2019

Collaborative Artificial Intelligence Algorithms for Medical Imaging Applications

Naji Khosravan
University of Central Florida

COLLABORATIVE ARTIFICIAL INTELLIGENCE ALGORITHMS FOR MEDICAL IMAGING APPLICATIONS

by

NAJI KHOSRAVAN
B.S. Amirkabir University of Technology, 2015

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2019

Major Professor: Ulas Bagci

ABSTRACT

In this dissertation, we propose novel machine learning algorithms for high-risk medical imaging applications. Specifically, we tackle current challenges in radiology screening process and introduce cutting-edge methods for image-based diagnosis, detection and segmentation. We incorporate expert knowledge through eye-tracking, making the whole process human-centered. This dissertation contributes to machine learning, computer vision, and medical imaging research by: 1) introducing a mathematical formulation of radiologists level of attention, and sparsifying their gaze data for a better extraction and comparison of search patterns. 2) proposing novel, local and global, image analysis algorithms. Imaging based diagnosis and pattern analysis are "high-risk" Artificial Intelligence applications. A standard radiology screening procedure includes detection, diagnosis and measurement (often done with segmentation) of abnormalities. We hypothesize that having a true collaboration is essential for a better control mechanism, in such applications. In this regard, we propose to form a collaboration medium between radiologists and machine learning algorithms through eye-tracking. Further, we build a generic platform consisting of novel machine learning algorithms for each of these tasks. Our collaborative algorithm utilizes eye tracking and includes an attention model and gaze-pattern analysis, based on data clustering and graph sparsification. Then, we present a semi-supervised multi-task network for local analysis of image in radiologists' ROIs, extracted in the previous step. To address missing tumors and analyze regions that are completely missed by radiologists during screening, we introduce a detection framework, *S4ND*: Single Shot Single Scale Lung Nodule Detection. Our proposed detection algorithm is specifically designed to handle tiny abnormalities in lungs, which are easy to miss by radiologists. Finally, we introduce a novel projective adversarial framework, *PAN*: Projective Adversarial Network for Medical Image Segmentation, for segmenting complex 3D structures/organs, which can be beneficial in the screening process by guiding radiologists search areas through segmentation

of desired structure/organ.

EXTENDED ABSTRACT

Radiology screening is proved to be a vital step for cancer detection in many applications. However, human errors stay as a significant issue in this process. Missing cases and over-diagnosis can have serious outcomes and increase mortality rate. Computer aided diagnosis (CAD) tools help radiologists to reduce diagnostic errors such as missing tumors and misdiagnosis. In this dissertation, we aim to develop a paradigm shifting CAD system, called collaborative CAD (C-CAD), that unifies CAD and eye-tracking systems in realistic radiology room settings. We propose a novel graph based analysis as our collaboration medium between the radiologist and our machine learning algorithms for medical image analysis.

We first developed an *eye-tracking* interface providing radiologists with a real radiology reading room experience. Further, we develop a graph based clustering and sparsification algorithm to transform eye-tracking data (gaze) into a graph model to interpret gaze patterns quantitatively and qualitatively. This algorithm will be used as a bridge between the radiologist and our machine learning algorithms.

Second, we develop a *local* image analysis algorithm. Once we extracted radiologists' ROIs using our graph formulation we incorporate our deep learning algorithm to locally analyze radiologists ROIs. We first show this process with a pilot study. Then, we develop a semi-supervised multi-task network to perform segmentation and diagnosis of abnormalities in the ROIs jointly and at the same time. The specific design of our algorithm, in this step, targets two critical challenges in medical image analysis: generalization and lack of large scale annotated data for training.

Finally, we introduce two *global* image analysis modules. The global image analysis modules will help for a better screening by handling the areas that are totally missed by radiologists during the screening. The goals of global modules are: 1) Capturing tiny abnormalities that can be

missed during the screening process, and 2) performing structure/organ segmentation to better guide the radiologists for high risk areas in case of abnormalities in organs with complex shape. Our first global module is *S4ND*: Single-Shot Single-Scale Lung Nodule Detection. This module is designed specifically to capture tiny abnormalities in the lung, which can be easily missed by radiologist during the screening process. Our second global module is *PAN*: Projective Adversarial Network for Medical Image Segmentation. *PAN* is a novel framework which captures 3D shape information through 2D projections and incorporates that in the segmentation using an adversarial learning process. Segmentation of 3D complex organs can be a vital first step to limit and guide their search space to more high-risk regions in medical images.

To my beloved parents,

whom I owe everything I am to,

my lovely sister,

and my beautiful girlfriend.

ACKNOWLEDGMENTS

I would like to thank my advisor, Prof. Ulas Bagci for his great support and advices throughout this journey. He taught me how to always maintain the highest expectations of myself. Also, I would like to thank Prof. Mubarak Shah, Prof. Gita Sukthankar, Prof. Abhijit Mahalanobis, and Prof. Mark Neider for accepting to be a part of my committee, and their helpful guidance throughout the process of proposing and defending my dissertation. I would also like to thank Dr. Amir Roshan Zamir, Dr. Afshin Dehghan, Dr. Subhabrata Bhattacharya, and Dr. Enrique Ortiz, whom despite very little or no overlap during my PhD in the lab I had the chance to work with, get to know and become friend with. Finally, I would like to thank all of the past and present members of the Center for Research in Computer Vision (CRCV), Tonya LaPrarie, Dr. Shervin Ardeshir, Dr. Nasim Souly, Shayan Modiri, Amir Mazaheri, Aidean Sharghi, Aliasghar Mortazi, Dr. Gonzalo Vaca, Dr. Khurram Soomro, Dr. Sarfaraz Hussein, Aisha Orooj Khan, Krishna Regmi, Dr. Dong Zhang, Dr. Yonatan Tariku, and Rodney Lalonde for their support, the good times, and the great memories.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

It has been long known that computer algorithms play a vital role in medical imaging applications, particularly in radiology field. There has been many machine learning and computer vision based algorithms developed to help radiologists for making their decision making process less hurdle, more efficient, and precise. However, many of such algorithms are, unfortunately, not adopted widely in clinical settings. This is due to the high-risk nature of such applications, requiring such computer algorithms to be extremely accurate, efficient and obey certain constraints of clinical routines. On the other hand, radiology errors are quite often happening. It has been shown that imaging screening in radiology has a huge impact on the mortality rate and successful diagnosis of several diseases. Lung cancer screening with low dose computed tomography (CT), for instance, was shown to reduce lung cancer mortality by 20% [1]. Yet, human error remains a significant problem to detect abnormalities. For instance, missing a tumor (*recognition error*) and misdiagnosing (*decision making error*) are perceptual errors [2]. It's reported that 35% of lung nodules are typically missed during the screening process [3]. Over-diagnosis is another significant perceptual bias leading to unnecessary treatment which can cause harm and unnecessary medical expenses.

Computer Science: To alleviate some of these errors, CAD systems have been developed [4, 5, 6]. CADs are often known as second opinion tools and they help to reduce false negative findings (i.e., missing tumors by radiologists). CADs also have serious limitations such as a large number of false positive findings and high execution times. Radiologists are expected to eliminate false positive findings generated by the CAD systems, which makes majority of CADs infeasible in routine practice.

Vision Science: Vision scientists have focused on exploring human errors in screening for more than three decades [2, 7, 8, 9, 10, 11, 12, 13, 14, 15]. One way to explore these perceptual errors is to use eye-tracking technology. It provides information about image interpretation by modeling perceptual and cognitive processes of radiologists.

In this dissertation, we introduce a paradigm-shift system that uses eye-tracking technology as a collaborative tool between radiologists and CAD systems. The general overview of proposed framework is illustrated in Figure 1.1. The rationale behind this idea comes from the fact that radiologists are good at eliminating false positives, which CADs often fail to achieve at a human level performance. On the other hand, CADs capture missing tumors better than the human observer. Because of this complementary properties, we call the proposed technology collaborative CAD (C-CAD), which puts human in center.

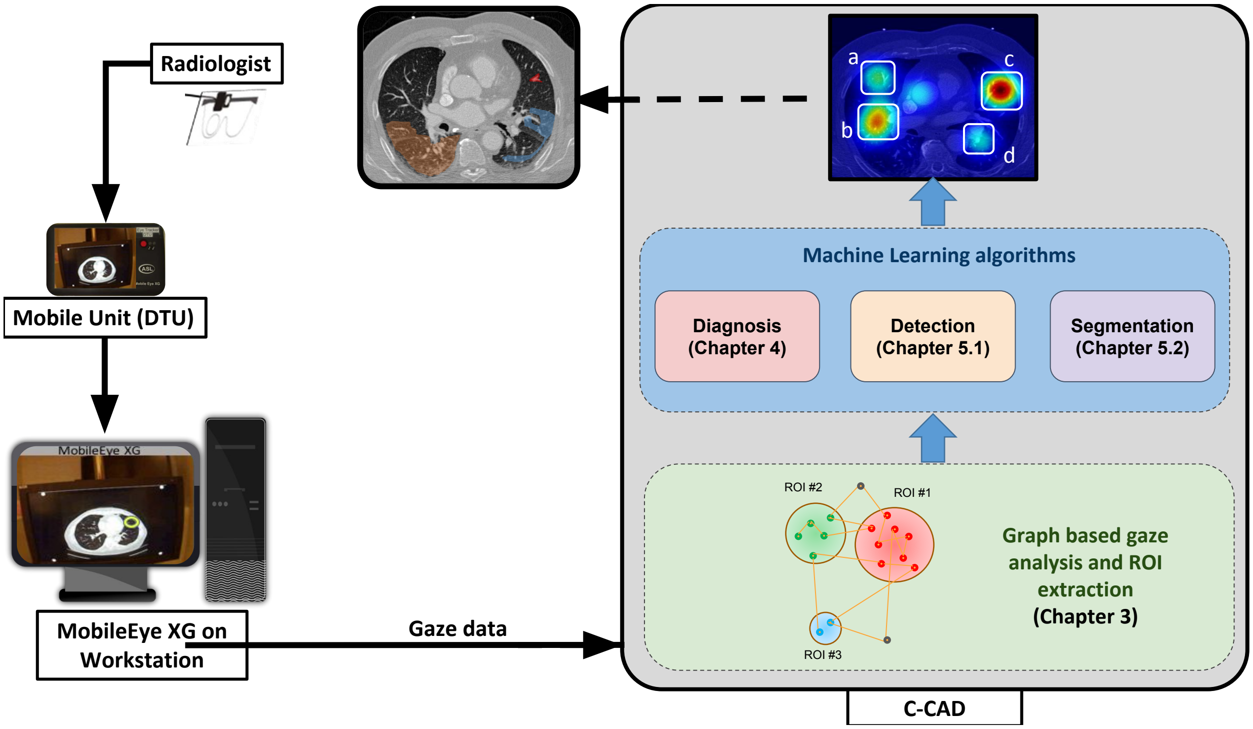
Radiologist
Mobile Unit (DTU)
MobileEye XG
MobileEye XG on Workstation
Gaze data
a
b
c
d
Machine Learning algorithms
Diagnosis (Chapter 4)
Detection (Chapter 5.1)
Segmentation (Chapter 5.2)
ROI #2
ROI #1
ROI #3
Graph based gaze analysis and ROI extraction (Chapter 3)
C-CAD

Figure 1.1: General overview of proposed framework and integration eye-tracking technology to CAD systems. Our proposed CAD has a graph based gaze analysis module to extract ROIs and uses that as a collaboration medium between radiologist and our other machine learning algorithms.

Contributions: This dissertation has significant broader impacts in radiology and imaging sciences and introduces several technical innovations, through the design of machine learning algorithms, as summarized below:

- A key aspect of any interactive system is to provide a natural feeling to the user. Having this natural feeling plays a crucial role specifically in the field of radiology, as imposing any constraint might affect the diagnosis accuracy. The majority of eye-tracking studies are conducted in the laboratory settings and no realistic eye-tracking based 3D screening experiment is reported in the literature. This dissertation fills this research gap and provides a natural (realistic) interaction framework.

- This dissertation proposes a new attention based data sparsification method applied to gaze patterns of radiologists, obtained through eye-tracking device during screening. The proposed approach allows local and global analysis of visual search patterns based on visual attention concepts, for the first time in the literature. More importantly, sparse representation of gaze patterns help interaction with the newly designed CAD system. Our system *truly* collaborates with its human counterpart (radiologists); therefore, it is fundamentally different than currently available second opinion tools.

- This dissertation further develops a new CAD system by proposing three state of the art deep learning algorithms integrated as *local* and *global* image analysis modules into the framework. The proposed system has been tested by radiologists with different years of experience, and robustness of the proposed C-CAD has been demonstrated.

- This dissertation extends the proposed graph analysis of C-CAD into a multi-parametric image analysis framework where users can utilize multiple screens (as in prostate screening with multi-parametric MRI), which is a challenge in practical clinical environments. To the best of our knowledge, this is the first study in the literature considering multiple screens in the eye-tracking platform, combining computer algorithms with screening operation.

1.1 Collaborative Computer Aided Diagnosis Through Eye-tracking.

Realistic radiology experience with eye-tracking is not achieved yet, mainly because of technical complexities of eye-tracking procedure in high-risk operations. Eye-tracking systems record data in two-dimensional (2D) and having a three-dimensional (3D) system needs an exact synchronization. Furthermore, one of the main challenges of quantitative modeling and comparison of gaze data stems from the difficulty of representing, analyzing, and interpreting dense eye-tracking data. This is not only technically challenging, but also computationally demanding [16]. The closest study addressing these problems was conducted by [17] who proposed the famous *scanning and drilling* paper analyzing gaze patterns at the global level. While *scanners* examine one slice of a radiology scan, more explicitly before moving to the next one, *drillers* keep going forward and backward in slices, moving through a 3D stack of the scan. **However, quantitative analysis of visual search patterns at the global and local level has never been addressed before**. We believe that such mapping will be extremely useful since it can serve as a natural interface between radiologists and CAD systems in that details of the gaze patterns will be used to guide a CAD system when it is necessary. In other words, it is not possible to benefit from human visual search in CAD systems when search patterns are represented only at the global level, which is the current practice at literature.

This dissertation addresses these challenges by (1) developing an eye-tracking interface that provides a real radiology reading room experience and (2) performing an attention based clustering and sparsification of dense eye-tracking data for building a C-CAD. The proposed algorithm preserves topological properties of the gaze data while reducing its size significantly. This allows us to quantitatively compare global search patterns of radiologists, extract radiologists' regions of interest (ROI) based on their level of attention during the screening process (local), and to combine this information with image content to do different image analysis tasks for each ROI. Radiologist's gaze data is represented as a graph and sparsified using the proposed attention based algorithm. Finally, a set of local and global image analysis modules are presented to perform processing of imaging data.

We initially, as a pilot study, thresholded the eye-tracking data by its time component to define potential attentional regions. Then, these regions (ROIs) were processed with computer vision based saliency models to remove some of the false positive regions from considerations. Final ROIs were used for image analysis, particularly for segmenting lung pathologies using attention. A 2D random walk algorithm [18] was utilized to segment those ROIs by combining visual saliency and visual attention information as seeds of the random walk algorithm. The average dice similarity of 86% was achieved [19]. In the current study, we significantly improve our design into a new level with multiple novel contributions.

1.2 Semi-supervised Multi-task Learning for Local Image Analysis.

One of the screening applications that we focus on is lung cancer screening. Lung cancer has the highest rate of mortality among the cancer related deaths [1]. Lung nodules are primary indicators of lung cancer. Early detection of lung nodules is of great importance in lung cancer screening. Existing research recognizes the critical role played by CAD systems in early detection and diagnosis of lung nodules. However, many CAD systems, which are used as cancer detection tools, produce a lot of false positives (FP) and require a further FP reduction step. Furthermore, guidelines for early diagnosis and treatment of lung cancer are consist of different shape and volume measurements of abnormalities. Segmentation is at the heart of our understanding of nodules morphology making it a major area of interest within the field of computer aided diagnosis systems.

This chapter set out to test the hypothesis that joint learning of false positive (FP) nodule reduction and nodule segmentation can improve the CAD systems' performance on both tasks. To support this hypothesis we propose a 3D deep multi-task CNN to tackle these two problems jointly. We tested our system on LUNA16 dataset and achieved an average dice similarity coefficient (DSC) of **91%** as segmentation accuracy and a score of nearly **92%** for FP reduction. As a proof of our hypothesis, we showed improvements of segmentation and FP reduction tasks over two baselines. Our results support that joint training of these two tasks through a multi-task learning approach improves system performance on both. We also showed that a semi-supervised approach can be used to overcome the limitation of lack of labeled data for the 3D segmentation task.

1.3 *S4ND*: Single-Shot Single-Scale Lung Nodule Detection

Successful diagnosis and treatment of lung cancer is highly dependent on early detection of lung nodules. Radiologists are analyzing an ever increasing amount of imaging data (CT scans) every day. CAD systems are designed to help radiologists in the screening process. However, automatic detection of lung nodules with CADs remains a challenging task. One reason is the high variation in texture, shape, and position of nodules in CT scans, and their similarity with other nearby structures. Another reason is the discrepancy between the large search space (i.e., entire lung fields) and respectively tiny nature of the nodules. Detection of tiny/small objects has remained a very challenging task in computer vision, which so far has only been solved using computationally expensive multi-stage frameworks. Current sate of art methods for lung nodule detection follow the same multi-stage detection frameworks as in other computer vision areas.

In this section of this chapter, we resolve the aforementioned issues by proposing a completely 3*D* deep network architecture designed to detect lung nodules in a single shot using a single-scale network. This detection algorithm is to be embedded complimentary to the eye-tracking screening applications in the radiology room. To the best of our knowledge, this is the first study to perform lung nodule detection in one step. Specific to the architecture design of the deep network, we make use of convolution blocks with dense connections for this problem, making one step nodule detection computationally feasible. We also investigate and justify the effect of different downsampling methods in our network due to its important role for tiny object detection. Lastly, we argue that lung nodule detection, as opposed to object detection in natural images, can be done with high accuracy using only a single scale network when network is carefully designed with its hyper-parameters.

1.4 *PAN*: Projective Adversarial Network for Medical Image Segmentation.

Segmentation is another important image analysis application that should be part of screening procedure. In this section of this chapter, we focus on the challenging problem of pancreas segmentation from CT images, although our framework is generic and can be applied to any 3D object segmentgation problem. This particular application has unique challenges due to the complex shape and orientation of pancreas, having low contrast with neighbouring tissues and relatively small and varying size. Pancreas segmentation were studied widely in the literature. Yu et al. introduced a recurrence saliency transformation network, which uses the information from previous iteration as a spatial weight for current iteration [20]. In another attempt, U-Net with an attention gate was proposed in [21]. Similarly, a two-cascaded-stage based method was used to localize and segment pancreas from CT scans in [22]. A prediction-segmentation mask was used in [23] for constraining the segmentation with a coarse-to-fine strategy. Furthermore, a segmentation network with RNN was proposed in [24] to capture the spatial information among slices. The unique challenges of pancreas segmentation (complex shape and small organ) shifted the literature towards methods with coarse-to-fine and multi-stage frameworks, promising but computationally expensive.

The current literature on segmentation fails to capture 3D high-level shape and semantics with a low-computation and effective framework. In this section, for the fist time in the literature, we propose a projective adversarial network (PAN) for segmentation to fill this research gap. Our method is able to capture 3D relations through 2D projections of objects, without relying on 3D images or adding to the complexity of the segmentor. Furthermore, we introduce an attention module to selectively integrate high-level, whole-image features from the *segmentor* into our adversarial network. With comprehensive evaluations, we will show that our proposed framework will achieve the state-of-the-art performance on publicly available CT pancreas segmentation dataset [25] even with a very simple encoder-decoder network as *segmentor*.

1.5 Summary

This Chapter started by defining the research gap and discussing the necessity of having an interactive framework for the radiology screening process. We discussed how and where this dissertation contributes to the defined research gap. In Section 1.1, we discussed the challenges of having a realistic eye-tracking based integration and how this dissertation addressed such challenges. In Section 1.2, we explained how multi-task learning and semi-supervised learning can be beneficial for our local image analysis module. In Section 1.3, we explained the challenges of detecting tiny abnormalities in the lung and how to tackle it. Finally, in Section 1.4, we argue the benefit of adversarial learning, and further, a novel projective framework that captures 3D information through 2D projections, for segmentation of complex 3D structures.

Rest of the dissertation is organized as follows: In Chapter 2 we go over the literature corresponding to each piece of our framework. In Chapter 3, we describe the proposed hardware and software integration, details of data acquisition parameters, and the proposed data representation technique with sparsification. Further, we introduce the potential of the proposed C-CAD system to handle multi-parametric images and multi-screen based eye-tracking and image analysis in general. In

Chapter 4 we go into the details of our semi-supervised multi-task learning algorithm for joint tumor diagnosis and segmentation. In Chapter 5, we introduce our two global image analysis modules. In Section 5.1, we describe our single-scale single-shot lung nodule detection. In Section 5.2, we propose a novel framework based on projective adversarial learning for complex 3D structure segmentation. We conclude this dissertation by discussion and the future work.

CHAPTER 2: LITERATURE REVIEW

In this chapter, we comprehensively review the literature on the effect of radiology screening and eye-tracking studies in Section . In Section 2.2 we study CADs for local image analysis. We then review current literature for lung nodule detection in Section 2.3 and medical image segmentation in Section 2.4, respectively.

