method is also useful for indoor RGB data where they have achieved 6.4 percent and 8.9 percent better 3D mAP than [47] and [25] on SUN-RGBD when it was run one to three times faster in orders of magnitude.

- **PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection**:

  In [49] they present a novel 3D object detection framework referred to as PointVoxel-RCNN(PV-RCNN), for detecting 3D object in point clouds.

  3D object detection has many applications such as robotics and autonomojus driving. Autonmous driving vechicles as well as robots rely on LiDAR sensors for getting 3D scene information as point clouds which are sparse and irregular. These point clouds are vital for 3D scene understanding. The crux behind this paper is to design new point-voxel integrated networks so as to learn 3D features from point clouds.

  Most of the existing 3D object detection techniques can be classified into two major categories with respect to point cloud representations, they are point based methods and grid-based methods.

  The grid-based methods transform point clouds which are irregular to regular representations. The regular representation could be either 3D voxels [54, 65, 61, 6, 51] or 2D bird view maps [5, 24, 63, 30, 27, 62, 29] , which are processed by 2D or 3D Convolutional Neural Networks(CNN) which are used to learn features for 3D detection.

  PointNet and its variants [40, 39] as well as the pointbased methods [42, 50, 59, 64] extract discriminative features from point clouds to perform 3D object detection. Usually the grid-based techniques have

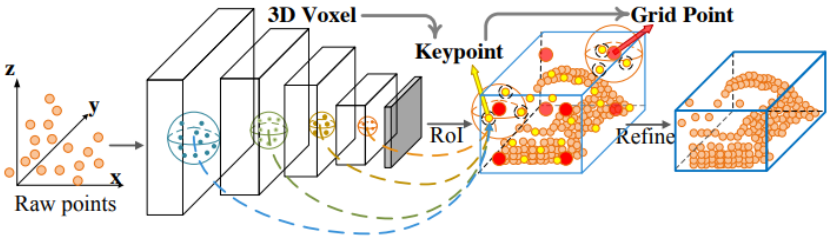


Figure 2: PVRCNN architecture [49]

less computational complexity. A major drawback of grid-based techniques is that inevitable information is lost but the localization accuracy is degraded. As compared to grid-based techniques, the point based techniques have higher computational complexity but achieves bigger receptive field by abstraction of the point set [39].

In this work they show that a unified framework can integrate the best of both the types of methods and can become better than prior state of the art techniques for 3D object detection.

They propose a new 3D object detection framework referred to as PVRCNN as shown in figure 2 that has advantages of both Voxel-based as well as Point-based learning methods. The core idea of PVRCNN is that the voxel-based operation encodes manifold scale feature representation and generates better 3D proposals, on the other hand PointNet-based architecture offers the facility of set abstraction operation which helps to preserve the accurate location information along with flexible rececptive fields. The authors assert that that integration of these two types of architectures facilitates to learn more discriminative features which in turn provides higher accuracy in fine-grained box refinement.

The biggest challenge is to combine both the feature learning schemes i.e. 3D voxel Convolutional Networks with sparse convolutions [6, 5] along with the PointNet-based set abstraction [39] . A solution to this problem would be to uniformly sample different grid points in each 3D proposal and use the set abstractions to get 3D voxel features that surround these grid points with a vow of proposal refinement. However this approach has more memory consumption since both number of

grid points and number of voxels is possibly quite large to achieve satisfactory performance.

So to integrate both the point cloud feature learning architectures , they came up with a two-step process, first step is voxel-to-keypoint encoding of the scene and the second step is keypoint-to-grid region of interest(ROI) feature abstraction process. A voxel Convolutional Neural Network(CNN) in addition to 3D sparse convolution is used to get voxel-wise feature learning as well as accurate proposal generation.

To get rid of the above mentioned problem of requiring a lot of voxels for encoding the entire scene, a set of keypoints are extracted using Furthest Point Sampling(FPS) which summarized the overall 3D information using the voxel-wise features. The features of every keypoint is accumulated by grouping the neighborhood voxel-wise features utilizing the PointNet based set abstraction to get the point cloud information which is mostly multiscale. The overall scene could be efficiently as well as effectively encoded using a small number of keypoints which are associated with multi-scale features.

In case of keypoint-to-grid Region of Interest(RoI) feature abstratction step, for each box proposal along with the grid point locations, a Region of Interest(RoI) grid pooling module is used, here keypoint set abstraction layer that has multiple radii is taken into consideration for each grid point so as to accumulate the features using the keypoints that has multi-scale context. All of the grid points features are then jointly used for the proposal refinement. PVRCNN hence takes advantage of both voxel-based and point-based networks so as to encode discriminative features in every box proposal to give accurate confidene prediction as well as fine-grained box refinement.

