UNIVERSITY OF CALIFORNIA,
IRVINE

Deep Learning in Medical Image Analysis

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Hao Tang

Dissertation Committee:
Professor Xiaohui Xie, Chair
Professor Charless C. Fowlkes
Professor Jing Zhang

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisor Professor Xiaohui Xie. for his continuous support of my Ph.D study, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study. It is my great honor to work with him.

I am extremely grateful to Professor Charless Fowlkes and Professor Jing Zhang, for their valuable suggestions on this dissertation. It is my great honor to have them in my final defense committee. Meanwhile, I would like to extend my sincere thanks to my advancement committe member Professor Stephan Mandt, Professor Roy Fox and Professor Daniel Chow, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

I am fortunate to collaborate with Daniel Kim from the School of Medicine during my first year at UCI, where we worked together on Tianchi competition and won the 1st place. I would like to thank him for his tremendous help, patience and advice to me when I was new to UCI and research.

I want to express my gratitude to my lab mates from my research group for maintaining an open and inclusive research environment and atmosphere, for the stimulating and encouraging discussions, and for all the fun we have had in the last couple of years. I would like to extend my sincere thanks to Xingwei Liu, Liangjian Chen, Haoyu Ma, Deying Kong, Xiangyi Yan, Yingxin Cao, and Shanlin Sun for their insightful input in tremendous research discussions among us.

I can hardly exhaust this list and hope my other friends can forgive me for not being able to mention their names here. It has been a luxury to have your support during my Ph.D. study. I would not be able to make this achievement without your support.

Lastly, but the most importantly, I would love to give the greatest thank to my parents. The education and encouragement from them always support me and shape who I am today. Their unconditional love, both spiritual and financial support is the bedrock of my life and help me through the hard times.

VITA

Hao Tang

EDUCATION

Doctor of Philosophy in Computer Science **2017**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Computational Sciences **2013**
Southeast University *Nanjing, Jiangsu*

INTERNSHIP EXPERIENCE

Machine Learning Intern **2020**
Facebook, Inc. *Seattle, WA*

Research Intern **2019**
Microsoft Corporation *Seattle, WA*

Research Intern **2018**
DeepVoxel, Inc. *Irvine, CA*

TEACHING EXPERIENCE

Teaching Assistant **2017–2021**
University of California, Irvine *Irvine, CA*

REFERED JOURNAL PUBLICATIONS

Clinically applicable deep learning framework for organs at risk delineation in CT images **2019**
Nature Machine Intelligence

REFERED CONFERENCE PUBLICATIONS

Automated pulmonary nodule detection using 3D deep convolutional neural networks **2018**
ISBI

Automatic pulmonary lobe segmentation using deep learning **2019**
ISBI

An end-to-end framework for integrated pulmonary nodule detection and false positive reduction **2019**
ISBI

Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation **2019**
MICCAI

Spatial Context-Aware Self-Attention Model For Multi-Organ Segmentation **2021**
WACV

Recurrent Mask Refinement for Few-Shot Medical Image Segmentation **2021**
ICCV

ABSTRACT OF THE DISSERTATION

Deep Learning in Medical Image Analysis

By

Hao Tang

Doctor of Philosophy in Computer Science

University of California, Irvine, 2021

Professor Xiaohui Xie, Chair

Developing algorithms to better interpret images has been a fundamental problem in the field of medical imaging analysis. Recent advances in machine learning, especially deep convolutional neural networks (DCNNs), have demonstrated great improvement to the speed and accuracy of many medical image analysis tasks, such as image registration, anatomical structures/tissue segmentation, and computer-aided diagnosis. Despite previous progress, these problems still remain challenging due to the limited amount of labeled data, large anatomical variance among patients, etc. In this dissertation, we propose various approaches to address the aforementioned challenges in order to achieve better accuracy, higher efficiency, and use fewer labeled data. First, to address the difficulty of accurately detecting pulmonary nodules in its early stage, we propose a novel CAD framework that consists entirely of 3D DCNNs for detecting pulmonary nodules and reducing false positives in chest CT images. Second, to avoid training several deep learning models to solve nodule detection, false-positive reduction, and segmentation separately which may be suboptimal and resource-intensive, we propose NoduleNet to solve the three tasks jointly in a multi-task fashion. To avoid friction between different tasks and encourage feature diversification, we incorporate two major design tricks: 1) decoupled feature maps for nodule detection and false positive reduction, and 2) a segmentation refinement subnet for increasing the precision of nodule segmentation. Third, to address the limitation in scope and/or scale of previous works on organs-at-risk

(OAR) delineation - with only a few OARs delineated and a limited number of samples tested, we propose a new deep learning model that can delineate a comprehensive set of 28 OARs in the head and neck area, trained with 215 CT samples collected and carefully annotated by experienced radiation oncologists with over ten years of experience. The accuracy of our model was compared to both previous state-of-the-art methods and a radiotherapy practitioner. Moreover, we deployed our deep learning model in actual RT planning of new patient cases, and evaluated the clinical utility of the model. Fourth, to reduce the information loss from cropping/downsampling 3D images due to limited GPU memory, we propose a new framework for combining 3D and 2D models, in which the segmentation is realized through high-resolution 2D convolutions, but guided by spatial contextual information extracted from a low-resolution 3D model. A self-attention mechanism is implemented to control which 3D features should be used to guide 2D segmentation. Last but not least, since DCNNs often require a large amount of data with manual annotation for training and are difficult to generalize to unseen classes, we propose a new few-shot segmentation framework RP-Net to address this issue. RP-Net has two important modules: 1) a context relation encoder (CRE) that uses correlation to capture local relation features between foreground and background regions, and 2) a recurrent mask refinement module that repeatedly uses the CRE and a prototypical network to recapture the change of context relationship and refine the segmentation mask iteratively.

Chapter 1

Introduction

1.1 Background

Medical imaging technologies, e.g. computed tomography (CT), magnetic resonance (MR), positron emission tomography (PET), mammography, ultrasound, X-ray and etc., have become a critical component for the early detection, diagnosis and treatment of disease over the last decades [9, 120]. In the clinic, the interpretation of medical images is mostly done by experienced human experts with many years' clinical experience. However, this process is often tedious and prone to human errors, due to the large variation of pathology or anatomical structures and potential fatigue of experts. Computational tools that can automatically conduct these image analysis can greatly alleviate the manual efforts of human experts, if these tools can achieve high accuracy within a reasonable amount of time.

Recent advances in artificial intelligence and machine learning, especially deep learning have brought dramatic changes to medical image analysis. Computational tools that are developed using deep learning have achieved huge accuracy and speed improvement, making it possible to be applied in clinical practice and help doctors to interpret various medical images. Despite

the extensive effort to improve the accuracy and efficiency of using deep learning based algorithm for automatic medical image analysis, it still remains challenging due to the limited amount of labeled data, large anatomical variance among patients, etc. In this dissertation, we first give a review on different challenges in different medical image tasks and then propose new approaches to address them.

1.2 Deep Learning in Medical Image Analysis

In recent years, deep learning has achieved substantial success and attention in the field of medical image analysis. In this dissertation, we focus on developing deep learning based methods to address challenges in several topics, including pulmonary nodule detection and segmentation using CT images, organs-at-risk (OAR) delineation for radiotherapy, and few-shot learning in the low data regime.

Lung cancer has become the leading cause of cancer death among men and women worldwide [124]. Detecting pulmonary nodules and screening lung cancer in early stages using Low-dose Computed Tomography (CT) have been demonstrated to be effective and can increase patients' 5-year survival rates [146]. However, traditional Computer aided diagnosis (CAD) system often detects too many false positive and thus hard to be applied in real clinical practice. Detecting pulmonary nodules with a low false positive rate while maintaining high sensitivity is challenging because of the variations in nodules' size, shape, and the abundance of tissues sharing similar appearance. This dissertation attempts to address these challenges in two ways. First, to reduce the false positive rate, we propose a novel CAD framework that consists entirely of 3D DCNNs for detecting pulmonary nodules and reducing false positives in chest CT images. Second, to avoid training several deep learning models to solve nodule detection, false positive reduction and segmentation separately which may be suboptimal and resource intensive, we propose NoduleNet to use multi-task learning by integrating the

three tasks into one network, and carefully remove frictions among them to maximize their performance.

Radiation therapy (RT) is an important treatment option for many types of cancers. It can be used as an adjuvant treatment before or after surgery, or as a radical treatment administered jointly with chemotherapy [69, 98, 5, 6, 81, 100, 130, 35, 62, 18]. An important step involved in RT planning is the delineation of organs-at-risk (OAR) from CT images so that they can be protected from irradiation. Existing methods mostly adopt atlas-based methods which cannot adequately account for anatomical variance. Although recent methods using deep learning on whole CT images show improved performance [44, 144, 182, 85], these studies were limited in scope and/or scale - with only a few OARs delineated and a limited number of samples tested. In this dissertation, we propose two new methods to address these issues in previous methods. First, we propose U_a-Net, an end-to-end DCNN, that first uses object detection to locate the potential OAR and then uses segmentation head for a finer-level segmentation. We apply U_a-Net to delineate 28 organs at a time, and deploy and evaluate its performance in real clinical practice. Second, to reduce the information loss from cropping/downsampling 3D images due to limited GPU memory, we propose a new framework for combining 3D and 2D models, in which the segmentation is realized through high-resolution 2D convolutions, but guided by spatial contextual information extracted from a low-resolution 3D model using self-attention.

Few-shot learning has been proposed as one of the potential solutions to addressing these challenges in the low data regime [128, 131, 153, 30, 70]. It is an attractive solution to address the expensive data labeling issue in medical image analysis. Although recent studies [113, 95] show promising result of using few-shot segmentation in medical image segmentation, they do not achieve satisfying accuracy. This is likely because they do not adequately capture the foreground and background relationship around the object being segmented. As a result, in this dissertation, we propose a new few-shot segmentation framework RP-Net

which iteratively uses correlation and mask refinement to better capture the foreground and background relationship.

1.3 Dissertation Outline and Contribution

The outline of the rest of the dissertation is as follows:

Chapter 2: In this chapter, we introduce a novel two-stage DCNN approach to detect pulmonary nodules. Our approach, unlike previous ones that are either 2D or 2.5D, is fully three-dimensional end-to-end and utilizes the state-of-the-art in object detection. First, nodule candidates are identified with a U-Net-inspired 3D Faster R-CNN trained using online hard negative mining. Second, false positive reduction is performed by 3D DCNN classifiers trained on difficult examples produced during candidate screening. Finally, we introduce a method to ensemble models from both stages via consensus to give the final predictions. By using this framework, we ranked first of 2887 teams in Season One of Alibaba's 2017 TianChi AI Competition for Healthcare. Part of this chapter is published in [135].

Chapter 3: In this chapter, we present an end-to-end framework for nodule detection, integrating nodule candidate screening and false positive reduction into one model, trained jointly. This work addresses the issues of most of the existing deep learning nodule detection systems that are constructed in two steps: a) nodule candidates screening and b) false positive reduction, using two different models trained separately. Although it is commonly adopted, the two-step approach not only imposes significant resource overhead on training two independent deep learning models, but also is sub-optimal because it prevents cross-talk between the two. We demonstrate that the end-to-end system improves the performance by 3.88% over the two-step approach, while at the same time reducing model complexity by one third and cutting inference time by 3.6 fold. Part of this chapter is published in [137].

Chapter 4: In this chapter, we release our manually labeled 50 CT scans which are randomly chosen from the LUNA16 dataset and explore the use of deep learning for pulmonary lobe segmentation. We propose pre-processing CT image by cropping region that is covered by the convex hull of the lungs in order to mitigate the influence of noise from outside the lungs. Moreover, we design a hybrid loss function with dice loss to tackle extreme class imbalance issue and focal loss to force model to focus on voxels that are hard to be discriminated. To validate the robustness and performance of our proposed framework trained with a small number of training examples, we further tested our model on CT scans from an independent dataset. Experimental results show the robustness of the proposed approach, which consistently improves performance across different datasets by a maximum of 5.87% as compared to a baseline model. Part of this chapter is published in [138].

Chapter 5: In this chapter, we propose a new end-to-end 3D deep convolutional neural net (DCNN), called NoduleNet, to solve nodule detection, false positive reduction and nodule segmentation jointly in a multi-task fashion. To avoid friction between different tasks and encourage feature diversification, we incorporate two major design tricks: 1) decoupled feature maps for nodule detection and false positive reduction, and 2) a segmentation refinement subnet for increasing the precision of nodule segmentation. Extensive experiments on the large-scale LIDC dataset demonstrate that the multi-task training is highly beneficial, improving the nodule detection accuracy by 10.27%, compared to the baseline model trained to only solve the nodule detection task. We also carry out systematic ablation studies to highlight contributions from each of the added components. Code is available at `https://github.com/uci-cbcl/NoduleNet`. Part of this chapter is published in [138]. Part of this chapter is published in [139].

Chapter 6: In this chapter, we present a deep learning model to automatically delineate OARs in head and neck, trained on a dataset of 215 CT scans with 28 OARs manually delineated by experienced radiation oncologists. On a hold-out dataset of 100 CT scans,

our model achieves an average Dice similarity coefficient (DSC) of 78.34% across the 28 OARs, significantly outperforming human experts and the previous state-of-the-art method by 10.05% and 5.18% respectively. Our model takes only a few seconds to delineate an entire scan compared to over half an hour required by human experts. These findings demonstrate the potential for deep learning to improve the quality and reduce the treatment planning time of radiation therapy. Part of this chapter is published in [134].

Chapter 7: In this chapter, we propose a new framework for combining 3D and 2D models, in which the segmentation is realized through high-resolution 2D convolutions, but guided by spatial contextual information extracted from a low-resolution 3D model. We implement a self-attention mechanism to control which 3D features should be used to guide 2D segmentation. Our model is light on memory usage but fully equipped to take 3D contextual information into account. Experiments on multiple organ segmentation datasets demonstrate that by taking advantage of both 2D and 3D models, our method is consistently outperforms existing 2D and 3D models in organ segmentation accuracy, while being able to directly take raw whole-volume image data as inputs. Part of this chapter is published in [136].

Chapter 8: In this chapter, we propose a new framework for few-shot medical image segmentation based on prototypical networks. Our innovation comes at the design of two important modules: 1) a context relation encoder (CRE) that uses correlation to capture local relation features between foreground and background regions; 2) a recurrent mask refinement module that repeatedly uses the CRE and a prototypical network to recapture the change of context relationship and refine the segmentation mask iteratively. Experiments on two abdomen CT datasets and an abdomen MRI dataset show the proposed method obtains substantial improvement over the state-of-the-art method by an average of 16.32%, 8.45% and 6.24% in terms of DSC, respectively.

Chapter 2

Automated Pulmonary Nodule Detection using 3D Deep Convolutional Neural Networks

2.1 Introduction

Lung cancer has been the leading cause of all cancer-related deaths, causing 1.3 millions death annually [123]. Detecting pulmonary nodules early is critical for a good prognosis of the disease, and low-dose computed tomography (CT) scans are widely used and very effective for this purpose. However, manually screening CT images is time-consuming for radiologists who are increasingly overwhelmed with data. Advanced computer-aided diagnosis systems (CADs) have the potential to expedite this process but the task is complicated by the variation in nodule size (from 3 to 50 mm), shape, density, and anatomical context, as well

as the abundance of tissues that resemble the appearance of nodules (e.g., blood vessels, chest wall).

Many approaches have been proposed for this challenge, often employing two stages: 1) nodule candidate screening, which identifies candidates with high sensitivity at the expense of accumulating many false positives, and 2) false positive reduction. Frameworks for the first stage commonly relied on techniques including voxel clustering and curvature computation [61, 90], while second stage methods carefully utilized low-level descriptors such as intensity, size, sphericity, texture, and contextual information [61, 90, 149]. These conventional methods had limited discriminative power due to their reliance on hand-crafted features. More recent efforts have focused on the use of convolutional neural networks (CNNs) and have produced encouraging results, but often use 2D or 2.5D networks in some components for inherently 3D data [21, 115, 117]; nodules can be impossible to discriminate from tissues such as blood vessels from axial slices. Moreover, frameworks incorporating the state-of-the-art models in object detection are still rare.

In this chapter, we propose a novel CAD framework that consists entirely of three-dimensional deep convolutional neural networks (3D DCNNs) end-to-end. Candidate detection is first performed by a U-Net [109]-like Faster Region-based CNN (Faster R-CNN) [105], which is the state-of-the-art model in object detection. The hard mimics identified by the detector are then used to train highly discriminative, deep 3D classifiers for false positive reduction. Both models heavily utilize residual shortcuts [46] that promote performance gains with deep architectures. The final prediction scores are generated by ensembling the detector and the classifiers, unifying contributions learned from both stages. We validated our proposed method on a dataset of 1000 low-dose CT images provided by the 2017 TianChi AI Competition for Healthcare organized by Alibaba [1], where our model ranked 1st in Season One.

2.2 Related work

Ding et al. (2017) proposed a CAD system using Faster R-CNN on 2D axial slices then false positive reduction with a 3D DCNN [21]. Dou et al. (2017) leverages 3D input but uses a binary classifier 3D CNN with online sample filtering for candidate screening, rather than a Faster R-CNN [27]. Our work utilizes an efficient 3D Faster R-CNN for detection and deep residual 3D classifiers for false positive reduction.

2.3 Proposed framework

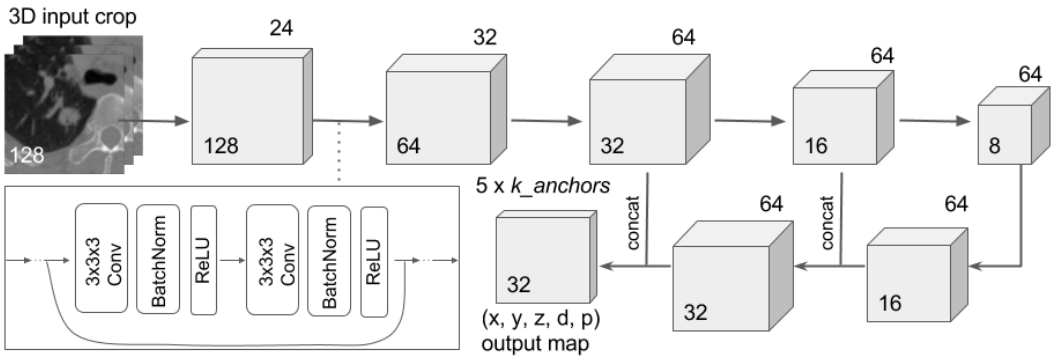

Figure 2.1: General architecture of the candidate-screening 3D Faster R-CNN.

Our proposed method for nodule detection roughly follows two stages: (1) candidate screening using a 3D Faster R-CNN, and (2) subsequent false positive reduction using 3D DCNN classifiers. The purpose of the Faster R-CNN in (1) is to identify nodule candidates while preserving high sensitivity, whereas the classifiers in (2) finely discriminate between true nodules and false positives. We find optimal results when models from both stages are ensembled for final predictions.

Rather than using one stage in which we heavily retrain the Faster R-CNN with hard examples, we believe the two-stage framework provides more flexiblity in adjusting the trade-off between sensitivity and specificity.

2.3.1 Candidate Screening Using 3D Faster R-CNN

The success of Faster R-CNN [105] and deep residual networks [46] in natural images, and U-Net [109] in medical images, has inspired the use of a deep residual 3D Faster R-CNN architecture with transposed convolutional layers, which is illustrated in Fig. 2.1. After a series of downsampling layers to encode high-level information, we concatenate early features with latter ones and feed them through several upsampling transposed convolutions, decoding high-resolution information regarding the nodule's location and diameter. Because we use over 30 convolutional layers, we use residual shortcuts extensively. Memory limits on 4 GPUs made it necessary to split the input image into overlapping $128 \times 128 \times 128$ input volumes, process them separately, and combine them.

The output is a $32 \times 32 \times 32$ map of (x, y, z) coordinates, diameter, and nodule probability corresponding to regions of the input volume. These five features are parameterized by three anchors whose sizes we set to 5, 10, and 30 mm based on the nodule size distribution in our dataset. Each input region is associated with output for each anchor, so the output map is of shape $32 \times 32 \times 32 \times 5 \times 3$.

We compute a classification loss L_{cls} for the predicted nodule probabilities and four regression losses L_{reg} associated with predicted nodule coordinates and diameters. The ground truth labels are determined for each anchor as follows. If an anchor i overlaps with a nodule with an intersection over union (IoU) equal or greater than a threshold of 0.5, we regard it as positive ($p_i^* = 1$). In contrast, if anchor i overlaps with a nodule with an IoU less than 0.2, we regard it as negative ($p_i^* = 0$). All other anchors do not contribute to the loss. Note also that only positive anchors contribute to the regression loss. The final loss for anchor i is

defined as

$$L(p_i, t_i) = \lambda L_{cls}(p_i, p_i^*) + p_i^* L_{reg}(t_i, t_i^*) \tag{2.1}$$

where p_i is the predicted nodule probability; t_i is the vector

$$t_i = \left(\frac{x - x_a}{d_a}, \frac{y - y_a}{d_a}, \frac{z - z_a}{d_a}, \log \frac{d}{d_a} \right) \tag{2.2}$$

of predicted relative coordinates and diameter, where x, y, z, d are the predicted nodule coordinates and diameter and
x_a, y_a, z_a, d_a are the coordinates and size of anchor i. Similarly, the ground truth nodule is expressed as the vector

$$t_i^* = \left(\frac{x^* - x_a}{d_a}, \frac{y^* - y_a}{d_a}, \frac{z^* - z_a}{d_a}, \log \frac{d^*}{d_a} \right) \tag{2.3}$$

where x^*, y^*, z^*, d^* are the coordinates and diameter of the ground truth box. We set λ to 1. We use binary cross entropy loss for L_{cls} and smooth $L1$ loss for L_{reg}.

Hard Negative Mining

Each input volume to the Faster R-CNN is dominated by numerous trivial negative locations (air). To make the negative samples as informative as possible, we used hard negative mining [121]. A pool of N predictions corresponding to condition negative anchors were randomly selected and ranked in descending order according to nodule probability. The top n samples

were chosen to be considered in the loss function.

2.3.2 False Positive Reduction Using 3D DCNN Classifier

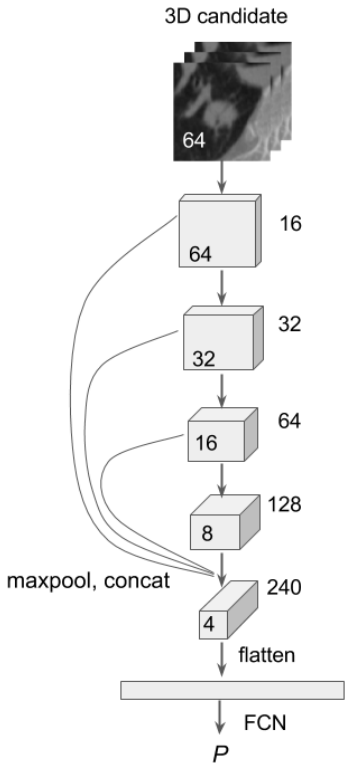

Figure 2.2: General architecture of the 3D false positive reduction classifier.

The predictions (x, y, z, d, p) from the Faster R-CNN are used to extract $64 \times 64 \times 64$ crops centered at (x, y, z) for input to a DCNN classifier, whose architecture is illustrated in Fig. 2.2. It begins with several residual blocks of Conv, BatchNorm, ReLU layers, which are ultimately fed to a fully-connected layer to calculate the final classification score. We integrate detailed local information about the nodule with more contextual features by adding shortcuts from the end of each block to the last feature map.

2.4 Experiments and results

We validated our framework on the large-scale TianChi competition dataset, which contains CT scans from 1000 patients from hospitals in China. The images were annotated by radiologists similarly to LUNA16 [117], i.e. with nodule location and size. We used 600 images for training (containing 969 annotated nodules), 200 for validation, and the remaining 200 comprised the test set.

The evaluation metrics included sensitivity and average number of false positives per scan (FPs/scan), where a detection is considered a true positive if the location falls within the radius of a nodule centroid. The competition ranked participants based on a CPM score defined as the average sensitivity at seven predefined FPs/scan rates: 1/8, 1/4, 1/2, 1, 2, 4, 8.

2.4.1 Training

The Faster R-CNN was trained with Adam for 150 epochs. The examples for each epoch were split such that 70% of the examples consisted of the entire training set of annotations (positive samples), and 30% consisted of random nodule-lacking cropped images from random scans (negative samples). The classifier was trained for 300 epochs with Adam using the same positive examples as the Faster R-CNN detector. These were balanced with hard negative samples, i.e. 969 samples for which the detector assigned a confidence score of 0.5 or greater. The input candidates for test set predictions were provided by the detector. For both detector and classifier, the checkpoint with the highest CPM on the validation set was used for prediction on the test set.

Table 2.1: Validation score comparison showing stepwise performance gains with hard negative mining and classifier ensembling.

Prediction Method	Validation CPM
Faster R-CNN	0.603
Faster R-CNN w/ hard negative mining	0.695
Average (Faster R-CNN, classifier)	0.723
Consensus (Faster R-CNN, 3 classifers)	0.758

2.4.2 Data Augmentation

We trained the Faster R-CNN with random x-, y-, and z-axis flips; random scaling; and large jitters to promote translational invariance and improve generalization. The classifiers were trained similarly along with random rotations. Interestingly, even though the nodule locations predicted by the detector are expected to be centers, minor regression errors made it necessary to add small jitters of up to 2 mm.

2.4.3 Faster R-CNN and Classifier Ensemble Results

The stepwise performance gains of the Faster R-CNN with hard negative mining and classifier ensembles are shown in Table 2.1. Hard negative mining substantially increased the validation CPM from 0.603 to 0.695, demonstrating the importance of using the most informative negative samples. The ensemble average of the Faster R-CNN and the classifier achieved an improved validation CPM of 0.723.

The validation CPM was increased further still, to 0.758, with a "consensus ensembling" method that worked as follows. Two additional classifiers with similar architecture were trained. If the three classifiers agreed with the ensemble average of the detector with the first classifier that a particular candidate location was the most probable nodule for that patient, then the probability score was increased such that "consensus" candidates would

Table 2.2: Top 5 submissions to Season One of the TianChi challenge, for which 2887 teams participated.

Team	Test CPM
Ours	0.815
Yi Yuan Smart HKBU	0.806
LAB2112 (qfpxfd)	0.780
Biana Information Technology	0.780
LAB518-CreedAI	0.769

rank higher than non-consensus ones across all test patients. Fig. 2.3 presents the free-response receiver operating characteristic (FROC) curves. Note the increase in sensitivity at 0.125 and 0.25 FPs/scan.

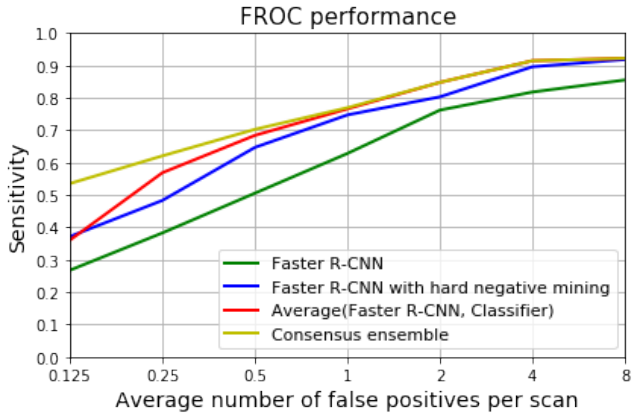

Figure 2.3: Free-response receiver operating characteristic (FROC) curves showing stepwise performance gains in validation with hard negative mining and classifier ensembling.

Ultimately, this consensus ensembling method was used in calculating our final test set predictions for the competition. The CPM score was 0.815, surpassing all other submissions for the TianChi challenge. Final rankings are shown in Table 2.2.

2.5 Conclusion

In this chapter, we present our fully three-dimensional framework of automatic pulmonary nodule detection. It consists of a U-Net-like 3D Faster R-CNN, trained with online hard negative mining, and a 3D classifier for false positive reduction. We introduce a consensus ensembling method to integrate both models for predictions. We validate our method in the 2017 TianChi Healthcare AI Competition, achieving superior performance (0.815 CPM). We believe our model is a powerful clinical tool that harnesses state-of-the-art architectures in a way that captures the spatial nature of CT data.