2.1 Radiology screening and eye-tracking studies

According to the American Cancer Society, lung and prostate cancers are the leading causes of death and also the fastest growing cancers in 2017 [1]. Medical imaging helps early detection of cancers, but in a recent lung cancer screening clinical trial, it was found that approximately 35% of lung nodules were missed during the screening process by radiologists [3]. Previous studies have shown that early diagnosis of cancers may have a greater impact on the population [1, 3, 26]. However, many open questions remain in screening examinations. For instance, definition of at-risk population affects the patient's inclusion in the study. Timing and intervals of screening are adjusted by clinical trials, but there is no optimal method yet to justify these selections. Nevertheless, even in these suboptimal conditions, exciting research is ongoing in this field, thanks to more advanced CT scanners and development of computerized image analysis methods. CAD systems have shown to be useful in reducing false negative (missed tumor) cases, but the main issue with all CAD systems is the presence of a high false positive rate [27]. For instance, when an automated lung nodule detection method was used in a study by [26], 84% of the missed lung cancers were marked by the computer. Despite this catch, the false-positive rate was very high: 28 false positive findings per scan.

A key aspect of biological vision studies is to understand perceptual and/or cognitive errors and how radiologists search radiology scans for finding abnormalities. These studies extensively benefit from different eye-tracking technologies [16]. For decades vision scientists have been studying this topic [2, 7, 8, 9, 10, 11, 12, 13, 14, 15]. Comparison of the visual search patterns of radiologists, and inferring local and global information from those patterns have accelerated the research in this field and led to a better understanding of the differences between expert and novice readers/radiologists, and general strategies for visual search in radiology scans. Some of these studies date back to the 1960s. In spite of decades of work, available methods in the literature fail to provide:

- A real radiology room experience for radiologists,
- A quantitative modeling and comparison of eye-tracking data,
- Exploration of eye-tracking tools' potential to compensate for CAD errors.

In particular, the interaction between radiologists and computers (either simple Picture Archiving and Communication System (PACS) or CAD systems) remain untouched except by a few seminal image analysis studies [17, 19, 28].

2.2 Review of CADs in deep learning era for local image analysis.

In recent years, deep learning based algorithms revolutionized many fields including medical image analysis applications. In conventional CAD systems (i.e., prior to the deep learning era), handcrafted feature design/extraction followed by a feature selection and classification scheme were the main steps. However, with the success of deep learning, this strategy has moved from *feature engineering* to *feature learning*. In very recent frameworks, Convolutional Neural Networks (CNN) have been used for feature extraction and off-the-shelf classification methods in most CAD systems [29, 30, 31, 32]. In this line of research, for instance, Hua et al. proposed using a Deep Belief Network and a CNN for lung nodule classification [33] while Kumar et al. used deep features extracted from an autoencoder to classify nodules into malignant and benign [34]. Deep learning based lung cancer detection has also been used as part of a screening strategy [35, 36].

Various deep learning networks were developed for lung cancer diagnosis in the literature [37, 38, 39]. In those works, authors have first incorporated shape information of lung nodules to improve diagnostic accuracy [37]. In another approach, they investigated Gaussian Process algorithms along with CNN to incorporate radiographical interpretations of nodule appearances to improve diagnostic decisions [39]. Later, the network was improved (called *TumorNET*) by converting the CNN into a multi-task deep learning strategy [38]. A multi-task 3D network for joint segmentation and false positive reduction of lung nodules in a semi-supervised manner was proposed in [40]. Meanwhile, many studies in the recent literature focused on false positive reduction in lung nodule detection. Some utilized multiple CNNs for multi-view lung nodule analysis [36], while some used 3D CNNs for a more efficient analysis [41, 42]. Furthermore, a multi-scale analysis of lung nodules using multiple 3D CNNs was proposed by [43]. The literature pertaining to lung nodule detection and characterization via CNN is vast. A brief overview of some network architectures related to lung cancer diagnosis can be found in [44, 45]. Specific to prostate cancer detection from radiology scans, recent works investigated the application of CNNs using multi-parametric MRI [46] and a semi-supervised approach for biopsy-guided cancer detection using a deep CNN [31]. Deep learning has also been used extensively as a feature learning tool for various applications such as MRI based prostate segmentation [47].

2.3 Review of Lung Nodule Detection Literature

The literature for lung nodule detection and diagnosis is vast. To date, the common strategy for all available CAD systems for lung nodule detection is to use a candidate identification step (also known as region proposal). While some of these studies apply low-level appearance based features as a prior to drive this identification task [48], others use shape and size information [49]. Related to deep learning based methods, Ypsilantis et al. proposed to use recurrent neural networks in a patch based strategy to improve nodule detection [50]. Krishnamurthy et al. proposed to detect candidates using a $2D$ multi-step segmentation process. Then a group of hand-crafted features were extracted, followed by a two-stage classification of candidates [49]. In a similar fashion, Huang et al. proposed a geometric model based candidate detection method which followed by a $3D$ CNN to reduce number of FPs [41]. Golan et al. used a deep $3D$ CNN with a small input patch of $5 \times 20 \times 20$ for lung nodule detection. The network was applied to the lung CT volume multiple times using a sliding window and exhaustive search strategy to output a probability map over the volume [51].

There has, also, been detailed investigations of high-level discriminatory information extraction using deep networks to perform a better FP reduction [36]. Setio et al. used 9 separate $2D$ convolutional neural networks trained on 9 different views of candidates, followed by a fusion strategy to perform FP reduction [36]. Another study used a modified version of Faster R-CNN, state of the art object detector at the time, for candidate detection and a patch based $3D$ CNN for FP reduction step [42]. However, all these methods are computationally inefficient (e.g., exhaustive use of sliding windows over feature maps), and often computed in 2D manner, not appreciating the 3D nature of the nodule space. It is worth mentioning that patch based methods are 3D but they suffer from the same computational burdens, as well as missing the entire notion of 3D nodule space due to limited information available in the patches.

2.4 Review of Segmentation Methods

Segmentation has been a major area of interest within the fields of computer vision and medical imaging for years. Owing to their success, deep learning based algorithms have become the standard choice for semantic segmentation in the literature. Most state-of-the-art studies model segmentation as a pixel-level classification problem [52, 53, 54]. Pixel-level loss is a promising direction but, it fails to incorporate global semantics and relations. To address this issue researchers have proposed a variety of strategies. A great deal of previous research uses a post-processing step to capture pairwise or higher level relations. Conditional Random Field (CRF) was used in [52] as an offline post-processing step to modify edges of objects and remove false positives in CNN output. In other studies, to avoid offline post-processing and provide an end-to-end framework for segmentation, mean-field approximate inference for CRF with Gaussian pairwise potentials was modeled through Recurrent Neural Network (RNN) [55].

In parallel to post processing attempts, another branch of research tried to capture this global context through multi-scale or pyramid frameworks. In [52, 53, 54], several spatial pyramid pooling at different scales with both conventional convolution layers and *Atrous* convolution layers were used to keep both contextual and pixel-level information. Despite such efforts, combining local and global information in an optimal manner is not a solved problem, yet.

Following by the seminal work by Goodfellow et al. in [56] a great deal of research has been done on adversarial learning [57, 58, 59, 60]. Specific to segmentation, for the first time, Luc et. al. [58] proposed the use of a discriminator along with a segmentor in an adversarial min-max game to capture long-range label consistencies. In another study *SegAN* was introduced, in which the segmentor plays the role of generator being in a min-max game with a discriminator with a multi-scale *L1* loss [59]. A similar approach was taken for structure correction in chest X-rays segmentation in [61]. A conditional GAN approach was taken in [60] for brain tumor segmentation.

2.5 Summary

This Chapter began with a literature review of radiology screening and eye-tracking studies in Section 2.1. We then provided a comprehensive survey on the existing Computer Aided Diagnosis systems for local image analysis in 2.2. Next, we discussed the literature on lung nodule detection in Section 2.3. Finally, we concluded this Chapter by a detailed review of medical image segmentation in Section 2.4.

CHAPTER 3: COLLABORATIVE COMPUTER AIDED DIAGNOSIS THROUGH EYE-TRACKING

The results of this Chapter have been published in the following paper:

- Naji Khosravan, Haydar Celik, Baris Turkbey, Elizabeth C Jones, Bradford Wood, Ulas Bagci, *"A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning,"* in Medical image analysis 51 (2019): 101-115 [62].

This work has also been highlighted in NIH Clinical Center director's speech in RSNA 2018.

In this chapter, we introduce a paradigm-shift system that uses eye-tracking technology as a collaborative tool between radiologists and CAD systems. The rationale behind this idea comes from the fact that radiologists are good at eliminating false positives, which CADs often fail to achieve at a human level performance. On the other hand, CADs capture missing abnormalities better than the human observer. Because of this complementary properties, we call the proposed technology C-CAD.

Briefly, we develop an accurate and efficient algorithm that accepts input from an eye-tracking device, and process this information to extract radiologists' pattern of search and attention regions, to be combined with image analysis modules. We hypothesize that combining the strength of radiologists and CAD systems will improve the screening/diagnosis performance. To test this hypothesis, we have conduct a lung cancer diagnosis experiment (with low dose CT scans) based on eye-tracking data recorded from multiple radiologists. We also show the applicability of our framework to multi-parametric MRI to conduct prostate cancer screening. We choose lung and prostate cancers due to their high rate of mortality and growth in 2017, thus confirming their clinical importance ([1]).

3.1 Data acquisition and Eye-Tracker:

In this study, a Fovio^{TM} Eye Tracker remote eye-tracker system (Seeing Machines Inc, Canberra, Australia) was used. This eye-tracker is a contact free eye-tracker as can be seen in Figure 3.1 and will be mounted on the desktop. The sampling rate of this device is $60Hz$. The divice has Binocular tracking and has 0.78 Degrees (Mean) 0.59 (Std. Dev.) angular error. This device has a Large head box ($31cm \times 40cm \times 65cm$) and is robust to glasses and ambient light, which is suitable for dark radiology room recordings. This device also supports multi-display tracking.

Figure 3.1: Fovio^{TM} Eye Tracker which was used in our experiments. This device is non-contact and desktop mounted and is robust to glasses and ambient light.

We collected eye tracking data using EyeWorks^{TM} Suite ($v.3.12$) on a DELL Precision $T3600$ using a Windows 7 operating system on an Intel Xeon CPU $E5-1603$ 0 @ $2.80GHz$ with $8GB$ of RAM. Figure 3.2 illustrates integration of Eye-Tracker into our system. Using EyeWorks^{TM}, eye movements were recorded by two synchronized, remote eye-trackers at $60Hz$. All stimuli were presented at a resolution of 1280×1024 on a DELL 19" LCD monitor. We utilized a $60Hz$ FOVIO eye-tracker and verified calibration through a five-point calibration procedure in EyeWorks^{TM} Record prior to the task. Calibration was considered sufficient if the dot following the eye movement trajectory was sustained (indicating that the eye movement monitor was not losing tracking) and if the calibration dot was accurate (falling on the calibration check targets at the center and corners of the screen when the participant was instructed to look at them, with inaccuracy of up to one centimeter for the upper two corner targets). The eye-tracker was located between 9.5 *cm* and 8 *cm* beneath the bottom of the viewing screen (eye-tracker was placed just under the viewing area and at a 25 degree angle with respect to the monitor.). Following calibration, participants completed the task as described above. After completing this task, the FOVIO was re-calibrated before moving on to a Smooth Pursuit task. Upon completion of screening, the experimental portion of the study was complete and subjects discussed the study with the experimenters before leaving. From consent to debriefing, the study duration spanned roughly 45 *minutes*. Custom made DICOM viewing software was built on Medical Image Processing, Analysis and Visualization software (MIPAV CIT, NIH, Bethesda, MD).

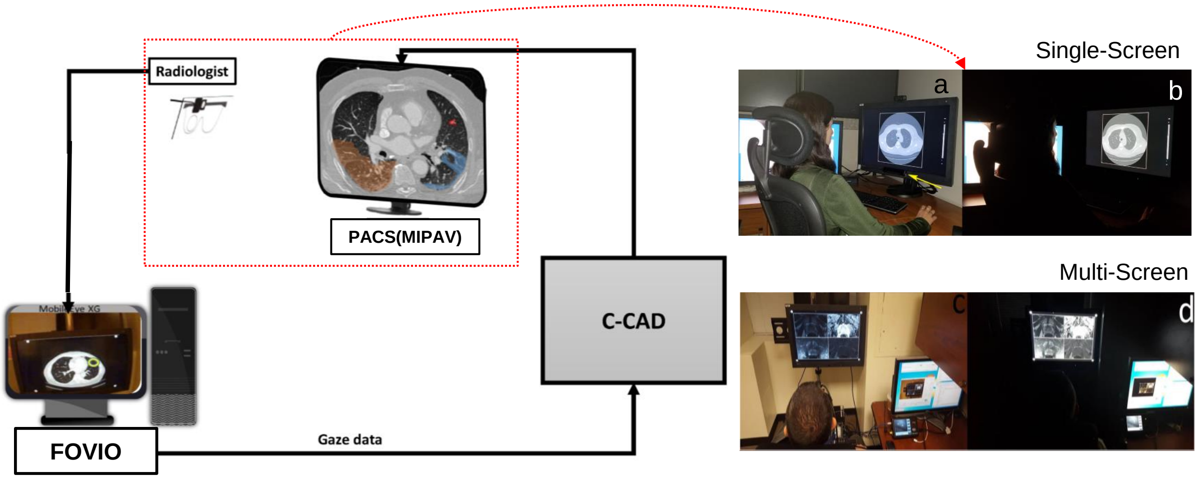

Figure 3.2: A representation of the Eye-Tracking system in a realistic radiology setting is illustrated. Eye-Tracking system, the connection to the workstation, and the C-CAD system are integrated into the PACS (MIPAV) system directly as shown on the left. Screening experiments in normal light (a,c) and dark (b,d) radiology rooms for single (a,b) and multi-screen (c,d) experiments are shown on the right.

3.2 Methodology

In our proposed framework, eye-tracking data goes through five steps to be converted from a dense graph into a set of diagnostic decisions and segmentation masks on the lesions inside ROIs: *Step 1)* The gaze data is represented with a graph. *Step 2)* A non-parametric clustering method is applied to the graph nodes. *Step 3)* A novel attention-based sparsification algorithm is applied to the graph to reduce redundant information. *Step 4)* Radiologists' Regions of Interest (ROIs) are extracted based on the level of attention, and *Step 5)* Two sets of deep learning based image analysis modules (local and global). (See Fig. 3.3 for the overview of the proposed system for an example application). In this chapter we focus on the first 4 steps of this process and modules which are developed for step 5 will be discussed in the following chapters of this Thesis.

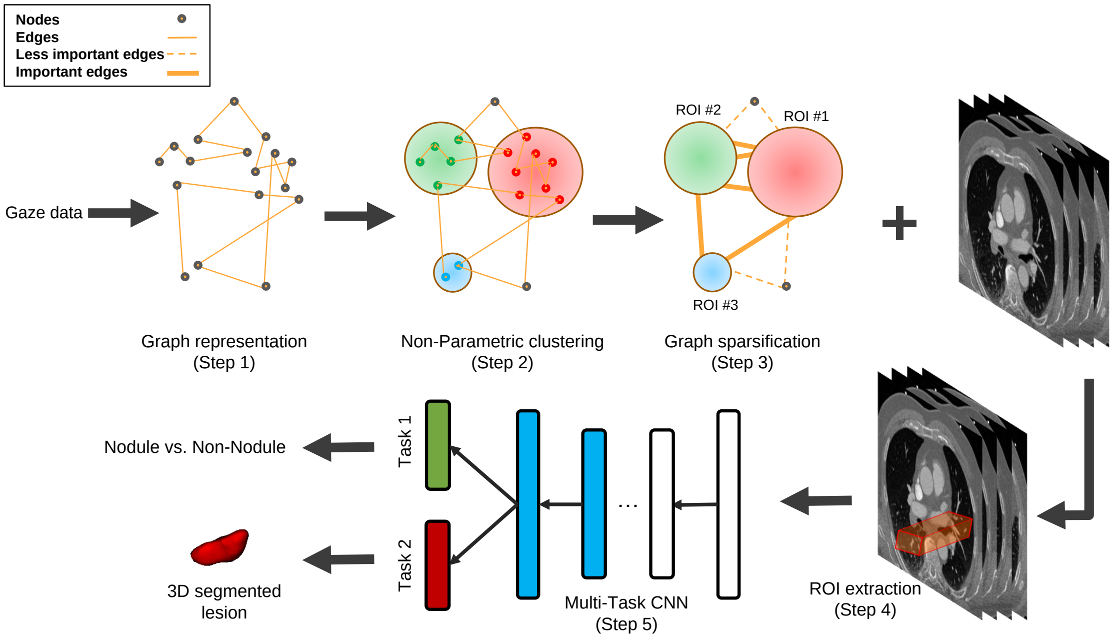

Figure 3.3: To extract radiologist ROIs the dense eye-tracking data goes through a clustering and sparsification algorithm. ROIs are then extracted and passed to the image analysis block along with the image.

3.2.1 Step 1: A graph representation of the eye-tracking data

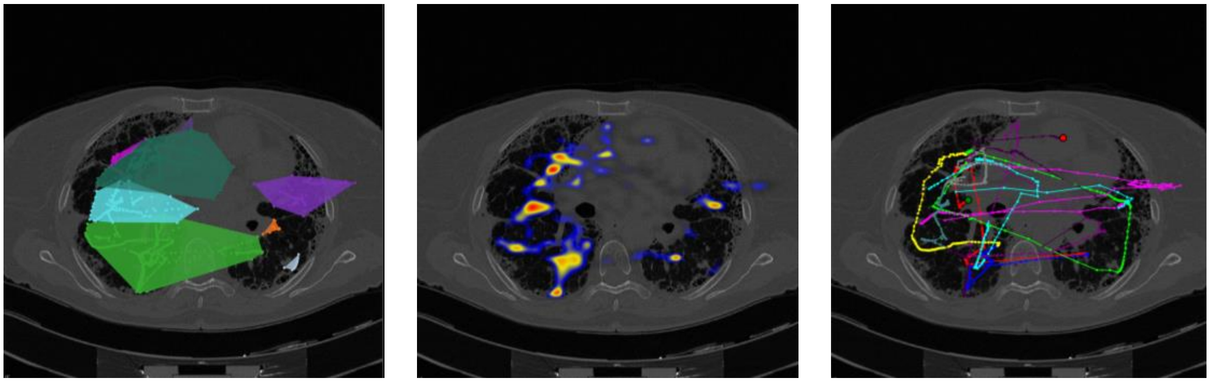

Figure 3.4: Eye-tracking data recorded from lung cancer screening. Low-dose CT is used in a single screen. Gaze patterns (right), heat maps of gaze patterns (middle), and coverage area of the gaze patterns (left) are illustrated.

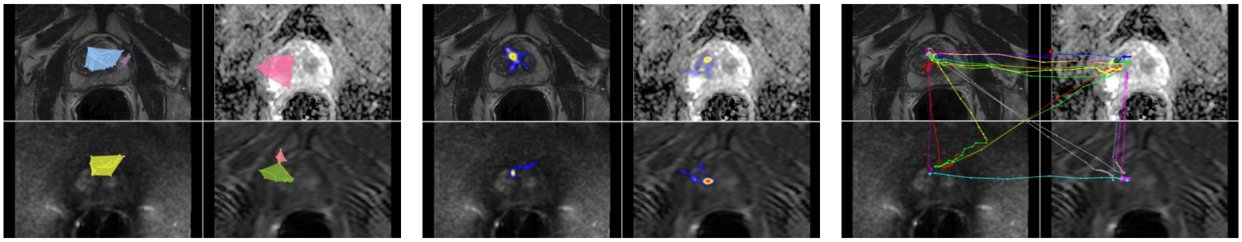

Figure 3.5: Eye-tracking data recorded from prostate cancer screening. Multi-Parametric MRI is used in four screens (left upper: T2-weighted (T2w), right upper: apparent diffusion coefficient (ADC) map, left lower: diffusion weighted imaging (DWI), right lower: dynamic contrast enhanced (DCE) maps). Gaze patterns across different screens and the paths are illustrated for an example screening task. Gaze patterns (right), heat maps of gaze patterns (middle), and coverage area of the gaze patterns (left) are illustrated.

An example of eye-tracking data recorded from two cancer screening tasks is shown in Fig. 3.4 and Fig. 3.5. For each experiment, 2D images are overlaid with the coverage area (left), heatmap (middle), and scanpaths (right) representations inferred from the gaze patterns. Once gaze patterns are recorded, they are dense, hence, difficult to analyze (See Fig. 3.6). The aim in data sparsification is to represent the data with far less parameters and without significantly losing its content. It is also desirable to process the data easily and efficiently when sparisified. To end this, we propose to represent eye-tracking data as a graph and reduce its size without distorting the topology of the data structure by utilizing clustering and sparsification algorithms.