The authors assert that PV-RCNN performs better than all previous techniques with remarkable ranks and margins. This model ranked 1st on KITTI 3D detection benchmark [10] and also outperforms other architectures on large-scale Waymo Open dataset.

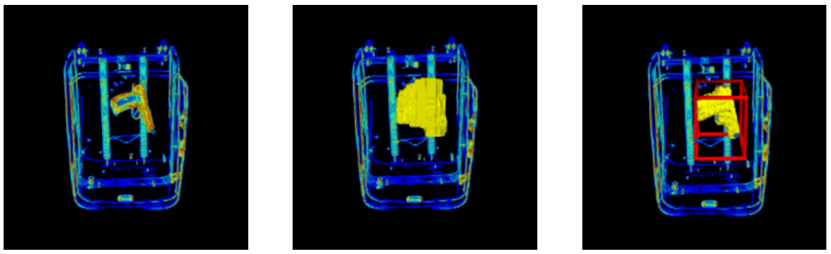
3D Object Detection from CT Scans using a Slice-and-fuse Approach

3D X-ray Computed Tomography screening is used not only in medical imaging [17, 22] but also in baggage screening in case of airport security [2, 11, 9, 12, 10, 37]. CT scans has several favorable properties in

comparison to other techniques of 3D scanning. Some of the favorable properties of CT scanning are listed below [36] :

- It is capable of high resolution even at 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 [10, 12]. In case of medical imaging the intra-class variability is quite less and the objects are mostly same shape. Hence detecting objects in medical images is quite easy as compared to baggage CT scans as 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 same shape such as keys,shoes, cell phones etc.



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

Figure 3: Classification diagram

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 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.

When the objects are large for eg. rifles which are long in just one dimension, then it should be trained with large 3D reception field as they used 3D fully convolutional network. However this would cause 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 slice and fuse approach first the 3D volume is sliced into multiple 2D slices. Then segmentation and detection is performed in individual 2D slices and 2D predictions in 3D space is pooled.

As shown in figure 4 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 is obtained. In 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.

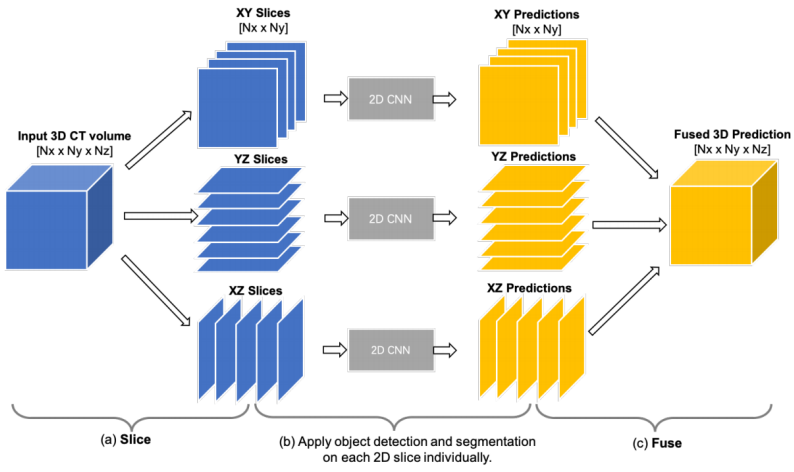


Figure 4: Slice and fuse diagram

Their strategy relies on two main observations. The first observation is

that if whole baggage CT scan is projected onto single two dimensional plane which causes severe occlusion among cluttered as well as targeted objects. A simple method to get rid of heavy occlusion .

The slicing method is quite suitable for task of object detection since no matter how huge the inpput volume is, one receptive field is focused only at a time. This provides them the flexibility to divide 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 [23].

A pistol could easily be recognized if grip panel as well as 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 are generated from 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 [33, 43, 32] and two stage strategy [15, 14, 46, 31] . The result of these algorithms is shown by the figure 3. In case of both the algorithms input volume is sliced and projected into 2D images.

In single stage object detector which is referred to as Retinal-SliceNet, each slice is utilized to predict the location of 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 3D object could be classified using 2D viewpoint. In case such as 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 IDSS 3D baggage CT dataset and for 3D semantic segmentation. In case of Real scan dataset Retinal-SliceNet

gives an accuracy of 95.26 percent and U-SliceNet gives an accuracy of 98.18 percent.

KIDNEY RECOGNITION IN CT USING YOLOV3 In [28]

Mask R-CNN In [19] 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 [14, 46] and Fully Convolutional Network (FCN) [34] . 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 correct detection of object along with 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 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 [46], 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 utiilizes the Faster R-CNN framework that facilitates multifarious range of architecture designs.