Chapter 3

An End-to-end Framework for Integrated Pulmonary Nodule Detection and False Positive Reduction

3.1 Introduction

Lung cancer has become the leading cause of cancer death among men and women worldwide [124]. Low-dose Computed Tomography (CT) has demonstrated to be an effective tool for detecting pulmonary nodules and screening lung cancer in early stages. Recent report suggests that detecting lung cancer in early stages can increase patients' 5-year survival rates by 63-75% [146]. However, locating nodules manually through CT scans is time-consuming. Over the past a few years, a lot of work has been done to automatically detect pulmonary nodules by using computer algorithms to read CT images. However, detecting pulmonary

nodules with a low false positive rate while maintaining high sensitivity is challenging because of the variations in nodules' size, shape, and the abundance of tissues sharing similar appearance.

In recent years, deep convolutional neural nets have shown great promise for automated nodule detection [26, 186, 135, 118, 22]. Most of the state-of-art nodule detection systems are constructed in two steps, composed of two separate subsystems: one used for generating nodule candidates, and the other for subsequent false positive reduction. The primary objective of the first subsystem is to generate a comprehensive list of candidate nodules with high sensitivity in mind, while the objective of the second subsystem is to remove false positives to improve specificity. Deep learning models have been proposed for both systems. The first subsystem usually uses segmentation-based methods or Region Proposal Network (RPN) [107] to generate candidates, while the second subsystem primarily uses classification models to distinguish nodules from non-nodules.

Although widely used, the two-step approach implemented in current deep learning systems has two major disadvantages. First, it is time-consuming and resource-intensive to construct and train two separate deep learning models. Although the objectives of the two subsystems are different, they share the commonality of extracting image features characterizing pulmonary nodules. As such, some of the model components can be shared and trained together. Second, the performance of the system may not be optimal because the two subsystems are trained separately without cross-talk between the two.

Here we propose an end-to-end framework for pulmonary nodule detection, integrating nodule candidate generation and false positive reduction into a single model with shared feature extraction blocks, trained jointly. The new end-to-end system substantially reduces model complexity by eliminating one third of the parameters of the corresponding two-step model. It simplifies the training process and cuts the inference time by 3.6 fold. Experiments show that the end-to-end system also improves performance, increasing nodule detection accuracy

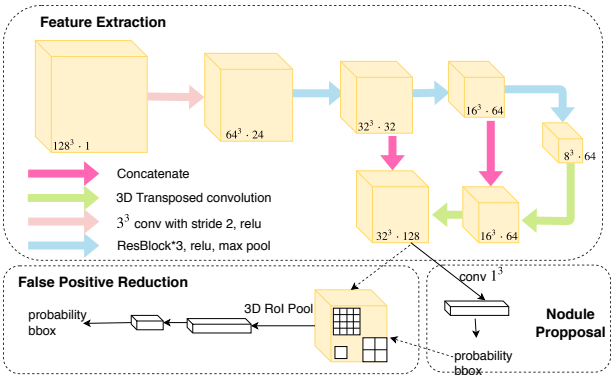

Figure 3.1: End-to-end pulmonary nodule detection framework

by 3.88% over the two-step approach.

Related Work Deep learning, especially deep convolutional neural net (DCNN), has shown great success in medical image analysis. Ding *et al.* [22] proposed a 2D regional proposal network for nodule candidate generation, followed by a 3D convolutional neural net for false positive reduction. Tang *et al.* [135] utilized 3D deep convolutional neural nets in both nodule candidate screening and false positive reduction. Zhu *et al.* [186] adopted 3D nodule candidate screening algorithm, and combined deep learning algorithm with a probabilistic model to explore the usage of weakly labeled clinical diagnosis data. There are also a few works focusing on false positive reduction, such as using multi-scale and model fusion to better classify nodules with various sizes [26] and using multi-view CNN for enhanced 3D information [116]. Recent work also explored using single stage nodule detection model, for instance Khosravan *et al.* [65] proposed using single scale and single shot detection model, which however, has performance limitation because of its single scale assumption and the use of classification instead of detection when approaching this problem.

3.2 Proposed method

The proposed framework largely follows the two stages strategy: (1) generating nodule candidates using 3D Nodule Proposal Network, and (2) subsequent nodule candidate classification for false positive reduction. Different from the aforementioned works where two 3D DCNNs need to be trained separately, we discover the underlying computation of feature extraction for both networks can be shared and forwarded only once. Different tasks can be done on top of the feature map using different branches. The nodule candidate screening branch uses 3D Region Proposal Network adapted from Faster-RCNN [107], and the predicted nodule proposal is then used to crop features of that nodule candidate using 3D Region of Interest (RoI) Pool layer, which are then fed as input to the nodule false positive reduction branch. The whole framework is shown in Figure 3.1.

In feature extraction network, we use 3D convolution layer with stride 2 as the very first layer to reduce GPU memory cost. The subsequent convolution blocks are built using residual blocks [49] with $3 \times 3 \times 3$ convolution followed by maxpooling to reduce spatial resolution.

3.2.1 Nodule Proposal Network

The output of feature extraction is a $32 \times 32 \times 32$ feature map where each pixel on feature map has 128 feature channels. Then a $1 \times 1 \times 1$ convolution layer is applied to this feature map to generate (z, y, x) coordinates, diameter and probability corresponding to the region of input volume. These five features are parameterized by five preset anchors of size 3, 5, 10, 20, 30.

We compute a classification loss and four regression losses associated with (z, y, x) and diameter for each of the anchor on each pixel on the feature map. We then use binary cross entropy loss with Online Hard-negative Example Mining (OHEM) for classification and $L1$

loss for four regressions.

Formally, our objective function is defined as:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i L_{reg}(t_i, t_i^*) \tag{3.1}$$

where i is the index of an anchor in one mini-batch and p_i is the probability that anchor contains a nodule candidate. p_i^* is 1 if an anchor is positive and 0 otherwise. λ is a hyper-parameter for balancing classification and regression losses and we set it to 1 in this case. N_{cls} is the total number of anchors considered for calculating the classification loss and N_{reg} is the total number of anchors considered for calculating regression losses. t_i is a vector representing the four parametrized coordinate offsets of the predicted bounded box and t_i^* is the ground truth of the four regression terms. More specifically, $t_i = (t_z, t_y, t_x, t_d)$ is defined as:

$$\begin{aligned} t &= (\frac{z-z_a}{d_a}, \frac{y-y_a}{d_a}, \frac{x-x_a}{d_a}, \log\frac{d}{d_a}) \\ t^* &= (\frac{z^*-z_a}{d_a}, \frac{y^*-y_a}{d_a}, \frac{x^*-x_a}{d_a}, \log\frac{d^*}{d_a}) \end{aligned} \tag{3.2}$$

where z, y, x, d denote square box's center coordinates and its diameter since we only need diameter to measure the size of a nodule. z, z_a, z^* denote the predicted box, anchor box and ground truth box respectively (likewise for y, x, d).

3.2.2 False Positive Reduction Network

The bounding box regression terms are applied to each anchor, representing the actual spatial location and diameter of nodule candidate, which we call nodule proposal. We then use 3D RoI Pool operation to extract a small feature map from each RoI (*i.e.*, $4 \times 4 \times 4$). These features contain all the information about this nodule candidate and they go through two

fully connected layers for predicting the probability that it is a nodule and four regression terms regarding its (z, y, x) coordinates and diameter.

We use nodule candidate whose probability is equal or greater than 0.5 for training this branch. A nodule candidate is considered as positive if it overlaps with a nodule with an intersection over union (IoU) larger than a threshold 0.5. In contrast, if it has an IoU less than 0.1 with a nodule, we consider it as negative. All other nodule candidates do not contribute to the classification loss and we only calculate regression losses for positive nodule candidates. Definitions of classification and regression losses are the same as Equation (4.1).

3.2.3 Training

We train the whole network in an end to end fashion. We first train the nodule proposal network using Stochastic Gradient Descent (SGD) for 60 epochs and then we train both network together for another 100 epochs. This is because, in the beginning the nodule proposal network predicts random nodule candidates which would be time-consuming for training the false positive reduction branch. Learning rate of SGD optimizer is scheduled as 0.01 initially, decreased to 0.001 after 80 epochs and 0.0001 after 120 epochs.

To improve the generalization ability of the network, input volume is randomly shifted, randomly flipped along all 3 axis, and randomly scaled between 0.9 and 1.1.

3.3 EXPERIMENTS AND RESULTS

We validated our framework on large-scaled Tianchi competition dataset[1]. It contains 800 CT scans from 800 patients with released ground truth label. The CT scans were annotated

[1] `https://tianchi.aliyun.com/competition/rankingList.htm?raceId=231601&season=0`

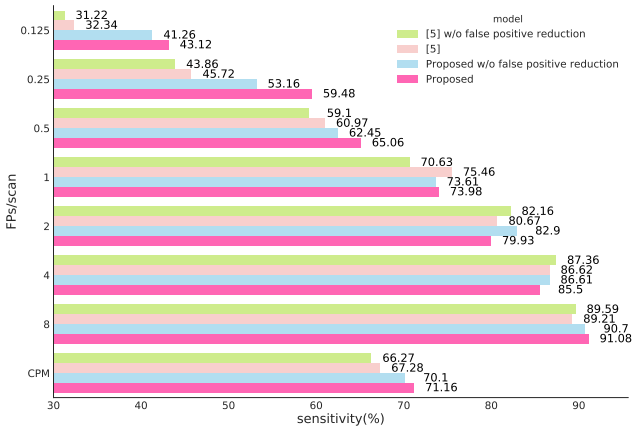

Figure 3.2: Performance comparison

	# Parameters	Inference Time
[135]	15903490	$10.2s$/CT
Proposed	9618523	$2.8s$/CT

Table 3.1: Comparison of number of parameters and time for inference between separate two stage framework [135] and the proposed framework

in a similar way to LUNA16 [118] with exact nodule location and diameter information. We used 600 CT scans for training and validation and another holdout 200 CT scans for reporting the performance of our model.

Free-Response Receiver Operating Characteristic (FROC) [68] analysis was adopted to quantify trade-off between sensitivity and specificity. We used the same evaluation metric as the LUNA16 challenge [146] and the evaluation was performed by measuring the detection sensitivity and false positives per scan (FPs/scan). A nodule detection is considered positive if and only if its predicted location falls within a distance R from the ground truth nodule's center, where R is one half of nodule's diameter. The final Competition Performance Metric (CPM) is defined as the average sensitivity at seven predefined FPs/scan rates: 1/8, 1/4, 1/2, 1,2,4, 8.

3.3.1 Performance comparison on holdout test set

We compared performance among single stage nodule detection framework ([135] w/o false positive reduction), a state-of-art separate two-stage framework [135] and the proposed end-to-end two-stage framework. The step-wise gains of using the end-to-end framework is summarized in Figure 3.2. As we can see, when training the nodule proposal network and false positive reduction network together, the proposed end-to-end framework not only improves nodule proposal performance by 3.73%, but further boosts the performance by 1.06% using false positive reduction, which yields a 3.88% improvement on CPM compared to previous state-of-art separate two-stage nodule detection model ([135]) without model ensemble.

Also, Table 3.1 shows the number of parameters used by the proposed framework, which is significantly lower than that of the previous two-stage model because of weight sharing. Moreover, since the proposed framework only needs to perform feature extraction once instead of forwarding the same patch of CT scan multiple times when inferring, it substantially reduces inference time for each CT scan from an average of $10.2s$ to $2.8s$ using single GPU.

3.3.2 Visualization

We randomly chose one patient from the holdout test set for visualizing performance gains of using the proposed framework in Figure 7.3. The end-to-end model yields more precise detection of nodule location and size and better probability score, which demonstrates the proposed end-to-end framework improves the quality of pulmonary nodule detection.

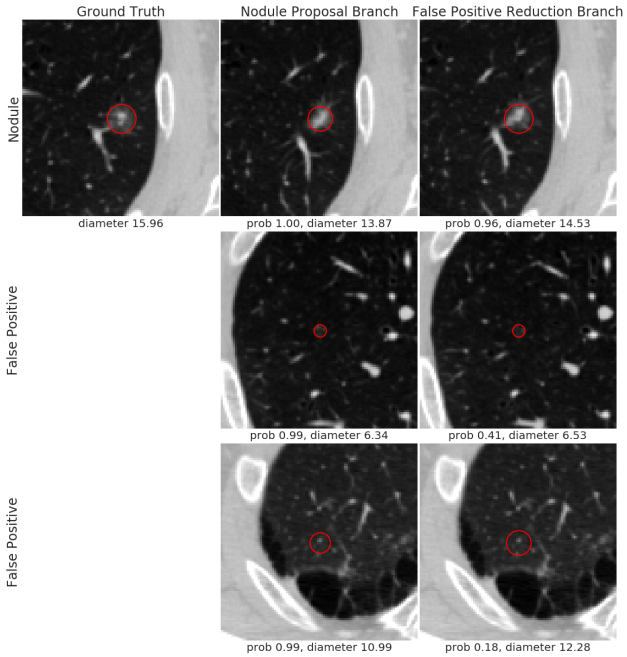

Figure 3.3: Visualization of predictions from different branches of the proposed end-to-end nodule detection framework. The first row is from a true nodule while the other rows are false positives. We only show the center slice of each nodule candidate. Note that the false positive reduction branch is able to refine nodule diameter for the true nodule and significantly reduce probabilities for false positives.

3.4 Conclusion

In summary, we have presented a novel end-to-end framework for pulmonary nodule detection integrating nodule candidate generations and false positive reduction. The new system substantially reduces model complexity and inference time, thereby simplifying the training process and reducing resource overhead. Additionally, it improves the nodule detection performance over the two-step approach commonly used in existing nodule detection systems. Altogether, our work suggests that an end-to-end framework is more desirable for constructing deep learning-based pulmonary nodule detection systems.

Chapter 4

Aotomatic Pulmonary Lobe Segmentation using Deep Learning

4.1 Introduction

Lung cancer has been the leading cause of all cancer-related disease during the past years [124]. Segmentation of pulmonary lobe based on Computed Tomography (CT) is an important task for Computer Aided Diagnosis systems (CADs). Pulmonary lobe segmentation is relevant in a wide range of clinical applications. The location and distribution of pulmonary disease can be a significant factor in determining the most suitable treatment. According to [42], locally distributed emphysema can be treated more effectively by lobar volume resection than homogeneously distributed emphysema. Another application is pulmonary nodule detection where detecting pulmonary nodule in its early stage is critical for a good prognosis of the disease. Recent success in deep learning especially the use of Deep Convolutional Neural Network (DCNN) has accelerated the development of automatic pulmonary nodule detection and classification system, such as [135, 184], which can be used to help radiologist

and reduce their labor work. Precise segmentation of lung lobes can be used to generate automatic electronic diagnosis report since the rough location of nodules are required and the precise coordinate information is rarely used in most medical institutes.

Human lungs are composed of five lobes (two in the left lung and three in the right lung). The upper lobe and lower lobe of left lung are separated by the major fissures (oblique fissure). In the right lung, there are three lobes, namely upper lobe, middle lobe and lower lobe. The upper lobe and middle lobe are divided by the minor fissure (horizontal fissure) while the major fissure (oblique fissure) separates the lower lobe from the rest of the lung.

Methods for pulmonary lobe segmentation have been focused on unsupervised models using classical computer vision techniques which usually include detecting fissures, locating bronchi and vessels, such as [67, 72]. More recently, FJS Bragman *et al.* applied probabilistic model in enhanced fissure detection using fissure prior which yields accurate results under various fissure incompleteness. However, the attempt of using deep learning in this task is still rare [31] because of the need for a large number of annotated training examples. Moreover, publicly available annotations for pulmonary lobe segmentation can hardly be found for supervised training of deep neural network.

In this chapter, we collaborate with our radiologist on manually creating and releasing reference annotations from a randomly chosen subset from the LUNA16 [118]. Next, we present a framework using DCNN that can be trained effectively and robustly with a small number of training examples. In order to further validate the generalization ability and robustness of the trained deep neural net, we annotated 10 more CT scans from Tianchi dataset as a holdout test set. Experimental results show the proposed framework generalize well to CT scans collected from different sources, which yields a maximum of 5.87% improvement as compared to a baseline model.

Our contributions of this work are summarized as below:

a). We propose pre-processing CT image by cropping region that is covered by the convex hull of the lungs in order to mitigate the influence of noise outside the lungs. We design a hybrid loss function with dice loss to tackle extreme class imbalance issue and focal loss to force model to focus on voxels that are hard to be discriminated. This achieves the state-of-art averaged dice coefficient of 91.48% on the LUNA16 test set and 94.17% on the Tianchi test set respectively.

b). We release our reference annotations on 50 CT scans randomly chosen from LUNA16 for supervised pulmonary lobe segmentation study, which is the first publicly available dataset with reference annotations on this task.

4.2 Data

In this section, we describe in details the source of CT scan and tool we used to generate reference annotations for pulmonary lobe segmentation task.

4.2.1 Data and annotation

We randomly chose 50 CT scans from LUNA16 [118] and collaborated with our radiologist in creating annotations for each CT scan. LUNA16 is a subset of LIDC-IDRI [3] for pulmonary nodules, which is the largest publicly available lung image dataset. LUNA16 was then created by removing CT scan that has a slice thickness greater than 3mm, inconsistent slice spacing or missing slices from LIDC-IDRI dataset to provide an evaluation framework for pulmonary nodule detection. LIDC-IDRI data uses the Creative Commons Attribution 3.0 Unported License.

Reference annotation for each CT scan was then manually delineated by radiologist using

Chest Image Platform[1]. This software platform is built on top of the 3D slicer and uses an interactive algorithm to perform lobe segmentation where user is required to mark points on three fissures [67, 72].

We used 40 of the annotated CT scans for training our model and 10 for testing on the LUNA16 dataset. In order to validate the robustness of our algorithm, we further annotated another 10 randomly chosen CT scans collected from a different source: Tianchi[2], which is also a large-scaled competition dataset.

4.2.2 Reference annotation availability

50 annotations created on the LUNA16 dataset will be made publicly available for supervised lung lobe segmentation study. However, the 10 annotations made on the Tianchi dataset will not be made publicly available at this time.

4.3 Method

We present in this section the framework using the deep convolutional neural network which includes pre-processing of removing regions outside the lung region, model architecture and hybrid loss function used to train the network.

4.3.1 Pre-processing

We propose pre-processing by cropping region covered by the convex hull of the lungs, which removes noise from different CT scans outside the lungs, as well as reducing the cost of GPU

[1] https://chestimagingplatform.org/about

[2] https://tianchi.aliyun.com/competition/rankingList.htm?raceId=231601&season=0

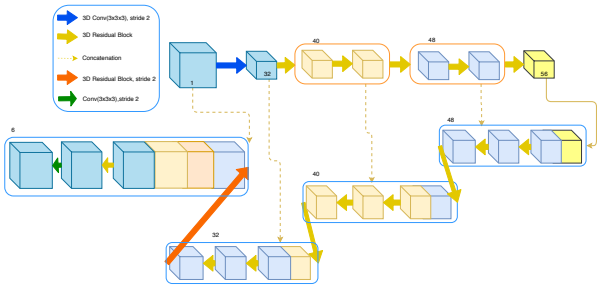

Figure 4.1: Neural network architecture. Each cube represents 3D image volume and the side number denotes the number of channels in that block. Different from original V-Net, we only use one down-sampling to balance the trade-off between feature representation capacity and GPU memory.

memory as the input volume is substantially smaller.

We start with normalizing the whole CT scan by truncating Hounsfield Unit (HU) values outside the range of $[-1000, 600]$. Next, we use OTSU to binarize the CT image. A binary morphology close is then used to remove regions outside the lungs and binary hole filling is applied to fill small isolated regions in the lung on a per slice base. The convex hull of the two lungs is computed and a binary morphology dilation using $5 * 5$ kernel is applied to preserve information near the border slice by slice.

4.3.2 Model architecture

The network architecture is illustrated in Figure 8.1. We use 3D residual block [49] as a basic building block which consists of two consecutive $3 * 3 * 3$ convolution layers followed by ReLU and Batchnorm. We only use one down sampling in this architecture to balance the trade-off between feature representation capacity and GPU memory, which is different from original V-Net [86] who employs the standard four down-samplings.

4.3.3 Loss

Dice loss is widely used for training a segmentation network using deep learning in the medical image. Dice loss performs relatively well when training samples are highly imbalanced as compared to cross entropy loss. However, dice loss fails to capture the difference of difficulty in classifying different voxels. For instance, voxels on the border are more difficult to be classified correctly than voxels are in the center of the lobe. As a result, we adopt a hybrid loss of both dice loss and focal loss [77] to address voxel-wise imbalance and force model to focus on those voxels that are hard to be correctly predicted. More formally, our hybrid loss function is defined as:

$$L = L_{dice} + \lambda L_{focal} \tag{4.1}$$

$$L_{dice} = \sum_{c}^{C} \sum_{i}^{N} \frac{p_{ic} * g_{ic}}{p_{ic} * g_{ic} + (1 - p_{ic}) * g_{ic} + p_{ic} * (1 - g_{ic})}$$

$$L_{focal} = -\frac{1}{N} \sum_{c}^{C} \sum_{i}^{N} \alpha_c * g_{ic} * (1 - p_{ic})^{\gamma} * log(p_{ic})$$

λ is a hyper-parameter controlling the balance between dice loss and focal loss, which is set to 1 in this chapter. N is the total number of voxels in each mini-batch and i is the index of each individual voxel. C denotes the total number of classes which is six in this task (one more class for background). p_{ic} is the predicted probability that i-th voxel is class c and g_{ic} is 1 if i-th voxel is class c and 0 otherwise. α and γ are parameters controlling weight for each class and adjusting the down-weighting of well-classified voxels respectively. We set α to be one and γ to be two as suggested in [77].

4.3.4 Data augmentation

Data augmentation is critical for training model that can generalize well across different datasets, especially when the number of training samples is small. The input volume is randomly shifted, z-axis flipped and XY-plane rotated in order to improve the generalization ability of the model.

4.4 Experiments

Dice coefficient was used to evaluate the performance of the model:

$$DC_c(P, G) = 2 * \frac{P_c \cup G_c}{P_c \cap G_c}, c \in C \tag{4.2}$$

$$DC_{avg}(P, G) = \frac{1}{C} \sum_{c}^{C} DC_c(P_c, G_c) \tag{4.3}$$

where P is the set of predictions for each voxel and G represents the set of ground truth label. We calculate dice coefficient for each lobe independently and averaged dice coefficient for all lobes as described in Equation (4.2) and Equation (4.3) respectively.

We split 50 annotated CT scans from LUNA16 into 40 for training and 10 for testing. We tested our model on a holdout 10 CT scans annotated from Tianchi dataset as well to illustrate the robustness of our proposed approach.

In order to assess the influence of hybrid loss and pre-processing by using convex hull, we added each component step by step and all models were trained using the same data augmentation with the same hyper parameters. We trained each model for 300 epochs using

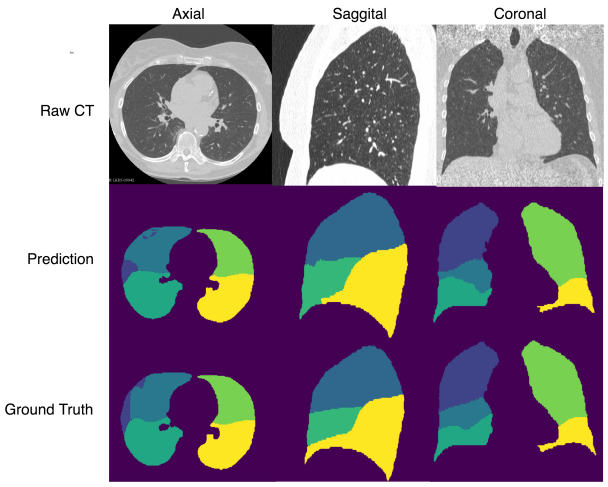

Figure 4.2: Comparison between prediction of the model and ground-truth in CT scan views: Axial, Sagittal and Coronal.

Adam as the optimizer and used the last epoch for predicting each test set. Batch size was set to one since the size of input volume might be different for each CT scan after pre-processing.

The step-wise gains of using hybrid loss and pre-processing by using convex hull are shown in Table 5.1. Also, we compared our best model with [31] which we found most relative to our work. As shown in the tables, model trained with hybrid loss increased the averaged dice coefficient of the baseline model by 3.87% on LUNA16 test set and 3.95% on Tianchi test set. By removing regions outside the lung using convex hull further increased the averaged dice coefficient by 0.54% on LUNA16 test set and 1.92% on Tianchi test set as compared to the model trained only with hybrid loss. Moreover, the comparison between [31] and our approach further validated the robustness and generalization ability of the proposed framework by improving the previous state-of-art result by a maximum of 2.39% on averaged dice coefficient.

We visualized in Figure 7.3 qualitative comparison between model prediction and reference annotation from three views, which illustrates the significance of focusing on hard negative examples and only regions inside and between two lungs.

LUNA16 test set						
	RU	RM	RL	LU	LL	AVG
[31]	**92.76**	**84.68**	**94.33**	88.10	94.78	90.93
DL	78.28	79.69	93.96	88.65	94.79	87.07
+ FL	90.58	78.41	93.95	96.01	95.77	90.94
+ CH	92.53	80.60	93.05	**96.10**	**95.30**	**91.48**
Tianchi test set						
	RU	RM	RL	LU	LL	AVG
[31]	93.11	86.43	94.54	89.30	95.40	91.76
DL	80.80	82.71	94.46	89.03	94.51	88.30
+ FL	92.59	84.75	93.08	95.94	94.88	92.25
+ CH	**95.11**	**87.92**	**95.15**	**97.21**	**95.46**	**94.17**

Table 4.1: Step-wise performance gains of using hybrid loss and pre-processing using convex hull as compared to a baseline model trained only with dice loss and previous state-of-art method [31]. RU, RM, RL, LU, LL and AVG represent the dice coefficient of right upper lobe, right middle lobe, right lower lobe, left upper lobe, left lower lobe and their average respectively. DL means model trained with dice loss and +FL means model trained with hybrid loss of focal loss and dice loss. +CH represents model trained with hybrid loss and training data cropped by the convex hull of the lungs.

4.5 Conclusion

In this chapter, we release our manual annotation by radiologist for 50 CT scans collected from the LUNA16 challenge and present a practical and robust framework for robust pulmonary lobe segmentation. We believe the public availability of those reference annotations will help the study of pulmonary lobe segmentation using supervised learning. Also, our proposed framework trained with a small number of training examples is proved to perform well across CT scans from different sources.

Chapter 5

NoduleNet: Decoupled False Positive Reduction for Pulmonary Nodule Detection and Segmentation

5.1 Introduction

Lung cancer has the highest incidence and mortality rates worldwide [8]. Early diagnosis and treatment of pulmonary nodules can increase the survival rate of patients. Computed tomography (CT) has been widely used and proved effective for detecting pulmonary nodules. However, manually identifying nodules in CT scans is often time-consuming and tedious, because a radiologist needs to read the CT scans slice by slice, and a chest CT may contain over 200 slices. Accurate and precise nodule segmentation can provide more in-depth assessment of the shape, size and change rate of the nodule. When nodule is identified, a follow up scan in 3 - 12 months is usually required to assess its growth rate [63]. The growth of the lung tumor may be an indicator for malignancy, and an accurate nodule segmentation

can be used for measuring the growth rate of the nodule.

In recent years, deep convolutional neural network has emerged as a leading method for automatically detecting and segmenting pulmonary nodules and have achieved great success. State-of-the-art frameworks for nodule detection often ustilize the 3D region proposal network (RPN) [106] for nodule screening [135, 185, 118, 76], followed by a 3D classifier for false positive reduction [23, 137]. Although single stage detector has also been proposed in [65], their hit criteria was different from what was more commonly adopted [118]. Moreover, the refinement provided by the extra classifiers may correct some errors made by the detectors. In terms of nodule segmentation, U-Net [110] and V-Net [87] like structure is predominantly used [160, 165, 2]. In practice, a computer aided diagnosis (CAD) system for pulmonary nodule detection and segmentation often consists of several independent subsystems, optimized separately.

There are some limitations on handling each task completely independent. First, it is time-consuming and resource intensive to train several deep convolutional neural networks. Although each component is designed for different purposes, they share the common procedure of extracting feature representations that characterize pulmonary nodules. Second, the performance of the whole system may not be optimal, because separately training several systems prevents communication between each other and learning intrinsic feature representations. Intuitively, the segmentation mask of the nodule should provide a strong guide for the neural network to learn discriminative features, which may in turn improve the performance of nodule detection.