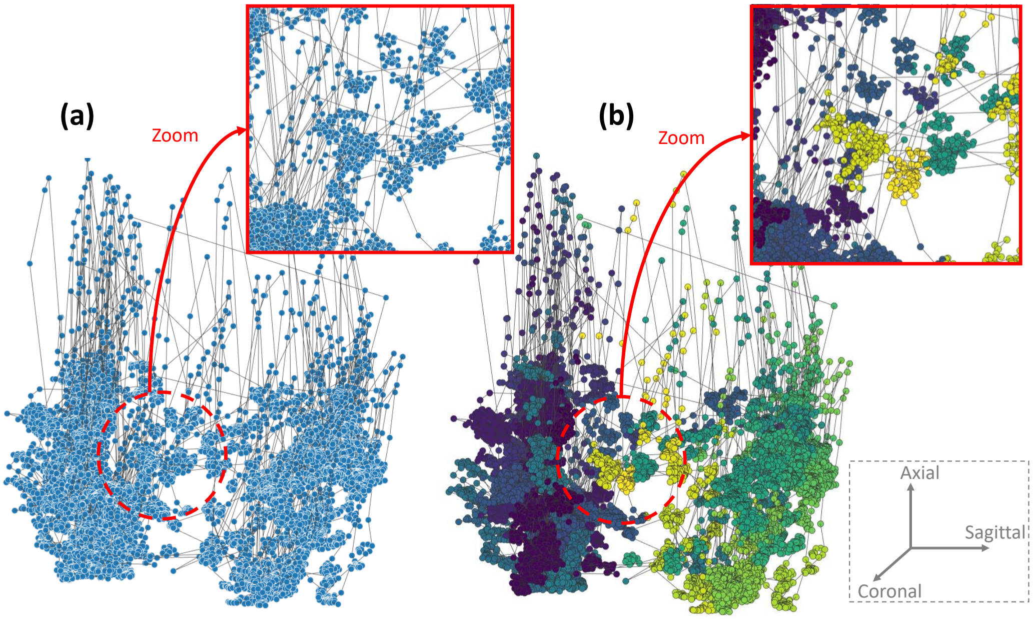

Figure 3.6: (a) 3D Graph representation of visual search patterns from a lung cancer screening experiment. (b) Clustering helps to group gaze points to define attention regions. Colors indicate different clusters.

Graph theory is concerned with a network of points (nodes or vertices) connected by lines (edges). It is a well-established branch of mathematics and it has numerous successful applications in diverse fields. Formally, a graph (G) refers to a set of vertices (V) and edges (E) that connect the vertices, and it is represented as $G = (V,E)$. In the current problem, a graph representation is a perfect choice for eye-tracking data because gaze locations (i.e., fixations) can easily be stored as vertices while path/directions (i.e., saccades) between gaze locations can be stored as edges in the graph. An example of a 3D graph representation of gaze patterns obtained from a lung cancer screening experiment using volumetric low-dose CT scans is illustrated in Fig. 3.6(a). A zoomed version of the graph indicates dense data points and edges between them. For simplicity, edges are shown as undirected.

Although graph representation allows parameterization of the patterns in the data, its analysis and interpretation are infeasible because the graph includes large amount of nodes and edges as exemplified in Fig. 3.6. Such graphs are called "dense", and *sparsification* operation can be considered to simplify the data to overcome challenges of the analysis of a dense graph. A graph sparsification algorithm reduces the graph density by omitting unnecessary edges. However, there are challenges unique to our problem which makes the conventional sparsification algorithms suboptimal in our case:

- *First*, the constructed graph is consecutive, meaning each edge in our graph connects only two distinct vertices (due to the nature of eye-tracking data). This causes a maximum vertex degree of 2 in our graph resulting in the failure of current sparsification algorithms to remove even a single edge from our graph. This is because all the edges in our graph are considered equally important for keeping the structure of the graph unchanged.
- *Second*, the radiologist's attention should be taken into account while sparsifying the data. This will make sure that the global visual search pattern and attention regions are both preserved after sparsification.

To overcome these challenges, we propose the next two steps of our algorithm to handle the consecutive nature of the data and encountering attention information, respectively.

3.2.2 Step 2: Non-Parametric clustering of the graph nodes

We propose to apply a non-parametric clustering algorithm to graph vertices of gaze data and reconstruct the graph from clustered vertices (i.e., one vertex for each cluster). There exists a great number of clustering methods in the literature due to its applicability in many fields. Each algorithm has advantages and disadvantages. Since gaze patterns are dense, it is desirable that the clustering algorithm is chosen from a time-efficient family of algorithms.

Non-Parametric clustering algorithms often have subjective measures for partitioning the data into distinct groups. Hence, many efforts are spent on designing such measures. The choice of this similarity measure (or called dissimilarity metric in the same fashion) is very important, as it has a strong effect on the resulting groupings. In the current problem, we make a domain-specific definition of *similarity* for gaze patterns. Simply, we use distance between gaze points as a similarity measure. This is because if radiologists spend more time in screening for a particular region (i.e., attention region), then the data collected from those regions are dense and in close vicinity of each other. Likewise, if the distances between gaze locations are far, it can be safely assumed that a different attention region is being examined. Given the fact that, there will be some heterogeneity in distance measurements, it is still reasonable to use it, as is common in most non-parametric clustering algorithms. However, any clustering algorithm with a distance based similarity metric will not be an optimal choice in our case since the data is extremely dense, and we desire the algorithm to run in linear time while processing large amounts of data.

We hypothesize that the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm can be a good fit for our purpose because it is time-efficient (linear time), non-parametric, and local ([63]). BIRCH uses a Cluster Feature $(CF) = (N, LS, SS)$ to make a large clustering problem tractable by concentrating on densely occupied portions. For N data points in a given cluster X_i, $LS = \sum_{i=1}^{N} X_i$, and $SS = \sum_{i=1}^{N} X_i^2$ are used to measure pairwise distances between data points. *N*, *LS*, and *SS* are basically representing norms of the cluster.

This step represents the dense data with a set of attention regions by clustering them into different groups allowing us to modify our graph as follows:

1. all the vertices pertaining to each cluster are removed except the centroid of the cluster,
2. all the edges that were connecting vertices from different clusters now connect the corresponding centroids, and
3. all the edges that were connecting vertices inside each cluster are modeled as self loops on the centroid (see Fig. 3.3).

This modification allows the maximum vertex degree of larger than 2 in the graph. The results of clustering on the graph vertices and the modified graph are shown in Fig. 3.6. (Note: the self loops on cluster centroids are not shown in the modified graph for the sake of visibility). Spectral sparsification algorithms can now be used to reduce the data further.

3.2.3 Step 3: Attention based graph sparsification

In the previous step, we simplify our graph by eliminating vertices inside each cluster and represent them by a centroid. This step will switch the attention mechanism from vertices to clusters, meaning that the attention region is now represented by clusters. Graph sparsification methods, on the other hand, eliminate edges and convert a dense graph *G* into a sparse graph *S*.

In our proposed method, we intend to preserve the structural similarity between the sparsified and original graphs; therefore, spectral sparsification algorithms are more suitable than the other kinds. It is mainly because many graphs can be characterized better with spectral estimations where spectral information is obtained by an adjacency matrix, which is normalized by its edges and subtracted from the identity matrix. In our problem, another constraint is to have a linear or nearly-linear sparsification algorithm so that the whole dense data analysis become efficient. Due to these two constraints (being fast and intention to preserve structural similarities), we adapted Spielmans's *nearly-linear time* spectral graph sparsification method ([64]) with a novel weight parameter, w, to reflect the attention regions inferred from eye-tracking data. This particular spectral sparsification algorithm forces the Laplacian quadratic form of the sparsified graph to be a $\sigma - spectral\ approximation$ of the original graph ([64]) and preserves the structural similarity between the sparsified and original graph. Note that the spectral sparsifier is defined as a subgraph of the original whose Laplacian quadratic form is approximately the same as that of the original graph on all real vector inputs, as proved by ([64]). That is, if the Laplacian (i.e., spectral properties) matrix is preserved, the structural similarity is preserved.

The Laplacian matrix of a weighted graph $G(V,E,w)$ is defined simply as:

$$L_G(i,j) = \begin{cases} -w_{i,j} & i \neq j \\ \sum_z w_{i,z} & i = j, \end{cases} \tag{3.1}$$

where $w_{i,j}$ represents the weight of edge between vertices i and j. The Laplacian quadratic form of G for $x \in \mathbb{R}^V$ is:

$$x^T L_G x = \sum_{i,j \in E} w_{i,j}(x(i) - x(j))^2. \tag{3.2}$$

Let $\hat{G}$ be a $\sigma - spectral\ approximation$ of G if for all $x \in \mathbb{R}^V$ such that:

$$\frac{1}{\sigma} x^T L_{\hat{G}} x \leq x^T L_G x \leq \sigma x^T L_{\hat{G}} x\ . \tag{3.3}$$

In the original implementation of the spectral sparsification algorithm ([64]), a weighted graph $G(V,E,w)$ is converted to a sparse graph $S(V,\hat{E},\hat{w})$ with $|\hat{E}| = O(n \log n/\alpha^2)$ in $\tilde{O}(m)$ time, for a graph with n vertices and m edges, such that

$$(1-\alpha)\sum_{i,j\in E} w_{i,j}(\delta x)^2 \leq \sum_{i,j\in \hat{E}} \hat{w}_{i,j}(\delta x)^2 \leq (1+\alpha)\sum_{i,j\in E} w_{i,j}(\delta x)^2, \tag{3.4}$$

where α is the sparsification parameter and δx denotes $x(i)-x(j)$. This method samples edges from G with a probability proportional to $w_{i,j}.r_{i,j}$, where $r_{i,j}$ represents effective resistance corresponding to the edge. Note that *effective resistance* is a distance measure, inspired by the homology (i.e., correspondence) between a graph and an electrical network.

We modify the spectral sparsification approach according to our unique problem by adding a novel weight function to it. We present the radiologists' level of attention on different regions by the edge weight between those regions. These weights are also used as a probability measure to define their importance. More important edges are defined as the edges connecting regions with more attention (indicating dense visual search patterns). We transfer these rationale into our graph using two primary parameters:

- *Number of nodes in each cluster (N)*: indicating a **global** representation of attention for a particular region. The more a radiologist spends time on a region, the denser the corresponding cluster is.
- *The amount of time spent on one cluster (C)*: indicating a **local** representation of attention for a cluster. The number of self-loops on each centroid can be considered to define such parameters.

We then define edge weights based on these two parameters as below:

$$w_{i,j} = exp(-N_i^2 \times C_{in})^{-1} \times exp(-N_j^2 \times C_{jn})^{-1}, \tag{3.5}$$

where N_i and N_j are number of nodes in clusters i and j, respectively, and C_{in} and C_{jn} are number of self-loops for clusters i and j. Each *exp* term can be considered as the attention level corresponding to each of the nodes. The major strength of our modified spectral sparsification algorithm is that we model both local and global visual patterns and their interactions through this weight function. The pseudo code of our sparsification method is given in Algorithm 1.

```
Input  : Dense graph: $G = (V, E)$, $\alpha$: Sparsification parameter
Output: Sparse graph: $S = (V, \hat{E}, \hat{w})$
for i and j in V do
    if There exists $e_{i,j}$ in E then
        Compute $N_i$ and $N_j$ (number of nodes in clusters corresponding to vertices i and j)
        Compute $C_{in}$ and $C_{jn}$ (number of self-loops corresponding to vertices i and j)
        Compute $w_{i,j}$ using eq.3.5
    end
end
return $G(V, E, w)$;
for $e_{i,j}$ in E do
    Sample edge $e_{i,j}$ form G with prob. proportional to $w_{i,j}.r_{i,j}$
    Add $e_{i,j}$ to S
end
return $S(V, \hat{E}, \hat{w})$;
```

Algorithm 1: Attention based Spectral Graph Sparsification.

To illustrate the effect of our sparsification method on very dense data, we applied the proposed method on synthetic data with different sparsification levels. Result is shown in Fig. 3.7. The data was created by randomly generating locations which were connected to each other consecutively to best mimic human gaze (in terms of consecutive connections). Progressively sparsified graphs are shown with respect to different levels of edge ratio. *Edge ratio* herein refers to the percentage of the total number of edges remaining after the sparsification; hence, the most sparse graph is represented on the last column where the edge-ratio is set to 0.2.

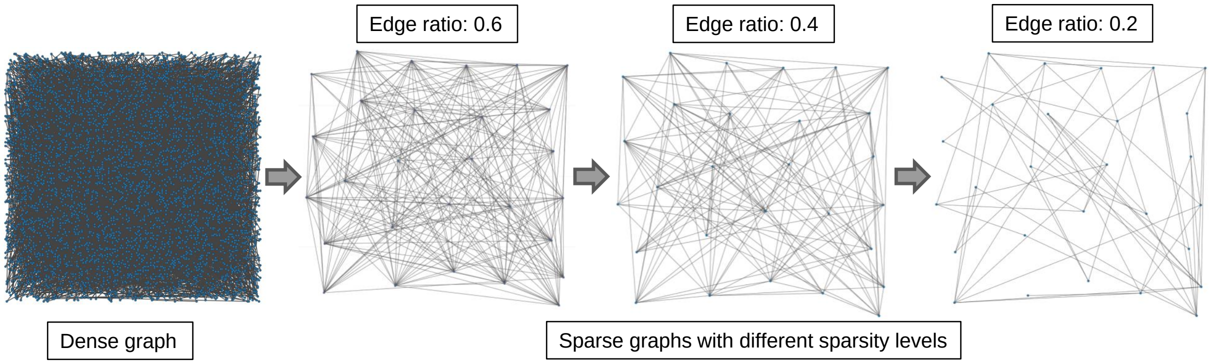

Figure 3.7: Results of applying proposed graph sparsification method on a 2D dense synthetic data. Edge ratio is the ratio of edges after applying the method to the original graph.

3.2.4 Step 4: Extracting attention-based ROIs

Having discussed how to construct the sparsified graph while keeping attention information, we now discuss how to extract the *Regions of Interest (ROIs)* from the graph. This step allows us to combine this information with image content, all in 3D space, and perform different analysis inside ROIs.

The attention level of each node inside our graph (a_i) is defined as a combination of global and local attention information on that node. This can be formulated as follows:

$$a_i = exp(-N_i^2 \times C_{in})^{-1}, \tag{3.6}$$

where N_i represents the number of nodes in cluster i (i.e., global attention level) and C_{in} represents the number of self loops on cluster i (i.e., local attention level). That is, higher values of a_i

correspond to higher focus of attention on a corresponding cluster centroid. The cluster centroid represents a location in the 3D space of image coordinates.

Our method enables us to perform any image analysis tasks on the ROIs including but not limited to segmentation, detection of particular pathology, and diagnosis. In the next section, we demonstrate how the search pattern and attention information from the radiologists' gaze can be combined with image content to perform radiological image analysis: nodule segmentation and false positive removal.

3.3 Experiments

In this section, we first explain the data used in our experiments as well as the experimental results of each component of proposed method.

3.3.1 Data

We tested our proposed method on synthetic data and two real screening experiments:

Synthetic data: included random generation of 5000 nodes with consecutive generated edges between these nodes, to better mimic the nature of eye-tracking data.

Lung cancer screening: the chest CT scans for this experiment were obtained from Lung Tissue Research Consortium (LTRC) (https://ltrcpublic.com/). The in-plane resolution of the CT images are 512×512 with a voxel size of $0.58 \times 0.58 \times 1.5\ mm^3$.

Participants: Three radiologists with different expertise levels participated in our experiments. Their reading experience levels were noted as 20, 10, and 3 years, respectively. After necessary adjustment and calibrations of eye-tracking equipment were done for each participant, they were given instructions about the screening process without letting them know the existence or absence of tumors in the scans. Mouse and other manipulations (zoom, scroll, contrast/brightness window selection) were recorded automatically by the software along with gaze locations.

3.3.2 Evaluation of Graph Sparsification

In order to show the effectiveness of our proposed graph sparsification method, we tested it first on a dense synthetic data. Then, a 3*D* lung cancer screening experiment was performed. To show that our algorithm is capable of analyzing multi-screen experiments, we applied the proposed algorithm on a 3D multi-parametric MRI prostate cancer screening as a feasibility study. All of our experiments were performed in a real radiology room setting without putting any restriction or limitations on the radiologist. The qualitative and quantitative results of the above-mentioned experiments are reported in the following sections.

3.3.2.1 Synthetic data experiment:

The goal of this experiment was to show the ability of our proposed method in dealing with very dense data. In the synthetic data, 5000 nodes were connected consecutively in the 3*D* space to create a dense data. The reason behind consecutive connections is to mimic data recorded from eye-trackers. Figure 3.8 illustrates the effect of our algorithm in sparsifying the data. The number of edges were reduced from 4269 to 524 in sparsification step, and the number of nodes were reduced from 5000 to 196 in the clustering step.

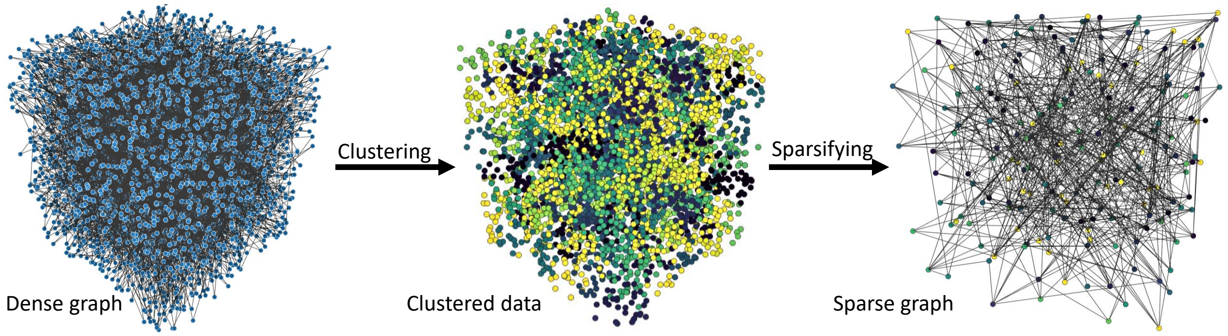

Figure 3.8: Sparsification results from synthetic data experiments. The number of graph nodes are reduced from 5000 to 196 in the clustering step, and the number of edges (after clustering) are reduced from 4269 to 524 in the graph sparsification step.

3.3.2.2 Lung cancer screening experiment:

In the lung cancer screening experiments, our participants examined volumetric chest CT scans and the corresponding data was recorded in 3D space. The qualitative results and comparison of visual search patterns of the three radiologists are reported in Fig. 3.9. Each column shows one step of the proposed algorithm and each row corresponds to a radiologist. As can be seen, dense graphs hardly reveal any comparisons between radiologists' visual search patterns. However, it is much easier to use sparsified graph (last column) for a global comparison of visual search pattern.

For a qualitative visualization on the image space (CT lungs), we showed the effect of our sparsification method on the dense eye-tracking data as well. Figure 3.10 shows the original gaze points, from 3 radiologists, on the volume renderings of corresponding lung images in the first row. The second row illustrates the timing component of visual search patterns on the whole scan as well as the selected regions (i.e., attention region is indicated by circles) for each radiologist. The third row shows sampled data points after the clustering algorithm is applied. This figure supports how successful a very dense data can be sparsified for any local/global image analysis task easily.

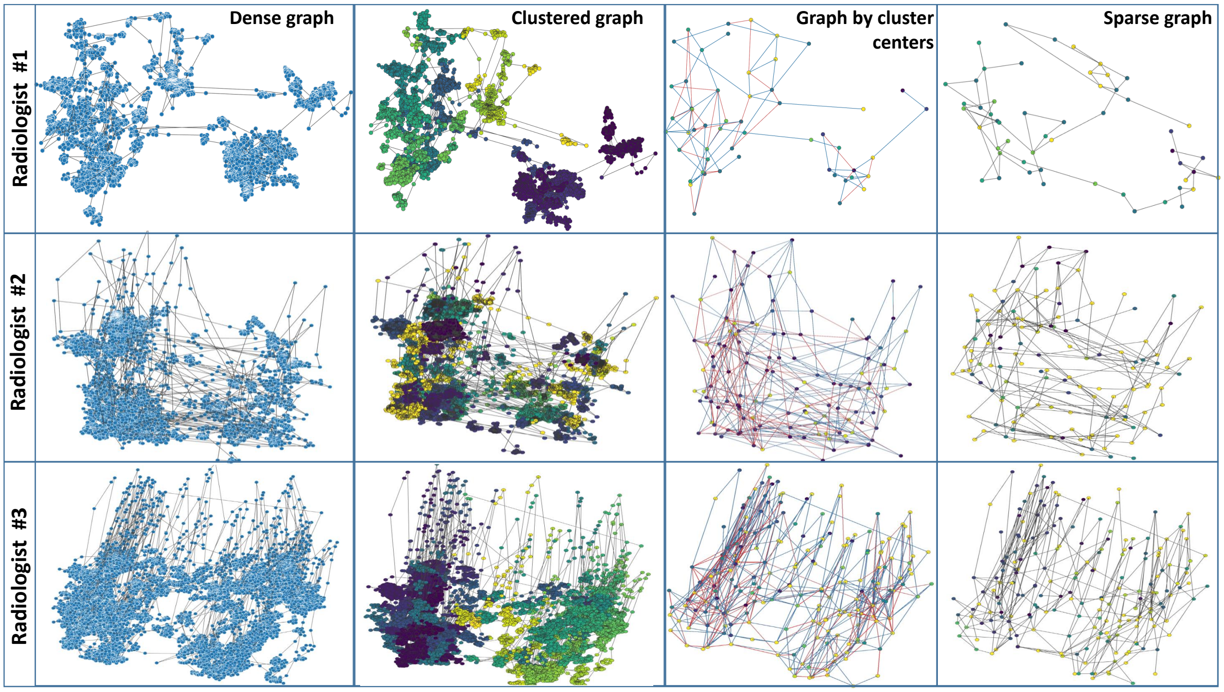

Figure 3.9: Lung cancer screening experiments with CT data. First column: dense gaze patterns. Second column: attention based clustering. Third column: nodes in clusters are reduced. Fourth column: sparse graph after further reducing edges.

3.3.2.3 Quantitative results for sparsification:

In order to compare the topology of the graphs before and after sparsification, we reported the *Diameter, Betweenness,* and *Mean Square Error (MSE)* of the graph Laplacian matrices. All these parameters are well established metrics representing structure of the graphs. *Diameter:* measures the length of maximum shortest path in a graph, *Betweenness:* is a measure of node centrality and counts the number of shortest paths that pass through a node. The Spearman's rank correlation coefficient is generally used to compare betweenness of the original and sparsified graph. *MSE:* relates the structure of the graph before and after sparsification based on the error in the Laplacian matrix of the graph. The results for lung screening data and synthetic data are plotted in Fig. 3.12 and Fig. 3.13, respectively. The above-mentioned metrics for 3 different radiologists are plotted in Fig. 3.12. Each point in the plot is computed corresponding to an edge ratio. The edge ratio is simply the ratio of edges in the graph after sparsification over the original graph. As expected, by removing more edges (edge ratio drops), betweenness and diameter metrics decrease and Laplacian MSE increases.