Even though Mask R-CNN is an extension of Faster R-CNN, but the construction of 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 [20, 14] , 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 quantization free layer which 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. [32] proposed Focal Loss for Dense Object Detection.

Present state-of-the-art techniques for object detection rely on two stage proposal driven mechanism. Popular R-CNN framework [16] has two stages, first stage creates a sparse set of object locations. In second stage each location is classified as 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 [43, 44] and SSD [33, 13] are faster and has 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 equivalent performance on COCO AP compared to the complex two stage detectors like Feature Pyramid Network(FPN) [31] , Mask R-CNN [19] or other variants of Faster R-CNN [46].

In R-CNN detectors class imbalance is addressed using two-stage cascade and sample heuristics. The proposal stage (e.g. Selective search [**?**] , RPN [46] narrows down the candidate object locations to quite small number (for e.g. 1-2k) by filtering out most of the background samples. In second stage which is classification stage the sampling heuristics like foreground to background ration(1:3) is performed to maintain balance between foreground and background.

In case of one stage detectors a large set of object locations (for 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 [57, 8, 52] bootstrapping [56, 48] .

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 which is dynamically scaled, the scaling factor in this loss function decays to zero as confidence increases on the correct class as shown in figure 1. A major advantage of this scaling factor is that

it can automatically down-weight the impact of 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 achieves similar results.