Although multi-task learning (MTL) and feature sharing offer an attractive solution to combine different tasks, a naive implementation may cause other problems [15]. First, because of the mismatched goals of localization and classification, it may be sub-optimal if these two tasks are performed using the same feautre map. Second, a large receptive field may integrate irrelevant information from other parts of the image, which may negatively affect and

confuse the classification of nodules, especially small ones. [15] decoupled localization and classification to address the problem in natural imaging. However, completely separating the two tasks without sharing any feature extraction backbone, still prevents cross-talk between two networks and may not be the most efficient. Therefore, a decoupled false positive reduction, that pools features from early scales of the feature extraction backbone, is proposed to address this problem, which allows learning both task-independent and task-dependant features.

Here, we propose a new end-to-end framework, called NoduleNet, for solving pulmonary nodule candidate screening, false positive reduction and segmentation jointly. NoduleNet consists of three parts: nodule candidate screening, false positive reduction and segmentation refinement (Figure 8.1). These three components share the same underlying feature extraction backbone and the whole network is trained in an end-to-end manner.

Our main contributions are summarized as follows:

- We propose a unified model to integrate nodule detection, false positive reduction and nodule segmentation within a single framework, trained end-to-end in a multi-task fashion.
- We demonstrate the effectiveness of the model, improving nodule detection accuracy by 10.27% compared to the baseline model trained only for nodule detection, and achieving a state-of-the-art nodule segmentation accuracy of 83.10% on Dice-Sørensen coefficient (DSC).
- We carry out systematic ablation studies to verify the contributions of several design tricks underlying NoduleNet, including decoupled features maps, segmentation refinement subnet, and multi-task training.

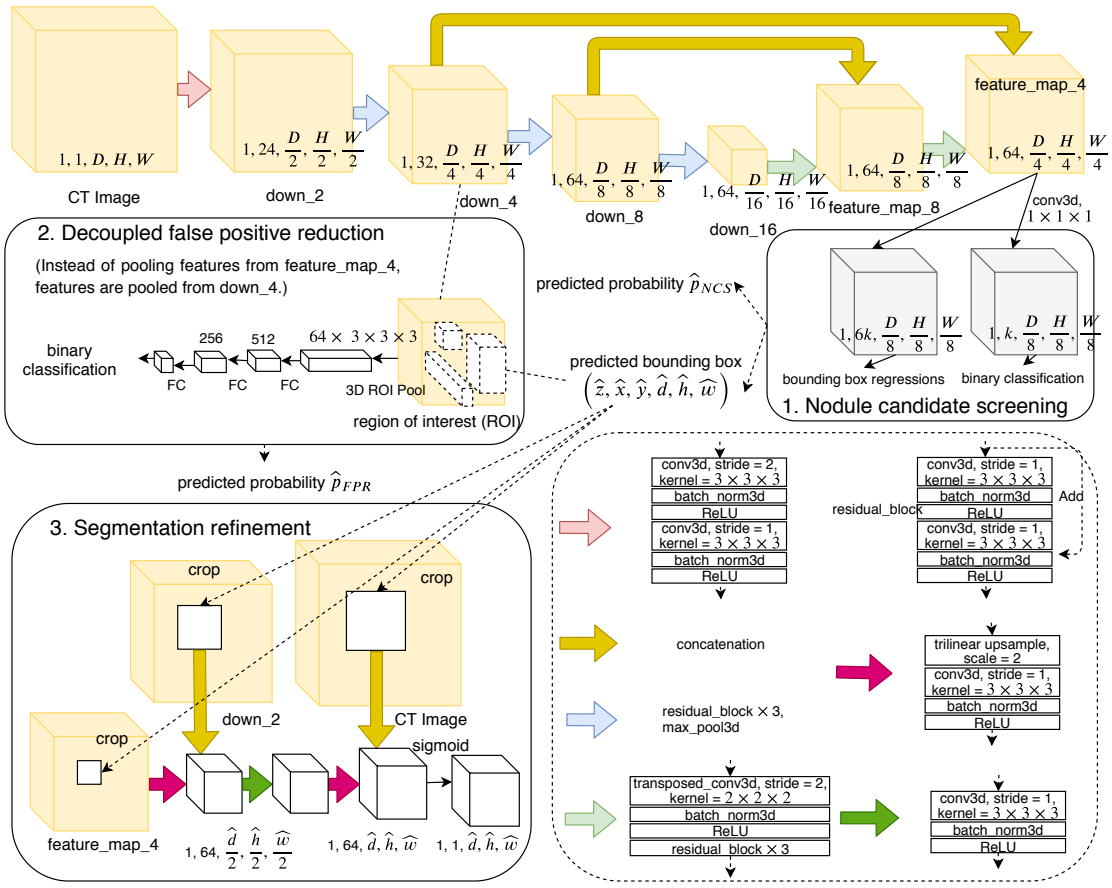

Figure 5.1: **Overview of NoduleNet. NoduleNet is an end-to-end framework for pulmonary nodule detection and segmentation, consisting of three sequential stages: nodule candidate screening, false positive reduction and segmentation refinement.** k **is the number of anchors. FC is short for fully connected layer.**

5.2 NoduleNet

Nodule candidate screening (NCS)

To generate nodule candidates, a $3 \times 3 \times 3$ 3D convolutional layer is applied to the feature map (feature_map_4 in Figure 8.1), followed by two parallel $1 \times 1 \times 1$ convolutional layers to generate classification probability and six regression terms associated with each anchor at each voxel on the feature map. An anchor is a 3D box, which requires six parameters to specify: central z-, y-, x- coordinates, depth, height and width. We chose cube of size 5, 10, 20, 30 and 50 as the 5 anchors in this chapter. Then, we minimize the same multi-task loss

function as [106].

Decoupled false positive reduction (DFPR)

Unlike [106] that performs classification using features pooled from the same feature map as RPN (feature_map_4). Learning using coupled feature map may lead to sub-optimal solutions of the two tasks. Instead, we use 3D region of interest (ROI) pooling layer to pool features from early feature map that has a small receptive field (down_4). This not only ensures the false positive reduction network has a small receptive field and can learn feature representations that are substantially different from nodule candidate screening network, but also allows sharing of a few feature extraction blocks. The false positive reduction network minimizes the same multi-task in loss function as the NCS .

Segmentation refinement (SR)

As shown in Figure 8.1, segmentation is performed at the same scale of the original input CT image, by progressively upsampling the cropped high-level feature map (feature_map_4) and concatenating them with low-level semantically strong features.

This approach is fundamentally different from the mask branch proposed in [48]. In [48], the authors perform segmentation by only using downsampled feature map and then resize the predicted mask back to the original image scale, which may lose precision due to bounding box regression errors and loss of more fine-grained local features.

Another advantage is that, only the regions have nodules are upsampled to the original image scale, which only accounts for a small area of the whole input image. This saves a large amount of GPU memory, making whole volume input feasible during training and testing, as compared to upsampling the whole feature map to original input scale in [87].

The segmentation refinement network minimizes the soft dice loss of the predicted mask sets $\{m\}$ and the ground truth mask sets $\{g\}$ of the input image.

5.3 Results

Data and experiment configurations

We used LIDC-LDRI [3] for evaluting the performance of NoduleNet. LIDC-LDRI is a large-scale public dataset for studying lung cancers, which contains 1018 sets of CT scans collected from multiple sites with various slice thickness. Nodules with diameter equal or greater than 3 mm in this dataset have contour outlined by up to four radiologists. We included only those CT scans met the selection criteria of LUNA16 [118] in this chapter. If the two segmentation masks provided by two radiologists have an intersection over union (IoU) larger than 0.4, we consider the two masks are referring to the same nodule. We consider nodules annotated by at least 3 out of 4 radiologists the ground truth, resulting in a total number of 586 CT scans with 1131 nodules. Note that the number of CT scans and nodules included in this chapter may be different from previous work [160, 165, 2], due to different inclusion criteria.

A six-fold cross validation was performed to demonstrate the performance of NoduleNet. All models in the experiment were trained using stochastic gradient descent (SGD) with initial learning rate 0.01, momentum 0.9 and $l2$ penalty 0.0001, for 200 epochs. The learning rate was scheduled to decrease to 0.001 after 100 epochs and to 0.0001 after another 60 epochs.

Free-Response Receiver Operating Characteristic (FROC) [68] analysis was adopted for evaluating the performance of nodule detection. We used the same hit criterium and competition performance metric (CPM) as in the LUNA16 [118]. Intersection over union (IoU) and Sørensen-Dice coefficient (DSC) were used for evaluating the performance of nodule segmentation.

Nodule detection performance

In order to fully verify and understand our aforementioned assumptions, we conducted extensive experiments using different network architectures and design choices. We use $\mathbf{N}_1$ to represent network that only has NCS branch, $\mathbf{N}_2$ for network has both NCS and FPR branches, and $\mathbf{N}_3$ for network has all NCS, FPR and SR branches. $\mathbf{F}_c$ represents the FPR branch is built on the same feature map as NCS, and $\mathbf{F}_d$ means the FPR branch is built on the decoupled feature map mentioned in previous section. **R** means the training data is extraly augmented with xy - plane rotation. **NCS** means the predicted probability comes from NCS branch, **FPR** means the predicted probability comes from FPR branch, and **FU** means the predicted probability is fused from NCS and FPR. Note that $\mathbf{N}_1$ is the widely used 3D RPN for nodule detection [135, 137, 185, 76], which was served as a strong baseline for evaluating the performance of each added component. The results are summarized in Table 5.1.

As seen from Table 5.1, the sensitivity at 8 false positives per patient rate has a consistent improvement of 1.0% to 1.5% by adding the segmentation refinement network ($\mathbf{N}_3$), which demonstrates the effectiveness of using the extra nodule segmentation information.

The average sensitivity of the NoduleNet using decoupled false positive reduction ($\mathbf{F}_d$) has around 3% to 4% improvement over the NoduleNet using coupled false positive ($\mathbf{F}_c$). Moreover, by adding rotation in data augmentation (**R**), the performance of **FPR** branch is further improved by around 2.5% while the performance of **NCS** branch remains almost the same. This verifies our assumption that classification should learn invariant features, while localization may learn co-variant features. Those findings demonstrate the importance of decoupling modules that are essentially learning different tasks.

By fusing the predicted probability from NCS and FPR, the performance was consistently improved by 0.7% - 1.0%, demonstrating that combining predictions from branches that

perceive different level of context information is important.

By adding false positive reduction and segmentation refinement network, the performance of the baseline detector (**NCS**) is correspondingly improved, showing the effectiveness of multi-task learning and feature sharing.

All together, NoduleNet outperforms a strong baseline single stage detector by 10.27%. Note that performance reported in LUNA16 may not be directly comparable to this work, because of different nodule selection criteria, and training and testing data splits. Also, this work focuses on the joint learning of nodule detection and segmentation, whereas the LUNA16 focuses only on nodule detection.

Method	0.125	0.25	0.5	1.0	2.0	4.0	8.0	Avg.
$\mathbf{N}_1$ **(NCS) [135, 137, 185, 76]**	52.17	62.51	71.09	80.46	87.27	91.07	94.43	77.00
$\mathbf{N}_2 + \mathbf{F}_c$ **(NCS)**	53.85	62.07	71.09	79.22	86.74	90.98	93.28	76.75
$\mathbf{N}_2 + \mathbf{F}_c$ **(FPR) [137]**	55.79	66.93	75.77	82.40	88.68	91.78	93.10	79.21
$\mathbf{N}_3 + \mathbf{F}_c$ **(NCS)**	53.67	63.84	74.62	83.20	88.51	92.04	94.96	78.69
$\mathbf{N}_3 + \mathbf{F}_c$ **(FPR)**	57.38	65.96	77.19	84.97	89.92	93.28	95.40	80.59
$\mathbf{N}_2 + \mathbf{F}_d$ **(NCS)**	56.15	66.93	74.54	82.23	88.59	92.22	95.05	79.39
$\mathbf{N}_2 + \mathbf{F}_d$ **(FPR)**	61.98	71.26	78.78	85.41	89.30	92.22	95.31	82.04
$\mathbf{N}_3 + \mathbf{F}_d$ **(NCS)**	61.45	70.20	78.16	84.62	90.27	93.63	96.20	82.08
$\mathbf{N}_3 + \mathbf{F}_d$ **(FPR)**	68.08	73.56	81.70	85.94	90.80	93.90	96.55	84.36
$\mathbf{N}_3 + \mathbf{F}_d$ **(FU)**	68.70	75.60	82.23	87.36	92.04	94.96	96.46	85.34
$\mathbf{N}_3 + \mathbf{F}_d + \mathbf{R}$ **(NCS)**	62.78	70.65	78.43	84.44	89.74	93.10	95.49	82.09
$\mathbf{N}_3 + \mathbf{F}_d + \mathbf{R}$ **(FPR)**	69.23	77.01	84.70	89.48	93.37	95.23	**96.55**	86.51
$\mathbf{N}_3 + \mathbf{F}_d + \mathbf{R}$ **(FU)**	**70.82**	**78.34**	**85.68**	**90.01**	**94.25**	**95.49**	96.29	**87.27**

Table 5.1: **CPM of different methods on the LIDC dataset based on six-fold cross validation. Shown are nodule detection sensitivities (unit: %) with each column denoting the threshold false positive rate per CT scan (FPs/scan). The last column denotes the average sensitivities across the seven pre-defined FPs/scan thresholds.**

Nodule segmentation performance

In Table 5.2, we compared the segmentation performance of NoduleNet to other deep learning based methods trained and tested on LIDC dataset [160, 165, 2]. NoduleNet outperformed

previous state-of-the-art deep learning based method by 0.95% on DSC, without the need to train a separate and dedicated 3D DCNN for nodule segmentation. We randomly selected several nodules for visualizing the segmentation quality (Figure 7.3).

Approach	**# Nodules**		**# Consensus**	**IoU** (%)	**DSC** (%)
	train	**test**			
Wu *et al.* [165]	1404	1404	3	N\A	73.89 ± 3.87
Aresta *et al.* [2]	1593	1593	3	55.00 ± 14.00	N\A
Wang *et al.* [160]	350	493	4	71.16 ± 12.22	82.15 ± 10.76
NoduleNet	1131	1131	3	69.98 ± 10.80	81.80 ± 8.65
NoduleNet	1131	712	4	**71.85 ± 10.48**	**83.10 ± 8.85**

Table 5.2: **IoU (%) and DSC (%) performance of nodule segmentation between different methods. "# Consensus" means each method includes nodules that are annotated by at least "# Consensus" experts.**

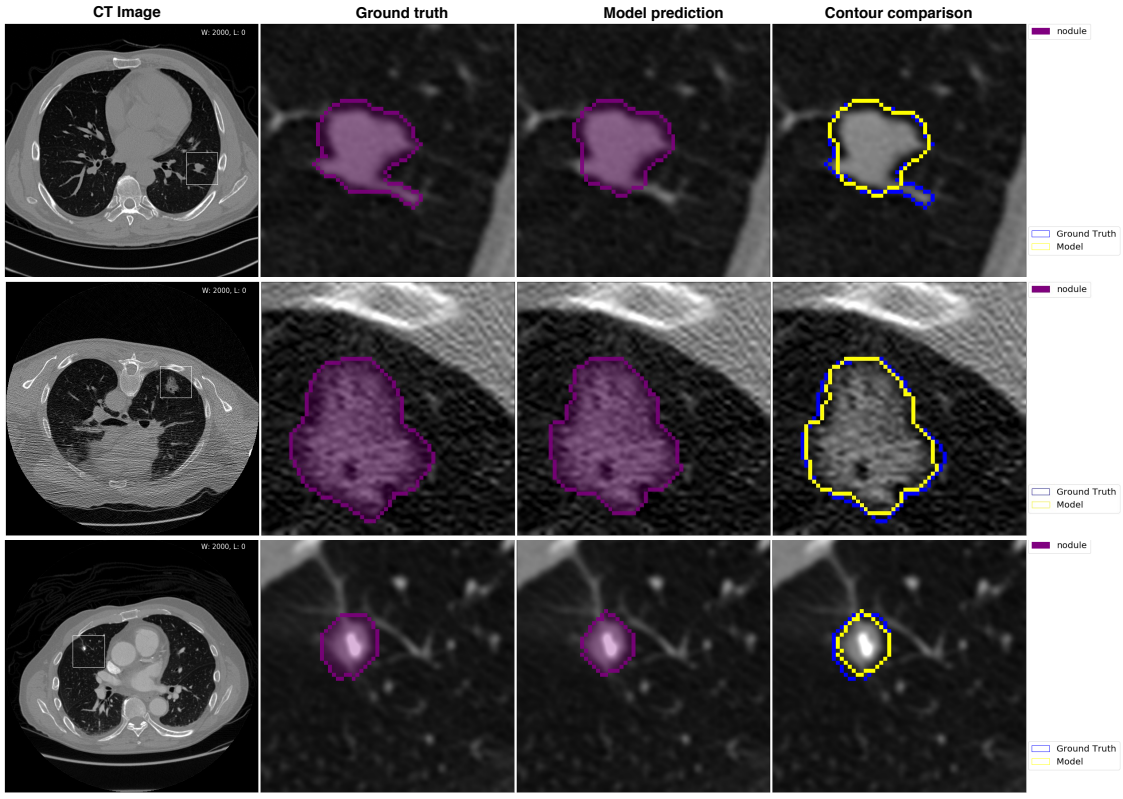

Figure 5.2: **Examples of nodule segmentation generated by NoduleNet.**

5.4 Conclusion

In this chapter, we propose a new end-to-end 3D DCNN, named NoduleNet, for solving pulmonary nodule detection, false positive reduction and segmentation jointly. We performed systematic analysis to verify the assumptions and intuitions behind the design of each component in the architecture. Cross validation results on LIDC dataset demonstrate that our model achieves a final CPM score of 87.27% on nodule detection and DSC score of 83.10% on nodule segmentation, representing current state-of-the-arts on this dataset. The techniques introduced here are general, and can be readily transferred to other models.

Chapter 6

Clinically Applicable Deep Learning Framework for Organs at Risk Delineation in CT images

6.1 Introduction

Radiation therapy (RT) is an important treatment option for many types of cancers. It can be used as an adjuvant treatment before or after surgery, or as a radical treatment administered jointly with chemotherapy [69, 98, 5, 6, 81, 100, 130, 35, 62, 18]. However, irradiation can lead to damage of normal organs ("organs at risk", OARs), especially in treating head and neck cancers, owing to the complex anatomical structures and dense distribution of important organs in this area. Damaging normal organs in the head and neck can result in a series of complications, such as xerostomia, oral mucosal damage, laryngeal edema, dysphagia, difficulty in opening mouth, visual deterioration, hearing loss, and cognitive impairment. Recently, as the efficacy of radiotherapy for head and neck cancers has been

steadily improving, irradiation complications affecting patients' quality of life have received increasing attention. Research on mitigating the toxic side effects of RT can have important clinical and social implications [169, 54, 29, 20, 32, 73, 84, 91].

A key step involved in mitigating the side effects of irradiation during RT planning is to accurately delineate all OARs so that they can be protected during radiotherapy. OARs delineation is typically done manually by radiation oncologists based on computed tomography (CT) scans, which can take significant time and efforts depending on the number of OARs to be delineated and the complexity of the local anatomical structures. In the case of head and neck, the treatment range of many tumors in the area, such as nasopharyngeal carcinoma, is relatively large and covers a large number of OARs. Manual delineation is especially daunting because of the anatomical complexity of the area. Computational tools that automatically segment anatomical regions can greatly alleviate clinicians' manual efforts, if these tools can delineate OARs accurately and within a reasonable amount of time.

A number of computational methods have been proposed to delineate OARs from CT or magnetic resonance imaging (MRI) images. Traditional OAR delineation methods are mostly atlas-based, producing OAR segmentation by aligning new images to a fixed set of manually labeled image templates. However, atlas-based methods have some limitations. First, they are computationally intensive and often take many minutes or even hours to complete, depending on hardware and implementation. Second, because of the reliance on templates, they cannot adequately account for anatomical variations naturally occurring among patients or due to the growth of tumors [19, 37, 28, 51, 74, 101, 127, 141, 143, 158, 154, 58, 39, 17, 152, 157, 36, 173].

Deep learning provides an alternative solution to the OAR delineation problem [66]. Deep convolutional neural networks (DCNNs) can learn to extract a hierarchy of complex image features directly from data and utilize these image features to segment anatomical structures, without mapping an image to templates. DCNNs have been successfully applied to segment

objects in natural images, as well as biomedical images from CT, MRI or microscopy. There are existing works of applying DCNN to head and neck OAR delineation. Most of these works apply deep learning to segment OARs from cropped local patches around regions of interest, which is achieved either manually or based on outputs from image registrations by mapping to templates [108, 88, 38, 57, 133, 166, 145, 97, 164, 12], or delineate OARs one slice at a time [92].

Recently, delineating OARs directly from whole-volume CT images has also been attempted and shown better performance than atlas-based methods [44, 144, 182, 85]. However, these studies were limited in scope and/or scale - with only a few OARs delineated and a limited number of samples tested. More importantly, previous studies have not examined the clinical utility of these methods and to what extent these methods might actually help clinicians.

In this study, we sought to address the questions described above by proposing a new deep learning model that can delineate a comprehensive set of 28 OARs in the head and neck area (Figure 6.1), trained with 215 CT samples collected and carefully annotated by experienced radiation oncologists with over ten years of experience. The accuracy of our model was compared to both previous state-of-the-art methods and a radiotherapy practitioner. Moreover, we deployed our deep learning model in actual RT planning of new patient cases, and evaluated the clinical utility of the model.

Our deep learning model addresses some key constraints of the existing deep learning methods used in this area. Most of the existing deep learning methods for OAR segmentation are based on U-Net-like network architectures [110, 87]. U-Net consists of an encoder and decoder architecture, with lateral connections concatenating low-level and high-level feature maps. The integrated image features infuse both low-level image information and high-level semantic content, which has proven very effective for precise voxel-wise classification. However, without constraining the rough location and size of an OAR, U-Net often produces many false positives (called pseudo hot-spots), scattered voxels lying outside the normal

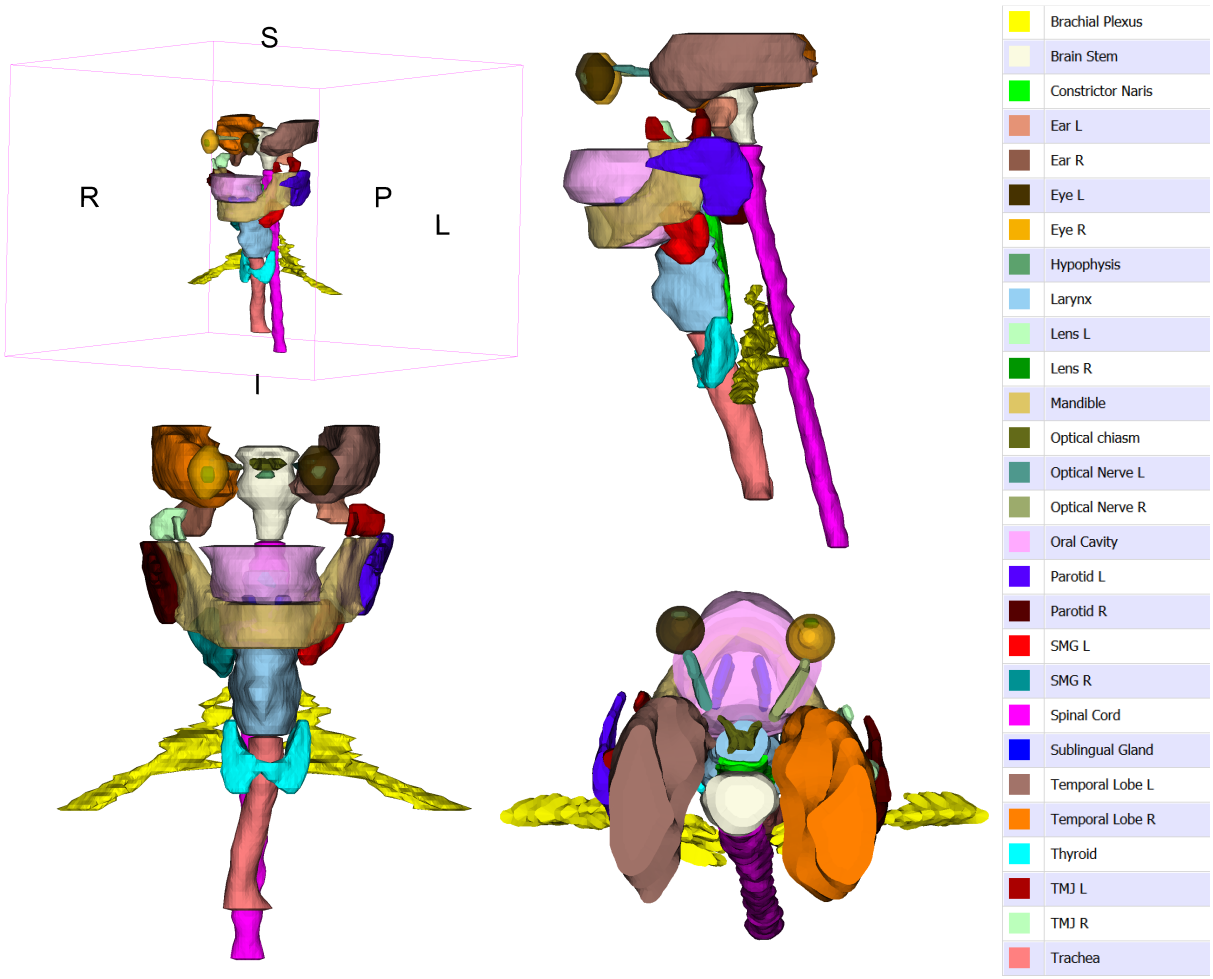

Figure 6.1: **An illustration of the 28 organs at risk (OARs) that we aim to delineate in the head and neck area. Views are from three different orientations.**

ranges of OARs (Supplementary Figure 1). These false positives are especially problematic in actual clinical applications, because it would take a significant amount of time and efforts to manually remove them from the delineation map, and, to a certain extent, it might offset the benefit received from correct predictions. In addition, there are also implementation issues when applying U-Net to whole-volume images. Training U-Net-like models based on whole volume input requires specialized graphics processing units (GPUs) with large memory. For these reasons, most of the existing deep learning methods have focused on segmenting OARs from local image patches or on a slice-by-slice basis.

To address the challenges outlined above, we propose a novel end-to-end deep learning frame-

work, named U_a-Net (stands for attention-modulated U-Net), to segment 28 OARs in the head and neck area from whole-volume CT images. The proposed framework consists of two stages: OAR detection and OAR segmentation. The first stage includes an OAR detection network, the objective of which is to identify the approximate location and size of each OAR [106], and to suppress false positive predictions outside the normal range of each OAR. The second stage includes an OAR segmentation network, which utilizes the results of the first stage as a guide and focuses on regions containing OARs to derive a fine-scale segmentation of each individual OAR. This attention mechanism allows us to perform local contrast normalization to enhance image features for improving the segmentation of OAR boundaries with low contrast. The networks of the two stages share the same underlying feature extraction backbone, based on 3D convolutions, allowing weight sharing and significantly reducing GPU memory cost. With the two-stage design, it is feasible to train and test the entire model end-to-end directly on whole-volume images, using easily accessible commodity GPUs.

In the following, we describe the detailed design of our model, the data that we have collected to train and test our model, and the evaluation results by comparing it to state-of-the-art methods and to a radiation oncologist. Furthermore, we present a study on the clinical utility of our method. In particular, we report how much time the clinicians can actually save when assisted by our model during RT planning. The work presented here provides a study of a clinically applicable deep learning model for OAR delineation in the head and neck area.

6.2 Data

We used three datasets in this study (Table 6.1). Dataset 1 contains an in-house collection of 175 CT scans with 28 OARs annotated (Figure 1) in each scan. We randomly split the data

into 145 cases for training and 30 cases for testing. Dataset 2 contains 140 CT scans from two different sources available at the Cancer Imaging Archive (TCIA) [16]. We manually delineated all 28 OARs in each of these scans, and randomly split the data with half reserved for training and the other half for testing. Altogether this brings the total number of samples used for training to 215 and the total number of samples for testing to 100. Dataset 3 is a publicly available dataset with a predetermined training (33 cases) and test (15 cases) set, each with 9 OARs annotated. It was used primarily for evaluation purposes.

Data source		# of OARs annotated	Train	Test
Dataset 1	In-house	28	145	30
Dataset 2	HNC [7]	28	18	17
	HNPETCT [147]	28	52	53
Dataset 3	PDDCA [104]	9	33	15
	Total		248	115

Table 6.1: **Datasets used in this study.**

Dataset 1 contains CT scans from the head and neck areas of patients and the delineation of OARs in each scan by radiation oncologists. The data include 175 CT scans from patients with head and neck cancer (Table 6.2), who undertook radiation therapy from 2016 to 2018. The CT scans were generated by GE Discovery 590 RT with the following scanning conditions: bulb voltage 120KV, current 300mA, scan layer thickness 2.5mm, SFOV 50cm, and scan range from top of the skull to tracheal carina.