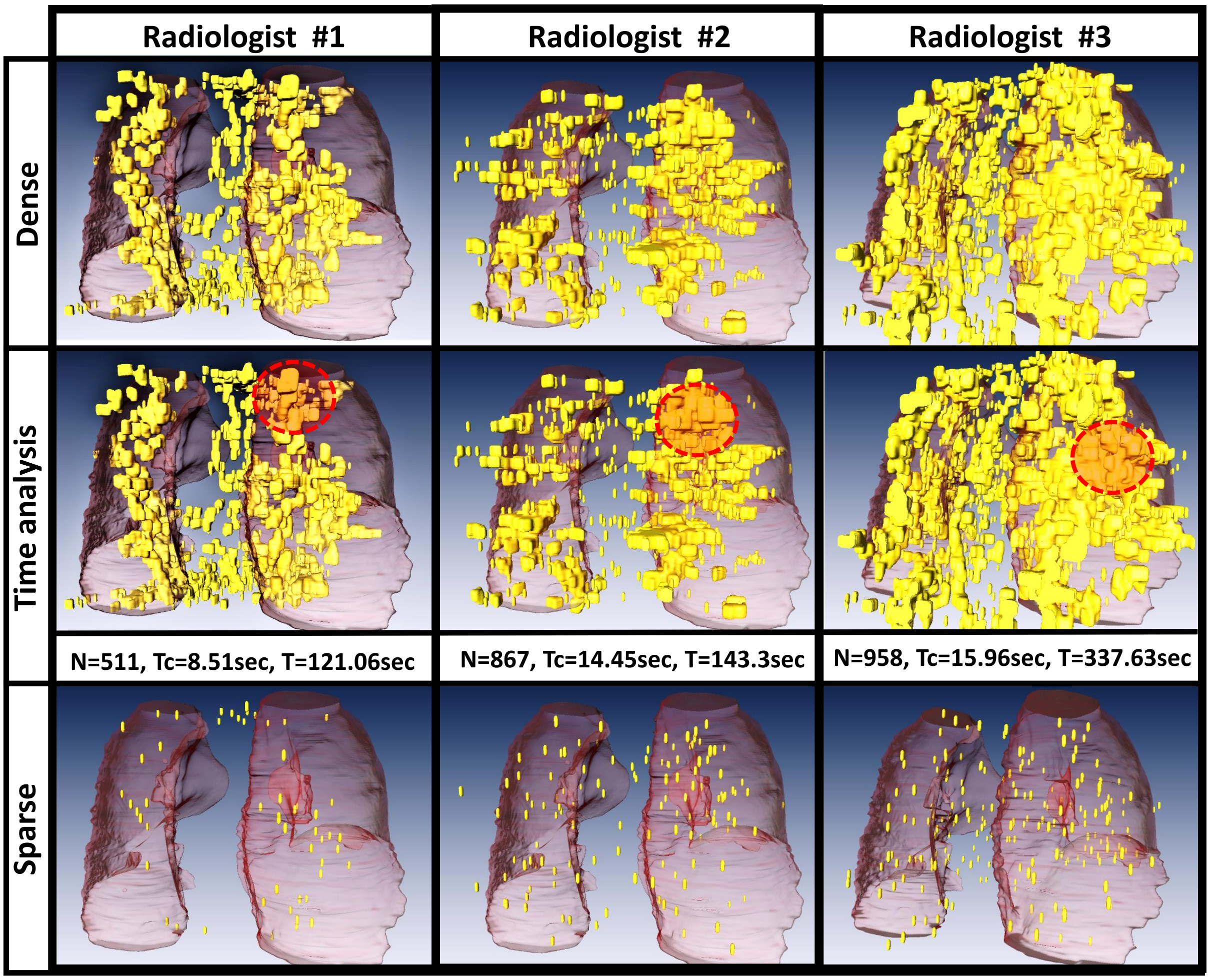
Radiologist #1
Radiologist #2
Radiologist #3
Dense
Time analysis
Sparse
N=511, Tc=8.51sec, T=121.06sec
N=867, Tc=14.45sec, T=143.3sec
N=958, Tc=15.96sec, T=337.63sec

Figure 3.10: Lung cancer screening experiment with CT data. Dense and sparse gaze points on 3D lung surface as well as time analysis. Number of nodes in the largest cluster (N), corresponding time spent by radiologist on that cluster (Tc) and overall screening time (T), with the eye-tracker frequency being $60Hz$, for each radiologist is computed.

3.3.3 Feasibility study of multi-screen eye-tracking:

As a proof-of-concept, we tested C-CAD on a multi-screen prostate MR screening experiment. Promising results show the flexibility and generalizability of our algorithm in dealing with more complex tasks including more than one screen.

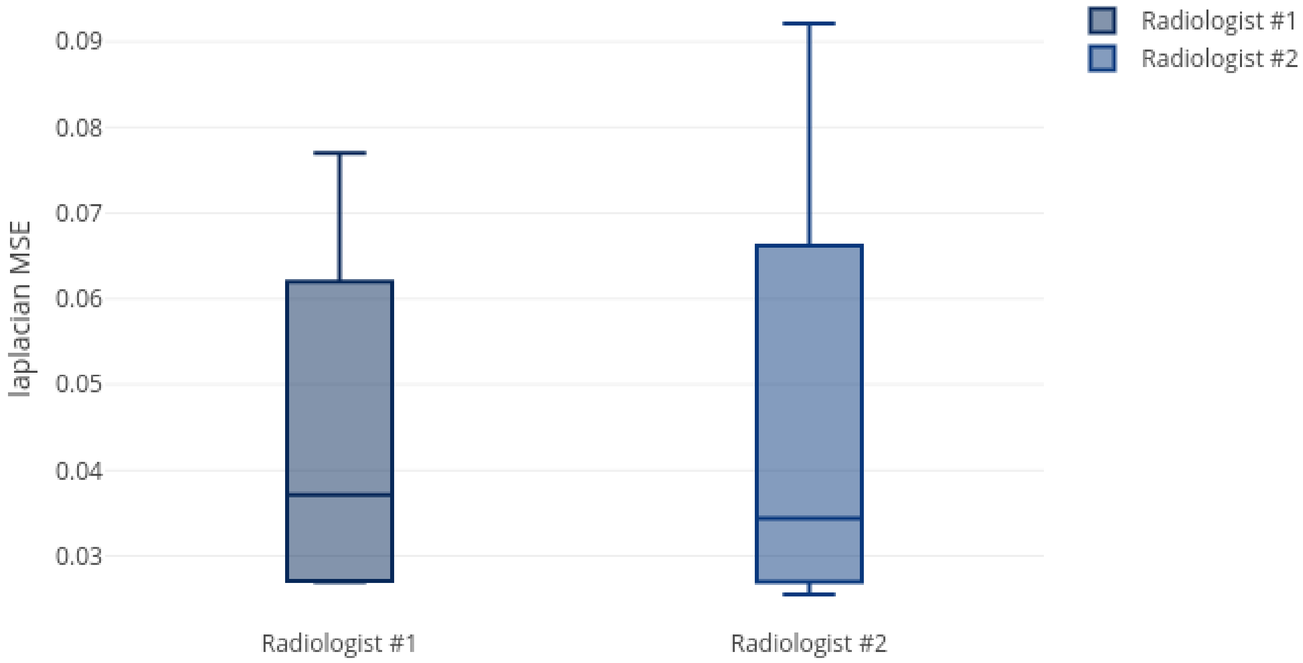

Figure 3.11: Inter-observer variation of MSE for 2 radiologists on 4 different scans.

Variation of MSE for the data recorded from two radiologists (who read 4 different scans) is plotted in Fig. 3.11. The MSEs are computed for the fixed edge ratio of 0.9 in this analysis. With a fixed edge ratio, higher MSE means that the original graph is more sparse. This further indicates that removing the same ratio of edges distorts the graph structure and, leads to a higher MSE. Hence, the radiologists' pattern of search can be compared within this variation. A higher average MSE means that the radiologist is performing a targeted search and most probably is more expert radiologist.

This experiment was performed on a multi-parametric MRI scan of a single subject. MRI characteristics are: axial T2 weighted (T2w), with FOV of 140×140 and resolution of $0.27 \times 0.27 \times 3\ mm^3$, Dynamic Contrast Enhanced (DCE) with FOV of 262×262 and resolution of $1.02 \times 1.02 \times 3$, $b = 2000 s/mm^2$, Diffusion Weighted Imaging (DWI) with FOV of 140×140 and resolution of $0.55 \times 0.55 \times 2.73 mm^3$. Apparent Diffusion Coefficient (ADC) map was derived from 5 evenly spaced b value $(0-750 s/mm^2)$ DWI.

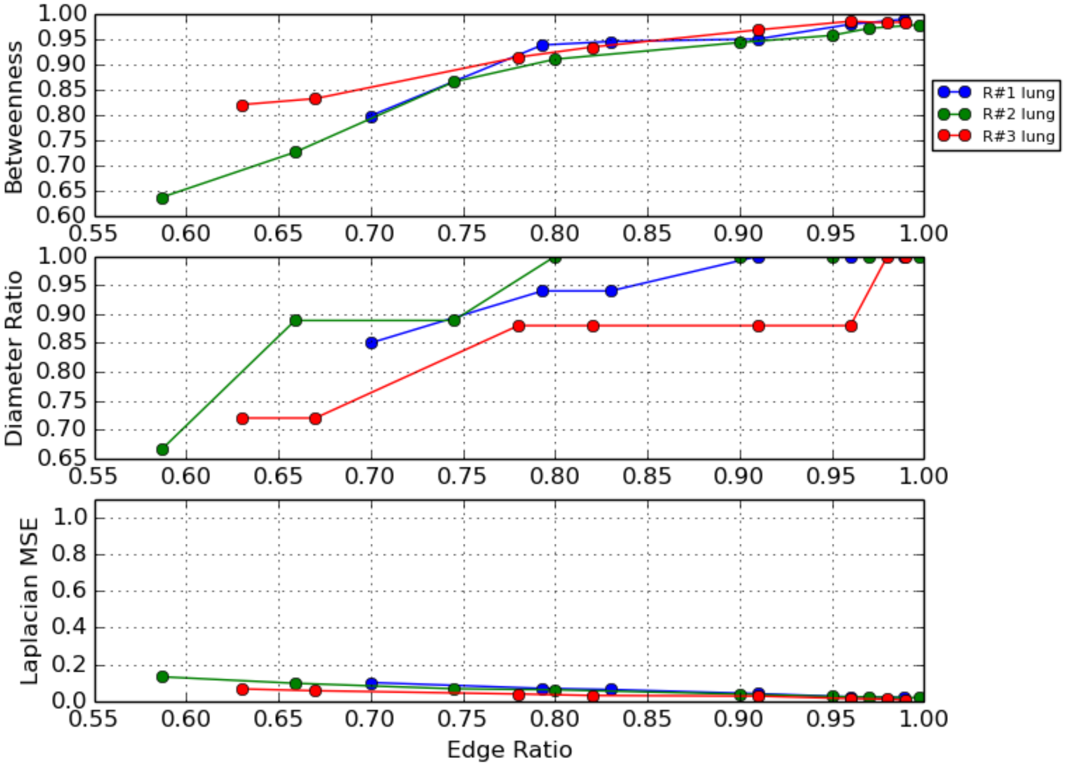

Figure 3.12: Quantitative parameters to compare graph topology between already clustered data and sparsified data with respect to the preserved edge ratio. *R#* indicates a particular radiologist (blue, green, red). (Lung cancer screening experiment)

One of our participating radiologists, an expert in prostate cancer screening, examined multi-parametric MRI (four 3D images) for routine prostate cancer screening. Based on the results reported in Fig. 3.14, it is evident by the sparsified graphs that the radiologist used axial T2-weighted images (anatomical information) and ADC maps (showing magnitude of diffusion) more frequently than other two images. This observation suggests that although all four modalities are being used for making a diagnostic decision T2-weighted and ADC map are more informative to the radiologists in the screening process. This observation can be useful in further developments of automatic computer-aided diagnosis systems.

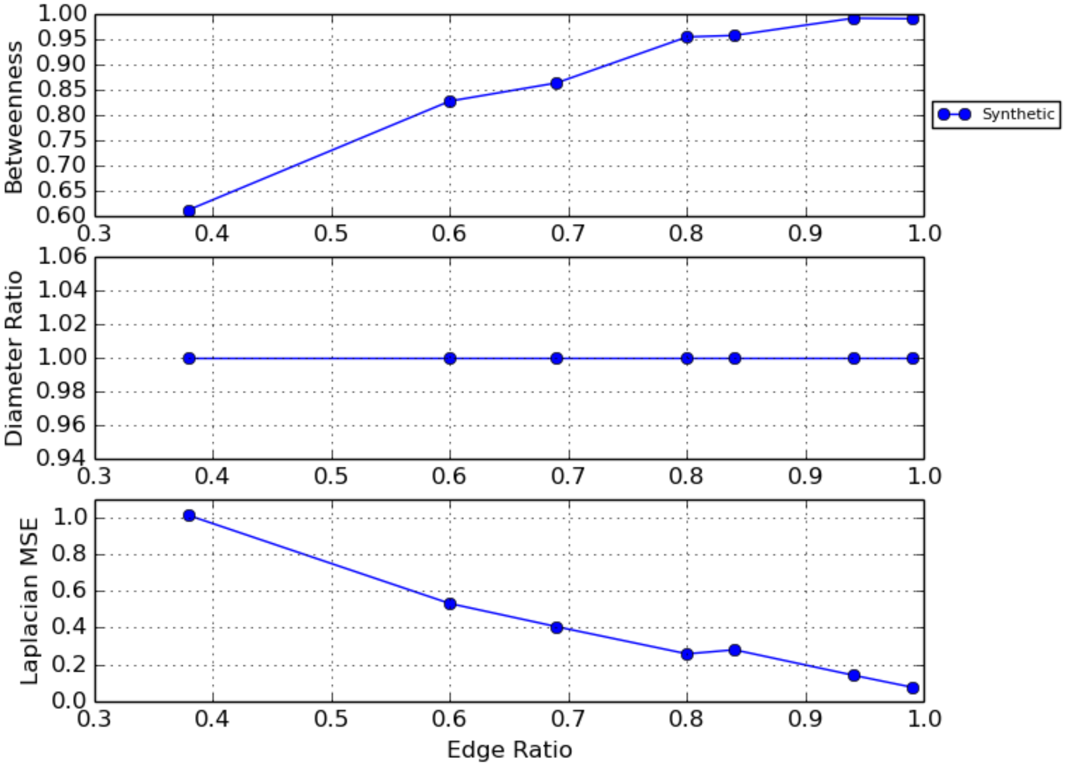

Figure 3.13: Quantitative parameters to compare graph topology between already clustered and sparsified data with respect to the preserved edge ratio. (Synthetic data experiment)

Quantitative results of our method for different modalities as well as the variation over these modalities are shown in Fig. 3.15 and Fig. 3.16, respectively. In screening, DWI and DCE modalities were used less frequently than T2-weighted and ADC modalities; therefore, the initial graph representations of the DWI and DCE are less dense compared to those of T2-weighted and ADC. For those less dense graphs, the sparsification algorithm achieved a similar MSE performance in most edge ratios larger than 0.5. From the reason that the sparsification algorithm keeps the graph in a $\sigma - spectral\ approximation$ of the original graph, we cannot remove large number of edges from less dense graphs. This situation is reflected in diameter ratio and betweenness metrics.

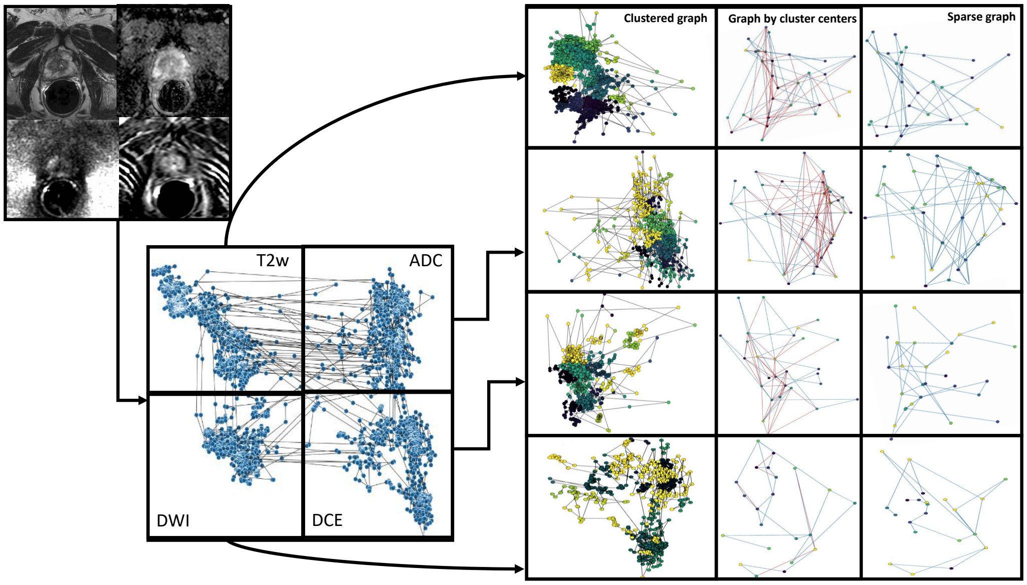

Figure 3.14: Prostate cancer screening experiments with multi-parametric MRI. Left: four MRI modalities and corresponding dense gaze patterns. Right: Clustered and sparsified gaze patterns corresponding to each modality. First column: clustered dense gaze patterns. Second column: attention based clustering. Third column: sparse graph after further reducing edges.

3.4 Summary

In this chapter, we proposed a framework that unifies eye-tracking into a CAD system. Our study offers a new perspective on eye-tracking studies in radiology because of its seamless integration into the real radiology rooms and collaborative nature of image analysis methods. First, the proposed system models the raw gaze data from eye-tracker as a graph. Then, a novel attention based spectral graph sparsification method is proposed to extract global search pattern of radiologist as well as attention regions. Later, we combine this information with our deep learning modules to perform a set of image analysis tasks. Our proposed sparsification method reduced 90% of data within seconds while keeping mean the square error under 0.1.

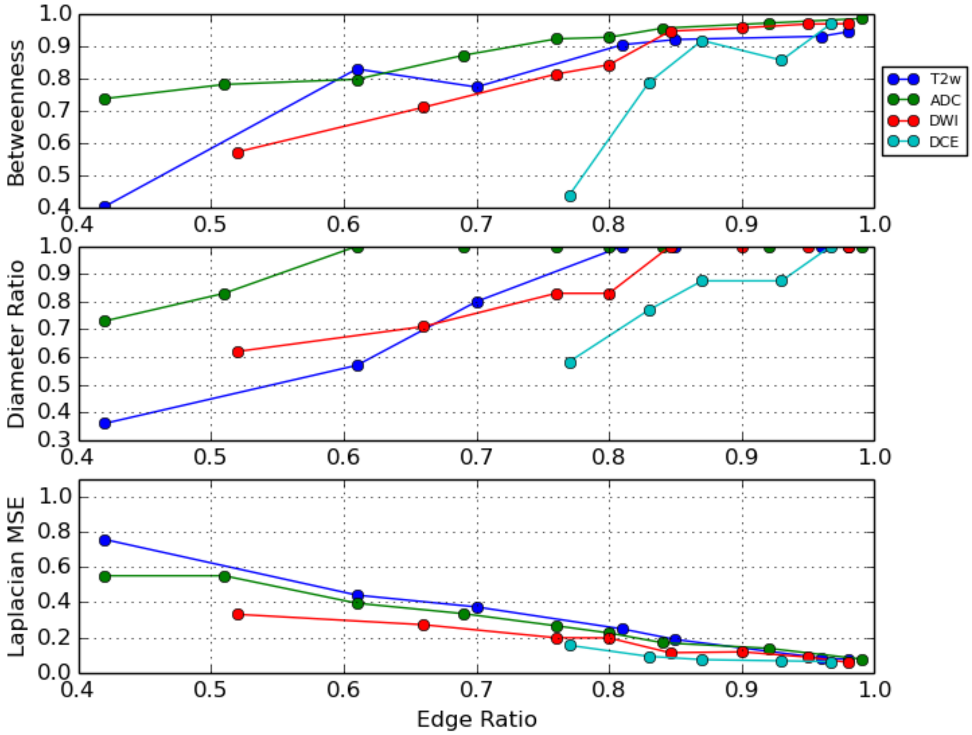

Figure 3.15: Prostate screening experiment quantitative results.

As can be interpreted from the lung screening experiment, the less experienced participant had more crowded visual search patterns and examined the most lung volume. Also, from the prostate cancer screening experiment, we observed that radiologists use anatomical/structural information more frequently than other modalities in screening (i.e., diffusion MRI). This potentially shows the importance of anatomical information in prostate cancer detection but at the same time we noticed that when anatomical information gives less clues to radiologists about potential abnormality, radiologists looks for complementary information from other imaging modalities to make inference. Our system provides visualization of this process for the first time in the literature. Scanpaths across different screens prove this observation as can be seen in Fig. 3.14.