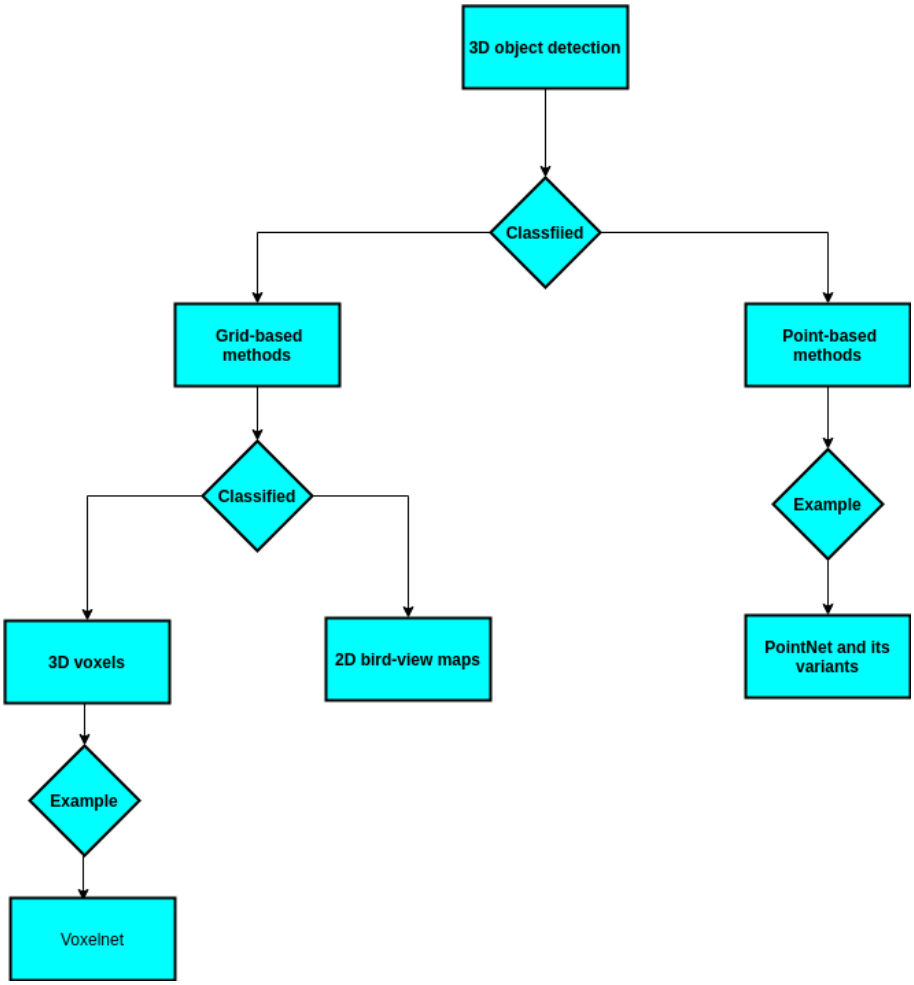


Figure 5: Classification diagram

Problem Statement

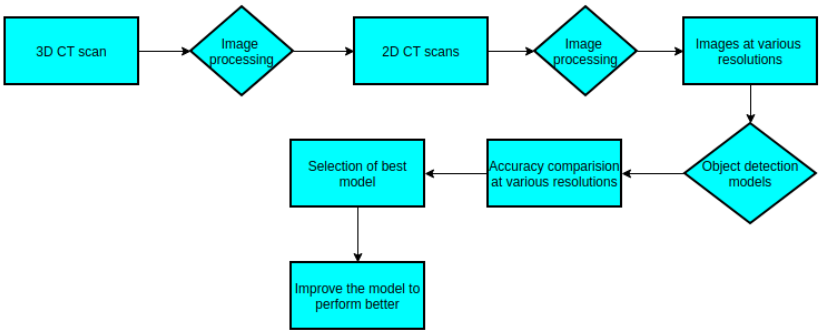


Figure 6: Classification diagram

- 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 2 models for object detection and select the corresponding dataset.
- To take the 3D dataset and convert it to 2D images
- To prepare various datasets at different resolutions
- To compare the accuracies of the models at different reolutions and select the best model
- To make some more changes to the model and try to improve the model
- To analyse the impact of resolution in accuracies of the model and frames per second the model can predict.

Expected Goals

- Minimum
  - To survey the various localization and classification networks on volume data.
  - To select two approaches
- Expected
  - Implementation of both the approaches
  - Comparison of accuracies of the model at different resolution
- Maximum
  - To improve the model's performance

Datasets

- **kits 19 Dataset (3D images, but for segmentation, bounding boxes not available)**
  - https://github.com/neheller/kits19
  - https://kits19.grand-challenge.org/data/
  - number of images = 300
- **RSNA pneumonia detection challenge (2D images)**
  - number of images = 26684
  - https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview
- **Covid 19 chest x-ray dataset(2D images)**
  - https://github.com/GeneralBlockchain/covid-19-chest-xray-lung-bounding-boxes-dataset
  - number of images = 1230
- **Covid Chest X-ray (2D images)**
  - https://github.com/ieee8023/covid-chestxray-dataset
  - number of images = 334
- **LIDC-IDRI (Image segmantation dataset used for competetion luna 2016)**

    - https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI
- **NIH Clinical Center -Lesion Dataset (3D images)**
    - https://nihcc.app.box.com/v/DeepLesion/folder/50715173939
    - number of images : 32000

References

[1] Kitti 3d object detection benchmark leader board. http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=3. Accessed on 01 September 2020.

[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] Kitti bird's eye view object detection benchmark leader board. http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=bev. Accessed on 01 September 2020.

[4] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. 2017.

[5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. 2017.

[6] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. 2019.

[7] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. 2017.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

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

[17] 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.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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

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

[21] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. 2018.

[22] 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.

[23] 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.

[24] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. Joint 3d proposal generation and object detection from view aggregation. 2018.

[25] J. Lahoud and B. Ghanem. 2d-driven 3d object detection in rgb-d images. 2017.

[26] Jean Lahoud and Bernard Ghanem. 2d-driven 3d object detection in rgb-d images. 2017.

[27] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019.

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

[29] Ming Liang*, Bin Yang*, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *In CVPR*, 2019.

[30] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *ECCV*, 2018.

[31] Tsung-Yi Lin, Piotr Dollar, 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.

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

[33] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Sin-

gle shot multibox detector. In *European conference on computer vision, pages 21–37. Springer*, 2016.

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

[35] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. 2015.

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

[37] 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.

[38] C. R. Qi, M. Nießner H. Su, M. Yan A. Dai, and L. Guibas. Volumetric and multi-view cnns for object classification on 3d data. 2016.

[39] C. R. Qi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. 2017.

[40] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. 2017.

[41] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. 2019.

[42] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, , and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgbd data. 2018.

[43] 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.

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

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2015.

[46] 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.

[47] Z. Ren and E. B. Sudderth. Three-dimensional object detection and layout prediction using clouds of oriented gradients. 2016.

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

[49] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. 2019.

[50] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn. 3d object proposal generation and detection from point cloud. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–779*, 2019.

[51] Shaoshuai Shi, Zhe Wang, Xiaogang Wang, and Hongsheng Li. Part-a2net: 3d part-aware and aggregation neural network for object detection from point cloud. 2019.

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

[53] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. 2015.

[54] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. 2016.

[55] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. 2015.

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

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

[58] Dominic Zeng Wang and Ingmar Posner. Voting for voting in online point cloud object detection. 2015.

[59] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In *In IROS. IEEE*, 2019.

[60] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. 2015.

[61] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. 2018.

[62] Bin Yang, Ming Liang, and Raquel Urtasun. Hdnet: Exploiting hd maps for 3d object detection. In *In 2nd Conference on Robot Learning (CoRL)*, 2018.

[63] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Realtime 3d object detection from point clouds. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7652–7660*, 2018.

[64] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: sparse-to-dense 3d object detector for point cloud. In *ICCV*, 2019.

[65] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. 2017.