The CT scans were manually delineated by a radiation oncologist with more than 10 years of experience, using Varian Eclipse 10.0 system for delineation and following the guidelines of [10], with assistance of MRI images from the same patients when necessary. The delineations were further reviewed and revised by a second expert with more than 30 years of clinical experience in radiation therapy. We call the manual delineations generated this way "gold standard" for both training and testing purposes, while being mindful of the caveat that there are potential subjective variants or even errors within the annotation.

A clinically relevant set of 28 OARs was delineated in each CT scan, including brachial

plexus, brain stem, constrictor naris, ears (left and right), eyes (left and right), hypophysis, larynx, lenses (left and right), mandible, optic chiasm, optic nerves (left and right), oral cavity, parotids (left and right), submandibular gland left (SMG L), submandibular gland right (SMG R), spinal cord, sublingual gland, temporal lobes (left and right), thyroid, temporomandibular joint left (TMJ L), temporomandibular joint right (TMJ R), and trachea (Figure 6.1).

		Train	Test
Number of patients (CTs)		145	30
Patient average age		61.0	65.5
Gender	Male	112	24
	Female	33	6
Tumor site	Nasopharynx	25	7
	Hypopharynx	14	7
	Larynx	18	2
	Nasal Cavity	3	1
	Brain	50	10
	Oropharynx	7	1
	Parotid Gland	4	0
	Other	24	2

Table 6.2: **Characteristics of the in-house collected data. The data consists of 175 CT scans randomly split to training and testing.**

We randomly split the dataset into a training set consisting of 145 CT scans and a test set consisting of 30 CT scans. We verified that the distributions of gender, age and tumor sites are roughly equal between the training and test sets (See Table 6.2). The training set was used to train our model, whereas the test set was used for evaluation and was not seen by the model during training.

Patient identities and other clinical information were removed from the data. Non-clinicians had only access to image data and corresponding OAR labels. The usage of the data for this study was reviewed and approved by an institutional review board (IRB) responsible for overseeing the human subject data, prior to the beginning of this study.

Dataset 2 consists of CT scans from two sources: Head-Neck Cetuximab (HNC) [7] and Head-

Neck-PET-CT (HNPETCT) [147], both available at TCIA [16]. HNC consists of image data from a clinical trial for stage III and IV head and neck carcinomas, while HNPETCT consists of imaging data from four different institutions in Québec with histologically proven head-and-neck cancer. We followed the same procedure as described in generating dataset 1 to annotate 28 OARs in each of CT scans.

Dataset 3 consists of CT scans from a public dataset called Public Domain Database for Computational Anatomy (PDDCA), used in the head and neck auto segmentation challenge at the 2015 MICCAI conference [104]. A total of 9 OARs were annotated in this dataset, including brain stem, mandible, optic chiasm, optic nerve (left and right), parotid (left and right) and submandibular gland (left and right).

Model performance

U_a-Net is an end-to-end deep learning model for OAR delineation, composed of two sub networks a) OAR detection, the objective of which is to identify approximate location and size of each OAR, and b) OAR segmentation, which extracts fine-scale image features and performs image segmentation with attention focused on individual OARs (Figure 8.1). It receives whole-volume images as input and outputs predicted masks of all 28 OARs at once. It follows the general U-Net-like structure for feature extraction, consisting of an encoder, a sequence of down-sampling blocks for extracting semantically more complex features, and a decoder, a sequence of up-sampling blocks for increasing the resolutions of image features for fine-scale segmentation. Each feature extraction block is composed of several residual sub-blocks, all based on 3D convolution to make use of three dimensional image features in volumetric CT images.

A major difference between U_a-Net and the conventional U-Net model [110] is that U_a-

Net utilizes the OAR detection module to first identify regions containing OARs, and then upsamples image features only within the detected OAR regions, instead of the whole-volume as in U-Net. The two-stage design enables the model to focus its attention on extracting high-resolution image features surrounding OARs, having the advantage of reducing false positive predictions outside the normal range of OARs and substantially cutting down the GPU memory consumption required for the up-sampling step. A detailed description of the model is provided in the Method section.

6.2.1 Performance metrics

We used volumetric Dice similarity coefficient (DSC) [104] and the 95th percentile Hausdorff distance (95% HD) [56, 104], two most commonly used metrics in this field, to evaluate the quality of OAR delineation. The OAR segmentation results of our model are represented by 28 binary masks, one for each OAR. Each binary mask is a three-dimensional array of the same size as the input CT images, with values of either 0 or 1, indicating whether the underlying voxel is part of the corresponding OAR. Let M_p and M_g be the set of voxels with value 1 in the predicted and gold standard masks, respectively. The DSC is defined as DSC$= 2|M_p \cap M_g|/(|M_p| + |M_g|)$, measuring the volumetric overlap between the two masks. In addition to DSC, we also measured Hausdorff distance between the boundaries of two masks. Let C_p and C_g denote the contours of the predicted and gold standard masks, respectively. Define max HD to be max $\{h(C_p, C_g),\ h(C_g, C_p)\}$, where $h(C_p, C_g) = \max_{a \in C_p} \min_{b \in C_g} \|a-b\|$. Because max HD is very sensitive to outliers, a more commonly used metric, called 95% HD, which measures the 95th percentile distance between two contours, is used instead.

To assess the clinical utility of our model, we calculated and compared the time that a radiation oncologist spent on delineating OARs, either from scratch or by modifying the

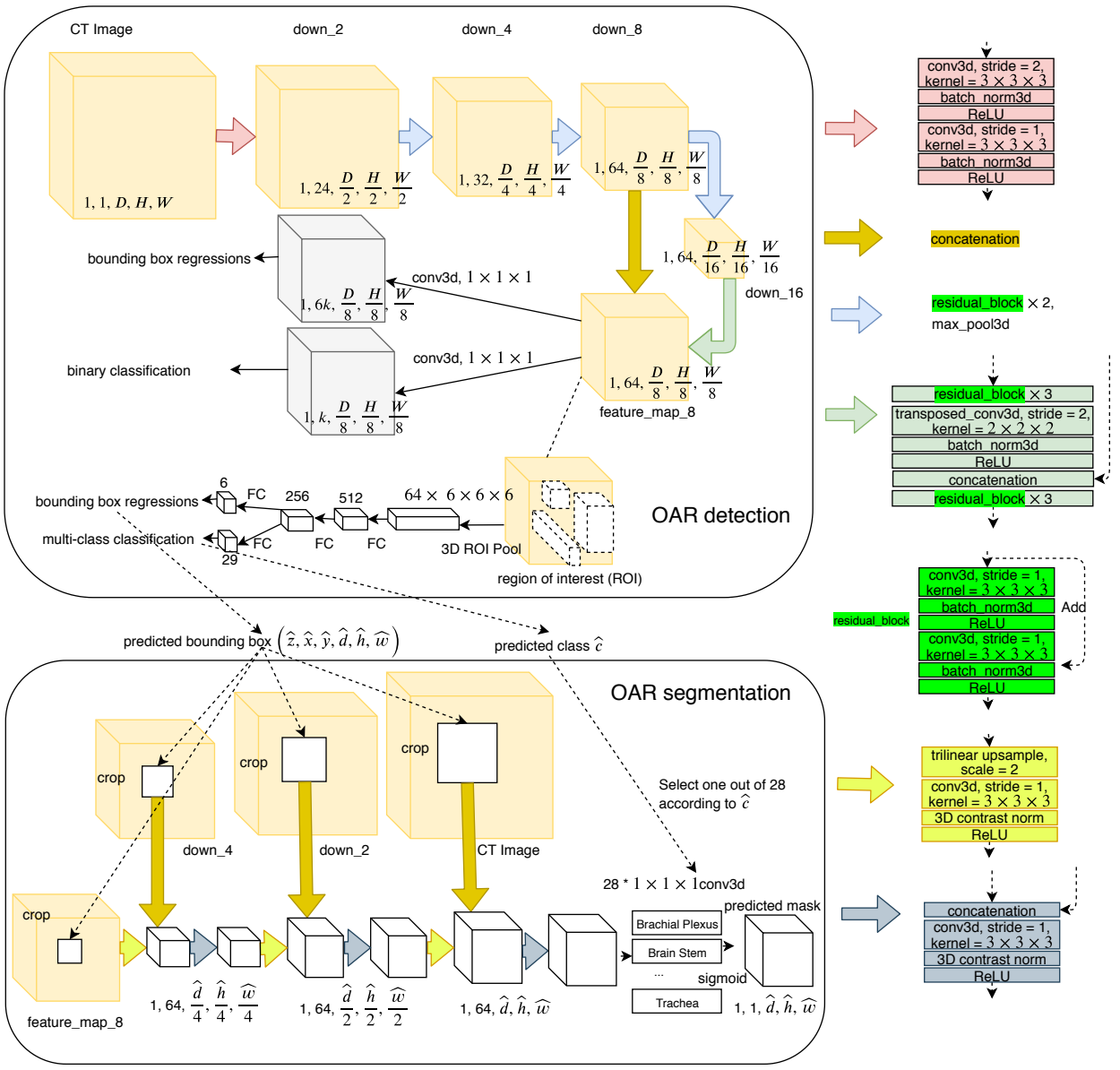

Figure 6.2: **Overview of U_a-Net. D, H, W represent the depth, height and width of the input CT image respectively. The size of each feature map is represented as (batch size, channel, depth, height, width). k is the number of anchors (12 in this study). FC is short for fully connected layer and ROI is short for region of interest. The proposed deep learning framework consists of two stages: OAR detection and OAR segmentation. OAR detection network outputs the locations and sizes of OARs in the CT image, which are then fed to the OAR segmentation network for finer grained segmentation using cropped CT image and feature maps. OAR detection and OAR segmentation network share the same underlying feature extraction backbone and can be jointly optimized. The whole framework is trained in an end-to-end fashion.**

delineation results from the model.

6.2.2 Comparison with state-of-the-art methods

The U_a-Net model was trained on 215 CT scans in the training set (from dataset 1 & 2) (See "Supplementary Methods" for model training details). Next, we evaluate its segmentation accuracy on the test set, which includes 30 CT scans from dataset 1 and 70 CT scans from dataset 2. Because the two test sets are from difference sources, the performance evaluations are reported in separate tables (Table 3 & 4 for dataset 1 and Supplementary Table 5 for dataset 2) for the purpose of assessing the robustness of our model across different data sites.

We first compared U_a-Net to a state-of-the-art deep learning model called AnatomyNet [182], which has been shown to be able to significantly outperform traditional atlas-based methods for OAR delineation in terms of both accuracy and speed. AnatomyNet is representative of U-Net like models for OAR segmentation, but stands out from the rest of the deep learning models in that it can segment OARs directly on whole-volume CT images instead of extracted local patches. To ensure consistent comparisons, we trained both models on the same training dataset using the same procedure.

In terms of DSC, U_a-Net outperformed AnatomyNet in 27 out of 28 OARs with an average improvement of 4.24% (Table 6.3) on dataset 1, and in 28 out of 28 OARs with an average improvement of 5.7% (Supplementary Table 5) on dataset 2. U_a-Net performed particularly better on anatomies that are difficult to delineate under normal contrast conditions, e.g., optic chiasm and sublingual gland, likely due to the local contrast normalization mechanism implemented in U_a-Net. U_a-Net performed slightly worse than AnatomyNet on the right ear; however, the difference is relatively small.

The advantage of U_a-Net over AnatomyNet is even more obvious when evaluated in terms

of the Hausdorff distance, decreasing the average 95% HD from 21.96mm to 6.21mm. As shown in (Table 6.4), AnatomyNet is prone to produce false positives outside the normal range of OARs, which is expected because its segmentation is done on whole-volume images instead of localized OAR regions as in U_a-Net. These false positives were small in terms of the number of voxels, having less negative effects on DSC that measures volumetric overlap, but they significantly increased the Hausdorff distance.

Next we compared U_a-Net to multi-atlas segmentation (MAS), a classical OAR delineation method based on image registration (See Supplementary Methods for details of the MAS method). In both datasets, MAS generated significantly lower scores (15.56% lower average DSC score on dataset 1 and 23.16% lower on dataset 2 compared to our model), indicating that the classical method is not as competitive as the deep learning based methods.

Finally we compared the performance of our model to previous state-of-the-art results on dataset 3 (PDDCA). Table 6.5 contains a summary of previously reported delineation results, evaluated in terms of DSC on 9 OARs from dataset 3 test set. U_a-Net obtained the best delineation results on 8 out of 9 OARs, achieving an average DSC score of 81.23% across the 9 OARs, higher than all previous methods.

6.2.3 Comparison with human experts

Having demonstrated that U_a-Net performed better than both the classical and state-of-the-art deep learning methods for OAR delineation, we proceeded to compare its performance to manual delineations produced by human experts. For this purpose, we enlisted a radiation oncologist with over 10 years of professional experience, who was not involved in annotating either the training or test datasets. The radiation oncologist manually delineated the 28 OARs on the 30 CT scans in the test set from dataset 1 in accordance with the normal professional procedure, but without consulting other professionals or seeking assistance from

additional data sources such as MRI images.

OAR	MAS	AnatomyNet	U_a-Net	Human	Human[1]
Brachial Plexus	30.38 ± 15.63	50.41 ± 8.08	**56.15 ± 10.83**	33.03 ± 7.83	33.03 ± 7.83
Brain Stem	82.25 ± 7.47	82.63 ± 4.57	**86.25 ± 3.86**	83.25 ± 4.63	83.47 ± 4.35
ConstrictorNaris	66.38 ± 8.21	73.68 ± 7.56	**75.46 ± 6.13**	62.34 ± 8.63	62.34 ± 8.63
Ear L	70.38 ± 14.94	76.68 ± 5.00	**77.28 ± 4.25**	43.57 ± 12.63	43.57 ± 12.63
Ear R	70.03 ± 15.57	**78.77 ± 5.77**	78.64 ± 6.35	39.71 ± 10.81	39.71 ± 10.81
Eye L	85.96 ± 10.99	88.41 ± 3.10	**92.51 ± 2.00**	90.71 ± 2.11	90.71 ± 2.11
Eye R	82.68 ± 17.38	89.25 ± 3.38	**92.49 ± 2.34**	91.51 ± 1.79	91.51 ± 1.79
Hypophysis	43.54 ± 18.45	56.18 ± 10.01	**63.86 ± 8.73**	59.26 ± 14.77	60.15 ± 15.07
Larynx	82.60 ± 8.19	83.06 ± 7.98	**89.25 ± 3.26**	68.60 ± 6.59	68.60 ± 6.59
Lens L	46.25 ± 24.29	77.25 ± 7.92	**81.90 ± 6.88**	64.27 ± 10.06	64.27 ± 10.06
Lens R	45.53 ± 23.94	78.06 ± 7.51	**83.04 ± 5.90**	71.79 ± 9.59	71.79 ± 9.59
Mandible	83.95 ± 11.48	91.97 ± 1.71	**93.12 ± 1.41**	90.97 ± 1.46	90.97 ± 1.46
Optic Chiasm	42.08 ± 17.52	60.55 ± 11.16	64.21 ± 16.39	28.61 ± 14.40	**69.66 ± 13.26**
Optic Nerve L	59.49 ± 14.61	72.55 ± 6.55	**75.73 ± 7.26**	65.10 ± 8.44	69.80 ± 8.87
Optic Nerve R	59.08 ± 16.53	72.95 ± 7.90	**76.06 ± 6.49**	66.14 ± 7.29	68.70 ± 8.85
Oral Cavity	86.10 ± 9.11	87.69 ± 5.67	**90.77 ± 2.32**	79.30 ± 3.59	79.30 ± 3.59
Parotid L	72.52 ± 15.57	82.28 ± 6.71	**84.86 ± 4.22**	78.46 ± 4.90	78.46 ± 4.90
Parotid R	71.20 ± 17.55	82.20 ± 7.26	**84.93 ± 3.99**	78.88 ± 4.41	78.88 ± 4.41
SMG L	60.89 ± 12.11	75.47 ± 8.93	**80.71 ± 7.32**	77.73 ± 6.25	77.73 ± 6.25
SMG R	63.70 ± 15.80	74.82 ± 14.69	**82.54 ± 7.47**	74.10 ± 16.92	74.10 ± 16.92
Spinal Cord	77.42 ± 16.70	80.32 ± 6.48	**85.64 ± 5.90**	84.59 ± 6.62	84.59 ± 6.62
Sublingual Gland	21.52 ± 16.34	39.94 ± 21.02	**45.99 ± 18.84**	35.16 ± 23.87	35.16 ± 23.87
Temporal Lobe L	80.05 ± 7.28	81.76 ± 5.33	**84.78 ± 2.62**	82.41 ± 5.01	82.79 ± 4.53
Temporal Lobe R	78.26 ± 7.40	72.97 ± 14.60	**84.13 ± 3.34**	80.90 ± 7.49	81.40 ± 7.62
Thyroid	63.68 ± 19.65	71.82 ± 11.40	**85.62 ± 4.63**	82.42 ± 6.16	82.42 ± 6.16
TMJ L	61.26 ± 19.86	86.65 ± 3.34	**87.96 ± 3.12**	84.67 ± 5.09	84.67 ± 5.09
TMJ R	63.45 ± 20.48	85.73 ± 3.69	**86.86 ± 3.60**	81.98 ± 8.59	81.98 ± 8.59
Trachea	65.86 ± 18.75	79.34 ± 7.75	81.29 ± 4.84	**91.05 ± 1.69**	**91.05 ± 1.69**
Average	64.87	76.19	**80.43**	70.38	72.17

Human[1] The same radiation oncologist modified his/her previous delineation by referencing the corresponding MRI images.

Table 6.3: **Dice similarity coefficient (DSC) comparison on the test set of dataset 1 (unit: %). L is short for left and R is short for right. SMG is short for submandibular gland and TMJ is short for temporomandibular joint.**

In terms of DSC, both U_a-Net and AnatomyNet performed better than the human expert. U_a-Net outperformed the human expert on 27 out of 28 OARs, with an average improvement of 10.15%. The human expert's delineation had lowest DSC scores on optic chiasm (28.61), brachial plexus (33.03) and sublingual gland (35.16), highlighting the challenge of manually delineating these organs, which are small in size and have a relatively low contrast in CT images. The gap between the human expert and our deep learning model delineation was smaller when the results were evaluated using Hausdorff distance. Both the human expert

OAR	MAS	AnatomyNet	U_a-Net	Human	Human[1]
Brachial Plexus	30.73 ± 25.99	37.97 ± 36.44	**18.27 ± 14.53**	43.20 ± 18.19	43.20 ± 18.19
Brain Stem	6.62 ± 2.99	5.30 ± 1.36	**4.75 ± 1.58**	5.04 ± 1.28	4.89 ± 1.22
ConstrictorNaris	9.49 ± 8.78	8.14 ± 6.20	**5.71 ± 3.34**	12.92 ± 7.57	12.92 ± 7.57
Ear L	9.81 ± 19.85	25.88 ± 76.28	**5.04 ± 1.35**	11.08 ± 3.49	11.08 ± 3.49
Ear R	9.01 ± 16.04	23.28 ± 68.48	**4.67 ± 1.42**	13.44 ± 3.31	13.44 ± 3.31
Eye L	4.73 ± 4.53	2.97 ± 0.54	**2.44 ± 0.75**	3.03 ± 0.67	3.03 ± 0.67
Eye R	6.37 ± 7.95	2.95 ± 0.54	**2.52 ± 1.16**	2.62 ± 0.61	2.62 ± 0.61
Hypophysis	5.25 ± 2.33	3.62 ± 0.74	3.23 ± 0.75	3.25 ± 0.80	**3.20 ± 0.98**
Larynx	11.53 ± 11.14	64.29 ± 89.51	**6.15 ± 5.83**	13.35 ± 7.26	13.35 ± 7.26
Lens L	5.69 ± 7.02	2.32 ± 0.82	**1.95 ± 0.80**	2.98 ± 0.63	2.98 ± 0.63
Lens R	6.77 ± 9.82	**2.00 ± 0.74**	2.07 ± 1.35	2.50 ± 0.76	2.50 ± 0.76
Mandible	8.19 ± 12.53	7.14 ± 21.95	**2.48 ± 0.83**	2.54 ± 0.30	2.54 ± 0.30
Optic Chiasm	7.34 ± 3.89	4.45 ± 1.19	4.58 ± 2.36	6.90 ± 1.81	**4.19 ± 1.70**
Optic Nerve L	6.94 ± 4.54	5.28 ± 3.68	**3.31 ± 0.91**	4.35 ± 1.29	3.75 ± 1.12
Optic Nerve R	7.53 ± 7.52	4.67 ± 1.84	4.11 ± 1.31	4.09 ± 1.24	**3.61 ± 1.05**
Oral Cavity	12.65 ± 9.60	12.79 ± 21.54	**7.38 ± 2.05**	11.83 ± 3.10	11.83 ± 3.10
Parotid L	11.49 ± 15.45	19.29 ± 65.05	**6.98 ± 3.83**	8.83 ± 3.65	8.83 ± 3.65
Parotid R	11.22 ± 12.84	20.67 ± 63.34	**6.26 ± 2.67**	8.54 ± 4.22	8.54 ± 4.22
SMG L	8.87 ± 3.47	**6.49 ± 4.00**	6.75 ± 4.65	6.86 ± 3.72	6.86 ± 3.72
SMG R	12.14 ± 14.42	6.59 ± 4.40	**5.65 ± 3.99**	6.40 ± 3.12	6.40 ± 3.12
Spinal Cord	**6.09 ± 5.54**	99.72 ± 80.01	6.86 ± 22.03	9.04 ± 18.95	9.04 ± 18.95
Sublingual Gland	13.00 ± 12.93	7.56 ± 3.29	**7.46 ± 3.07**	9.19 ± 3.88	9.19 ± 3.88
Temporal Lobe L	13.19 ± 4.85	12.03 ± 4.34	11.32 ± 3.73	**8.32 ± 2.71**	9.79 ± 4.69
Temporal Lobe R	14.73 ± 5.27	16.17 ± 6.21	13.58 ± 4.52	**9.64 ± 4.06**	10.22 ± 5.57
Thyroid	12.52 ± 13.92	128.99 ± 83.86	**3.94 ± 2.43**	5.93 ± 6.44	5.93 ± 6.44
TMJ L	7.74 ± 6.78	12.21 ± 25.25	**2.79 ± 1.09**	2.96 ± 1.30	2.96 ± 1.30
TMJ R	10.29 ± 18.13	28.13 ± 79.13	**2.74 ± 0.68**	2.88 ± 0.98	2.88 ± 0.98
Trachea	33.68 ± 18.21	43.91 ± 56.11	20.85 ± 8.99	**2.57 ± 1.26**	**2.57 ± 1.26**
Average	10.84	21.96	**6.21**	8.01	7.94

Human[1] The same radiation oncologist modified his/her previous delineation by referring to the corresponding MRI images.

Table 6.4: **Average 95th percentile Hausdorff distance comparison on the test set of dataset 1 (unit: mm). L is short for left and R is short for right. SMG is short for submandibular gland and TMJ is short for temporomandibular joint.**

and U_a-Net did substantially better than AnatomyNet, for the reasons explained above. U_a-Net did better than the human expert (with smaller 95% HD) on most of the OARs (22 out of 28), lowering the average 95% HD from 8.01 mm to 6.28 mm - a 21% reduction. Because the Hausdorff distance is very sensitive to outliers, it was a more challenging metric for models than human experts, whose mistakes were mostly confined in regions around OARs.

In real clinical practice, clinicians may also reference MRI images during the delineation process. To benchmark the delineation quality of clinicians in a real clinical setting, we also

asked the same clinician to update the delineation results guided by inputs from both CT and MRI images. We observed a noticeable improvement in delineation quality of several OARs, especially those with low CT image contrast such as optic chiasm and optic nerves. This led to the increase of the average DSC to 72.17%. We should note that the lower score is likely contributed by multiple factors, including inter-observer variation [10, 148, 66, 158], a common issue in OAR delineation, as well as skills and experiences.

Altogether, the experimental results described above suggested that 1) the two-stage model implemented in U_a-Net was beneficial for improving the performance of deep learning models in terms of DSC, and substantially so in terms of the Hausdorff distance; and 2) the model was also capable of providing better delineation performance than human expert. Note that our model can complete the entire delineation process of a case within a couple of a second (Supplementary Table 7). By contrast, the human expert took on average 34 minutes to complete one case, highlighting the significant advantage of the deep learning model (Table 6.6).

Study	Brain Stem	Mandible	Optic Chiasm	Optic Nerve		Parotid		SMG	
				L	R	L	R	L	R
Raudashl *et al.* [104]	88.0	93.0	55.0	62.0	62.0	84.0	84.0	78.0	78.0
Fritscher *et al.* [38]			49.0 ± 9.0			81.0 ± 4.0	81.0 ± 4.0	65.0 ± 8.0	65.0 ± 8.0
Ren *et al.* [108]			58.0 ± 17.0	72.0 ± 8.0	70.0 ± 9.0				
Wang *et al.* [164]	**90.0 ± 4.0**	94.0 ± 1.0				83.0 ± 6.0	83.0 ± 6.0		
Zhu *et al.* [182]	86.7 ± 2.0	92.5 ± 2.0	53.2 ± 15.0	72.1 ± 6.0	70.6 ± 10.0	88.1 ± 2.0	87.3 ± 4.0	81.4 ± 4.0	81.3 ± 4.0
Tong *et al.* [144]	87.0 ± 3.0	93.7 ± 1.2	58.4 ± 10.3	65.3 ± 5.8	68.9 ± 4.7	83.5 ± 2.3	83.2 ± 1.4	75.5 ± 6.5	81.3 ± 6.5
Nikolov *et al.* [92]	79.5 ± 7.8	94.0 ± 2.0		71.6 ± 5.8	69.7 ± 7.1	86.7 ± 2.8	85.3 ± 6.2	76.0 ± 8.9	77.9 ± 7.4
U_a-Net (Our)	87.5 ± 2.5	**95.0 ± 0.8**	**61.5 ± 10.2**	**74.8 ± 7.1**	**72.3 ± 5.9**	**88.7 ± 1.9**	**87.5 ± 5.0**	**82.3 ± 5.2**	**81.5 ± 4.5**

Table 6.5: **Dice similarity coefficient (DSC) comparison with previous published results on PDDCA (unit: %). L is short for left and R is short for right. SMG is short for submandibular gland.**

6.2.4 Clinical performance

Having demonstrated that our model can do better than both the state-of-the-art method and human expert, we next sought to find out its clinical utility, i.e., to what extent the model can actually help clinicians. For this purpose, we conducted a study to compare the time spent by radiation oncologists in delineating OARs under two modes - without or with

assistance from our model. In the first mode, the delineation was performed completely manually from scratch. In the second mode, the delineation results of all 28 OARs from our model were provided to the clinician, who would then verify the results and revise incorrect delineations when necessary. The overall work time in this case includes the time spent on verifying the results, as well as the time spent on modifying model's predictions.

Ten new CT scans from real-life radiotherapy planning were studied. We recorded the time spent by an experienced radiation oncologist to delineate 28 OARs in each of these ten cases, operating under the two modes described above. To ensure the quality of delineation, all delineation results were checked and confirmed by a second radiation oncologist.

Without assistance from the model, the radiation oncologist spent on average 33.6 ± 2.55 minutes to delineate one case. By contrast, when assisted by our model, the delineation time was substantially reduced, reaching an average value of 13.1 ± 3.14 minutes (Table 6.6), representing roughly a 61% reduction in time. A paired t-test confirmed that the differences were statistically significant (p-value = 4.5e-9). We note that the radiation oncologist accepted the model delineation results on most OARs without any need of modification. Most of the modification time was spent on brachial plexus and temporal lobes, two OARs with relatively large volumes.

This study confirmed the clinical utility of our model, demonstrating that the model can generalize well on previously unseen cases, and can save clinician's time by as much as 61%, when offered as a tool to assist the clinician's manual delineation.