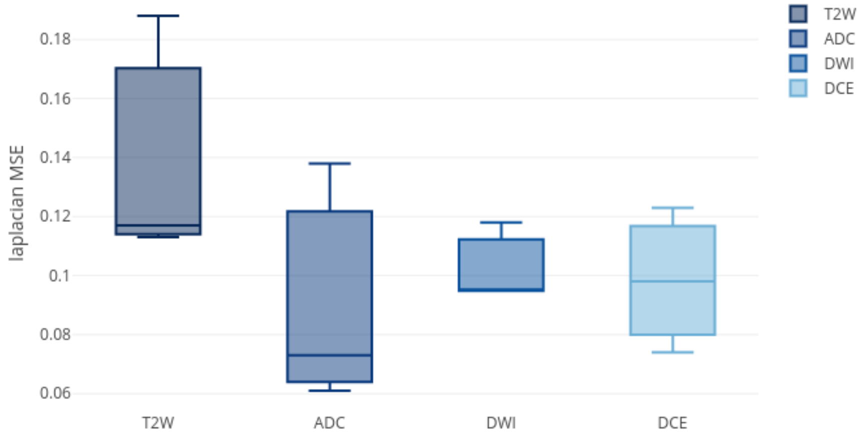

Figure 3.16: Variation of MSE on different prostate images per modality.

Our work has some limitations that should be noted. For lung cancer screening with C-CAD, our system has the assumption that the radiologists are examining only lung regions and the ROIs fall into the lung regions. If the radiologist starts focusing on some other areas, outside the lungs, the segmentation results might not be as desired, because of non-lung regions. To solve this problem, one may include a simple segmentation step into the C-CAD to restricts the ROI definition into the lungs only. However, this procedure may affect analysis of incidental findings too.

In conclusion, CAD systems are often prone to high number of false positives findings, which is one of the main drawbacks in such systems. Missing tumors, especially in their early stages, is also very common in screening. To increase the efficacy of the lung cancer screening process, we propose a novel computer algorithm, namely collaborative CAD (C-CAD). Our proposed method takes into account the gaze data of radiologists during the screening process, and incorporates this information into the CAD system for better accuracy in reducing false positive findings in particular. In return, C-CAD has the capability of improving true positive findings as well as reducing missing cases.With our proposed attention based graph sparsification method, qualitative comparison and analysis of different radiologists' visual search patterns (both locally and globally) has become feasible. Our proposed system is a promising step toward combining the efficiency of searching strategy from the expert radiologist and accuracy of analysis from deep learning methods.

Our framework is capable of integrating any image analysis block. In the following chapters we will design a set of local and global image analysis modules to be integrated intp our framework.

CHAPTER 4: LOCAL IMAGE ANALYSIS

The results of this Chapter have been published in the following papers :

- Naji Khosravan, Ulas Bagci, *“Semi-supervised multi-task learning for lung cancer diagnosis,”* in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 710-713), IEEE [40]. (This work was selected as an *Oral* presentation at the 40th IEEE EMBC conference, 2018.)
- Naji Khosravan, Haydar Celik, Baris Turkbey, Ruida Cheng, Evan McCreedy, Matthew McAuliffe, Sandra Bednarova, Elizabeth Jones, Xinjian Chen, Peter Choyke, Bradford Wood, Ulas Bagci, *“Gaze2Segment: A Pilot Study for Integrating Eye-Tracking Technology into Medical Image Segmentation,”* In Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging, pp. 94-104. Springer, Cham, 2016 [19].

In the previous chapter we developed a framework to analyze radiologists’ gaze data and extract their regions of interest (ROIs) based on the level of attention, to be integrated to a set of image analysis modules. This chapter includes our local image analysis module. We begin this chapter by describing a simple local module called *Gaze2Segment* as a proof of concept. Then, we further describe a novel deep 3D multi-task network trained in a semi-supervised manner, as our advanced processing module.

4.1 *Gaze2Segment*: A Pilot Study for Integrating Eye-Tracking Technology into Medical Image Segmentation

In this section, As a proof of concept, radiologist visual attention map is combined with a computer-derived saliency map, extracted from the gray-scale CT images. The visual attention map is used as an input for indicating roughly the location of a region of interest. With computer-derived saliency information, on the other hand, we aimed at finding foreground and background cues for the object of interest found in the previous step. These cues are used to initiate a seed-based delineation process.

4.1.1 Methodology

4.1.1.1 Local Saliency Computation for sampling Foreground/Background Cues

In biological vision, humans tend to capture/focus on most salient regions of an image. In computer vision, many algorithms have been developed to *imitate* this biological process by defining a *saliency* concept with different context. The mostly used definition of saliency is based on the distinctiveness of regions with respect to their both local and global surroundings. Although this definition is plausible for many computer vision tasks, it alone may not be suitable for defining salient regions in radiology scans where object of interests are not often as distinctive as expected. In addition, radiologists use high level knowledge or contextual information to define regions of interest. Due to all these reasons, we propose to use a *context-aware saliency* definition that aims at detecting the image regions based on contextual features[65]. In our implementation, we extracted image context information by predicting which point attracts the most attention. This step combines radiologist's knowledge with image context. The context-aware saliency explains the visual attention with feature-driven four principles, three of which were implemented in our study:

(1) local low-level considerations, (2) global considerations, (3) visual organization rules, and (4) high-level factors.

(1) For local low-level information, image was divided into local patches (p_u) centered at pixel u, and for each pair of patches, their distance ($d_{position}$) and normalized intensity difference ($d_{intensity}$) were used to assess saliency of a pixel u, as formulated below:

$$d(p_u, p_v) = d_{intensity}/(1 + \lambda d_{position}), \tag{4.1}$$

where λ is a weight parameter. Pixel u was considered *salient* when it was highly dissimilar to all other image patches, $d(p_u, p_v)$ is high $\forall v$.

(2) For global considerations, a scale-space approach was utilized to suppress frequently occurring features such as background and maintain features that deviate from the norm. Saliency of any pixel in this configuration was defined as the average of its saliency in M scales $\{(r_1, r_2, ..., r_M), r \in R\}$ as $\bar{S_u}$:

$$\bar{S_u} = (1/M) \sum_{r \in R} S_u^r \tag{4.2}$$

$$S_u^r = 1 - exp\{-(1/K) \sum_{k=1}^{K} d(p_u^r, p_v^r)\} \quad \text{for} \quad (r \in R). \tag{4.3}$$

This scale-based global definition combined K most similar patches for the saliency definition and indicated more salient pixel u when S_u^r was large.

(3) For visual organization rules, saliency was defined based on the Gestalt laws suggesting areas that were close to the foci of attention should be explored significantly more than far-away regions. Hence, assuming $d_{foci}(u)$ is the Euclidean distance between pixel u and the closest focus

of attention pixel, then the saliency of the pixel was defined as $\hat{S_u} = \bar{S_u}(1 - d_{foci}(u))$. A point was considered as a focus of attention if it is salient.

(4) High-level factors such as recognized objects can be applied as a post processing step to refine saliency definition. In our current implementation, we did not apply this consideration.

Since we inferred *where* information of object of interest from visual attention map (Step 3), we only explored *what* part of object of interest from saliency definition. Once saliency map is created, we confined our analysis into the regions indicated by corresponding visual attention maps ($a(u)$). Since saliency map includes object of interest information, we extracted foreground information from this map (called foreground cues/seeds) by simply setting the most salient pixel in this region as a foreground cue. This step helped relocating the attention gaze exactly to the center of the closest most salient object and allowed a perfect seed selection.

Furthermore, we defined the background cues for a given local region indicated by the visual saliency map as follows. We first computed the gradient information ∇I from a gray-scale CT image I. For a given visual attention map $a(u)$ and saliency map $S(u)$ at a pixel u, we employed a search starting from $\nabla I(u)$ and moving into 4 perpendicular directions. Our search was stopped soon after we passed through a high intensity value on gradient image because object boundary locations show high gradient values due to abrupt intensity changes. Those four pixels defined outside the object boundary are considered as background cues. This process is illustrated in Figure 4.1.

4.1.1.2 Lesion Segmentation

After identifying background and foreground seeds, any seed-based segmentation algorithm such as graph-cut, random walk (RW), and fuzzy connectivity, can be used to determine precise spatial extent of the object of interest (i.e., lesion). In our work, we choose to implement RW as it is fast and robust, and offers optimal image segmentation for a given set of seed points. Details of the conventional RW image segmentation algorithm can be found in [18].

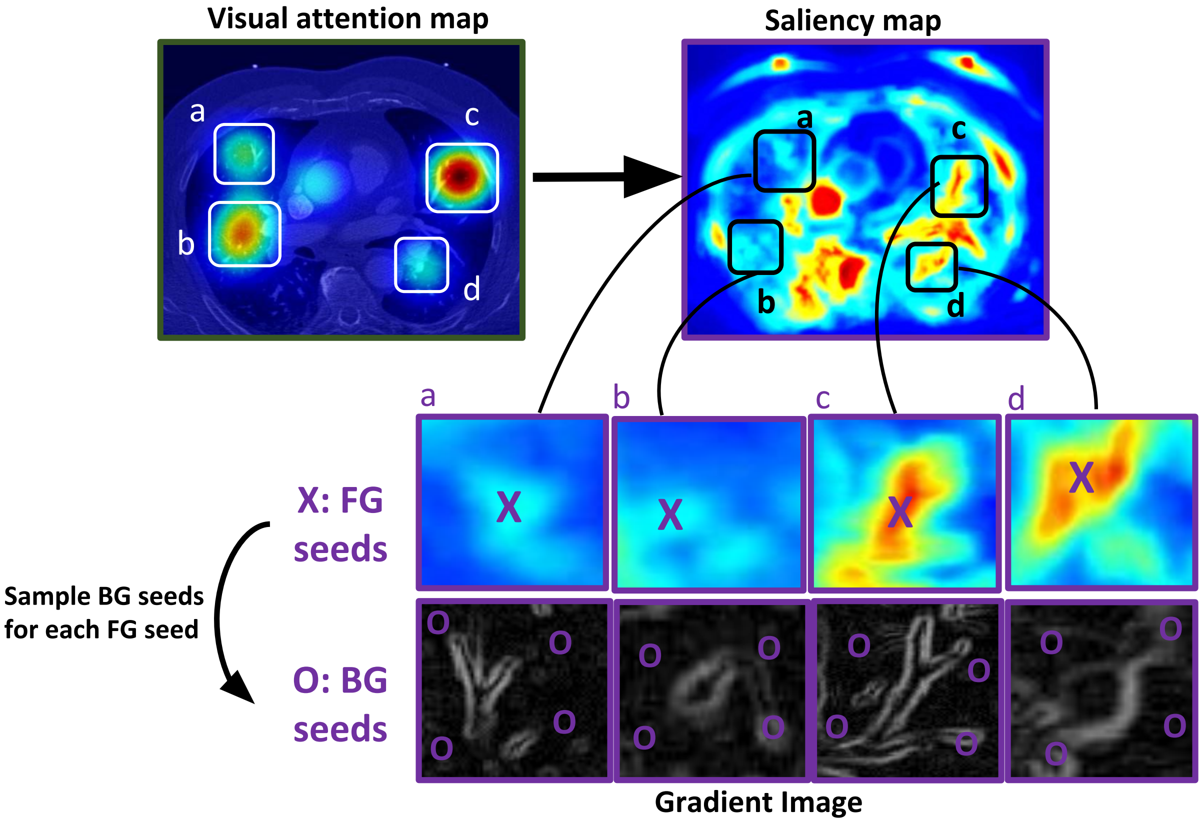

Figure 4.1: Foreground (FG) regions are obtained from visual attention maps processed from gaze information. After this *recognition* step, we identify most distinct FG seed by using the corresponding regions of saliency map. Once FG seeds are allocated, background (BG) seeds are found by using gradient information of the gray-scale CT image. For each FG seed, four perpendicular directions are searched and edge locations indicating the intensity value changes are used to select BG seeds.

4.1.2 Experiments

We tested our system on four chest CT volumes pertaining to patients diagnosed with lung cancer, evaluated by three radiologists having different levels of expertise. In-plane resolution of the image is 512×512 with a voxel size of $0.58 \times 0.58 \times 1.5\ mm^3$. Imaging data and corresponding lesion labels as well as annotations were obtained from Lung Tissue Research Consortium (LTRC) (`https://ltrcpublic.com/`) with an institutional agreement. Blind to diagnostic information of the chest CT scans, the radiologists read the scan once, and interpret the results in routine radiology rooms. Participating radiologists have more than 20, 10, and 3 years of experiences, respectively. This variability in experience levels allowed us to test robustness our system. As shown by results regardless of user experience and pattern of gaze and attention, our system perfectly captured the attention gaze locations and performed the segmentation successfully.

Figure 4.2 shows the proposed system's visual attention map, local saliency map, foreground/background seed samples, and segmentation results at different anatomical locations. Quantitatively, we used reference standards from LTRC data set and independently re-evaluated by one of the participating radiologists. We have used dice similarity coefficient (DSC) and Haussdorff Distance (HD) to evaluate accuracy of segmentation results over two reference standards. The average DSC was found to be 86% while average HD found to be 1.45 mm. We did not find statistically significant difference between segmentation results when manual seeding and interactive RW were used ($p > 0.05$).

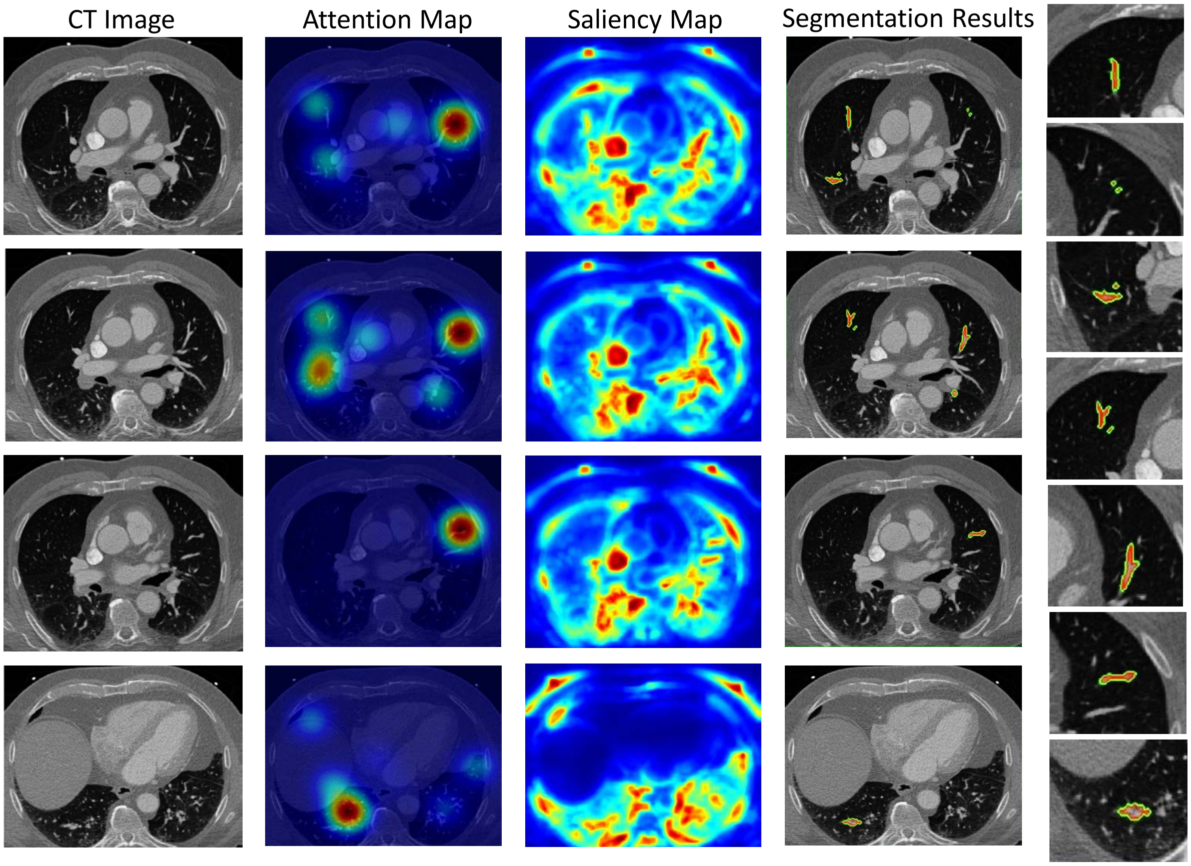

Figure 4.2: Qualitative evaluation of medical image segmentation through ***Gaze2Segment*** system is illustrated. Last column shows the segmentation results zoomed in for better illustratoin.

4.2 Semi-supervised Multi-task Learning for Local Image Analysis

Existing automated lung nodule detection systems produce a lot of false positives (FP). Hence, there is an additional step needed to further reduce these FPs. This is a fundamental component of nearly all available CAD systems in the literature [66]. In the FP reduction step, candidates are being classified as *nodule* or *non-nodule* using discriminative features. Segmentation, on the other hand, is of interest as it is the first step toward quantification and different shape/size and volume measurements. In this chapter, we argue about the use of segmentation within the FP removal step. Since a good 3D segmentation of lung nodules leads to accurate volume/shape measurement analysis in cancer screening and treatment planning, it can be used as a discriminator information for FP identification. Although some studies used different nodule attributes in a multi-task manner with pretrained networks to do nodule characterization [38], till now, none of previous studies used segmentation within a FP reduction jointly.

This chapter proposes a new methodology for addressing both *FP reduction* and *segmentation* problems, jointly. We propose a general model (Figure 5.2) that can perform both tasks with high accuracy through a multi-task learning (MTL) strategy. Our proposed model has a novel 3D deep encoder-decoder CNN architecture. We also exploit a semi supervised approach for training our model to avoid the need for large number of manual annotations for 3D segmentation masks. **Our contributions**, in this Section, can be specified as: **1)** This is the first study to propose joint segmentation and FP reduction of lung nodules through a fully 3D CNN, which is a critical step toward using CAD systems efficiently in clinical applications. **2)** Our work opens a door to possible improvements of CAD systems via a MTL approach. **3)** This work will generate fresh insight on how to tackle the problem of lack of available annotated medical image data through a semi-supervised learning method, which is more efficient if used along with MTL.

4.2.1 Methodology

The proposed 3D deep MTL algorithm is based on Convolutional Neural Network (CNN) and learns segmentation and FP reduction through some *shared and task specific layers*. The proposed architecture along with the training strategy is illustrated in Fig.5.2. In the rest of this section, we explain the proposed framework step by step.

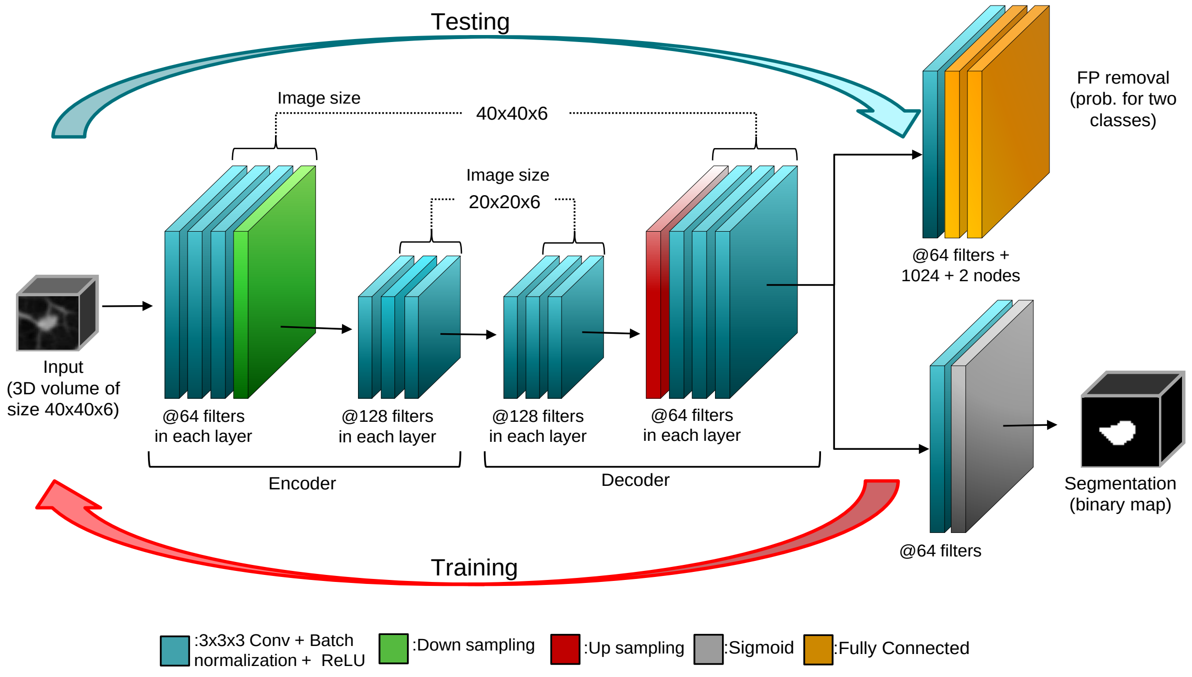

Figure 4.3: The 3D deep multi-task CNN architecture. The size of all convolution kernels is set to $3 \times 3 \times 3$ with a stride of 1 in each dimension. The downsampling and upsampling operators are performed only in the xy plane. All convolution layers are $3D$. The network has 14 shared layers, 3 FP removal specific layers, and 2 segmentation specific layers. Red and blue arrows show the semi-supervised learning paradigm to train the proposed network.

4.2.1.1 Multi-Task Learning

MTL allows solving multiple learning tasks at the same time by optimizing multiple loss functions instead of one [67]. MTL can be beneficial in multiple senses: (1) *Generalization ability:* in MTL, a single model can be used to perform multiple tasks at the same time. Such as, in our case, it is desirable to have one general model, with the same accuracy if not better, instead of having multiple separate models. (2) *Highlighting underlying features:* depending on the selection of the tasks, features learned from one task can act discriminative for other tasks as well. These features might not always be easy to learn by a single task network due to their complexity or more discriminating effect of other features. However, learning multiple tasks jointly can strengthen the effect of these underlying features and boost the performance on one or all tasks. (3) *Dealing with lack of data:* in radiology field, it is not easy to gather large number of annotated data for training deep networks. An MTL model can benefit each task during training due to actively sharing features in relevant tasks.

The problem of jointly learning multiple tasks can be formulated as follows. Assume that we have N supervised tasks. The training set for each task can be considered as $D_n = (x_{in}, y_{in})$. In which $i = 1 : k_n$, where k_n is the number of training samples for the n_{th} task. With $x_{in} \in X^{(n)}$ and $y_{in} \in Y^{(n)}$ the problem of learning multiple tasks, jointly, can be narrowed down to the optimization problem of:

$$\min_{w} \sum_{n=1}^{N} L(Y^{(n)}, f(X^{(n)}) + \lambda \|f\|, \tag{4.4}$$

where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{+}$ is the loss function measuring the per-task prediction error, f is the multi-task model and w is the model's parameter set. In our study, we use cross entropy as loss function for both tasks. Cross entropy, also known as negative log likelihood, measures the similarity between two probability distributions and conventionally defined as:

$$L(Y^{(n)}, f(X^{(n)}) = \sum_{i=1}^{k_n} -y_i \log(f(x_i)), \tag{4.5}$$

where y_is are the true labels and $f(x_i)$s are the predictions for each task. To optimize equation 4.4, ADAM optimizer was used with an initially selected learning rate of 10^{-3}.

Since morphology (i.e., size, volume, and shape) information plays a key role in screening, diagnosis, and prognosis, we earlier postulate that this information can be effectively used for FP rejection, which is a significant challenge for most CADs. There is a strong need for reducing those findings (FPs) because it tremendously increases the workload of radiologists. We proved in the following that an MTL based CAD system can solve these two problems jointly: segmenting nodules while deciding whether they are FP or not. We believe that once the shape and appearance information can be highlighted in the shared layers of a network, other task specific layers can also learn if the nodule is a true nodule or not. In other words, features for classification and segmentation are combined through shared layers of the proposed network. To our best, this is the first study conducting this for both FP removal and segmentation.

4.2.1.2 Architecture

The inputs to our network are 3D volumes and the outputs are probabilities of each volume belonging to class of nodules or non-nodules. Our second output is a binary segmentation mask for those nodules. Our network has 19 layers: the first 14 layers are trained on both tasks, 5 task specific layers (2 for segmentation, 3 for classification) are trained only on one of the tasks. Each convolution layer in the architecture consists of a set of 3D convolution kernels (with size of $(3,3,3)$ and a stride of 1) following by a batch normalization (BN) and a rectified linear unit activation (ReLU). Number of kernels in each layer is depending on its location in the architecture. A max-pooling layer with the kernel size of $(2,2)$ is used to perform down-sampling in the encoder. A bilinear interpolation is used for the up-sampling images in the decoder.

Our network *forks* after $14th$ layer into two branches (see Fig.5.2). Segmentation specific branch contains a convolution layer following by a sigmoid layer, which produces binary masks. FP reduction branch contains a convolution layer followed by two fully connected layer. The fully connected layers have 1024 and 2 nodes, respectively, and output the probability of each patch belonging to each one of classes (nodule vs. non-nodule).

4.2.1.3 Semi-supervised training

Due to the large number of parameters, deep CNNs need a large amount of annotated data to be trained efficiently. However, finding a large amount of such data is very challenging and expensive, specifically in the field of medical imaging. Semi-supervised learning methods are one way to address such issues. In semi-supervised methods, the model is initially trained on the part of data set which has labels. This model is then used to estimate labels for unlabeled data, which will be used to refine the model. The algorithm for semi-supervised learning strategy is illustrated in Algorithm 2. It can be argued that semi-supervised approach, if utilized naively, can lead to error propagation in the model and even cause worse performance. This problem, however, can be solved by iteratively performing prediction and training on small portions of unlabeled data and improving performance step by step. Constant improvements of results in our case supports that our algorithm perfectly handles error propagation and outperforms the baseline.

```
Input  : labeled data: $(X_l, Y_l)$, unlabeled data: $X_u$
Train model $f$ on $(X_l, Y_l)$
for $x$ in $X_u$ do
    Predict on $x \in X_u$
    Add $(x, f(x))$ to labeled data
    Retrain model $f$
end
Return refined model $f$;
```

Algorithm 2: Semi-Supervised training algorithm.

4.2.2 Experiments

Data: To evaluate our network we used Lung Nodule Analysis (LUNA16) Challenge dataset [68]. This dataset is gathered from the largest publically available LIDC-IDRI dataset. Scans with a slice thickness greater than 2.5 mm were excluded from the dataset leaving a total of 888 chest

CT scans. The dataset contains the location of nodules accepted by at least 3 out of 4 radiologists leading to a total of 1186 nodule annotations. We performed our experiments on a total number of more than $500,000$ candidate locations provided by the dataset for the FP reduction task, which are a combination of outputs of candidate generation methods in the literature. This dataset is divided into 10 subsets by the provider. We performed 10-fold cross validation to evaluate our method. To handle the unbalance ratio between nodules and non-nodules we performed data augmentation on the nodules (shift in 6 directions). It should be mentioned that the number of segmentation masks available for this study was only **270** out of 1186 total nodule annotations and the masks for the rest (916 nodules) was created using the proposed semi-supervised strategy.

Segmentation: We used Dice Similarity Coefficient (DSC) as the metric to measure segmentation accuracy. To show the improvements, we compared the final model to 2 baselines of our model. Learning curves are plotted in Fig 4.4. In first baseline, we trained the model as a single task model using only the portion of annotated data which is available (depicted as single-manual ground truth (GT) in the plot-green). In second baseline, we trained the model jointly on both segmentation and FP reduction tasks as a MTL network with the same manual GT (depicted as joint-manual GT in the plot-pink). This multi-task model was used to generate annotations for the rest of the dataset. Next, we trained the model using the semi-supervised approach (depicted as joint-combined GT in the plot-blue). Note that we trained all models from scratch.

As shown, MTL based network outperforms single task based methods and semi-supervised approach improves results of MTL further. Our network reaches a DSC of **91%** compared to the baseline which does not go beyond **82%**.

FP reduction: To observe the effect of proposed semi-supervised MTL method on FP reduction performance, we compared the learning curves of three training strategies as follows. Single task network trained to only perform FP reduction (depicted as Single-green), Multi-Task using only manual GT available for the segmentation (depicted as Joint-Manual GT in the plot-pink) and Multi-Task using semi-supervised approach (depicted as Joint-Combined GT in the plot-blue). Figure 4.4 shows sensitivity through training epochs. As expected improvements are observed in the classification results (from **88%** to **98%**).

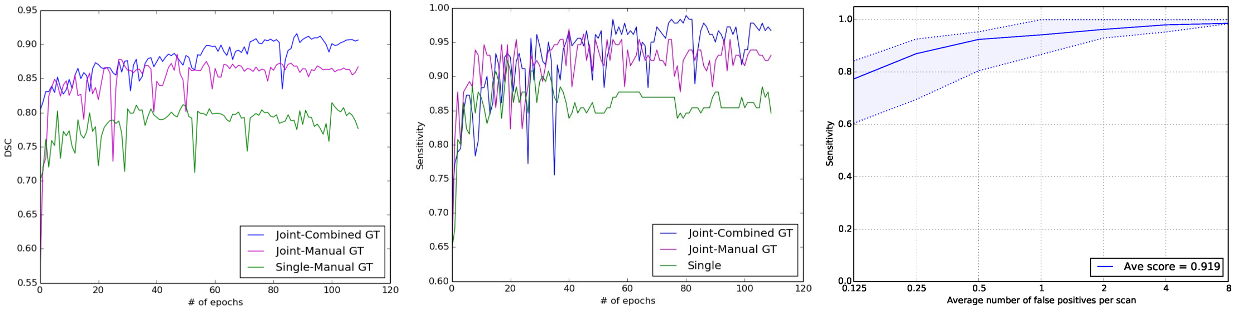

Figure 4.4: Comparison of two baselines with proposed method. First baseline is single task network, second is semi-supervised MTL. *Left*: Dice similarity coefficient over first 100 learning epochs is shown. *middle*: Showing sensitivity for FP reduction task over the first 100 epochs. Improvement of segmentation through different training strategies are depicted. *Right*: is showing the FROC curve.

Our results also show that, beside improving segmentation results, using a semi-supervised approach benefits FP reduction task as well (Joint-Combined GT in the plot-blue). This supports our rationale behind proposing a multi-task network strongly by showing that a better segmentation, which is highlighting shape and appearance information better, helps the other relevant task (FP reduction). Summary of the best performance on each task using different learning strategies is illustrated in Table 5.1.

Table 4.1: Dice similarity coefficient and sensitivity for three different learning methods is shown.

Training strategy	DSC%	Sensitivity%
Single task	82%	88%
Multi task (manual GT)	86%	95%
Semi-Supervised multi task	**91%**	**98%**

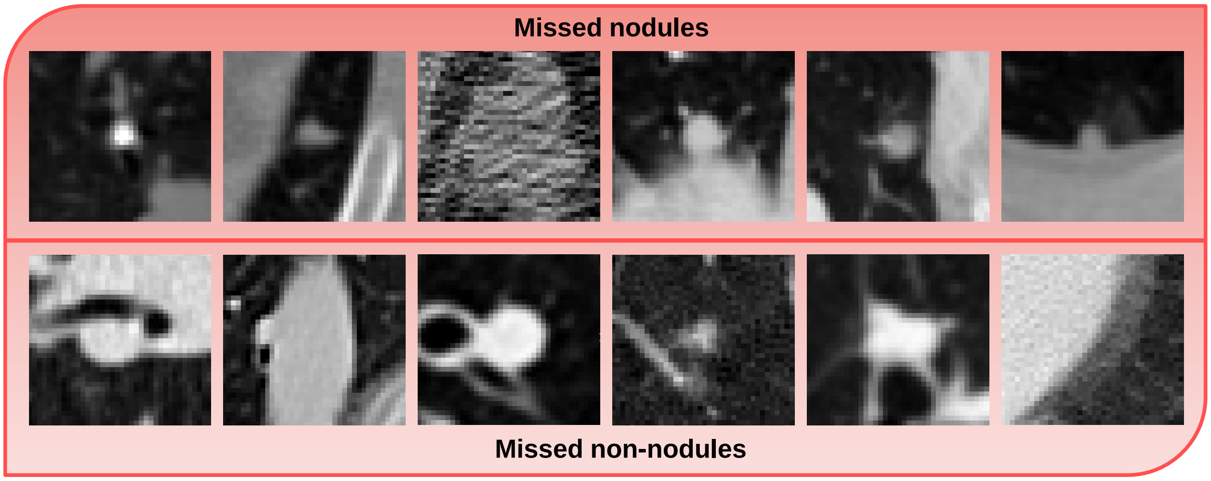

Figure 4.5: Limitation of our system/failing cases: The first row shows 6 examples of missing nodules. The bottom row shows some examples of non-nodules which are mistakenly considered as nodules.

Furthermore, to have a more accurate evaluation of our system, we used Free-Response Receiver Operating Characteristic (FROC) analysis [69]. Sensitivity at 7 FP/scan rates (i.e. $0.125, 0.25, 0.5, 1, 2, 4, 8$) is computed and the corresponding results are plotted in Fig. 4.4. The overall *score* of system is defined as the average sensitivity for these 7 FP/scan rates. Our network achieved an average score of $\sim$**92%** (see Table.4.2).

Table 4.2: System performance in terms of sensitivity based on number of FPs/scan.

FPs/scan	0.125	0.25	0.5	1	2	4	8	Average
Sensitivity	0.773	0.870	0.924	0.941	0.962	0.980	0.986	**0.919**

4.3 Summary

In this chapter, we proposed a 3D deep multi-task CNN for simultaneously performing segmentation and FP reduction. We showed that sharing some underlying features for these tasks and training a single model using shared features can improve the results for both tasks, which are critical for lung cancer screening. Furthermore, we showed that a semi-supervised approach can improve the results without the need for large number of labeled data in the training. It should be also note that there are some cases that our algorithm missed for FP reduction task. We illustrated some of those rarely seen examples of missing cases in Fig.4.5. One reason seems to be the small size of the missed nodule. Alternatively, very similar appearance of missing cases to other abnormalities and normal lung parenchyma. As an alternative direction to semi-supervised approach, one may use GAN to generate realistic data. One recent study created realistic nodules to support this idea [70].

The module introduced in this chapter will analyze images locally, in the radiologists' ROIs, and help them with diagnosis decisions regarding the abnormalities in their plain sight. In the next chapter we will introduce our global image analysis modules. The global modules will analyze image as a whole and will help the radiologists with the missing cases.

CHAPTER 5: GLOBAL IMAGE ANALYSIS

The results of this Chapter have been published in the following papers:

- Naji Khosravan and Ulas Bagci, *"S4ND: Single-Shot Single-Scale Lung Nodule Detection,"* in Proceedings of the 21st International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2018) (pp. 794-802)[71] (This work has received significant attention and was featured in more than 25 press articles). A corresponding US patent on this work was filed, and is pending.
- Naji Khosravan, Aliasghar Mortazi, Michael Wallace, Ulas Bagci, *"PAN: Projective Adversarial Network for Medical Image Segmentation,"* in Proceedings of the 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2019)[72]

In the previous chapter, we introduce a multi-task local image analysis for analyzing radiologists' ROIs. Other than analysis of ROIs, it's necessary to analyze medical images globally for the abnormalities that might be missed by the radiologists sue to small size, similarity to neighbouring areas or unusual location.

In this chapter, we'll introduce two other image analysis modules, which will handle this global analysis task. The chapter starts with our detection module, which will handle detection of small abnormalities that might be missed by the radiologist. Further, we continue the chapter by introducing a module for segmentation of complex 3D organs, which will help radiologists in guiding their screening in more high risk areas.

5.1 *S4ND*: Single-Shot Single-Scale Lung Nodule Detection

The most recent lung nodule detection studies rely on computationally expensive multi-stage frameworks to detect nodules from CT scans. To address this computational challenge and provide better performance, in this section of this chapter, we propose S4ND, a new deep learning based method for lung nodule detection. Our approach uses a single feed forward pass of a single network for detection. The whole detection pipeline is designed as a single $3D$ Convolutional Neural Network (CNN) with dense connections, trained in an end-to-end manner. S4ND does not require any further post-processing or user guidance to refine detection results. Experimentally, we compared our network with the current state-of-the-art object detection network (SSD) in computer vision as well as the state-of-the-art published method for lung nodule detection (3D DCNN). We used publicly available 888 CT scans from LUNA challenge dataset and showed that the proposed method outperforms the current literature both in terms of efficiency and accuracy by achieving an average FROC-score of 0.897. We also provide an in-depth analysis of our proposed network to shed light on the unclear paradigms of tiny object detection.

5.1.1 Methodology

Fig. 5.4 shows the overview of the proposed method for lung nodule detection in a single shot. The input to our network is a $3D$ volume of a lung CT scan. The proposed $3D$ densely connected Convolutional Neural Network (CNN) divides the input volume into a grid of size $S \times S \times T$ cells. We model lung nodule detection as a cell-wise classification problem, done simultaneously for all the cells. Unlike commonly used region proposal networks, our proposed network is able to reason the presence of nodule in a cell using global contextual information, based on the whole 3D input volume.

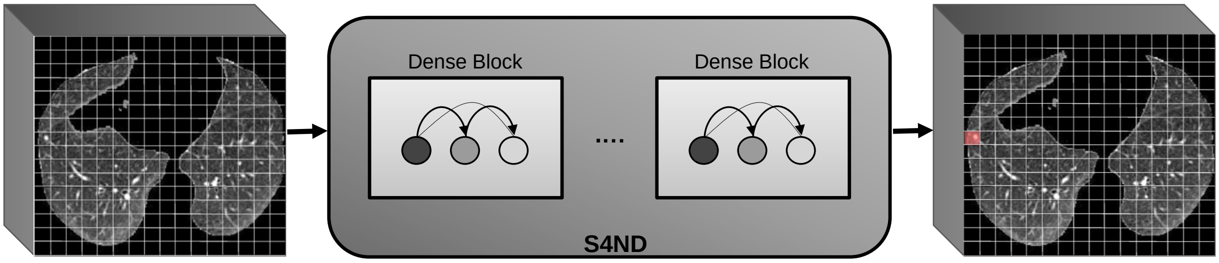

Figure 5.1: Our framework, named S4ND, models nodule detection as a cell-wise classification of the input volume. The input volume is divided by a $16 \times 16 \times 8$ grid and is passed through a newly designed 3*D* dense CNN. The output is a probability map indicating the presence of a nodule in each cell.

5.1.1.1 Single-Scale Detection

As opposed to object detection in natural scenes, we show that lung nodule detection can be performed efficiently and with high accuracy in a single scale. Current literature reports the most frequently observed nodule sizes fall within 32 *mm* by 32 *mm* [68], most of which are less than 9 *mm* and are considered as small (def. American Thoracic Society). Nodules less than 3 *mm* in size are the most difficult to detect due to their tiny nature and high similarities to vessels. Based on the statistics of nodule size and the evidence in literature, we hypothesize that a single scale framework with the grid size that we defined ($16 \times 16 \times 8$ leading to the cell sized of $32 \times 32 \times 8$ on a volume of size $512 \times 512 \times 8$) is sufficient to fit all the expected nodule sizes and provide good detection results without the need to increase the algorithmic complexity to multi-scale. This has been partially proven in other multi-scale studies [43].

5.1.1.2 Dense and Deeper Convolution Blocks Improve Detection

The loss of low-level information throughout a network causes either a high number of false positives or low sensitivity. One efficient way that helps the flow of information in a network and keeps this low-level information, combining it with the high level information, is the use of dense connections inside the convolution blocks. We empirically show that deeper densely-connected blocks provide better detection results. This, however, comes with the cost of more computation. In our experiments we found that dense blocks with 6 convolution layers provide a good balance of detection accuracy and computational efficiency.

5.1.1.3 Max-Pooling Improves Detection

As we go deeper in a CNN, it is desired to pick the most descriptive features and pass only those to the next layers. Recently, architectures for object detection in natural images preferred the use of convolutions with stride 2 instead of pooling [73]. In the context of tiny object detection, this feature reduction plays an important role. Since our objects of interest are small, if we carelessly pick the features to propagate we can easily lose the objects of interest through the network and end up with a sub-optimal model. In theory, the goal is to have as less pooling as possible. Also, it is desired to have this feature sampling step in a way that information loss is minimized. There are multiple approaches for sampling information through the network. Average pooling, max pooling and convolutions with stride 2 are some of the options. In our experiments, we showed that max pooling is the best choice of feature sampling for our task as it selects the most discriminative feature in the network. Also, we showed that convolution layers with stride of 2 are performing better compared to average pooling. The reason is that convolution with stride 2 is very similar in its nature to weighted averaging with the weights being learned in a data driven manner.

5.1.1.4 Proposed 3D Deep Network Architecture

Our network architecture consists of 36, 3*D* convolution layers, 4 max-pooling layers and a sigmoid activation function at the end. 30 of convolution layers form 5 blocks with dense connections and without pooling, which enhance low-level information along with high-level information, and the remainder form the transition layers. The details of our architecture can be seen in Fig. 5.2. The input to our network is $512 \times 512 \times 8$ and the output is a $16 \times 16 \times 8$ probability map. Each cell in the output corresponds to a cell of the original image divided by a $16 \times 16 \times 8$ grid and decides whether there is a nodule in that cell or not.

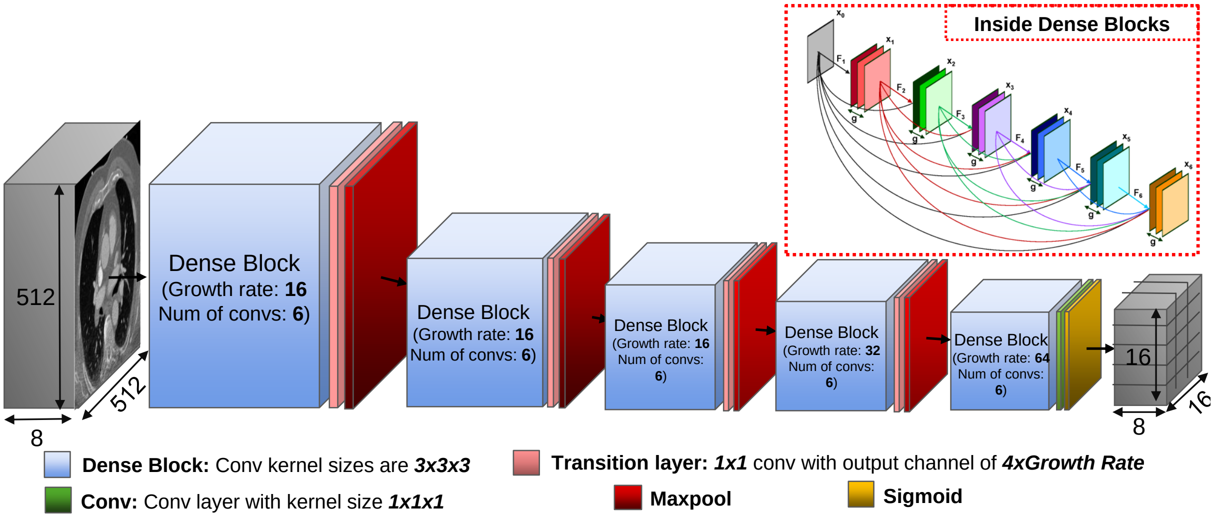

Figure 5.2: Input to the network is a $512 \times 512 \times 8$ volume and output is a $16 \times 16 \times 8$ probability map representing likelihood of nodule presence. Our network has 5 dense blocks each having 6 conv. layers. The growth rates of blocks 1 to 5 is $16, 16, 16, 32, 64$ respectively. The network has 4 transition layers and 4 max-pooling layers. The last block is followed by a convolution layer with kernel size $1 \times 1 \times 1$ and output channel of 1 and a sigmoid activation function.