Visualization

We randomly selected two representative CT scans from the holdout test set for visualizing the delineation quality of the proposed method. Figure 6.3 shows the model prediction and

	Tumor category	Mode one (min)	Mode two (min)	t	p
PA1	NPC	35	13		
PA2	NPC	33	11		
PA3	NPC	30	13		
PA4	NPC	38	20		
PA5	HD	36	11		
PA6	BM	34	12		
PA7	NPC	32	8		
PA8	NPC	33	14		
PA9	BM	35	15		
PA10	BM	30	14		
Average		33.60 ± 2.55	13.10 ± 3.14	21.68	4.5e-9

Table 6.6: **Time comparison for oncologist delineating using two approaches. NPC is short for nasopharyngeal carcinoma, HD for hodgkin disease and MD for brain metastasis. Mode one refers to delineation performed completely manually, without any assistance of computational tools, and mode two refers to delineation done by modifying the delineation result generated by the proposed method. t denotes the paired t-test statistic, and p denotes the p-value.**

manual delineation results of 28 OARs in axial planes. Figure 6.4 shows an unusual case where the head is tilted. In both cases, the model was able to generate delineations that matched well to the results produced by human experts.

6.3 Discussion

In this study, we presented a new deep learning model to automatically delineate OARs in the head and neck area. We demonstrated that the new model improved the delineation accuracy over the state-of-the-art method by 5.18% in terms of the DSC score, and substantially more in terms of the Hausdorff distance. In addition, we showed that the model also performed better than a radiation oncologist, achieving 10.15% higher DSC and 1.80 mm lower 95%HD, when averaged across 28 OARs.

The success of the model can be attributed to its two-stage design. Different from most existing deep learning models in this field, which are based on U-Net or its variants, U_a-Net

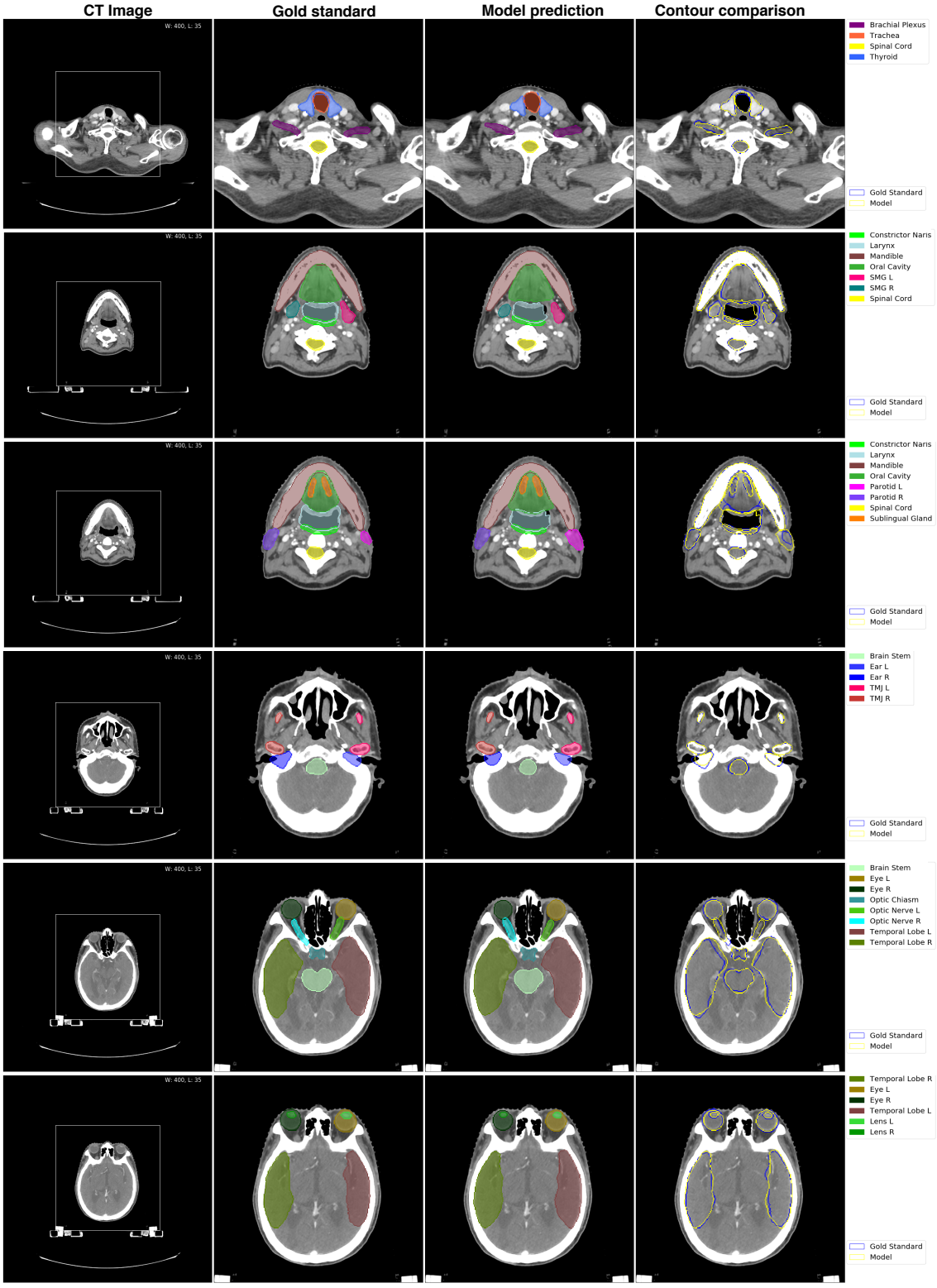

Figure 6.3: **Visualization of one randomly selected CT scan from test set. The six rows are six representative axial slices in the CT image showing the 28 OARs. The first column is the original CT scan, and the second to the fourth columns are zoomed in version of the white square box on the original CT image. The second column shows the delineation of oncologists, the third column shows prediction generated by the proposed deep learning method and the fourth column shows comparison of the two contour outlines.**

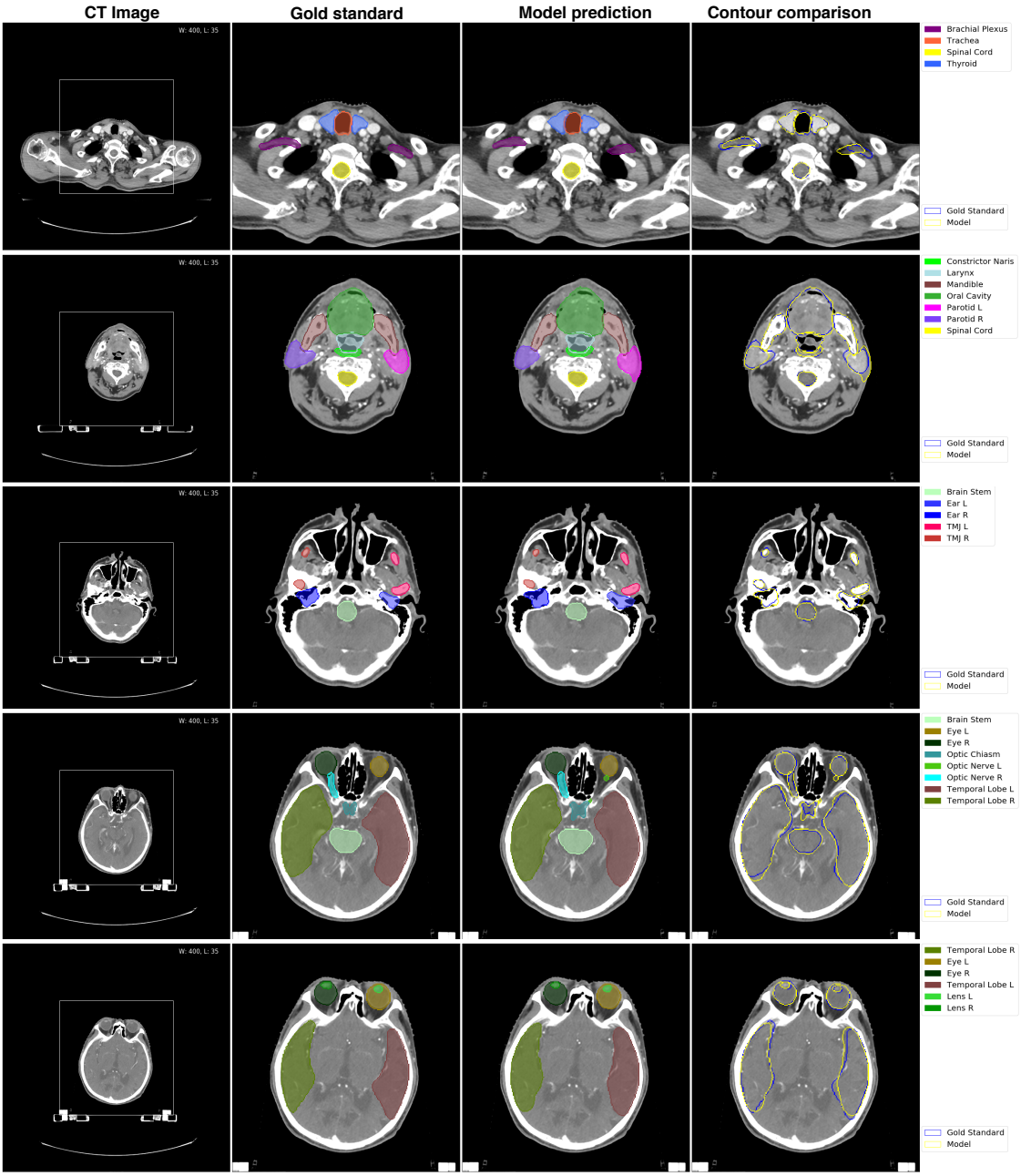

Figure 6.4: **Visualization of a second randomly selected CT scan from test set. The five rows are five representative axial slices in the CT image. The first column is the original CT scan, and the second to the fourth columns are zoomed in version of the white square box on the original CT image. The second column shows the delineation of oncologists, the third column shows prediction generated by the proposed deep learning method and the fourth column shows comparison of the two contour outlines. This CT was taken when the head of the patient was tilted. However, the proposed framework can still generate accurate segmentation, demonstrating the robustness of the framework.**

first identifies regions containing OARs and then focuses on extracting image features around these focal areas. There are several advantages with this design. First, it allows the model to perform local contrast normalization within each detected anatomy, which we notice has significant impact on delineating anatomies with low CT image contrast. In addition, it becomes more efficient to train and force the model to learn better features to segment boundaries of OARs, because the segmentation loss function is now confined to the local OAR regions. Second, the design significantly reduces false positives, effectively eliminating outliers away from the normal range of OARs. This is reflected in the improvement of the Hausdorff scores. Third, the design cuts down GPU memory consumption and is more computationally efficient. Training deep neural nets on volumetric CT images is computationally intensive and requires hardware with large GPU memory, which has become a bottleneck in many deep learning applications in this field. In models developed for OAR delineation, most of the GPU memory consumption occurs at the last few layers, where the image feature maps were upsampled to have the same spatial resolution as the original CT image. Our model only upsamples feature maps containing OARs, and thus is able to drastically cut down GPU memory consumption. With our model, it becomes feasible to delineate all 28 OARs from whole-volume CT images using only commodity GPUs (e.g., with 11 Gb memory), which we believe is important for the method to be able to be deployed in actual clinics.

This study has several important limitations. First, only CT images were used by our model to delineate OARs. Some anatomies, such as the optic chiasm, have low contrast on CT and are difficult to delineate based on the CT modality alone. It is important to integrate images from other modalities (e.g., MRI) into the deep learning framework to further improve the delineation accuracy. Second, although we have taken great care in generating gold standard annotations (See Data subsection), these annotations were still carried out manually by human experts, with the caveat of potential subjective variations and even errors. We showed that an independent human expert only reached an average DSC score of 70.38 (72.17% if the

human experts reference MRI images in addition to CT) on the test of dataset 1 (Table 6.3). Some of these discrepancies can be attributed to the inter-observer variation among experts [10, 148, 66, 158], although they have been trained to follow the same delineation guidelines and procedures. In this regard, further improving the annotation quality and generating an industry-wide standardized dataset will be necessary in the future. Nonetheless, our study suggested that the deep learning model provides an attractive solution for standardizing the delineation process and ensuring consistent results across institutions and individuals. Third, the dataset used here is relatively small for deep network training. In dealing with this constraint, we have limited the number of layers and the number of free parameters to control the complexity of our model, and have augmented the training data through affine and elastic transformation (Supplementary Methods). However, more data collected from a more diverse set of sources will be needed to improve the cross-domain adaptation and generalization of the model.

In summary, we demonstrate that our proposed deep learning approach can accurately delineate OARs in head and neck with an accuracy comparable to an experienced radiation oncologist. It is clinically applicable and can already save approximately 2/3 of a clinician's time spent on OAR delineation. With further improvements on model and data, it is conceivable that the time-consuming OAR delineation process critical for radiation therapy planning may be fully automated by deep learning methods.

6.4 Method

U_a-Net consists of two submodules - one for OAR detection and the other for OAR segmentation. The goal of the OAR detection module is to identify the location and size of each OAR from the CT images, while the goal of the OAR segmentation module is to further segment OAR within each detected OAR region. The overall diagram of the network architecture is

shown in Figure 8.1.

6.4.1 OAR detection module

The OAR detection module receives a whole-volume CT image as input (with dimension $D \times H \times W$ denoting depth, height and width respectively), and extracts images features through a sequence of down-sampling blocks, followed by an up-sampling block. Each down-sampling block is composed of two residual sub-blocks, all based on 3D convolution, reducing resolution by one-half along each axis after each down-sampling. The last down-sampling block is up-sampled to a final feature map of 64 channels with size $D/8 \times H/8 \times W/8$ (feature_map_8 in Figure 2) through transposed convolution and by concatenating feature maps of the same size from the corresponding down-sampling block. OAR candidate screening is carried out based on this final feature map, with one head for bounding box regression and one head for binary classification (to be detailed in the following). Detected OAR candidate bounding boxes further undergo a 3D ROI-pooling [41] step to generate feature maps of fixed sizes, which are then used for further bounding box regression and for multi-class classification to identify the class label associated with each OAR.

To generate OAR candidates, we branch the final feature map of the detection module (feature_map_8 in Figure 8.1) into two separate heads - one for bounding box regression and the other for binary classification, with each head undergoing $3 \times 3 \times 3$ convolution followed by $1 \times 1 \times 1$ convolution. Each bounding box is represented by a rectangular cuboid, defined by six parameters $\mathbf{t} = (x, y, z, d, h, w)$ with (x, y, z) denoting its center and (d, h, w) denoting its depth, height and width in original CT image coordinates. Overall 12 anchors are used to generate OAR candidates at each sliding window. The selection of the anchors and their sizes are described in Supplementary Methods.

The anchors produce a list of candidate bounding boxes. Let $\mathbf{t}_i \in R^6$ be the bounding box

parameter associated with the i-th anchor, predicted by the regression head, and p_i be the probability of the anchor being an OAR, predicted by the classification head. We minimize a multi-task loss function

$$\mathcal{L}_d(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i \mathcal{L}_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i \mathcal{L}_{reg}(\mathbf{t}_i, \mathbf{t}_i^*) \tag{6.1}$$

where the first term is a classification loss, the second term is a regression loss, and λ is a hyper parameter balancing the two losses (set to 1 in this study). N_{cls} and N_{reg} are the total number of anchors included in classification and regression loss calculation, respectively. p_i^* is 0 if the i-th anchor does not contain any OAR and 1 otherwise. $\mathbf{t}_i^*$ is the ground truth box parameter. Both $\mathbf{t}_i$ and $\mathbf{t}_i^*$ are parameterized relative to the size of the anchor box (See Supplementary Methods for details). We use weighted binary focal loss for $\mathcal{L}_{cls}$ ([79]) and smooth ℓ_1 loss for $\mathcal{L}_{reg}$.

To assign a class label to each OAR proposal, we apply an ROI-pooling step [41] to image features extracted from the feature_map_8 in regions specified by its predicted bounding box to derive a feature map with fixed dimensions. Two fully connected layers are subsequently applied to classify each OAR proposal into one of 29 classes (28 anatomies plus 1 background) and to further regress coordinates and size offsets of its bounding box. We minimize a similar multi-task loss function as Equation (6.1), with $\mathcal{L}_{cls}$ replaced by a weighted cross entropy loss of 29 classes, while the regression loss remains the same. The final output of the OAR detection network is the predicted bounding box $(\hat{x}, \hat{y}, \hat{z}, \hat{d}, \hat{h}, \hat{w})$ in the original image coordinates, and the corresponding class label $\hat{c}$ for each OAR.

6.4.2 OAR segmentation module

The goal of the segmentation module is to segment each of the OAR regions returned by the detection module. The module takes the bounding box and the class label of each OAR as

input, and produces a binary mask to delineate the OAR in the original image resolution. It starts by cropping feature maps from the feature_map_8, the location and size of which are specified by the predicted bounding box. The cropped feature maps are subsequently up-sampled by a sequence of up-sampling blocks to derive a final set of feature maps in the original CT resolution (i.e., from 1/8x to 1x resolution). Each up-sampling block is composed of a trilinear up-sampling, followed by a 3D convolution and local contrast normalization. To incorporate fine-scale local image features, we also crop image features from the feature maps derived by the down-sampling blocks (in the detection module) in regions specified by the bounding box, and concatenate them into the feature maps of the corresponding up-sampling blocks. The final segmentation feature map consists of 64 channels with size $\hat{d} \times \hat{h} \times \hat{w}$, the same as the dimension of the predicted bounding box. Finally we apply a $1 \times 1 \times 1$ 3D convolution (chosen according to the class label $\hat{c}$) to this final feature map, followed by sigmoid transformation, to generate the predicted mask $\mathbf{m}$, a set indexed by voxel coordinates with $\mathbf{m}_i$ denoting the probability of voxel i being the foreground of the OAR. The same procedure is applied to each detected OAR within a CT scan. The final binary mask $\mathbf{m}^c$ associated with OAR c is taken to be the union of all $\mathbf{m}_i$ whose predicted OAR class label is c.

The segmentation loss associated with one CT scan is defined to be

$$\mathcal{L}_s = \sum_{c=1}^{28} \mathrm{I}(c)(1 - \phi(\mathbf{m}^c, \mathbf{g}^c)) \tag{6.2}$$

where $\mathrm{I}(c)$ is an indicator function, taking 1 if OAR c is detected by the detection module and zero otherwise. $\mathbf{g}^c$ denotes the ground truth binary mask of OAR c: $\mathbf{g}^c_i = 1$ if voxel i is within the OAR and zero otherwise. $\phi(\mathbf{m}, \mathbf{g})$ computes a soft Dice score between the predicted mask $\mathbf{m}$ and the ground truth $\mathbf{g}$,

$$\phi(\mathbf{m}, \mathbf{g}) = \frac{\sum_{i=1}^{N} \mathbf{m}_i \mathbf{g}_i}{\sum_{i=1}^{N} \mathbf{m}_i \mathbf{g}_i + \alpha \sum_{i=1}^{N} \mathbf{m}_i (1 - \mathbf{g}_i) + \beta \sum_{i=1}^{N} (1 - \mathbf{m}_i) \mathbf{g}_i + \epsilon} \tag{6.3}$$

where i is a voxel index and N denotes the total number of voxels. The terms $\sum_{i=1}^{N} \mathbf{m}_i(1-\mathbf{g}_i)$ and $\sum_{i=1}^{N}(1-\mathbf{m}_i)\mathbf{g}_i$ can be understood as soft false positives (FPs) and soft false negatives (FNs), respectively. Parameters α and β control the weights of penalizing FPs and FNs, and were set to be 0.5 in this study. The ϵ term was added to ensure the numerical stability of the loss function.

Local contrast normalization

To facilitate the training of the segmentation module, we apply a response normalization step that we refer to as "local contrast normalization" to feature maps of the up-sampling blocks. It standardizes each 3D response map to zero mean and unit variance across all voxels. More specifically, let $x \in R^{C\times D\times H\times W}$ be a feature map with C channels and of dimension $D \times H \times W$. The local contrast normalization step transforms the map to $y \in R^{C\times D\times H\times W}$ with

$$y_{cijk} = \frac{x_{cijk} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} \tag{6.4}$$

where μ_c and σ_c are the mean and standard deviation of the voxel intensities within the c-th channel of the feature map x. We found that the local contrast normalization step can not only facilitate the training by making it converge faster, but also improve segmentation accuracy (See Supplementary Discussion for details.)

At a high level, U_a-Net shares similarities with Mask-RCNN [48] and the feature pyramid network (FPN) [78]. However, the overall objectives and implementation details are quite different. First, the ultimate goal of U_a-Net is on segmentation, while Mask-RCNN is on both object detection and segmentation. As a result, Mask-RCNN has two equally important parallel head - one for detection and the other for segmentation after ROI aligning. By contrast, U_a-Net is a two-stage model with only segmentation head in the second stage.

Second, U_a-Net is designed to perform segmentation on the original image resolution. It is different from FPN in that a) it does not perform multi-scale segmentation as would be in the case of FPN, and b) its bottom-up (up-sampling) path and lateral feature concatenation from the top-down path only involve cropped image/feature map regions containing the detected OARs, different from FPN where both bottom-up and top-down paths are on whole images/feature maps. These design considerations allow the model to focus its attention on fine and detailed segmentation of each individual OAR.

Chapter 7

Spatial Context-Aware Self-Attention Model For Multi-Organ Segmentation

7.1 Introduction

Segmentation of organs or lesions from CT images has great clinical implications. It can be used in multiple clinical workflows, including diagnostic interventions, treatment planning and treatment delivery [40]. Organ segmentation is an importance procedure for computer-assisted diagnostic and biomarker measurement systems [150]. Organ-at-risk (OAR) segmentation and tumor segmentation are also crucial to the planning of radiation therapy [132]. Moreover, the segmentation-based models of anatomical structures can support surgical planning and delivery [53].

Organ segmentation is typically done manually by experienced doctors. However, manually segmenting CT image by doctors is often time consuming, tedious and prune to human error, as a typical CT scan can contain up to hundreds of 2D slices. Computational tools that automatically segment organs from CT images can greatly alleviate the doctors' manual

effort, given a certain amount of accuracy is achieved.

There is a vast volume of work on organ segmentation using CT or magnetic resonance (MR) image. Traditional segmentation methods are mostly atlas-based. These methods rely on a set of accurate image templates with manual segmentation, and then use image registration to align the new image to the templates. Because of the reliance on the pre-computed templates, these methods may not adequately account for the anatomical variance due to variations in organ shapes, removal of tissues, growth of tumor and differences in image acquisition [167]. Also, registration is computationally intensive and may take up to hours to complete [19, 37, 28, 51, 74, 101, 127, 141, 143, 158, 154, 58, 39, 17, 152, 157, 36, 173].

Deep learning-based methods provide an alternative solutions with substantial accuracy improvement and speed-up. With recent advances in deep learning especially deep convolutional neural network, automatic segmentation using computer algorithm has shown great promise in achieving near human performance [4, 177, 76, 134], and various applications have been deployed in clinical practice.

Fully convolutional network [82] and U-Net [110] are two of the most widely used deep learning-based segmentation algorithms for this purpose. Many its variants have been proposed in recent years, including V-Net [87] and Attention U-Net [94]. These methods can use either 2D or 3D convolutions as its basic building component. 2D methods usually operate on a slice by slice basis, while 3D methods often operate on a 3D block or a stack of multiple 2D slices [80, 179]. The whole volume prediction can be obtained by predicting each slice or block using a sliding window. Additionally, some may stack multiple 2D slices in the input channel and use 2D convolution as a way to include some 3D features, and this is often referred as 2.5D model.

However, a CT image is inherently 3D. Cutting the images into slices or blocks often ignores the rich information and relation within the whole image volume. A big challenge in de-

veloping algorithm for consuming the whole image volume is the GPU memory limitation. Simply storing the tensors of the image features would require huge amount of GPU memory. One way is to adopt a coarse-to-fine strategy [187, 171, 163, 176, 161, 83, 168, 187, 171, 181, 11, 111], where in the first stage the organs of interest are roughly located, and in the second stage the segmentation masks are further refined by using a smaller input based on the localization. This usually requires training multiple CNNs for different stages and organs. Recently, several methods have been proposed to use the whole CT image volume as input, and achieve state-of-the-art accuracy and inference speed [134, 183, 187, 139, 43]. Despite their successes, there exists several disadvantages. First, to reduce the GPU memory consumption, these methods usually directly downsample the input in the very first convolution layer, which may lead to loss of local features. Moreover, they require carefully tailored image input size in order to fit the whole-volume image. However, they will still face GPU memory limitation if the image resolution becomes higher, because the memory requirement grows quickly with the size of the image volume. This makes previous whole-volume algorithm less effective when adapted to new dataset. Second, some of them make strong assumption on the organs/region they segment, thus lacking the ability to generalize well to other parts of the CT image.

We seek to incorporate 3D whole volume information into 2D model in a scalable way. We hypothesize that the benefits of using 3D convolution on the whole-volume image may come from its capability of modeling the shapes and relationships of the 3D anatomical structures. However, to model such shapes and relationships, we do not have to use very high-resolution image. 3D convolution on downsampled image volume may suffice to extract such information and can save a lot of computation and GPU memory. We can use 2D convolution on the original image slice to compensate for the loss of resolution. To fuse both 3D context features and 2D features, we implement a new module called multi-slice feature aggregation based on self-attention [151], which treats the 2D feature map as query and 3D context map as key, and uses self-attention to aggregate the rich 3D context information.

In this chapter, we propose a new deep learning framework named **S**patial **C**ontext-**A**ware Self-**A**ttention Model (SCAA). Our main contributions are: i) a new framework for combining 3D and 2D models that takes the whole-volume CT image as input; ii) a self-attention mechanism to filter and aggregate 3D context features from the whole volume image to guide 2D segmentation iii) the proposed method can scale to larger input volume without concerning the GPU memory limitation that common 3D methods face. Experiment results on a head and neck (HaN) dataset of 9 organs and an abdomen dataset of 11 organs show the proposed model consistently outperforms state-of-the-art methods in terms of organ segmentation accuracy, while being able to take the whole-volume CT image as input.

7.2 Method

Figure 8.1 shows the details of the proposed method. The proposed model consists of four parts: a 3D context feature encoder f^{3D}, a 2D feature encoder f^{Enc}, a multi-slice feature aggregation (MSFA) module f^{MSFA}, and at last a 2D decoder f^{Dec}. The input to the model $f : x \rightarrow f^{Dec}(f^{Enc}(f^{3D}(x)))$ is the whole CT image volume $\mathbf{I} \in R^{D \times H \times W}$, and the outputs are D 2D segmentation masks for C classes $\mathbf{m} \in R^{D \times H \times W \times C}$. D, H, W are the depth, height and width of the image volume.