Densely connected convolution blocks: As stated, our network consists of 5 densely connected blocks, each block containing 6 convolution layers with an output channel of g, which is the growth rate of that block. Inside the blocks, each layer receives all the preceding layers' feature maps as inputs. Fig. 5.2 (top right) illustrates the layout of a typical dense block. Dense connections help the flow of information inside the network. Assume x_0 is the input volume to the block and x_i is the output feature map of layer i inside the block. Each layer is a non-linear function F_i, which in our case is a composition of convolution, batch normalization (BN) and rectifier linear unit (ReLU).

With dense connections, each layer receives a concatenation of all previous layers' feature maps as input $x_i = F_i([x_0, x_1, ..., x_{i-1}])$, where x_i is the output feature map from layer i and $[x_0, x_1, ..., x_{i-1}]$ is the channel-wise concatenation of previous layers' feature maps.

Growth rate (GR): is the number of feature maps that each layer F_i produces in the block. This number is fixed for each block but it can change from one block to the other. Assume the number of channels in the input layer of a block is c_0 and the block has i convolution layers with a growth rate of g. Then the output of the block will have $c_0 + (i-1)g$ channels.

Transition layers: as can be seen in the above formulations, the number of feature maps inside each dense block increases dramatically. Transition layers are $1 \times 1 \times 1$ convolution layers with $4 \times g$ output channels, where g is the growth rate of previous block. Using a convolution with kernel size of $1 \times 1 \times 1$ compresses the information channel-wise and reduces the total number of channels throughout the network.

Training the network: The created ground truths for training our network are $3D$ volumes with size $16 \times 16 \times 8$. Each element in this volume corresponds to a cell in the input image and has label 1 if a nodule exists in that cell and 0 otherwise. The design of our network allows for an end-to-end training. We model detection as a cell wise classification of input which is done in one feed forward path of the network in one shot. This formulation detects all the nodules in the given volume simultaneously. The loss function for training our network is weighted cross-entropy defined as:

$$L(Y^{(n)}, f(X^{(n)}) = \sum_{i=1}^{k_n} -y_i \log(f(x_i)), \tag{5.1}$$

where Ys are the labels and Xs are the inputs.

5.1.2 Experiments

Data and evaluation: To evaluate detection performance of S4ND, we used Lung Nodule Analysis (LUNA16) Challenge dataset (consisting of a total of 888 chest CT scans, slice thickness<2.5 mm, with ground truth nodule locations). For the training, we performed a simple data augmentation by shifting the images in 4 directions by 32 pixels. We sampled the 3D volumes for training so that nodules appear in random locations to avoid bias toward location of nodules. We performed 10-fold cross validation to evaluate our method by following the LUNA challenge guidelines. Free-Response Receiver Operating Characteristic (FROC) analysis has been conducted to calculate sensitivity and specificity [69]. Suggested by the challenge organizers, sensitivity at 7 FP/scan rates (i.e. $0.125, 0.25, 0.5, 1, 2, 4, 8$) was computed. The overall *score* of system (Competition Performance Metric-CPM) was defined as the average sensitivity for these 7 FP/scan rates.

Building blocks of S4ND and comparisons: This subsection explains how we build the proposed S4ND network and provides a detailed comparison with several baseline approaches. We compared performance of S4ND with state-of-the-art algorithms, including SSD (single-shot multibox object detection) [73], known to be very effective for object detection in natural scenes. We show that SSD suffers from low performance in lung nodule detection, even though trained from scratch on LUNA dataset. A high degree of scale bias and known difficulties of the lung nodules detection (texture, shape, etc.) in CT data can be considered as potential reasons. To address this poor performance, we propose to replace the convolution layers with *dense* blocks to improve the information flow in the network. Further, we experimentally tested the effects of various down sampling techniques. Table 5.1 shows the results of different network architectures along with the number of parameters based on these combinations. We implemented the SSD based architecture with 3 different pooling strategies: (1) average pooling (2D Dense Avepool), (2) replacing pooling layers with convolution layers with kernel size 3×3 and stride 2 (2D Dense Nopool) and (3) max pooling (2D Dense Maxpool). Our experiments show that max pooling is the best choice of feature sampling for tiny object detection as it selects the most discriminating feature in each step. *2D Dense Nopool* outperforms the normal average pooling (*2D Dense Avepool*) as it is in concept a learnable averaging over 3×3 regions of our network, based on the way we defined kernel size and stride.

Table 5.1: Comparison of different models with varying conditions.

	Model	Sensitivity%	Num of parameters	CPM
Randomly selected 1-fold	2D SSD	77.8%	59,790,787	0.649
	2D Dense Avepool	84.8%	67,525,635	0.653
	2D Dense Nopool	86.4%	70,661,955	0.658
	2D Dense Maxpool	87.5%	67,525,635	0.672
	3D Dense	93.7%	694,467	0.882
	3D Increasing GR	95.1%	2,429,827	0.890
	3D Deeper Blocks	94.2%	1,234,179	0.913
	Proposed (S4ND)	**97.2%**	4,572,995	**0.931**
10-fold	3D DCNN [42]	94.6%	11,720,032	0.891
	Proposed (S4ND)	**95.2%**	4,572,995	**0.897**

3D Networks, growth rate (GR), and comparisons: We implemented S4ND in a completely 3D manner. Growth rate for all the blocks inside the network was initially fixed to 16 (3D Dense). However, we observed that increasing the growth rate in the last 2 blocks of our network, where the computational expense is lowest, (from 16 to 32 and 64, respectively) improved the performance of detection (3D Increasing GR in Table 5.1). Also, having deeper blocks, even with a fixed growth rate of 16 for all the blocks, help the information flow in the network and improved the results further (3D Deeper Blocks in Table 5.1). The final proposed method benefits from both deeper blocks and increasing growth rate in its last two blocks. Fig. 5.3 (left) shows the FROC comparison of proposed method with the baselines. The 10-fold cross validation results were compared with the current state of the art lung nodule detection method (3D DCNN which is the best published results on LUNA dataset) [42]. Our proposed method outperformed the best available results both in sensitivity and FROC score, while only using as less as a third of its parameters, and without the need for multi-stage refinements.

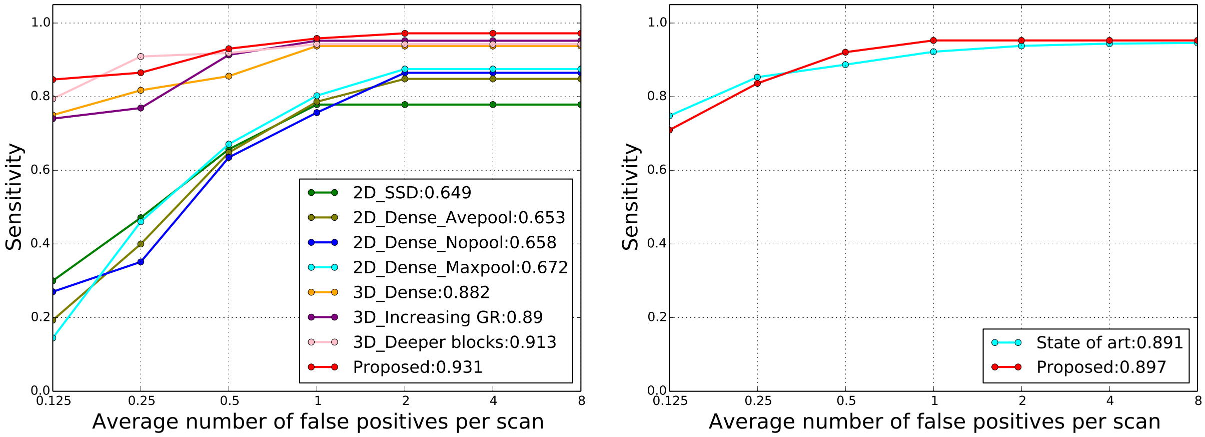

Figure 5.3: Comparison of base line as well as comparison with the state of the art. Numbers in front of each method in the legend show Competition Performance Metric (CPM).

Major findings: (1) We obtained 0.897 FROC rate in 10-fold cross validation, and consistently outperformed the state of the art methods as well as other alternatives. (2) SSD (the state of the art for object detection in natural images) resulted in the lowest accuracy in all experiments. Proposed S4ND, on the other hand, showed that single scale single shot algorithm performs better and more suited to tiny object detection problem. (3) The proposed method achieved better sensitivity, specificity, and CPM in single fold and 10-fold throughout experiments where S4ND used less than the half parameters of 3D DCNN (current state of the art in lung nodule detection). (4) A careful organization of the architecture helps avoiding computationally heavy processing. We have shown that maxpooling is the best choice of feature selection throughout the network amongst current available methods. (5) Similarly, dense and deeper connections improve the detection rates through better information flow through layers. It should be noted that the runtime of our algorithm for the whole scan, on the test phase, varies from 11 *secs* to 27 *secs* based on the number of slices in the scan on a single NVIDIA TITAN Xp GPU workstation with RAM of 64 GBs.

5.1.3 Summary

This section of this chapter introduces a single-shot single-scale fast lung nodule detection algorithm without the need for additional FP removal and user guidance for refinement of detection process. Our proposed deep network structure is fully 3D and densely connected. We also critically analyzed the role of densely connected layers as well as maxpooling, average pooling and fully convolutional down sampling in detection process. We present a fundamental solution to address the major challenges of current region proposal based lung nodule detection methods: candidate detection and feature resampling stages. We experimentally validate the proposed network's performance both in terms of accuracy (high sensitivity/specificity) and efficiency (less number of parameters and speed) on publicly available LUNA data set, with extensive comparison with the natural object detector networks as well as the state of the art lung nodule detection methods. A promising future direction will be to combine diagnosis stage with the detection.

In the rest of this chapter we will go into the details of our segmentation module, which deals with segmentation of complex organs. Finding abnormalities in some organs are more challenging compare to other. This challenge is normally due to complex 3D shape of the organ and it's high similarity to the neighbouring tissues. Pancreas is a good example of such organs. finding cysts in pancreas is highly dependant of finding the boundaries of pancreas in medical images. Our proposed segmentation module will perform an accurate segmentation of pancreas to help radiologists guide their screening in the high risk regions and avoid missing cases.

5.2 *PAN*: Projective Adversarial Network for Medical Image Segmentation

Adversarial learning has been proven to be effective for capturing long-range and high-level label consistencies in semantic segmentation. Unique to medical imaging, capturing 3D semantics in an effective yet computationally efficient way remains an open problem. In this section of this chapter, we address the aforementioned challenges by proposing a novel projective adversarial network, called PAN, which incorporates high-level 3D information through 2D projections. Furthermore, we introduce an attention module into our framework that helps for a selective integration of global information directly from our *segmentor* to our adversarial network. For the clinical application we chose pancreas segmentation from CT scans. Our proposed framework achieved state-of-the-art performance without adding to the complexity of the segmentor.

5.2.1 Methodology

Our proposed method is built upon the adversarial networks. The proposed framework's overview is illustrated in Figure 5.4. We have three networks: a segmentor (S in Figure 5.4), which is the only network being used during the test phase, and two adversarial networks (D_s and D_p in Figure 5.4), each with a specific task to guide the training of S. The first adversarial network (D_s) captures high-level *spatial* label contiguity while the second adversarial network (D_p) enforces the *3D semantics* through a 2D projection learning strategy. The adversarial networks were used only during the training phase to boost the performance of the segmentor without adding to its complexity.

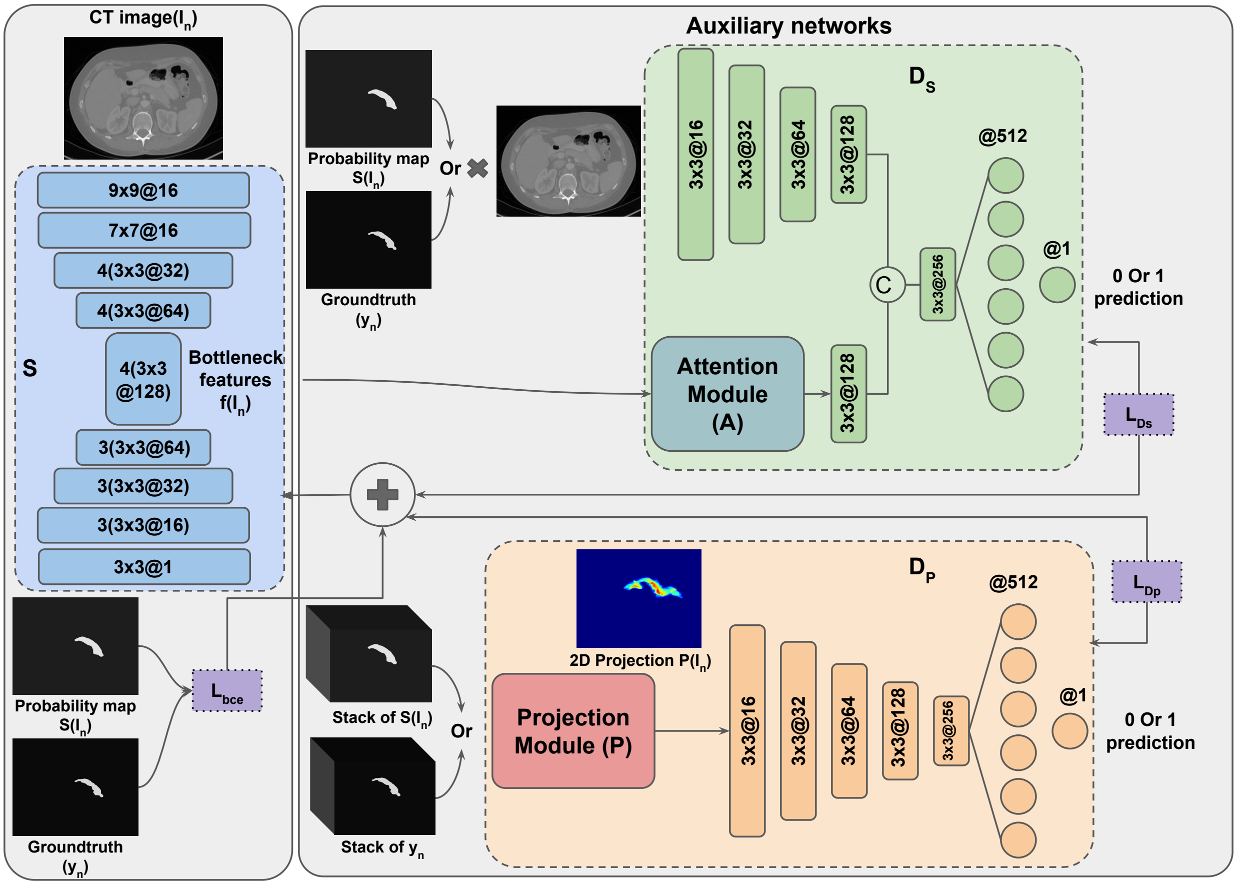

Figure 5.4: The proposed framework consists of a segmentor S and two adversarial networks, D_s and D_p. S was trained with a hybrid loss from D_s, D_p and the ground-truth.

5.2.1.1 *Adversarial training*

To train our framework, we use a hybrid loss function, which is a weighted sum of three terms. For a dataset of N training samples of images and ground truths (I_n, y_n), we define our hybrid loss function as:

$$l_{hybrid} = \sum_{n=1}^{N} l_{bce}(S(I_n), y_n) - \lambda l_{D_s} - \beta l_{D_p}, \tag{5.2}$$

where l_{D_s} and l_{D_p} are the losses corresponding to D_s and D_p and $S(I_n)$ is the segmentor's prediction. The first term in Equation 5.2 is a weighted binary cross-entropy loss. This loss is the state-of-the-art loss function for semantic segmentation and for a grey-scale image I with size $H \times W \times 1$ is defined as:

$$l_{bce}(\hat{y}, y) = -\sum_{i=1}^{H \times W} (w y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)), \tag{5.3}$$

where w is the weight for positive samples, y is the ground-truth label and $\hat{y}$ is the network's prediction. Equation 5.3 encourages S to produce predictions similar to ground-truth and penalizes each pixel *independently*. High-order relations and semantics cannot be captured by this term.

To account for this drawback, the second and third terms are added to train our auxiliary networks. l_{D_s} and l_{D_p} are defined below, respectively:

$$l_{D_s} = l_{bce}(D_s(I_n, y_n), 1) + l_{bce}(D_s(I_n, S(I_n)), 0), \tag{5.4}$$

$$l_{D_p} = l_{bce}(D_p(P_{I_n}, Py_n), 1) + l_{bce}(D_p(P_{I_n}, P_{S(I_n)}), 0). \tag{5.5}$$

Here P is the projection module, l_{bce} is the binary cross-entropy loss with $w = 1$ in Equation 5.3 corresponding to a single number (0 or 1) as the output. Minimizing Equation 5.2 while training D_s and D_p with l_{D_s} and l_{D_p} will put the whole framework into a min-max adversarial game.

5.2.1.2 Segmentor (S)

Our base network is a simple fully convolutoinal network with an encoder-decoder architecture. The input to the segmentor is a 2D grey-scale image and the output is a pixel-level probability map. The probability map shows probability of presence of the object at each pixel. We use a hybrid loss function (explained in details in Section 5.2.1.1) to update the parameters our segmentor (S). This loss function is composed of three terms enforcing: (1) pixel-level high-resolution details, (2) spatial and high-range label continuity, (3) 3D shape and semantics, through our novel projective learning strategy.

As can be seen in Figure 5.4, the segmentor contains 10 conv layers in the encoder, 10 conv layers in the decoder and 4 conv layers as the bottleneck. The last conv layer is a 1×1 conv layer with the channel output of 1, combining channel-wise information in the highest scale. This layer is followed by a sigmoid function to create the notion of probability.

5.2.1.3 Adversarial Networks

Our adversarial networks are designed with the goal of compensating for the missing global relations and correcting higher-order inconsistencies, produced by a single pixel-level loss. Each of these networks produces an adversarial signal and apply it to the segmentor as a term in the overall loss function (Equation 5.2). The details of each network is described below:

Spatial semantics network ($\mathbf{D_s}$): This network is designed to capture spatial consistencies within each frame. The input to this network is either the segmented object by the ground-truth or by the segmentor's prediction. The Spatial semantics network (D_s) is trained to discriminate between these two inputs with a binary cross-entropy loss, formulated as in Equation 5.4. The adversarial signal produced by the negative loss of D_s to S forces S to produce predictions closer to ground-truth in terms of spatial semantics.

As illustrated in Figure 5.4 top right, D_s has a two-branch architecture with a late fusion. The top branch processes the segmented objects by ground-truth or segmentor's prediction. We propose an extra branch of processing, getting the bottleneck features corresponding to the original gray-scale input image, and passing them to an attention module for an information selection. The processed features are then concatenated with the first branch and passed through the shared layers. We believe that having the high-level features of whole image along with the segmentations improves the performance of D_s.

Our attention module learns where to attend in the feature space to have a more discriminative information selection and processing. The details of the attention module are described in the following.

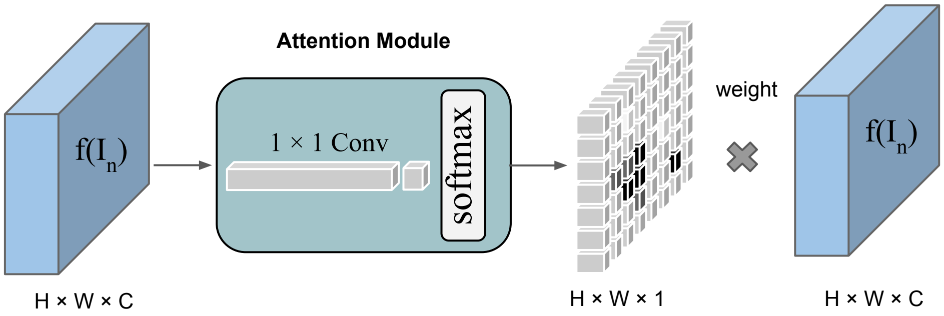

Figure 5.5: Attention module assigns a weight to each feature allowing for a soft selection of information.

Attention module (A): We feed the high-level features form the segmentor's bottleneck to D_s. These features contain global information about the whole frame. We use a soft-attention mechanism, in which our attention module assigns a weight to each feature based on its importance for discrimination. The attention module gets the features with shape $w \times h \times c$, as input, and outputs a weight set with a shape of $w \times h \times 1$. A is composed of two 1×1 convolution layers followed by a softmax layer (Figure 5.5). The softmax layer introduces the notion of *soft selection* to this module. The output of A is then multiplied to the features before being passed to the rest of the network.

Projective network ($\mathbf{D_p}$): Any 3D object can be projected into 2D planes from specific viewpoints, resulting in multiple 2D images. The 2D projection contains 3D semantics information, to be retrieved. In this section, we introduce our projective network (D_p). The main task of D_p is to capture 3D semantics without relying on 3D data and from the 2D projections. Inducing 3D shapes form 2D images has previously been done for 3D shape generation [74]. Unlike existing notions, however, in this paper we propose 3D semantics induction from 2D projections, to benefit segmentation for the first time in the literature.