7.2.1 3D context feature encoder and 2D encoder

f^{3D} first downsamples the input 3D volume $\mathbf{I}$ to $\mathbf{I}' \in R^{D^{3D} \times H^{3D} \times W^{3D}}$. D^{3D}, H^{3D}, W^{3D} are the depth, height, and width of the downsampled 3D volume. We use a downsample factor of two in this chapter. Note we can also downsample the volume to other resolutions, e.g. isotropic 4mm resolution. It then applies 3D convolution blocks three times, where each convolution block consists of two residual blocks followed by one $2 \times 2 \times 2$ max pooling,

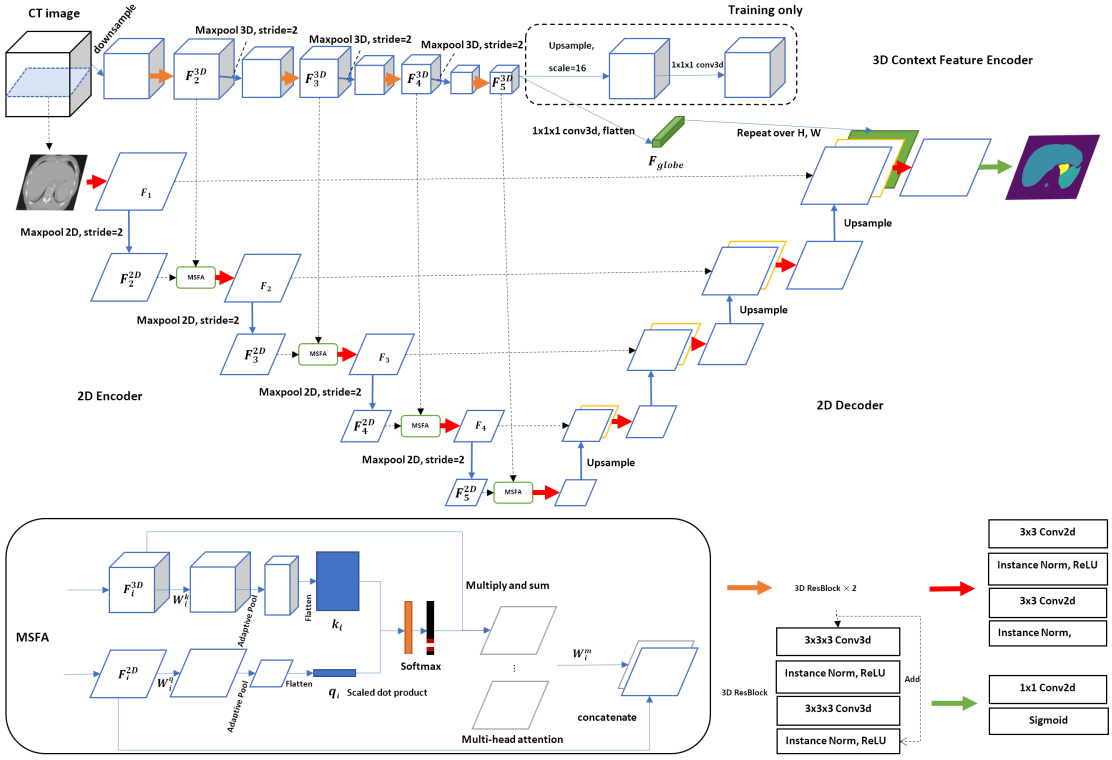

Figure 7.1: **Overview of spatial context-aware self-attention model (SCAA.)** SCAA consists of a 3D context feature encoder, a 2D encoder, a 2D decoder and a multi-slice feature aggregation (MSFA) module. SCAA starts with extracting 3D features from the downsampled CT image using 3D convolutions. Then the 2D encoder extracts 2D features and uses MSFA module to fuse 2D and 3D features hierarchically. Lastly, the 2D decoder decodes the fused 2D and 3D features and outputs 2D segmentation masks of each organ. The numbers of feature channels in F_i are 64, 96, 128, 192, 256 for $i = 1, 2, 3, 4, 5$ respectively. The number of feature channels in F_i^{2D} is 96, 128, 192, 256 for $i = 2, 3, 4, 5$ respectively. The numbers of feature channels in F_i^{3D} are 24, 32, 64, 64 for $i = 2, 3, 4, 5$ respectively. W_i^k is implemented by using $1 \times 1 \times 1$ 3D convolution with 2, 2, 4, and 4 feature channels for $i = 2, 3, 4, 5$ respectively. W_i^q is implemented by using 1×1 2D convolution with 2, 2, 4, and 4 feature channels for $i = 2, 3, 4, 5$ respectively. The xy spatial resolution of the output of the adaptive pooling are 16×16, 8×8, 4×4, 4×4 for scale $i = 2, 3, 4, 5$ respectively. The number of attention heads for scale $i = 2, 3, 4, 5$ is 2, 2, 4, and 4 respectively.

aiming at extracting 3D context features in the whole CT image.

The output of f^{3D} are four feature maps at different scales, denoted as $F_i^{3D} \in R^{C_i^{3D} \times D_i^{3D} \times H_i^{3D} \times W_i^{3D}}$, where $i = 2, 3, 4, 5$. This means the feature map F_i^{3D} is downsampled by a factor of 2^i compared to the original image. D_i^{3D}, H_i^{3D} and W_i^{3D} are the depth, height and width of the feature map at scale i, the values of which depend on the size of input image. C_i^{3D} equals 24, 32, 64 and 64 for $i = 2, 3, 4, 5$ respectively. After F_5^{3D}, we flatten the channel, depth, height and width dimension into a vector and regard it as a global descriptor F_{globe} for the 3D volume.

f^{Enc} is similar to U-Net encoder. It consumes one axial slice of the CT image $\mathbf{S} \in R^{H \times W}$ and applies 2D convolution blocks five times, where each block consists of two convolutions followed by instance normalization and ReLU activation, and a max pooling at the end. The 2D feature encoder outputs five sets of feature maps at different scales, denoted as $F_i^{2D} \in R^{C_i^{3D} \times H_i^{2D} \times W_i^{2D}}$, where $i = 1, 2, 3, 4, 5$. C_i^{2D} equals 64, 96, 128, 192 and 256 for $i = 2, 3, 4, 5$ respectively.

7.2.2 Multi-scale feature aggregation

Inspired by Transformer [151, 103], we implement a self-attention mechanism to filter and extract useful 3D context features from our 3D feature maps F_i^{3D}, and we name this module as multi-scale feature aggregation (MSFA). We regard 3D features as values, and generate queries from the 2D features and keys from the 3D features. Based on feature similarities of current 2D features and all slices in the 3D feature map (along z dimension), the MSFA will generate an attention vector $\mathbf{a_i} \in R^{D_i^{3D}}$ the same length of the depth of the 3D feature map. This attention then is applied to the 3D feature map to generate a 2D feature map that is considered as the aggregated 3D context features.

We start by mapping our 2D feature map F_i^{2D} and 3D feature map F_i^{3D} $(i = 2, 3, 4, 5)$, to one query and D_i^{3D} keys, of the same embedding space. We use a weight metric W_i^q $(1 \times 1$ 2D convolution) to generate our query $q \in R^{C_{embed} \times H_i \times W_i}$. We use a weight metric W_i^k $(1 \times 1 \times 1$ 3D convolution) to generate our keys $\{k_j\}$ of size $C_i^{embed} \times H_i^{3D} \times W_i^{3D}$, where $j = 1, 2, ..., D_i^{3D}$. An adaptive pooling operation is used to reduce the spatial resolution of the query and keys to $H_i' \times W_i'$, followed by a flatten operation to make them one dimensional. As a result, the embed dimension of the query and keys is now of size $C_i^{embed} \cdot H_i' \cdot W_i'$:

$$\begin{aligned}
(F_i^{3D'})_{c'dhw} &= (F_i^{3D})_{cdhw}(W_i^k)_{c'c} \\
(F_i^{2D'})_{c'hw} &= (F_i^{2D})_{chw}(W_i^q)_{c'c} \\
k_i &= Flatten(AdaptivePool_{D_i^{3D} H_i' W_i'}(F_i^{3D'})) \\
q_i &= Flatten(AdaptivePool_{H_i' W_i'}(F_i^{2D',}))
\end{aligned} \tag{7.1}$$

$(\cdot)_{(\cdot)}$ is the Einstein summation convention.

A scaled dot product is used to compute the response of the 2D feature map with the 3D feature map $r_j = \frac{q \cdot k_j}{\sqrt{C_i^{embed} \cdot H_i' \cdot W_i'}}$. A softmax is followed to generate our attention $\mathbf{a_i} = \text{softmax}(\mathbf{r})$. We then multiply the attention $\mathbf{a_i}$ over the depth dimension of F_i^{3D} and sum over the depth dimension to generate our aggregated context feature map $F_i^{agg} \in R^{C_i^{3D} \times H_i \times W_i}$:

$$(F_i^{agg})_{chw} = (F_i^{3D})_{cdhw}(\mathbf{a_i})_d \tag{7.2}$$

A multi-head attention mechanism is also used. We generate m_i such fused 2D feature map and then use a weight metric W_i^m $(1 \times 1$ 2D convolution) to aggregate multiple self-attention output. m_i is 2, 2, 4 and 4 for $i = 2, 3, 4, 5$ respectively. This multi-head attention allows our model to focus on different parts of the 3D context volume to extract features required

by different classes. Two 2D convolution blocks on the concatenated 2D feature map of F_i^{2D} and F_i^{agg} are used to better combine the 2D and 3D context features. F_i denotes our final 2D feature map for scale i. Note that F_1 is the same as F_1^{2D}.

7.2.3 2D decoder

f^{Dec} is similar to the U-Net decoder. Starting from F_5, a 2D upsample is used first to increase the spatial resolution by 2. Then we concatenate the upsampled features with the corresponding encoder feature map and apply one 2D convolution block. The last upsampled feature map is of the same resolution as our input image. We concatenate our 3D global descriptor F_{globe} to each pixel's feature vector and use a 1×1 convolution to obtain the final axial segmentation mask for each class $\{\mathbf{m}_c^{2D}\}$, where $c \in Z_{<C}^*$ and C is the number of classes.

7.2.4 Loss function and implementation details

The loss function is defined as:

$$L^{2D} = \sum_{c}^{N} 1 - \phi(\mathbf{m}^c, \mathbf{g}^c) \tag{7.3}$$

$\mathbf{g}$ is the ground truth segmentation for the axial slice. $\phi(\mathbf{m}, \mathbf{g})$ computes a soft Dice score between the predicted mask $\mathbf{m}$ and the ground truth $\mathbf{g}$:

$$\phi(\mathbf{m}, \mathbf{g}) = \frac{\sum_i^N \mathbf{m}_i \mathbf{g}_i}{\sum_i^N \mathbf{m}_i \mathbf{g}_i + \alpha \sum_i^N \mathbf{m}_i (1 - \mathbf{g}_i) + \beta \sum_i^N (1 - \mathbf{m}_i) \mathbf{g}_i + \epsilon} \tag{7.4}$$

N is the number of total pixels in the batch. α and β are two hyper parameters controlling the penalty for false positive and true negative respectively, and we set them to both 0.5. ϵ

is used for numerical stability.

To facilitate the training of 3D context feature encoder, we add an auxiliary 3D segmentation loss. We first upsample F_5^{3D} by a factor of 16, so it has the same spatial resolution as the downsampled 3D image volume. A $1 \times 1 \times 1$ 3D convolution is used to obtain the 3D segmentation mask $\{m_c^{3D}\}$. We use the same dice loss to get our 3D supervision loss L^{3D}. The final loss is then $L = L^{3D} + L^{2D}$.

We use one CT image for each batch during training. For each batch, we generate one 3D image volume and randomly sample 16 axial slices (batch size 16 for the 2D network). We only need to forward the 3D context encoder once per batch. We use Adam with initial learning rate 10^{-4} as optimizer for a total of 150 epochs. We applied elastic transformation and random jitter for data augmentation.

For testing, we only need to forward the 3D context feature encoder once, and we segment each 2D slice using the 2D decoder/encoder and MSFA.

7.3 Experiments

7.3.1 Datasets

Two datasets were used for evaluation: i) MICCAI 2015 head and neck (HaN)organ-at-risk (OAR) segmentation challenge dataset [104], containing a training of 33 CT images and a test of 15 CT images. The dataset contains manually labeled organ segmentation mask for 9 organs: brain stem, mandible, optic nerve left and right, optic chiasm, parotid left and right, submandibular gland (SMG) left and right; ii) an in-house abdomen multi-organ segmentation dataset* (ABD-110) containing 110 contrast enhanced CT images and

*Use of this dataset has been approved by an institutional review board (IRB).

Anatomy	U-Net	CA	C-CA	SCAA	SCAA*
Large Bowel	80.5 ± 9.4	79.6 ± 10.2	81.5 ± 9.0	81.5 ± 10.0	**82.5 ± 9.2**
Duodenum	63.4 ± 18.6	67.6 ± 17.4	69.9 ± 17.2	**71.4 ± 17.2**	70.7 ± 17.5
Spinal Cord	90.3 ± 3.8	90.4 ± 3.8	**91.0 ± 3.7**	90.7 ± 4.0	90.8 ± 3.5
Liver	95.5 ± 1.9	96.0 ± 1.9	96.2 ± 1.4	96.4 ± 1.1	**96.4 ± 1.2**
Spleen	94.6 ± 3.1	95.2 ± 2.3	95.4 ± 2.0	95.6 ± 2.3	**95.9 ± 1.4**
Small Bowel	72.2 ± 16.2	72.4 ± 16.0	75.4 ± 16.0	76.1 ± 15.1	**76.5 ± 15.3**
Pancreas	79.8 ± 9.1	79.9 ± 10.6	81.8 ± 9.4	81.8 ± 8.5	**82.1 ± 9.1**
Kidney L	95.7 ± 1.2	95.7 ± 1.8	95.8 ± 1.4	**96.0 ± 1.4**	96.0 ± 1.5
Kidney R	95.3 ± 3.0	95.5 ± 2.8	95.6 ± 3.3	95.6 ± 2.7	**95.7 ± 2.5**
Stomach	84.2 ± 16.7	85.0 ± 15.8	86.1 ± 15.6	86.8 ± 13.6	**87.5 ± 14.3**
Gallbladder	78.6 ± 19.5	78.2 ± 20.1	81.4 ± 18.1	**82.7 ± 17.2**	82.2 ± 17.7
Average	84.6	85.0	86.4	86.8	**86.9**

Anatomy	U-Net	CA	C-CA	SCAA	SCAA*
Large Bowel	9.5 ± 7.8	9.7 ± 8.3	8.9 ± 8.7	7.1 ± 4.6	**6.6 ± 5.0**
Duodenum	7.8 ± 4.9	7.4 ± 4.9	6.6 ± 4.7	6.2 ± 4.8	**5.7 ± 4.1**
Spinal Cord	1.8 ± 2.6	1.9 ± 2.8	1.8 ± 2.5	1.9 ± 3.0	**1.6 ± 2.3**
Liver	3.9 ± 3.9	2.5 ± 2.6	2.6 ± 3.4	2.1 ± 1.5	**1.9 ± 1.4**
Spleen	6.5 ± 12.8	2.4 ± 4.7	2.5 ± 7.3	1.7 ± 4.6	**1.2 ± 0.7**
Small Bowel	7.8 ± 7.3	8.1 ± 7.8	9.0 ± 11.4	8.3 ± 8.4	**7.4 ± 7.1**
Pancreas	4.1 ± 3.3	3.9 ± 3.9	3.6 ± 3.4	3.5 ± 3.5	**3.3 ± 3.7**
Kidney L	1.5 ± 1.4	1.5 ± 1.1	1.9 ± 5.4	1.2 ± 0.6	**1.2 ± 0.4**
Kidney R	1.8 ± 3.2	1.5 ± 1.6	**1.3 ± 1.0**	1.7 ± 2.4	1.3 ± 1.1
Stomach	7.2 ± 8.1	6.6 ± 7.4	**5.7 ± 7.4**	8.2 ± 10.5	5.9 ± 7.9
Gallbladder	7.0 ± 11.5	6.0 ± 7.6	6.0 ± 11.4	4.9 ± 9.1	**3.1 ± 4.6**
Average	5.4	4.7	4.5	4.3	**3.6**

Table 7.1: **Left:** DSC (unit: %). Higher the better. **Right:** 95%HD (unit: mm). Lower the better. Bold numbers represent the best performance. CA stands for context-aware model, which does not progressively integrate 3D features from the 3D model. C-CA for center context-aware model, which only integrates the corresponding center slice from the 3D feature maps. SCAA for spatial context-aware self-attention model, which uses the MSFA module to aggregate 3D features from the whole 3D volume. SCAA* for model without concatenating F_{globe} to the last feature map.

11 organs (large bowel, duodenum, spinal cord, liver, spleen, small bowel, pancreas, left and right kidney, stomach and gallbladder). The 110 CT scans were collected from 110 patients who had radiotherapy during the past three years. The CT scans were manually delineated by one experienced doctor and then manifested by another. We use the official split of training set to train the model and test on the official test set on MICCAI 2015 challenge dataset, following the same protocol as previous work [134, 92, 104, 43]. All experiments on ABD-110 dataset was conducted using 4-fold cross validation.

We report the segmentation performance using dice similarity coefficient (DSC) in percentage and 95% hausdorff distance (HD) in mm following previous work [104]. DSC measures the overlap between the predicted mask $\mathbf{m}$ and ground truth mask $\mathbf{g}$:

$$\mathbf{DSC} = \frac{2|\mathbf{m} \cup \mathbf{g}|}{|\mathbf{m} \cap \mathbf{g}|} \tag{7.5}$$

7.3.2 Ablation study on ABD-110

To compare different ways of integrating 3D features and demonstrate the contribution of each of the add-on modules in the proposed model, we conducted ablation studies with the

following different settings: 1) model with only a 3D global descriptor F_{globe} concatenated to the last feature map (CA), to demonstrate the importance of including 3D features progressively during feature extraction. 2) model without MSFA module, but only uses the corresponding center slice feature from the 3D context feature maps (C-CA), to show the effectiveness and importance of using self-attention to aggregate 3D features. 3) SCAA model without concatenating F_{globe} to the last feature map (SCAA*), to show whether F_{globe} is crucial to improve the segmentation accuracy.

As we can see from Table 8.2, by adding the 3D global descriptor to the 2D U-Net, CA outperforms the 2D U-Net by 0.4% and lowers the 95%HD by 0.7 mm, showing the importance of integrating the 3D holistic information. Next, by progressively integrating 3D features to 2D feature extractor, C-CA outperforms CA by 1.4% in DSC and 0.2 mm in 95%HD, showing the importance of integrating 3D context features hierarchically. Finally, by adding the MSFA module based on self-attention, SCAA outperforms C-CA by 0.4%, demonstrating the effectiveness of the implemented self-attention mechanism to weigh the information from different adjacent slices. To find out which part of the 3D features contribute the most, we then compare the performance of SCAA and SCAA*. SCAA* achieves average DSC of 86.9% and average 95%HD 3.6 mm, while SCAA achieves 86.8% and 4.3 mm. They are very close in terms of DSC and SCAA* wins on 95%HD. This demonstrates that F_{globe} has very little contribution, and progressively integrating 3D features with 2D encoder achieves the most performance gain.

For all following comparisons with other methods, we use the best performing configuration SCAA*.

7.3.3 Comparison with previous methods on ABD-110

To compare with previous methods of multi-organ segmentation on the ABD-110 dataset, we ran the following representative algorithms: U-Net [110], Attention U-Net [94], U_a-Net [134], and nnU-Net [60]. U-Net is a well-established medical image segmentation baseline algorithm. Attention U-Net is a multi-organ segmentation framework that uses gated attention to filter out irrelevant response in the feature maps. U_a-Net is a state-of-the-art end-to-end two-stage framework for multi-organ segmentation in the head and neck region. nnU-Net is a self-adaptive medical image semantic segmentation framework that wins the first in the Medical Segmentation Decathlon (MSD) challenge [126]. nnU-Net mainly consists of three main deep learning-based segmentation methods: a 2D U-Net (slice-wise), a 3D U-Net (patch-wise) and a coarse-to-fine cascade framework consisting of two 3D U-Nets. Its final model is an ensemble of the three methods. The above-mentioned works cover a wide range of algorithms for multi-organ segmentation and should provide a comprehensive and fair comparison to our proposed method on the in-house dataset. For 3D Attention U-Net, we followed the same preprocessing as in its original paper, to downsample the image to isotropic 2mm resolution due to GPU memory limitation. However, for all other methods, we feed the original CT image with its original image spacing.

The results are shown in table 8.1. First, by comparing 2D and 3D methods, we can see that the performance of 2D methods is on par with 3D methods on kidneys, spinal cord and liver, which is likely because those organs are usually large and have regular shapes. However, for organs like stomach, small and large bowels, 3D methods generally perform better. This may be because those organs often have more anatomical variance, and a 3D holistic understanding of the context is beneficial. Next, U_a-Net was 1.4% lower than 2D U-Net and 3.5% lower than SCAA. This may be because U_a-Net was designed mainly for the head and neck region where organs are relatively small and do not overlap too much with each other. The abdomen region, on the other hand, is more complicated as the bounding boxes

Study	Brain Stem	Mandible	Optic Chiasm	Optic Nerve		Parotid		SMG		Avg.
				L	R	L	R	L	R	
Raudashl *et al.* [104]	88.0	93.0	55.0	62.0	62.0	84.0	84.0	78.0	78.0	76.0
Fritscher *et al.* [38]			49.0 ± 9.0			81.0 ± 4.0	81.0 ± 4.0	65.0 ± 8.0	65.0 ± 8.0	-
Ren *et al.* [108]			58.0 ± 17.0	72.0 ± 8.0	70.0 ± 9.0					-
Wang *et al.* [164]	**90.0 ± 4.0**	94.0 ± 1.0				83.0 ± 6.0	83.0 ± 6.0			-
Zhu *et al.* [182]	86.7 ± 2.0	92.5 ± 2.0	53.2 ± 15.0	72.1 ± 6.0	70.6 ± 10.0	88.1 ± 2.0	87.3 ± 4.0	81.4 ± 4.0	81.3 ± 4.0	79.2
Tong *et al.* [144]	87.0 ± 3.0	93.7 ± 1.2	58.4 ± 10.3	65.3 ± 5.8	68.9 ± 4.7	83.5 ± 2.3	83.2 ± 1.4	75.5 ± 6.5	81.3 ± 6.5	77.4
Nikolov *et al.* [92]	79.5 ± 7.8	94.0 ± 2.0		71.6 ± 5.8	69.7 ± 7.1	86.7 ± 2.8	85.3 ± 6.2	76.0 ± 8.9	77.9 ± 7.4	-
Tang *et al.* [134]	87.5 ± 2.5	95.0 ± 0.8	61.5 ± 10.2	74.8 ± 7.1	72.3 ± 5.9	88.7 ± 1.9	87.5 ± 5.0	82.3 ± 5.2	81.5 ± 4.5	81.2
Guo *et al.* [43]	87.6 ± 2.8	95.1 ± 1.1	**64.5 ± 8.8**	75.3 ± 7.1	74.6 ± 5.2	88.2 ± 3.2	88.2 ± 5.2	**84.2 ± 7.3**	**83.8 ± 6.9**	82.4
SCAA (proposed)	89.2 ± 2.6	**95.2 ± 1.3**	62.0 ± 16.9	**78.4 ± 6.1**	**76.0 ± 7.5**	**89.3 ± 1.5**	**89.2 ± 2.3**	83.2 ± 4.9	80.7 ± 5.2	**82.6**

Table 7.3: **Comparison of DSC with previous methods on the MICCAI 2015 9 organs segmentation challenge.** Numbers are the higher the better (best in bold).

of some organs overlap a lot with each other (e.g. large bowel and small bowel), which makes U_a-Net less effective. Finally, comparing SCAA to nnU-Net, we find SCAA outperform nnU-Net by 1.2%. The best configuration of nnU-Net on ABD-110 was ensemble of a 2D U-Net (slice by slice) and a 3D U-Net (patch-wise). Both nnU-Net and SCAA consider the fusion of 2D model and 3D model, but they implement it in different ways - nnU-Net uses ensemble to combine 2D and 3D models while SCAA integrates 3D model into 2D model in an end-to-end fashion and jointly optimizes both models. This improvement then is likely due to the soft attention mechanism that allows SCAA to filter and extract relevant features from the large 3D context and better fuse the 2D and 3D models. Altogether, we demonstrated the effectiveness of the proposed method, which achieves an average DSC of 86.9% on the in-house dataset.

7.3.4 Performance on MICCAI2015

A second multi-organ segmentation dataset from MICCAI 2015 organ-at-risk (OAR) segmentation challenge[104] was used for evaluation. First, as we can see from Table 7.3, SCAA outperforms [92] by 4.2%. [92] used a combination of 3D and 2D convolution on 21 stacked slices for OAR segmentation. This shows the use of larger context information is beneficial for a good segmentation accuracy. Next, by comparing SCAA to AnatomyNet [182] which is a 3D model that takes the whole-volume CT as input, SCAA was 2.4% higher. This is likely due to the attention mechanism that helps the model to filter irrelevant features from

Method	Batch size	Estimate (GB)	Actual (GB)	# of parameters
2D U-Net [110]	4	2.86	3.35	34.51 M
3D U-Net [94]	1	27.96	out of memory	5.88 M
3D Attention U-Net [94]	1	17.31	out of memory	6.40 M
SCAA (3D encoder)	1	3.22	-	-
SCAA (2D U-Net & MSFA)	4	2.13	-	-
SCAA (total)		5.35	6.44	7.82 M

Table 7.4: **GPU memory consumption comparison.** We used PyTorch as the deep learning framework to measure the actual GPU memory cost.

the entire volume. Also, SCAA outperforms U_a-Net [134] by 1.4%. U_a-Net is an end-to-end two-stage model that first detects bounding box of OARs and then segments organs within the bounding box. SCAA performed better, partly because SCAA did not enforce a 'hard' attention (bounding box) but rather use 'soft' attention to enable the model focus on a smaller region. This keeps SCAA away from potential bounding box regression error and missing detection. Finally, SCAA outperforms previous state-of-the-art method [43] in 5 out of 9 organs and achieves an average DSC of 82.6%, 0.2% higher than the state-of-the-art method. Also note that [43] is a two-stage segmentation framework, which consists of two DCNNs. Our proposed method (SCAA), however, is a one-stage end-to-end solution for multi-organ segmentation, requiring less training time and computation, as well as fewer parameters.

7.3.5 Memory consumption

One advantage of the proposed method is that it significantly reduces the GPU memory while at the same time preserves the large 3D context features. To demonstrate GPU memory consumption when using whole-volume as input, we estimated and measured the actual GPU memory cost (using PyTorch as framework) for different 3D models during training in Table 7.4. We made several assumptions: i) the input image volume is of size 256 $\times$ 256 $\times$ 256. This is the size used for whole-volume input with original image spacing. ii)

we only take the memory cost of storing tensor and its gradient after each convolution and batch/instance normalization layer into consideration, because they consume the most GPU memory. iii) each number is a floating point number (32 bits). For 2D U-Net, the numbers of channel for the five scales are 64, 128, 256, 512 and 1024 respectively, as in the original implementation [110]. For the 3D U-Net, as the network has more parameters in convolution kernels, fewer channels are used in practice (16, 32, 64, 128 and 256) [94]. The memory cost and number of parameters of [110] and [94] were computed by running their released code. The GPU memory cost is for the training phase, and that of inference is approximately half of the values in Table 7.4. The actual cost is computed by running the algorithm on a GTX 1080 Ti GPU card (12 GB memory).

As seen from Table 7.4, compared to most 3D U-Net based methods, we only require 6.44 GB total memory for a batch size of four during training, which is approximately 35.1% of the 3D Attention U-Net, demonstrating the efficiency of the proposed method. Moreover, our method supports distributing batches among multiple GPU devices, which is more scalable than previous 3D methods.

7.3.6 Visualization

We visualize the attention vector $\mathbf{a_i}$ learnt for F_i^{3D} Figure 7.2 and the prediction of a random CT image from the ABD-110 dataset Figure 7.3. As seen from Figure 7.2, the 3D slice features that are useful when segmenting each 2D slice are mostly its adjacent slices. This accords with the intuition that the most prominent and useful 3D information should be mostly from its neighbouring slices. But it is also important to incorporate full 3D context information progressively. We have demonstrated the effectiveness of the self-attention mechanism in Section 7.3.2 by comparing to C-CA that only integrates the corresponding center slice feature from the 3D feature map.

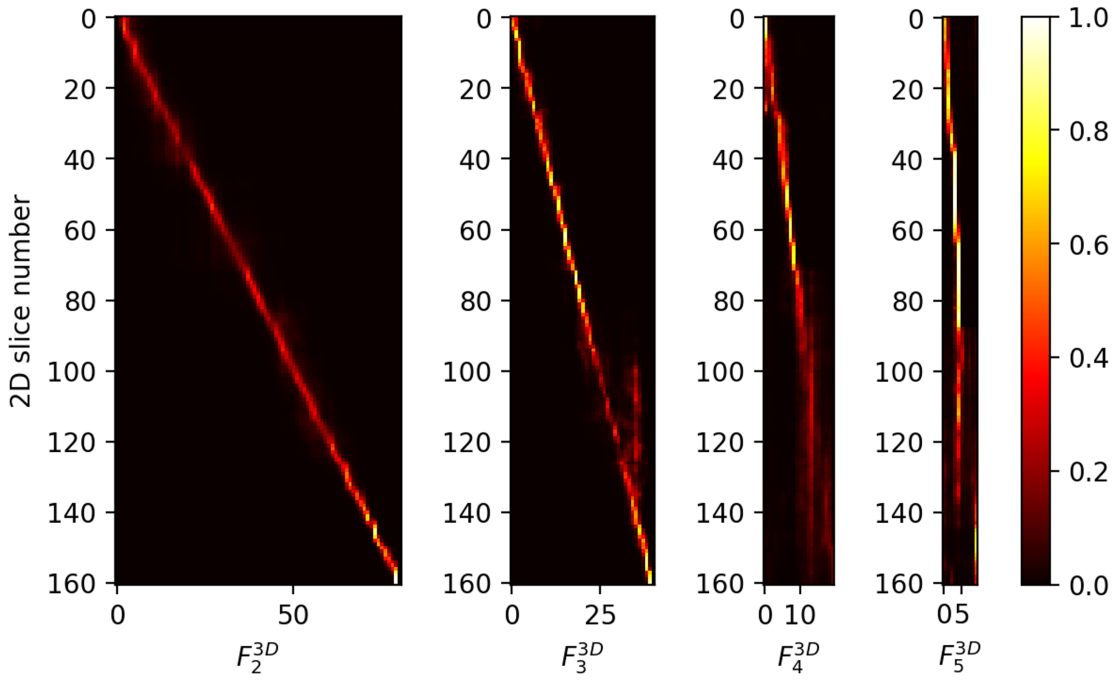

Figure 7.2: **Attention vector learnt by the proposed method.** Y-axis is the slice number of the CT image, and the X-axis is the slice number (depth) of the 3D feature map F_i^{2D}.