The projection module (P) projects any 3D volume (V) on a 2D plane as:

$$P((i,j),V) = 1 - \exp\left(-\sum_k V(i,j,k)\right), \tag{5.6}$$

where each pixel in the 2D projection $P((i,j),V)$ gets a value in the range of $[0,1]$ based on the number of voxel occupancy in the third dimension of corresponding $3D$ volume (V). For the sake of simplicity, we refer to the projection of a 3D volume V as $P(V)$. We pass each 3D image through our segmentor (S) slice by slice and stack the corresponding prediction maps. Then, these maps are fed to the projection module (P) and are projected in the axial view.

The input to D_p is either the projected ground-truth or projected prediction map produced by S. D_p is trained to discriminate these inputs using the loss function defined in Equation 5.5. The adversarial term produced by D_p in Equation 5.2 forces S to create predictions which are closer to ground-truth in terms of 3D semantics. Incorporating D_p as an adversarial network to our segmentation framework helps S to capture 3D information through a very simple 2D architecture and without adding to its complexity in the test time.

5.2.2 *Experiments*

We evaluated the efficacy of our proposed system with the challenging problem of pancreas segmentation. This particular problem was selected due to the complex and varying shape of pancreas and relatively more difficult nature of the segmentation problem compared to other abdominal organs. In our experiments we show that our proposed framework outperforms other state-of-the-art methods and captures the complex 3D semantics with a simple encoder-decoder. Furthermore, we have created an extensive comparison to some baselines, designed specifically to show the effects of each block of our framework.

5.2.2.1 *Data and evaluation*

We used the publicly available TCIA CT dataset from NIH [25]. This dataset contains a total of 82 CT scans. The resolution of scans is $512 \times 512 \times Z$, $Z \in [181, 466]$ is the number of slices in the axial plane. The voxels spacing ranges from $0.5mm$ to $1.0mm$. We used a randomly selected set of 62 images for training and 20 for testing to perform a 4-fold cross-validation. Dice Similarity Coefficient (DSC) is used as the metric of evaluation.

5.2.2.2 *Comparison to baselines*

To show the effect of each building block of our framework we designed an extensive set of experiments. In our experiments we start from only training a single segmentor (S) and go to our final proposed framework. Furthermore, we show comparison of encoder-decoder architecture with other state-of-the-art semantic segmentation architectures.

Table 5.2 shows the results adding of each building block of our framework. The eccoder-decoder architecture is the one showed in Figure 5.4 as S, while the Atrous pyramid architecture is similar to the recent work of [54]. This architecture is currently state-of-the-art for semantic segmentation. In which an Atrous pyramid is used to capture global context. We added an Atrous pyramid with 5 different scales: 4 Atrous convolutions at rates of $1,2,6,12$, with the global image pooling. We also replaced the decoder with 2 simple upsampling and conv layers similar to the main paper [54]. We refer the readers to the main paper for more details about this architecture due to space limitations [54]. We found out having an extensive processing in the decoder improves the results compared to the Atrous pyramid architecture (possibly a better choice for segmentation of objects at multiple scales). This is because our object of interest is relatively small.

Table 5.2: Comparison with baselines.

1-fold	**Model**	**DSC%**
	Encoder-decoder (S)	57.7
	Atrous pyramid	48.2
	$S+D_s$	85.0
	$S+D_s+A$	85.9
	$S+D_s+A+D_p$	**86.8**

Moreover, we showed that adding a spatial adversarial notwork (D_s) can boost the performance of S dramatically, in our task. Introducing attention (A) helps for a better information selection (as described in section 5.2.1.3) and boosts the performance further. Finally, our best results is achieved by adding the projective adversarial network (D_p), which adds integration of 3D semantics into the framework. This supports our hypothesis that our segmentor has enough capacity in terms of parameters to capture all this information and with proper and explicit supervision can achieve state-of-the-art results.

5.2.2.3 Comparison to the state-of-the-art

We provide the comparison of our method's performance with current state-of-the-art literature on the same TCIA CT dataset for pancreas segmentation. As can be seen from experimental validation, our method outperforms the state-of-the-art with dice scores, provides better efficiency (less computational burden). Of a note, the proposed algorithm's least achievement is consistently higher than the state of the art methods.

Table 5.3: Comparison with state-of-the-art on TCIA dataset.

	Approach	**Average DSC%**	**Max DSC%**	**Min DSC%**
4-fold cross validation	Roth et al.[25]	71.42 ± 10.11	86.29	23.99
	Roth et al.[75]	78.01 ± 8.20	88.65	34.11
	Roth et al.[22]	81.27 ± 6.27	88.96	50.69
	Zhou et al.[23]	82.37 ± 5.68	90.85	62.43
	Cai et al.[24]	82.40 ± 6.70	90.10	60.00
	Yu et al.[20]	84.50 ± 4.97	**91.02**	62.81
	Ours	$\mathbf{85.53 \pm 1.23}$	88.71	**83.20**

5.2.3 Summary

In this section, we proposed a novel adversarial framework for 3D object segmentation. We introduced a novel projective adversarial network, inferring 3D shape and semantics form 2D projections. The motivation behind our idea was that integration of 3D information through a fully 3D network, having all slices as input, is computationally infeasible. Possible workarounds are: 1) down-sampling the data or 2) sacrificing number of parameters, which are sacrificing information or computational capacity, respectively. We also introduced an attention module to selectively pass whole-frame high-level feature from the segmentor's bottleneck to the adversarial network, for a better information processing. We showed that with proper and guided supervision through adversarial signals a simple encoder-decoder architecture, with enough parameters, achieves the state-of-the-art performance on the challenging problem of pancreas segmentation. We achieved a dice score of **85.53%**, which is state-of-the art performance on pancreas segmentation task, outperforming previous methods. Furthermore, we argue that our framework is general and can be applied to any 3D object segmentation problem and is not specific to a single application.

CHAPTER 6: CONCLUSION AND FUTURE WORK

Here we highlight the concluding remarks on this dissertation, and expand on potential future directions of this research.

6.1 Conclusion

In this dissertation, we propose a collaborative framework for high-risk AI medical applications. Our study offers a new perspective on eye-tracking studies in radiology because of its seamless integration into the real radiology rooms and collaborative nature of image analysis methods. First, the proposed system models the raw gaze data from eye-tracker as a graph. Then, a novel attention based spectral graph sparsification method is proposed to extract global search pattern of radiologist as well as attention regions. Later, we combine this information with our local and global image analysis modules for screening.

We propose a 3D deep multi-task CNN for simultaneously performing segmentation and diagnosis. We show that sharing some underlying features for these tasks and training a single model using shared features can improve the results for both tasks, which are critical for lung cancer screening. Furthermore, we show that a semi-supervised learning approach can improve the results without the need for large number of labeled data in the training.

Further, we introduce a single-shot single-scale detection framework for tiny abnormalities. For application we chose lung nodule detection. Our algorithm does not need additional FP removal and user guidance for refinement of detection process. Our proposed deep network structure is fully 3D and densely connected. We also critically analyze the role of densely connected layers as well as maxpooling, average pooling and fully convolutional down sampling in detection process. We present a fundamental solution to address the major challenges of current region proposal based lung nodule detection methods: candidate detection and feature resampling stages. We experimentally validate the proposed network's performance both in terms of accuracy (high sensitivity/specificity) and efficiency (less number of parameters and speed) on publicly available LUNA data set, with extensive comparison with the natural object detector networks as well as the state of the art lung nodule detection methods.

In the last chapter, we propose a novel adversarial framework for 3D object segmentation. Our framework introduced a novel projective adversarial network, which inferres 3D shape and semantics form 2D projections. Also, an attention module is introduced to selectively pass whole-frame high-level feature from the segmentor's bottleneck to the adversarial network, for a better information processing. We show that with proper and guided supervision through adversarial signals a simple encoder-decoder architecture, with enough parameters, achieves the state-of-the-art performance on the challenging problem of pancreas segmentation.

6.2 Future Work

In this dissertation, we propose a generic collaborative framework for computer aided diagnosis, detection and segmentation of radiology images. The design of our framework allows integration of multiple modules for different purposes. Depending on the desired application modifying current modules or adding new ones can be a future direction to extend the scope of this framework. Furthermore, other than medical applications our framework can be used for other high-risk applications of AI including but not limited to pilot training, autonomous driving, Augmented Reality (AR) or Virtual Reality (VR) applications.

Our work has some limitations that should be noted. These limitations can be a good future direction for the research community. One of the limitations regarding Chapter 3 is the common problem of lack of large amount of medical data for conducting scanpath analysis (with several radiologists). A good future direction can be to address this limitation and explore the validity of the proposed methods in different settings, incorporating the behavioural patterns into screening experiments such as cognitive fatigue of the radiologists.

As for our local image analysis module, an alternative direction to semi-supervised approach for handling lack of data, can be using Generative Adversarial Networks (GAN) to generate realistic data for training. One recent study created realistic nodules to support this idea [70]. Also, exploring which tasks can be auxiliary and can be learned jointly in a multi-task framework can be a very promising research direction. Knowing and combining different tasks through different modalities can also be a very interesting future direction which can help integration of clinical data into visual information for a more accurate decision making.

Our detection module, S4ND, currently does the localization as well as nodule non-nodule classification. A promising future direction to improve this module will be to combine diagnosis stage with the detection. Joint detection and diagnosis eliminates the need for a further processing step after detection. This will lead to a more compact and computationally efficient framework for both detection and diagnosis. Also, the proposed framework can also be extended to be used in similar applications such as tiny object detection in natural or areal images.

A promising direction to improve our last module, PAN, for segmentation is to not only project the 3D volume but to learn the view point automatically. Each pancreas is unique in its shape and orientation compared to others. Due to this characteristic the performance of segmentation network can be improved if the projection viewpoint is being learned and be different for each individual scan. This can be done by including a viewpoint detection network, which learns the most informative viewpoint to project the 3D object and then pass it to the adversarial network.

This viewpoint detecting network can be trained end-to-end in the framework and learn the best viewpoints given an input volume.

LIST OF REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2017,” *CA: A Cancer Journal for Clinicians*, vol. 67, no. 1, pp. 7–30, 2017. [Online]. Available: http://dx.doi.org/10.3322/caac.21387

[2] H. L. Kundel, C. F. Nodine, and D. Carmody, “Visual scanning, pattern recognition and decision-making in pulmonary nodule detection.” *Investigative radiology*, vol. 13, no. 3, pp. 175–181, 1978.

[3] C. Caroline, “Lung cancer screening with low dose ct,” *Radiologic clinics of North America*, vol. 52, no. 1, p. 27, 2014.

[4] M. Firmino, A. H. Morais, R. M. Mendoça, M. R. Dantas, H. R. Hekis, and R. Valentim, “Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects,” *Biomedical engineering online*, vol. 13, no. 1, p. 41, 2014.

[5] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, “Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: A review,” *Computers in biology and medicine*, vol. 60, pp. 8–31, 2015.

[6] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, “Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review,” *Clinical imaging*, vol. 37, no. 3, pp. 420–426, 2013.

[7] C. S. Lee, P. G. Nagy, S. J. Weaver, and D. E. Newman-Toker, “Cognitive and system factors contributing to diagnostic errors in radiology,” *American Journal of Roentgenology*, vol. 201, no. 3, pp. 611–617, 2013.

[8] G. McCreadie and T. Oliver, “Eight ct lessons that we learned the hard way: an analysis of current patterns of radiological error and discrepancy with particular emphasis on ct,” *Clinical radiology*, vol. 64, no. 5, pp. 491–499, 2009.

[9] M. O. Al-Moteri, M. Symmons, V. Plummer, and S. Cooper, “Eye tracking to investigate cue processing in medical decision-making: A scoping review,” *Computers in Human Behavior*, vol. 66, pp. 52–66, 2017.

[10] E. M. Kok and H. Jarodzka, “Before your very eyes: The value and limitations of eye tracking in medical education,” *Medical education*, vol. 51, no. 1, pp. 114–122, 2017.

[11] A. C. Venjakob, “Visual search, perception and cognition when reading stack mode cranial ct,” 2015.

[12] T. Drew, K. Evans, M. L.-H. Võ, F. L. Jacobson, and J. M. Wolfe, “Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images?” *Radiographics*, vol. 33, no. 1, pp. 263–274, 2013.

[13] D. Manning, S. Ethell, T. Donovan, and T. Crawford, “How do radiologists do it? the influence of experience and training on searching for chest nodules,” *Radiography*, vol. 12, no. 2, pp. 134–142, 2006.

[14] S. Littlefair, P. Brennan, W. Reed, and C. Mello-Thoms, “Does expectation of abnormality affect the search pattern of radiologists when looking for pulmonary nodules?” *Journal of digital imaging*, vol. 30, no. 1, pp. 55–62, 2017.

[15] G. Tourassi, S. Voisin, V. Paquit, and E. Krupinski, “Investigating the link between radiologists’ gaze, diagnostic decision, and image content,” *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1067–1075, 2013.

[16] A. C. Venjakob and C. R. Mello-Thoms, “Review of prospects and challenges of eye tracking in volumetric imaging,” *Journal of Medical Imaging*, vol. 3, no. 1, pp. 011 002–011 002, 2016.

[17] T. Drew, M. L.-H. Vo, A. Olwal, F. Jacobson, S. E. Seltzer, and J. M. Wolfe, “Scanners and drillers: Characterizing expert visual search through volumetric images,” *Journal of vision*, vol. 13, no. 10, pp. 3–3, 2013.

[18] L. Grady, “Random walks for image segmentation,” *PAMI, IEEE Trans. on*, vol. 28, no. 11, pp. 1768–1783, 2006.

[19] N. Khosravan, H. Celik, B. Turkbey, R. Cheng, E. McCreedy, M. McAuliffe, S. Bednarova, E. Jones, X. Chen, P. Choyke *et al.*, “Gaze2segment: A pilot study for integrating eye-tracking technology into medical image segmentation,” in *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging*. Springer, 2016, pp. 94–104.

[20] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, “Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8280–8289.

[21] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, “Attention u-net: learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.

[22] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers, “Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation,” *Medical image analysis*, vol. 45, pp. 94–107, 2018.

[23] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, “A fixed-point model for pancreas segmentation in abdominal ct scans,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 693–701.

[24] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, “Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function,” *arXiv preprint arXiv:1707.04912*, 2017.

[25] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, “Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2015, pp. 556–564.

[26] I. Samuel G. Armato, F. Li, M. L. Giger, H. MacMahon, S. Sone, and K. Doi, “Lung cancer: Performance of automated lung nodule detection applied to cancers missed in a ct screening program,” *Radiology*, vol. 225, no. 3, pp. 685–692, 2002.

[27] B. Al Mohammad, P. Brennan, and C. Mello-Thoms, “A review of lung cancer screening and the role of computer-aided detection,” *Clinical Radiology*, vol. 72, no. 6, pp. 433–442, 2017.

[28] A. C. Venjakob, T. Marnitz, P. Phillips, and C. R. Mello-Thoms, “Image size influences visual search and perception of hemorrhages when reading cranial ct an eye-tracking study,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, p. 0018720816630450, 2016.

[29] B. van Ginneken, A. A. Setio, C. Jacobs, and F. Ciompi, “Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans,” in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 2015, pp. 286–289.

[30] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, and B. van Ginneken, “Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box,” *Medical image analysis*, vol. 26, no. 1, pp. 195–202, 2015.

[31] Y. Tsehay, N. Lay, X. Wang, J. T. Kwak, B. Turkbey, P. Choyke, P. Pinto, B. Wood, and R. M. Summers, “Biopsy-guided learning with deep convolutional neural networks for prostate cancer detection on multiparametric mri,” in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 642–645.

[32] M. Le, J. Chen, L. Wang, Z. Wang, W. Liu, K. Cheng, and X. Yang, “Automated diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks.” *Physics in medicine and biology*, 2017.

[33] K.-L. Hua, C.-H. Hsu, S. C. Hidayati, W.-H. Cheng, and Y.-J. Chen, “Computer-aided classification of lung nodules on computed tomography images via deep learning technique,” *OncoTargets and therapy*, vol. 8, 2015.

[34] D. Kumar, A. Wong, and D. A. Clausi, “Lung nodule classification using deep features in ct images,” in *Computer and Robot Vision (CRV), 2015 12th Conference on*. IEEE, 2015, pp. 133–138.

[35] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, “Improving computer-aided detection using convolutional neural networks and random view aggregation,” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1170–1181, 2016.

[36] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, “Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks,” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[37] M. Buty, Z. Xu, M. Gao, U. Bagci, A. Wu, and D. J. Mollura, “Characterization of lung nodule malignancy using hybrid shape and appearance features,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 662–670.

[38] S. Hussein, K. Cao, Q. Song, and U. Bagci, “Risk stratification of lung nodules using 3d cnn-based multi-task learning,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 249–260.

[39] S. Hussein, R. Gillies, K. Cao, Q. Song, and U. Bagci, “Tumornet: Lung nodule characterization using multi-view convolutional neural network with gaussian process,” in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 1007–1010.

[40] N. Khosravan and U. Bagci, “Semi-supervised multi-task learning for lung cancer diagnosis,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 710–713.

[41] X. Huang, J. Shan, and V. Vaidya, “Lung nodule detection in ct using 3d convolutional neural networks,” in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 379–383.

[42] J. Ding, A. Li, Z. Hu, and L. Wang, “Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 559–567.

[43] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, “Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1558–1567, 2017.

[44] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, “Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning,” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[45] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts *et al.*, “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge,” *Medical image analysis*, vol. 42, pp. 1–13, 2017.

[46] Y. K. Tsehay, N. S. Lay, H. R. Roth, X. Wang, J. T. Kwaka, B. I. Turkbey, P. A. Pintob, B. J. Woodc, and R. M. Summersa, “Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images,” in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2017, pp. 1 013 405–1 013 405.

[47] Y. Guo, Y. Gao, and D. Shen, “Deformable mr prostate segmentation via deep feature learning and sparse patch matching,” *IEEE transactions on medical imaging*, vol. 35, no. 4, pp. 1077–1089, 2016.

[48] E. Lopez Torres, E. Fiorina, F. Pennazio, C. Peroni, M. Saletta, N. Camarlinghi, M. Fantacci, and P. Cerello, “Large scale validation of the m5l lung cad on heterogeneous ct datasets,” *Medical physics*, vol. 42, no. 4, pp. 1477–1489, 2015.

[49] S. Krishnamurthy, G. Narasimhan, and U. Rengasamy, “An automatic computerized model for cancerous lung nodule detection from computed tomography images with reduced false positives,” in *International Conference on Recent Trends in Image Processing and Pattern Recognition*. Springer, 2016, pp. 343–355.

[50] P.-P. Ypsilantis and G. Montana, "Recurrent convolutional networks for pulmonary nodule detection in ct imaging," *arXiv preprint arXiv:1609.09143*, 2016.

[51] R. Golan, C. Jacob, and J. Denzinger, "Lung nodule detection in ct images using deep convolutional neural networks," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 243–250.

[52] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[53] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[54] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[55] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1529–1537.

[56] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[57] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *arXiv preprint arXiv:1809.07294*, 2018.

[58] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.

[59] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, no. 3-4, pp. 383–392, 2018.

[60] M. Rezaei, K. Harmuth, W. Gierke, T. Kellermeier, M. Fischer, H. Yang, and C. Meinel, "A conditional adversarial network for semantic segmentation of brain tumor," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 241–252.

[61] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, and E. P. Xing, "Scan: Structure correcting adversarial network for organ segmentation in chest x-rays," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 263–273.

[62] N. Khosravan, H. Celik, B. Turkbey, E. C. Jones, B. Wood, and U. Bagci, "A collaborative computer aided diagnosis (c-cad) system with eye-tracking, sparse attentional model, and deep learning," *Medical image analysis*, vol. 51, pp. 101–115, 2019.

[63] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: an efficient data clustering method for very large databases," in *ACM Sigmod Record*, vol. 25, no. 2. ACM, 1996, pp. 103–114.

[64] D. A. Spielman and N. Srivastava, "Graph sparsification by effective resistances," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1913–1926, 2011.

[65] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *PAMI, IEEE Trans. on*, vol. 34, no. 10, pp. 1915–1926, 2012.

[66] I. Sluimer, A. Schilham, M. Prokop, and B. van Ginneken, “Computer analysis of computed tomography scans of the lung: a survey,” *IEEE transactions on medical imaging*, vol. 25, no. 4, pp. 385–405, 2006.

[67] R. Caruana, “Multitask learning,” in *Learning to learn*. Springer, 1998, pp. 95–133.

[68] A. A. A. Setio and et al, “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge,” *Medical Image Analysis*, vol. 42, no. Supplement C, pp. 1 – 13, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841517301020

[69] H. Kundel, K. Berbaum, D. Dorfman, D. Gur, C. Metz, and R. Swensson, “Receiver operating characteristic analysis in medical imaging,” *ICRU Report*, vol. 79, no. 8, p. 1, 2008.

[70] M. J. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, “How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis,” *arXiv preprint arXiv:1710.09762*, 2017.

[71] N. Khosravan and U. Bagci, “S4nd: Single-shot single-scale lung nodule detection,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 794–802.

[72] N. Khosravan, A. Mortazi, M. Wallace, and U. Bagci, “Pan: Projective adversarial network for medical image segmentation,” *arXiv preprint arXiv:1906.04378*, 2019.

[73] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[74] M. Gadelha, S. Maji, and R. Wang, “3d shape induction from 2d views of multiple objects,” in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 402–411.

[75] H. R. Roth, L. Lu, A. Farag, A. Sohn, and R. M. Summers, “Spatial aggregation of holistically-nested networks for automated pancreas segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 451–459.