7.4 Conclusion

In this chapter, we propose a Spatial Context-aware Self-Attention model for multi-organ segmentation. The proposed model uses a self-attention mechanism to filter useful 3D contextual information from the large 3D whole-volume CT image to guide the segmentation of 2D slice. It addresses the GPU memory concerns that common whole volume-based 3D methods confront. Experiments on two multi-organ segmentation datasets demonstrate the state-of-the-art performance of the proposed model.

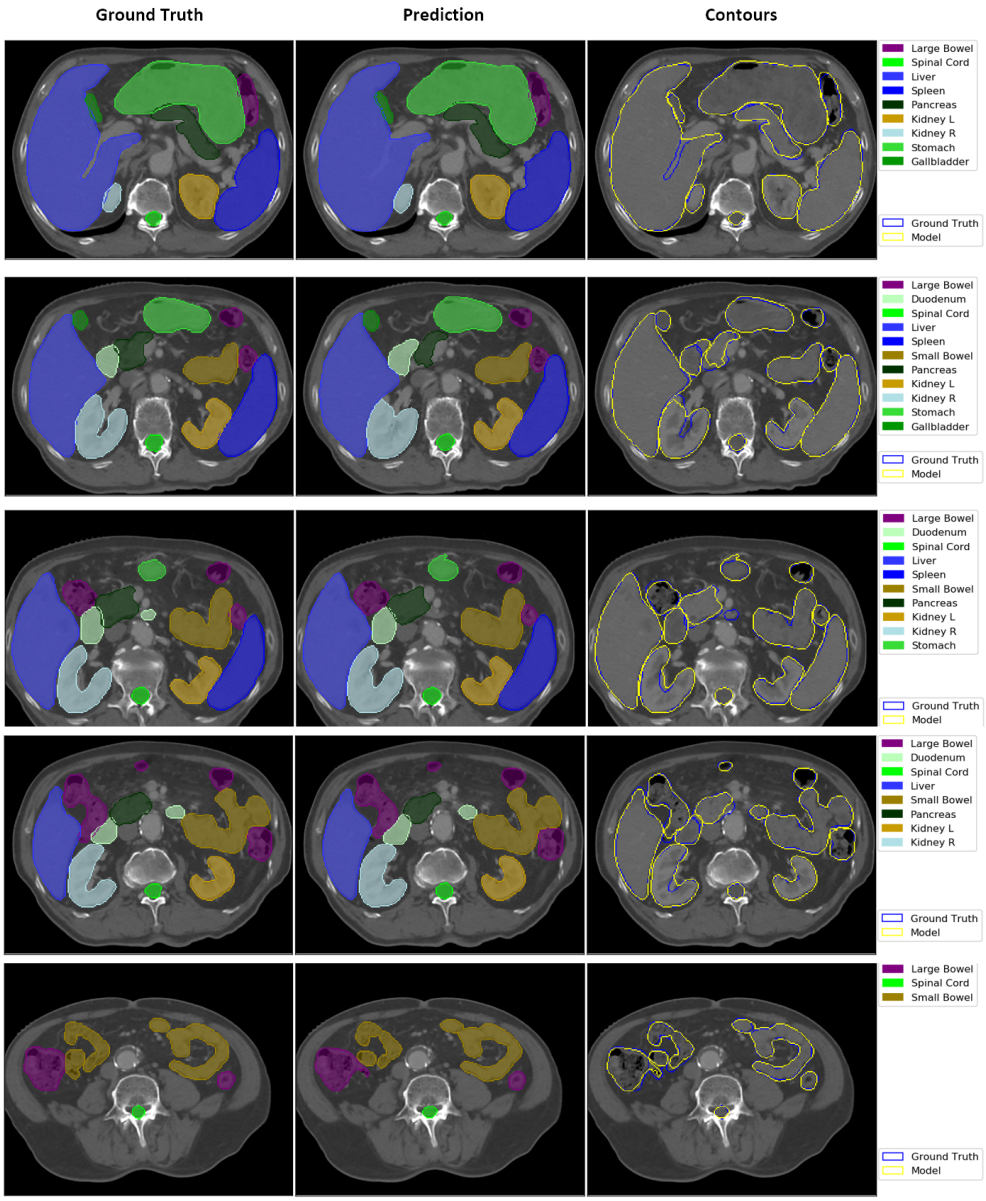

Figure 7.3: **A CT image from ABD-110 dataset.** The first and second columns are the ground truth and predicted mask overlayed on the original CT image slice, respectively. The third column shows the comparison of contours of the ground truth and predicted mask on the same slice.

Chapter 8

Recurrent Mask Refinement for Few-Shot Medical Image Segmentation

8.1 Introduction

Medical image segmentation is a fundamental task in medical image analysis. It is used in many clinical applications, including disease diagnosis, treatment planning and treatment delivery. Segmentation of anatomical structures or lesions is usually done manually by experienced doctors, which is often tedious and labor-intensive. With the recent use of deep convolutional neural networks, automated segmentation tools using computer programs can achieve near human accuracy on multiple tasks with very short processing time. However, in order to achieve good performance, these systems are usually trained in a fully supervised fashion with large amounts of annotated data. Acquiring a dataset with abundant manual labels is often very expensive and time-consuming as it requires experts with many years'

clinical experience. Moreover, the differences in image acquisition protocols among different medical equipment and institutes pose great challenges to the generalization ability of the learning based systems.

Few-shot learning has been proposed as one of the potential solutions to addressing these challenges in the low data regime [128, 131, 153, 30, 70]. The main few-shot image segmentation approach forms the problem as meta learning [33, 34, 52] and uses supervised learning to train few-shot learning models. A few-shot learning model is trained to extract class-specific features from the set of support images with annotations, and then perform segmentation on the query images by using distilled knowledge from the support images. During test time, by extracting features from a set of new support images (unseen classes), the model is able to segment novel classes. Many few-shot learning methods have been proposed and achieved great performance on natural image segmentation tasks [102, 119, 25, 122, 159, 174, 172]. However, applying few-shot learning models for medical image segmentation is still in early stages [96, 113].

Few-shot segmentation in medical images is different than that in natural images. Many approaches are based on prototypical networks [128], and often apply masked average pooling [25, 159, 174] to extract class prototypes from feature maps within the foreground mask. This step usually assumes the masked region contains sufficient features to distinguish different classes, especially foreground and background. However, this may not always be true in medical images. Distinct local appearances and context information are more critical in determining the boundary for foreground and background. A clear boundary to separate regions of interest from the background is of critical importance in medical image segmentation. Moreover, the background is usually large and spatially inhomogeneous while the foreground is small and homogeneous [95], and there exists the abundance of tissues that share very similar appearance to each other, all of which add ambiguity to define the foreground and background regions. To address this issue, we encourage the network to explicitly model the

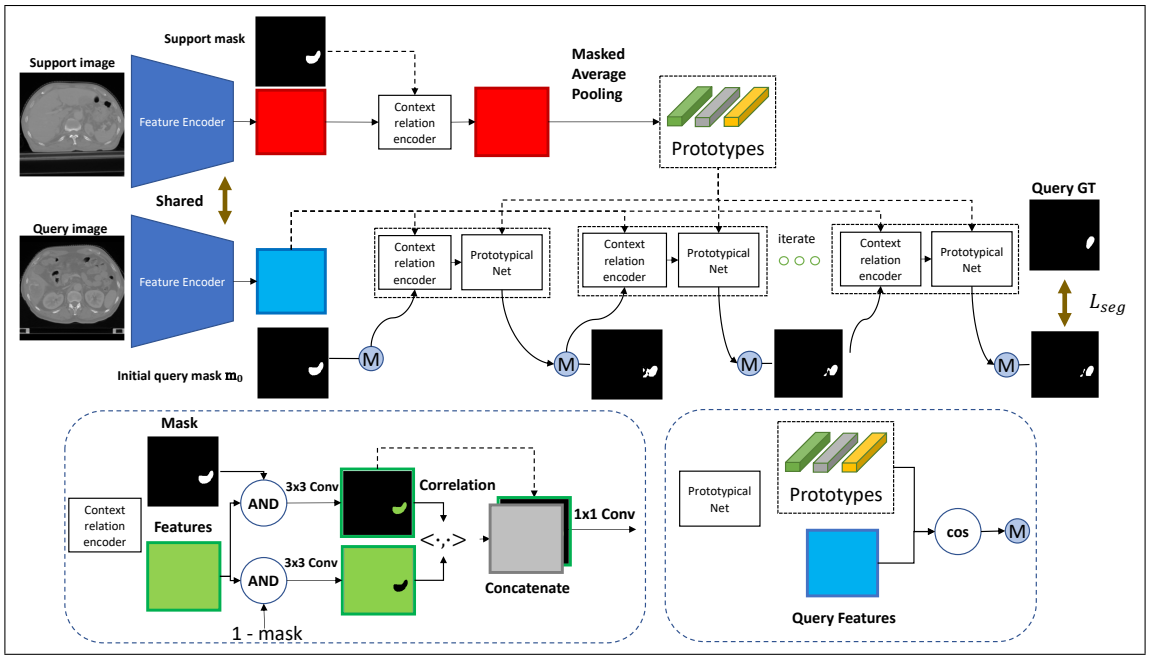

Figure 8.1: RP-Net consists of three main components: (1) A feature encoder that extracts features from both support and query images; (2) A context relation encoder (CRE) that use correlation to enhance the local context relationship features; (3) A recurrent mask refinement module that iteratively uses CRE and a prototypical network to recaptures the change of local context features and refines the mask.

context relationship between foreground and background pixels, especially pixels around the boundary.

In this work, we introduce a new network framework for few shot medical image segmentation using prototypical network (RP-Net: **R**ecurrent **P**rototypical **Net**works). First, we propose a context relation encoder (CRE) on top of the extracted features, to explicitly model the relation between foreground and background feature maps. The relationships between foreground and background regions are more important in defining the boundary of the regions of interest in medical image segmentation. To force the model to distill and utilize the local context relation information, CRE uses correlation to capture the differences in the foreground and background regions. Pixel features are augmented with the context relation features. The explicit extraction of the context relationship poses a strong constraint to the features the model would learn and forces it to focus on the boundary of the

region of interest. A prototypical network is followed to produce predicted masks using these augmented features.

Second, we propose a recurrent mask refinement module that iteratively refines the segmentation using CRE and prototypical networks. This design draws inspiration from recent works [140, 99, 55] that employ iterative refinement. More importantly, the prediction mask modifies the mask in the previous step, which results in updated local context relationship. The recurrent module serves the purpose to recapture the updated context relationship and recompute its context relationship based on new prediction. Starting from the segmentation mask from the previous step, the model uses the refined prediction mask in the previous step to compute new context features using CRE, and then feeds it to the same prototypical network. The weights of the module are shared among multiple iterations so it is fully recurrent. This recurrent module facilitates the learning and forces the model to learn to gradually refine the segmentation.

Our contributions are summarized as:

- A context relation encoder (CRE) that uses correlation between foreground and background to enhance context relationship features around the object boundary.

- A new framework for few-shot medical image segmentation that iteratively refines the prediction mask through a recurrent module that uses CRE and prototypical networks.

- We conducted experiments on two abdomen CT datasets and one abdomen MRI dataset. Experiments show that the proposed framework outperforms the SOTA few-shot framework for medical image segmentation by an average of 16.32% on ABD-110 dataset [136], 8.45% on MICCAI15 Multi-Atlas Abdomen Labeling challenge dataset [71] and 6.24% on ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge [64] in terms of DSC.

8.2 Related work

8.2.1 Medcial image segmentation

The development of the deep convolutional neural networks has lead to various successful applications in medical image segmentation, including segmentation of tissue [129, 156, 93], organ [134, 13, 144, 180, 40, 24] and lesions [139, 45, 178, 155, 75, 114]. One of the most famous and widely used network architecture is U-Net [110]. U-Net uses lateral connection to fuse features from encoders and decoders. Many its variants were proposed, with different focus on their designs. V-Net [87] extends the use of U-Net to 3D volume data. Attention U-Net [94] proposes to use gated mechanism to filter features. nnUNet [59] combines different U-Net like network architectures and automatically configure the optimal setting for different tasks, which is the best out of box U-Net. These SOTA methods require abundant manual annotations for their specific tasks to achieve good performance. They are designed to fully utilize the power of annotated dataset, and is limited when segmenting novel classes.

8.2.2 Few-shot learning

Few-shot learning can be categorized into three main focuses: data, model and algorithm [162]. One main stream of few-shot segmentation in natural image that focuses on the model is prototypical networks [128]. Prototypical network uses the idea of meta learning [33, 34, 52] and applies averaged mask pooling to pool class-specific features from the support set, which is called prototypes. Then, segmentation for the query image is done by computing the cosine distance with each class prototype . PANet [159] further improves upon this idea by proposing a prototype alignment network to better utilize the support set, by also predicting on support images using query images as support set.

In few-shot medical image segmentation, most works focus on generating new training data to enlarge the training set given only a few labels [175, 89, 96, 170]. However, this still requires retraining the model when a new class needs to be segmented. More recently, a few works focus on designing network architecture that does not require retraining the model. Squeeze and excite [113] first proposes a few-shot learning architecture specifically designed for medical image segmentation. They propose to use squeeze and excite modules to fuse information from support image on to query image to guide the segmentation arm. [95] proposes local prototypes to enrich the representation of class prototypes and a self-supervised training strategy using super pixels. Likewise, we focus on few-shot medical image segmentation without retraining the model, and we propose a new framework that uses CRE and recurrent mask refinement module to better capture local feature and shape differences around foreground object boundary.

8.3 Method

We first describe the formal definition of few-shot medical image segmentation. Next, we introduce the architecture of RP-Net, especially the context relation encoder (CRE) and recurrent mask refinement module.

8.3.1 Problem definition

In few-shot medical image segmentation task, the model is trained using images and a set of semantic labels C_{tr} drawn from a training dataset D_{tr}. During inference, the model segments a new set of semantic classes C_{te} from test images D_{te}, given a few labeled examples of C_{te}. Note that $C_{tr} \cap C_{te} = \emptyset$. For example, the model is trained using semantic labels $C_{tr} = \{\text{liver, left and right kidney}\}$ and during testing time the model needs to segment new

semantic classes $C_{te} = \{\text{spleen}\}$. Let N be the number of semantic classes in C_{te}, and K be the number of examples for each semantic class in C_{te}. The few-shot learning problem is also referred to as N-way K-shot learning. In medical image segmentation, most works usually consider 1-way 1-shot learning [113, 95].

To achieve the goal of segmenting unseen classes in inference time, an episodic training strategy is used widely [159, 95, 113]. To simulate the situation in testing time where only K examples for each class are provided, the episodic training schema randomly draws each training example in the form of a support and query data pair $[(\mathbf{x}_s, \mathbf{y}_s), (\mathbf{x}_q, \mathbf{y}_q)]$ from D_{tr}. The model is trained to distill knowledge about a semantic class from the support set $(\mathbf{x}_s, \mathbf{y}_s)$ and then apply this knowledge to segment query set $\mathbf{x}_q$. In inference time, only the K support images $\mathbf{x}_s$ and their corresponding labels $\mathbf{y}_s$ are given, and the model performs segmentation on query images $\mathbf{x}_q$.

8.3.2 Proposed method

We now introduce RP-Net for few-shot learning in medical images. For the rest of this section, we consider a 1-way K-shot learning problem. The architecture of RP-Net is shown in Figure 8.1. Our approach consists of three steps: 1) extracting image features, 2) enhancing context relation features using CRE, 3) iteratively applying CRE and prototypical network to refine the segmentation mask. All stages are differentiable and can be trained end-to-end.

Feature extraction

The input to the network is a set of K support images $\mathbf{x}_s \in R^{H \times W \times 1}$ and a query image $\mathbf{x}_q \in R^{H \times W \times 1}$, padded to the same height H and width W. The support and query images are first aligned globally using affine transformation, which is a common step in many medical

image tasks.

The model first uses the same feature encoder f_θ to extract support features $\mathbf{F}_s \in R^{H' \times W' \times Z}$ and query features $\mathbf{F}_q \in R^{H' \times W' \times Z}$ respectively. H′ and W′ are the height and width of the feature map, and C is the number of feature channels. An adapted version of the U-Net backbone was used as the feature encoder f_θ. Instead of upsampling the feature maps to the original resolution as implemented in the original U-Net, we remove the last two upsampling blocks in the U-Net to save GPU memory and computation. This results in the resolution of the support and query features being 1/4 of the image resolution ($H' = H/4, W' = W/4$).

Context relation encoder (CRE)

In medical image segmentation, the local context features are important to determine the boundary of foreground and background. To strengthen and emphasize these features, we propose the context relation encoder to enhance context features and force the model to focus on the shape and context of the region of interest rather than pixels themselves.

CRE takes the extracted features $\mathbf{F}$ (we drop subscript q and s for convenience) and foreground mask $\mathbf{m}$ as input and outputs augmented features $\mathbf{F}_{cre} = f_{cre}(\mathbf{F}, \mathbf{m}) \in R^{H' \times W' \times Z}$. $\mathbf{m}$ is the mask of the foreground class from the support image ($\mathbf{y}_s$), or the proposed foreground mask of a query image. Features of foreground and background are first extracted by masking $\mathbf{F}$ using the mask $\mathbf{m}$: $\mathbf{F}_f = \phi_f(\mathbf{F} \odot \mathbf{m})$ and $\mathbf{F}_b = \phi_b(\mathbf{F} \odot (1 - \mathbf{m}))$. ϕ_f and ϕ_b denote 3×3 convolution. Next, a correlation computation is applied to acquire the context relation features between foreground and background feature vectors at each spatial location (x, y) of $\mathbf{F}_b$ and $(x - i, x - j)$ of $\mathbf{F}_f$ with offset i and j:

$$\mathbf{C}^{(x,y,i,j)} = \sum_z \mathbf{F}_f^{(x,y,z)} \mathbf{F}_b^{(x-i,x-j,z)} \tag{8.1}$$

Instead of computing correlation between every pair of pixels on $\mathbf{F}_f$ and $\mathbf{F}_b$, we limit the maximum displacement d for comparison at each location (x, y). Given a maximum displacement d, we only compute correlation $C^{(x,y,i,j)}$ in a neighborhood of size $2d + 1$ by limiting the range of (i, j). As a result, the context relation feature $\mathbf{C}$ is of size $H' \times W' \times (2d + 1)^2$. $\mathbf{C}^{(x,y)}$ effectively captures information of how a background pixel is related to foreground when it is close to the object boundary. Finally, we concatenate $\mathbf{C}$ and $\mathbf{F}_f$ along channel dimension and apply a 1×1 convolution to fuse foreground features and context relation features to obtain $\mathbf{F}_{cre}$. d is set to 5 based on empirical results (see Table 8.2 for details).

Compared to directly computing correlation between feature maps, separating feature map into foreground and background features is important. Correlation calculated this way is sparse and has only non-zero values around the boundary, which captures the shape of the object and clearly differentiate a pixel from the background. Correlation calculated between full feature maps is not able to achieve this because it does not have the sense of boundary of the region.

Prototypical networks

Following [95, 159], we use a relative simple method for calculating the prototypes, averaging feature vectors within the mask and across support images. Given the enhanced image features of support set $\mathbf{F}_{cre,s}$, we first compute the prototype of class c via masked average pooling:

$$\mathbf{p}_c = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{x,y} \mathbf{F}_{cre,s}^{(k,x,y)} \mathbf{y}_s^{(k,x,y,c)}}{\sum_{x,y} \mathbf{y}_s^{(k,x,y,c)}} \tag{8.2}$$

where (x, y) is the index of pixels on the feature map, (x, y, c) indexes the spatial locations of the binary mask of class c and K is the number of support images.

Segmentation is done using a non-parametric metric learning method. Prototypical network calculates the distance between the query feature vector and the computed prototypes $P = \{\mathbf{p}_c | c \in C\}$. A softmax over the distances is applied to produce a probabilistic output over all classes. Formally, for each pixel at location (x, y) of query feature map $\mathbf{F}_{cre,q}$, we have:

$$\begin{aligned} \mathbf{m}_{soft} &= \cos(\mathbf{F}_{cre,q}, P), \text{and} \\ \cos(\mathbf{F}_{cre,q}, P)^{(x,y,c)} &= \frac{\exp(-\alpha d(\mathbf{F}_{cre,q}^{(x,y)}, \mathbf{p}_c))}{\sum_{\mathbf{p}_j \in P} \exp(-\alpha d(\mathbf{F}_{cre,q}^{(x,y)}, \mathbf{p}_j))} \end{aligned} \tag{8.3}$$

where the distance function d is a commonly used cosine distance and α is a scaling factor for this distance function to work best with the softmax function. α is set to 20 [159]. The class prediction can be obtained by:

$$\mathbf{m}^{(x,y)} = \arg\max_c \mathbf{m}_{soft}^{(x,y,c)} \tag{8.4}$$

Recurrent mask refinement

Since the mask $\mathbf{m}$ used to compute context relation features would change every time the network makes a prediction, we propose a recurrent mask refinement module to recapture this change and compute new context relation features based on the previous prediction.

The recurrent mask refinement module estimates a sequence of mask predictions $\{\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_n\}$ from an initial mask which is the union of all support masks: $\mathbf{m}_0 = \bigcup_{i=1}^{K} \mathbf{y}_s^i$. At each iteration t, it produces a new segmentation mask $\mathbf{m}_t$ based on $\mathbf{m}_{t-1}$. The design of this architecture mimics the steps of an optimization algorithm. For this purpose, all the weights in the recurrent module are shared across multiple iterations. The model is trained to learn to modify the mask gradually so that the final output mask $\mathbf{m}_n$ converges to an optimum solution. Note that, in this work the $\mathbf{m}_0$ is initialized using the average of support masks since images are affine aligned, but it is also possible to better initialize $\mathbf{m}_0$ using other methods.

This recurrent mask refinement module takes in support features $\mathbf{F}_s$, query features $\mathbf{F}_q$ and the mask $\mathbf{m}_{t-1}$ in previous step, uses CRE to enhance query features, and applies prototypical network to output a segmentation mask $\mathbf{m}_t$.

$$\mathbf{m}_{soft,t} = \cos(f_{cre}(\mathbf{F}_q, \mathbf{m}_{soft,t-1}), P) \tag{8.5}$$

We apply 4 iterations of the recurrent mask refinement module during training to save memory and computation cost. In inference time, we apply 10 iterations. We show in Figure 8.2 the performance at each iteration during inference time and 10 iterations are sufficient to obtain a stable result. The final prediction is obtained by upsampling $\mathbf{m}_n$ to the same resolution of the $\mathbf{x}_q$ using bilinear interpolation.

Loss function

We supervise our network using dice loss and cross entropy between the final predicted mask $\mathbf{m}_{soft,n}$ and ground truth segmentation mask $\mathbf{y}_q$:

$$\begin{aligned} L_{seg} &= \beta L_{dice} + L_{ce} \\ L_{dice} &= 1 - \frac{2\sum_{i,j,c} \mathbf{m}_{soft,n}^{(i,j,c)} \mathbf{y}_q^{(i,j,c)}}{\sum_{i,j,c} \mathbf{m}_{soft,n}^{(i,j,c)} + \sum_{i,j,c} \mathbf{y}_q^{(i,j,c)}} \\ L_{ce} &= -\frac{1}{HWC} \sum_{i,j,c} \mathbf{y}_q^{(i,j,c)} \log(\mathbf{m}_{soft,n}^{(i,j,c)}) \end{aligned} \tag{8.6}$$

where β is a constant controlling the strength of the two loss terms and is set to 1. Note that the use of the sum of dice loss and cross entropy is widely used in medical image segmentation

Dataset	Method	Spleen	Kidney L	Kidney R	Liver	mean
ABD-110	PANet-init [159]	30.95	19.24	17.64	49.91	29.43
	PANet [159]	35.89	40.22	41.54	52.36	42.50
	SE-Net [113]	29.48	37.48	37.53	19.09	30.89
	SSL-ALPANet [95]	64.90	61.58	64.05	71.83	65.59
	Affine	50.42	53.04	52.025	66.99	55.62
	RP-Net (Ours)	**78.77**	**81.89**	**85.12**	**81.88**	**81.91**
	Fully supervised [136]	95.9	95.7	95.7	96.4	95.92
ABD-30	SE-Net [113]	0.23	32.83	14.34	0.27	11.91
	PANet [159]	25.59	32.34	17.37	38.42	29.42
	SSL-ALPNet [95]	60.25	63.34	54.82	73.65	63.02
	Affine	48.99	43.44	45.67	68.93	51.75
	RP-Net (Ours)	**69.85**	**70.48**	**70.00**	**79.62**	**72.48**
	Fully supervised [180]	96.8	95.3	92.0	97.4	95.4
ABD-MR	SE-Net [113]	51.80	62.11	61.32	27.43	50.66
	PANet [159]	50.90	53.45	38.64	42.26	46.33
	SSL-ALPNet [95]	67.02	73.63	78.39	73.05	73.02
	Affine	62.87	64.70	69.10	65	65.41
	RP-Net (Ours)	**76.35**	**81.40**	**85.78**	**73.51**	**79.26**
	Fully supervised [60]	-	-	-	-	94.6

Table 8.1: DSC comparison with other methods on ABD-110, ABD-30 and ABD-MR (unit: %).

tasks, such as [60].

8.4 Experiment

8.4.1 Setup

Dataset We conducted experiments using two abdomen CT datasets and one MR dataset:

- ABD-110 is an abdomen dataset from [136] that contains 110 3D CT images from patients with various abdomen tumors and these CT scans were taken during the treatment planning stage.

Experiment	Method	Spleen	Kidney L	Kidney R	Liver	mean
Added components	Affine	50.42	53.04	52.025	66.99	55.62
	Affine + Grabcut	57.93	64.17	64.25	65.27	62.91
	Affine + Concat	56.41	52.39	54.99	70.87	58.66
	Affine + CRE	57.73	58.05	60.62	73.53	62.48
	Affine + Concat + Recurrent	59.99	60.65	62.31	83.03	66.50
	Affine + CRE + Recurrent	78.77	81.89	85.12	81.88	81.91
Backbone	VGG16	73.57	67.49	56.81	72.04	67.48
	Res18	72.39	79.13	81.61	80.89	78.50
	U-Net	78.77	81.89	85.12	81.88	81.91
Correlation radius	$d = 0$	78.40	81.90	82.12	83.89	81.58
	$d = 1$	80.03	81.87	82.09	82.1	81.52
	$d = 3$	79.12	81.79	83.41	81.32	81.41
	$d = 5$	78.77	81.89	85.12	81.88	81.91
	$d = 7$	77.56	80.25	81.77	80.22	79.95
Initialization	Affine	50.42	53.04	52.02	66.99	55.62
	Demons	63.60	63.89	61.89	73.59	65.74
	RP-Net (Affine)	78.77	81.89	85.12	81.88	81.91
	RP-Net (Demons)	80.31	83.55	85.01	82.86	82.93

Table 8.2: Ablation study on ABD-110 (unit: %). Underlined is the final configuration used in RP-Net.

- ABD-30 is an abdomen dataset from the MICCAI 2015 Multi-Atlas Abdomen Labeling challenge [71]. It contains 30 3D abdominal CT scans (ABD-30) from patients with various pathologies and has variations in intensity distributions between scans.

- ABD-MR is an MRI dataset from ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge [64]. It contains 20 3D T2-SPIR MRI scans.

We perform the same 5-fold cross validation and consider only 1-way 1-shot learning, following the same protocol as previous work setting 2 [95]. Liver, spleen and left and right kidney are used as semantic classes. Within each fold, one organ is considered as unseen semantic class for testing while the rest are used for training. Moreover, to reduce the variance by choosing only one support image during inference, following [159], for each query image in the test set we randomly sample one support image from the test set, repeat this process for 5 times and the final result is obtained by averaging the 5 runs.

Evaluation metric We use the same evaluation metric Sørensen–Dice coefficient (DSC) as in previous work [95, 113]. DSC measures the overlap of the prediction mask $\mathbf{m}$ and ground truth mask $\mathbf{g}$, and is defined as:

$$\mathrm{DSC}(\mathbf{m}, \mathbf{g}) = \frac{2|\mathbf{m} \cup \mathbf{g}|}{|\mathbf{m}| + |\mathbf{g}|} \tag{8.7}$$

Implmentation details All images are resampled to have the same xy-plane spacing of 1.25mm $\times$ 1.25mm. For segmenting 3D volume data, we follow the same protocol used in [95, 113] by dividing the support and query images into 12 chunks and segmenting all slices in the query chunk by using the center slice in the corresponding chunk of the support image. During training, a pair of support and query images and their labels are both cropped to have a fixed size of 256 $\times$ 256 around the image center. Support and query images are aligned online using affine transformation before feeding into the network. RP-Net is trained from scratch using Adam as optimizer with initial learning rate 0.0001 for 50 epochs and the learning rate is reduced by a factor of 10 every 20 epochs. We also add the alignment loss to train RP-Net as in [159].

8.4.2 Comparison with the state-of-the-art methods

Table 8.1 shows the performance comparison of RP-Net with previous work on ABD-110, ABD-30, ABD-MR respectively. PANet [159] is an extended version of the widely used prototypical network [128] designed for natural image segmentation. PANet-init means directly using the pretrained VGG16 feature extraction backbone without any finetuning on the few-shot setting. SE-Net [113] is the first specifically designed architecture for few-shot medical image segmentation. SSL-ALPANet [95] is the state-of-the-art few-shot medical image seg-

mentation framework that uses self-supervised learning and prototypical networks. Affine is the result of the accuracy after globally aligning the support and query image using affine transformation, which we use as an initial mask. [95] reported performance for PANet-init, PANet, SE-Net and SSL-ALPNet on ABD-30 and ABD-MR, so these numbers are directly quoted. We ran these algorithms using public available code to report their performance on ABD-110.

First, compared to PANet, RP-Net outperforms PANet by 39.49%, 43.06% and 21.75% on the three datasets ABD-110, ABD-30 and ABD-MR respectively. Second, compared to SE-Net, RP-Net outperforms SE-Net by 51.02%, 60.57% and 27.42% on ABD-110, ABD-30 and ABD-MR respectively. Third, compared to the state-of-the-art method SSL-ALPANet, RP-Net outperforms SSL-ALPANet by an average of 16.32%, 9.46% and 6.24% on ABD-110, ABD-30 and ABD-MR respectively.

These experiments demonstrate our approach can achieve the SOTA accuracy on medical image datasets with different image modalities (CT and MRI). Also, we focus on designing a new framework for few-shot medical image segmentation, which outperforms other approaches of the same motivation, e.g. SE-Net by a large margin. Moreover, it is worth noting that the performance gain to SSL-ALPANet decreases from dataset ABD-110 to ABD-MR. The main reason may be that the number of training data for ABD-30 and ABD-MR is small and RP-Net benefits more from more training data. Additional gain may be obtained by combining our method with the self-supervised training schema proposed in SSL-ALPANet.

8.4.3 Ablation study

Ablation experiments are conducted using the ABD-110 dataset, because it has more data compared to the other two. Table 8.2 shows the results for the following experiments.

Effect of each component To verify the contribution of the two added components - context relation encoder and recurrent module, we conducted experiments by adding one component at a time: 1) model trained and tested without the CRE. To make use of the support mask which is used in CRE, we concatenate the mask to the feature map from backbone and apply a 3×3 convolution for a fair comparison (denoted as concat). 2) model trained without recurrent module. Note that if we remove both CRE and recurrent training, the model becomes the PANet [159]. Moreover, we compare with Grabcut [112] which is an unsupervised method that uses iterated Graphcut. Grabcut can be seen as an unsupervised version of our algorithm.

First, we verify the effect of using CRE. Affine + Concat is a naive way of integrating support masks by concatenating it directly to the feature maps, which outperforms the Affine by 3.04%. Affine + CRE implements the more sophisticated way of exploring local feature differences using CRE, which outperforms the Affine + Concat by 3.82%. This shows the CRE better captures the local difference via the use of correlation. However, the performance improvement is still not significant and the reason is that the mask prediction is changed each time and it lacks a mechanism to recapture this change and recompute the new local differences. The recurrent mask refinement module serves this purpose and we discuss its effect in the next paragraph.

Second, we compare the performance of using the recurrent mask refinement module. Affine + Concat + Recurrent means we apply the recurrent module to the concatenated feature map, which performs 7.84% better than not using the recurrent module (Affine + Concat). This shows that the recurrent training indeed helps the model to find the right mask prediction because the initial mask from support is a very rough estimation of the location of the region of interests. If we combine the two added components together (Affine + CRE + Recurrent), we can achieve a big improvement by 15.39% compared to Affine + Concat + Recurrent. This demonstrates that the integration of recurrent module to recapture local

changes in the CRE is very important and can greatly boost the performance.

Third, we compare with Grabcut. Our method is in some sense similar to Grabcut - we both use an iterative update to refine the segmentation mask. Grabcut outperforms the baseline Affine by 7.29%, showing that iteratively refining a mask is indeed beneficial. RP-Net (Affine + CRE + Recurrent) outperforms Grabcut by 19%. There are mainly three reasons for this large improvement. First, Grabcut only uses one image, thus only image intensity is used to separate foreground and background region. On the contrary, RP-Net uses the support images to extract knowledge about the relationship between the foreground and background region, and utilize this knowledge to guide the segmentation of the new image. Second, Grabcut only refines the mask in the probable foreground region which is a human defined boundary and lacks the flexibility to attend other areas in the image, as well as the ability to correct error in the sure foreground region. RP-Net does not have these constraints and can potentially use information from the whole image. Third, RP-Net uses training data to train the feature extractor, while Grabcut is not a learning-based method and only uses information directly derived from pixel intensity.

Effect of feature extraction backbone We also experimented with three different feature extraction backbones - VGG16 [125], Res18 [50] and U-Net [110]. To make sure the output feature map is 1/4 of the original image resolution for a fair comparison, we only kept the first two downsampling operations in both VGG16 and Res18 backbones and the rest of the network architecture remained the same. As seen from Table 8.2, VGG16 backbone performs the worst among the three backbones, which is 8.03% lower than Res18. U-Net backbone outperforms Res18 backbone by an average of 2.32% which is mainly because of the lateral connection in U-Net that fuses both low-level and high-level features. This demonstrates that RP-Net is compatible with different backbones, and backbones that perform better on medical image segmentation task, such as U-Net, would result in similar gain when combined with RP-Net.

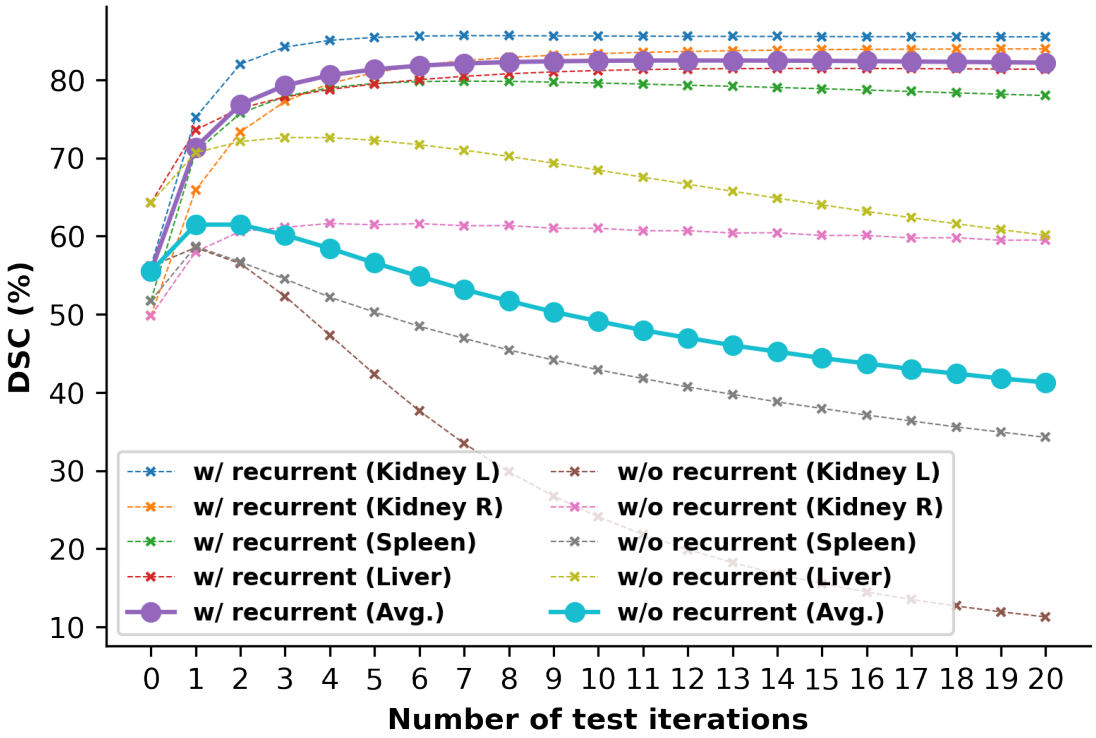

Figure 8.2: DSC at each refinement iteration. This figure shows the DSC performance of the proposed model per iteration. DSC of four organs and an average is shown for two models: one w/ recurrent training (purple) and one w/o recurrent training (cyan).

Effect of correlation radius We conducted experiments with different radius $d = 0, 1, 3, 5, 7$ in the correlation layer, which controls how many neighbouring pixels are included when computing correlation. $d = 0$ means the correlation computation is carried out only at a single point. Note that even with $d = 0$, the model is able to use features from the surrounding pixels because ϕ_f and ϕ_b are used to extract foreground and background specific features. Table 8.2 shows our approach is not very sensitive to the radius, and this is likely because RP-Net is designed to focus on a small region around the object boundary at a time, a larger context may not necessarily bring more benefits.

Effect of number of inference iterations We show in Figure 8.2 the performance at each inference iteration from one fold in ABD-110. Although the model is trained using 4 iterations of recurrent module, we can apply more iterations during inference. As seen

from this figure, a model without recurrent training diverges after the 1st iteration, while a model with recurrent training quickly converges and does not diverge after 20 epochs. It demonstrates that with the recurrent training, the model learns to gradually refine its prediction and converges to a stable solution.

Effect of initialization Demons [142] is a medical image registration method that uses deformable registration, which performs 10.12% better than a simple affine transformation. As shown in Figure 8.2, using a better initialization (Demons), RP-Net achieves a 1.02% improvement. Although better initialization improves the result, the improvement is small compared to that of the initialization itself, and our network is less sensitive to the initial mask as long as it roughly locates the foreground region. For this reason, we only use initialization mask from Affine transformation for its simplicity.

8.4.4 Qualitative result

We show in Figure 8.3 how the segmentation mask converges to the optimum solution in multiple iterations. In general, we can observe that RP-Net refines the initial mask gradually, finds a better segmentation mask at each iteration, and finally converges to an optimum solution. RP-Net is able to learn to distill knowledge about the relation between the foreground and background from the support image, and apply it to segment query images by comparing local differences and modifying its prediction to conform to the shape and boundary. Moreover, RP-Net generates satisfying segmentation masks that have a clear boundary along the object boundary, demonstrating the successful design of the CRE and recurrent module.

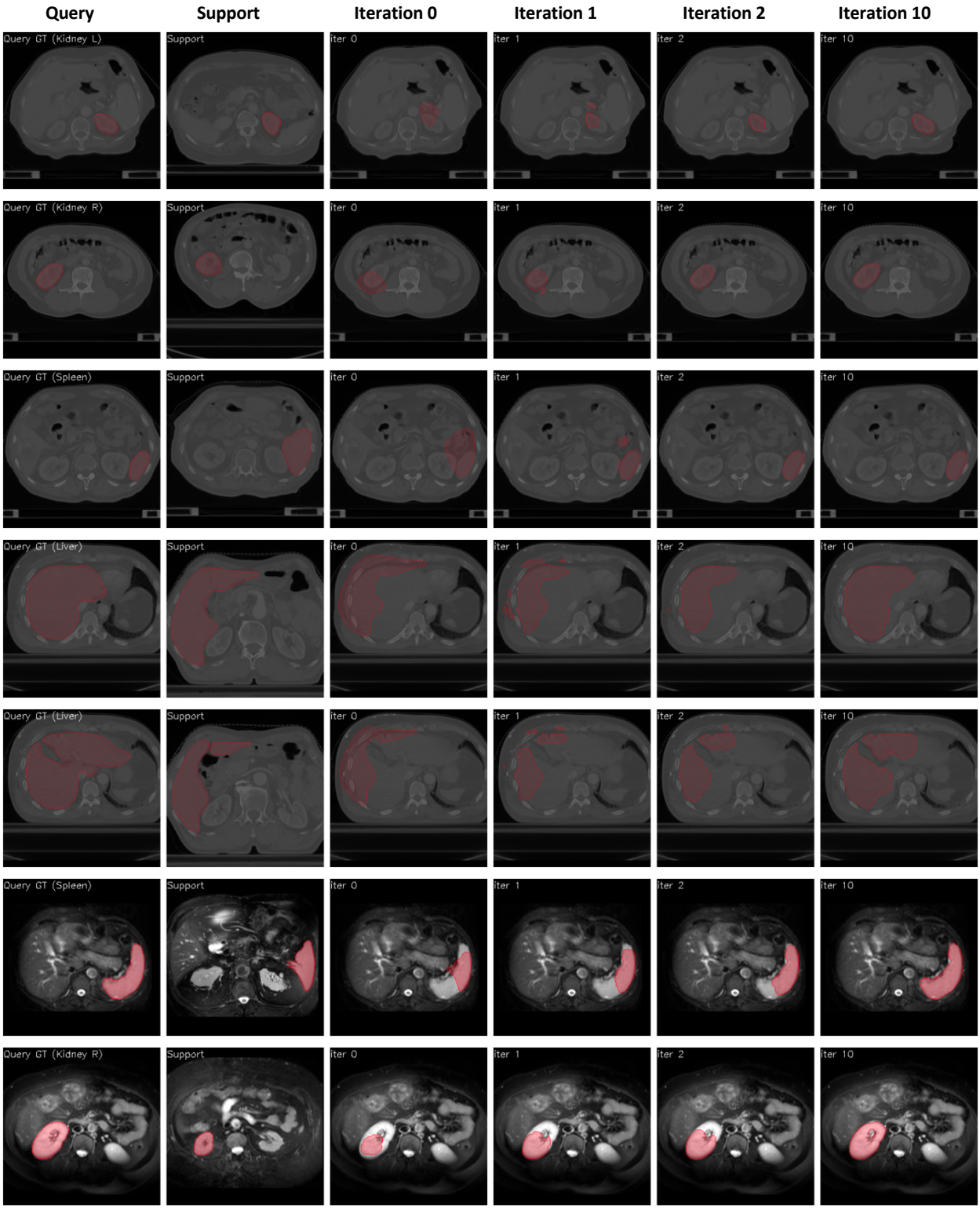

Figure 8.3: Examples of predication of RP-Net at different iterations. Each row represents one slice of the a test scan (row 1-5 are CT images, row 6-7 are MR images).

8.5 Conclusion

In this work, we present a new few-shot medical image segmentation framework that refines the segmentation mask iteratively using a context relation encoder and a recurrent module. The proposed model learns to incrementally refine the segmentation mask to better align the object boundary. Experiments on three organ segmentation datasets demonstrate that RP-Net outperforms the previous state-of-the-art approach by as much as 16% in terms of DSC. Moreover, the proposed CRE and recurrent module are generic and can also be integrated into other types of network to enhance context relationship features.

Chapter 9

Conclusion and Outlook

9.1 Conclusion and Contributions

Recent advances in machine learning, especially deep convolutional neural networks (DCNN), are changing and automating the medical image analysis in clinical practice, making it faster and more accurate. However, challenges remain due to the limited data and high accuracy requirement. In this dissertation, we propose several methods to resolve these challenges in different applications, including pulmonary nodule detection and segmentation, organs-at-risk (OAR) delineation, and few-shot medical image segmentation.

In chapter 2, we introduce how to design nodule detection framework that fully utilizes 3D convolutions. In chapter 3, we discuss how to reduce the model complexity by integrating nodule detection and false positive reduction into an end-to-end solution. In chapter 4, we release a new dataset for pulmonary lobe segmentation and propose a new framework for accurate lobe segmentation. In chapter 5, we introduce an unified framework for nodule detection, false positive reduction and segmentation. In chapter 6, we introduce 3D DCNN to significantly improve the accuracy and speed for OAR delineation in clinical practice. In

chapter 7, we introduce how to design a DCNN for medical image segmentation that greatly reduces the GPU memory consumption. In chapter 8, we introduce a new few-shot medical segmentation network that requires less labeled data.

These methods contribute to improve the accuracy, achieve better efficiency, and deal with limited labeled data. They also have the potential to be applied in clinical practice to help doctors in their daily routine.

9.2 Future Directions

Besides from the exciting achievement and progress mentioned above, there are still many challenging and interesting topics in the deep learning based medical image analysis. First, the state-of-the-art nodule detection method still reports many false positives and is limited when applied in clinical practice. Whether using self-attention and transformer based models that are capable of capturing long range relationships would be helpful is an interesting topic. Second, one important factor contributing to the success of deep learning is the availability of a large labeled dataset. However, acquiring labels is expensive in medical image analysis while at the same time we have access to a large amount of CT/MRI images that are not annotated. How to learn and extract useful features from the large amount of unlabeled data has drawn more and more attention in the field of computer vision [47, 14].In our previous work, we adopt the few-shot learning idea, but this may be a too extreme case and it does not solve the more general question: how to learn useful features from unlabeled data. On the other hand, self-supervised learning that uses contrastive learning strategy show promising results on natural images. How to effectively use the self-supervised learning to extract meaningful features in medical images would be very interesting and challenging. Moreover, previous works mostly focus on the classification task. However, segmentation is also a very important task in medical image analysis. How to design a pretraining strategy

that would benefit the segmentation remains an interesting problem.

Bibliography

[1] Tianchi medical ai contest: Intelligent diagnosis of pulmonary nodules. `https://tianchi.aliyun.com/competition/rankingList.htm?raceId=231601&season=0`.

[2] G. Aresta, C. Jacobs, T. Araújo, A. Cunha, I. Ramos, B. van Ginneken, and A. Campilho. iw-net: an automatic and minimalistic interactive lung nodule segmentation deep network. *arXiv preprint arXiv:1811.12789*, 2018.

[3] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[4] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, et al. Human-level cmr image analysis with deep fully convolutional networks. *arXiv preprint arXiv:1710.09289*, 2017.

[5] J. Bernier, C. Domenge, M. Ozsahin, K. Matuszewska, J.-L. Lefèbvre, R. H. Greiner, J. Giralt, P. Maingon, F. Rolland, M. Bolla, et al. Postoperative irradiation with or without concomitant chemotherapy for locally advanced head and neck cancer. *New England Journal of Medicine*, 350(19):1945–1952, 2004.

[6] J. A. Bonner, P. M. Harari, J. Giralt, R. B. Cohen, C. U. Jones, R. K. Sur, D. Raben, J. Baselga, S. A. Spencer, J. Zhu, et al. Radiotherapy plus cetuximab for locoregionally advanced head and neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between cetuximab-induced rash and survival. *The lancet oncology*, 11(1):21–28, 2010.

[7] W. R. Bosch, W. L. Straube, J. W. Matthews, and J. A. Purdy. Data from head-neck cetuximab, 2015.

[8] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[9] H. Brody. Medical imaging. *Nature*, 502(7473):S81–S81, 2013.

[10] C. L. Brouwer, R. J. Steenbakkers, J. Bourhis, W. Budach, C. Grau, V. Grégoire, M. van Herk, A. Lee, P. Maingon, C. Nutting, et al. Ct-based delineation of organs

at risk in the head and neck region: Dahanca, eortc, gortec, hknpcsg, ncic ctg, ncri, nrg oncology and trog consensus guidelines. *Radiotherapy and Oncology*, 117(1):83–90, 2015.

[11] 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.

[12] C. E. Cardenas, R. E. McCarroll, L. E. Court, B. A. Elgohari, H. Elhalawani, C. D. Fuller, M. J. Kamal, M. A. Meheissen, A. S. Mohamed, A. Rao, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. *International Journal of Radiation Oncology* Biology* Physics*, 101(2):468–478, 2018.

[13] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 170:446–455, 2018.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[15] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018.

[16] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

[17] O. Commowick, V. Grégoire, and G. Malandain. Atlas-based delineation of lymph node levels in head and neck computed tomography images. *Radiotherapy and Oncology*, 87(2):281–289, 2008.

[18] J. S. Cooper, T. F. Pajak, A. A. Forastiere, J. Jacobs, B. H. Campbell, S. B. Saxman, J. A. Kish, H. E. Kim, A. J. Cmelak, M. Rotman, et al. Postoperative concurrent radiotherapy and chemotherapy for high-risk squamous-cell carcinoma of the head and neck. *New England Journal of Medicine*, 350(19):1937–1944, 2004.

[19] J.-F. Daisne and A. Blumhofer. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. *Radiation oncology*, 8(1):154, 2013.

[20] A. Dietz, V. Rudat, J. Dreyhaupt, M. Pritsch, F. Hoppe, R. Hagen, L. Pfreundner, U. Schröder, H. Eckel, M. Hess, et al. Induction chemotherapy with paclitaxel and cisplatin followed by radiotherapy for larynx organ preservation in advanced laryngeal and hypopharyngeal cancer offers moderate late toxicity outcome (delos-i-trial). *European Archives of Oto-Rhino-Laryngology*, 266(8):1291–1300, 2009.

[21] J. Ding et al. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. *CoRR*, abs/1706.04303, 2017.

[22] J. Ding, A. Li, Z. Hu, and L. Wang. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. *CoRR*, abs/1706.04303, 2017.

[23] 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*, pages 559–567. Springer, 2017.

[24] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE transactions on medical imaging*, 38(5):1116–1126, 2018.

[25] N. Dong and E. P. Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018.

[26] 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*, 64(7):1558–1567, 2017.

[27] Q. Dou et al. Automated pulmonary nodule detection via 3d convnets with online sample filtering and hybrid-loss residual learning. *CoRR*, abs/1708.03867, 2017.

[28] H. Duc, K. Albert, G. Eminowicz, R. Mendes, S.-L. Wong, J. McClelland, M. Modat, M. J. Cardoso, A. F. Mendelson, C. Veiga, et al. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. *Medical physics*, 42(9):5027–5034, 2015.

[29] E. Esteller, X. León, M. de Juan, and M. Quer. Delayed carotid blow-out syndrome: a new complication of chemoradiotherapy treatment in pharyngolaryngeal carcinoma. *The Journal of Laryngology & Otology*, 126(11):1189–1191, 2012.

[30] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[31] F. T. Ferreira, P. Sousa, A. Galdran, M. R. Sousa, and A. Campilho. End-to-end supervised lung lobe segmentation. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2018.

[32] M. Fesneau, Y. Pointreau, S. Chapet, L. Martin, P. Pommier, M. Alfonsi, B. Laguerre, N. Feham, C. Berger, P. Garaud, et al. Concomitant chemoradiotherapy using carboplatin, tegafur-uracil and leucovorin for stage iii and iv head-and-neck cancer: results of gortec phase ii study. *International Journal of Radiation Oncology* Biology* Physics*, 76(1):154–163, 2010.

[33] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[34] C. Finn, K. Xu, and S. Levine. Probabilistic model-agnostic meta-learning. *arXiv preprint arXiv:1806.02817*, 2018.

[35] A. A. Forastiere, H. Goepfert, M. Maor, T. F. Pajak, R. Weber, W. Morrison, B. Glisson, A. Trotti, J. A. Ridge, C. Chao, et al. Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer. *New England Journal of Medicine*, 349(22):2091–2098, 2003.

[36] V. Fortunati, R. F. Verhaart, W. J. Niessen, J. F. Veenland, M. M. Paulides, and T. van Walsum. Automatic tissue segmentation of head and neck mr images for hyperthermia treatment planning. *Physics in Medicine & Biology*, 60(16):6547, 2015.

[37] V. Fortunati, R. F. Verhaart, F. van der Lijn, W. J. Niessen, J. F. Veenland, M. M. Paulides, and T. van Walsum. Tissue segmentation of head and neck ct images for treatment planning: a multiatlas approach combined with intensity modeling. *Medical physics*, 40(7), 2013.

[38] K. Fritscher, P. Raudaschl, P. Zaffino, M. F. Spadea, G. C. Sharp, and R. Schubert. Deep neural networks for fast segmentation of 3d medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 158–165. Springer, 2016.

[39] K. D. Fritscher, M. Peroni, P. Zaffino, M. F. Spadea, R. Schubert, and G. Sharp. Automatic segmentation of head and neck ct images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. *Medical physics*, 41(5), 2014.

[40] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018.

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

[42] N. E. T. T. R. Group. Patients at high risk of death after lung-volume–reduction surgery. *New England Journal of Medicine*, 345(15):1075–1083, 2001.

[43] D. Guo, D. Jin, Z. Zhu, T.-Y. Ho, A. P. Harrison, C.-H. Chao, J. Xiao, and L. Lu. Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2020.

[44] A. Hänsch, M. Schwier, T. Gass, T. Morgas, B. Haas, J. Klein, and H. K. Hahn. Comparison of different deep learning approaches for parotid gland segmentation from ct images. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 1057519. International Society for Optics and Photonics, 2018.

[45] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[46] K. He et al. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[47] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[48] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[49] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[50] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[51] M. Hoogeman, X. Han, D. Teguh, P. Voet, P. Nowak, T. Wolf, L. Hibbard, B. Heijmen, and P. Levendag. Atlas-based auto-segmentation of ct images in head and neck cancer: What is the best approach? *International Journal of Radiation Oncology• Biology• Physics*, 72(1):S591, 2008.

[52] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[53] R. D. Howe and Y. Matsuoka. Robotics for surgery. *Annual review of biomedical engineering*, 1(1):211–240, 1999.

[54] K.-Y. Hsiao, S.-A. Yeh, C.-C. Chang, P.-C. Tsai, J.-M. Wu, and J.-S. Gau. Cognitive function before and after intensity-modulated radiation therapy in patients with nasopharyngeal carcinoma: a prospective study. *International Journal of Radiation Oncology* Biology* Physics*, 77(3):722–726, 2010.

[55] J. Hur and S. Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019.

[56] D. P. Huttenlocher, W. J. Rucklidge, and G. A. Klanderman. Comparing images using the hausdorff distance under translation. In *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 654–656. IEEE, 1992.

[57] B. Ibragimov and L. Xing. Segmentation of organs-at-risks in head and neck ct images using convolutional neural networks. *Medical physics*, 44(2):547–557, 2017.

[58] A. Isambert, F. Dhermain, F. Bidault, O. Commowick, P.-Y. Bondiau, G. Malandain, and D. Lefkopoulos. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. *Radiotherapy and oncology*, 87(1):93–99, 2008.

[59] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[60] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[61] C. Jacobs et al. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical Image Analysis*, 18(2):374–384, 2014.

[62] A. S. Jones, B. Fish, J. E. Fenton, and D. J. Husband. The treatment of early laryngeal cancers (t1–t2 n0): surgery or irradiation? *Head & Neck: Journal for the Sciences and Specialties of the Head and Neck*, 26(2):127–135, 2004.

[63] J. Kalpathy-Cramer, B. Zhao, D. Goldgof, Y. Gu, X. Wang, H. Yang, Y. Tan, R. Gillies, and S. Napel. A comparison of lung nodule segmentation algorithms: methods and results from a multi-institutional study. *Journal of digital imaging*, 29(4):476–487, 2016.

[64] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[65] N. Khosravan and U. Bagci. S4nd: Single-shot single-scale lung nodule detection. *arXiv preprint arXiv:1805.02279*, 2018.

[66] M. Kosmin, J. Ledsam, B. Romera-Paredes, R. Mendes, S. Moinuddin, D. de Souza, L. Gunn, C. Kelly, C. Hughes, A. Karthikesalingam, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiotherapy and Oncology*, 135:130–140, 2019.

[67] J.-M. Kuhnigk, V. Dicken, S. Zidowitz, L. Bornemann, B. Kuemmerlen, S. Krass, H.-O. Peitgen, S. Yuval, H.-H. Jend, W. S. Rau, et al. New tools for computer assistance in thoracic ct. part 1. functional analysis of lungs, lung lobes, and bronchopulmonary segments. *Radiographics*, 25(2):525–536, 2005.

[68] H. Kundel, K. Berbaum, D. Dorfman, D. Gur, C. Metz, and R. Swensson. Receiver operating characteristic analysis in medical imaging. *ICRU Report*, 79(8):1, 2008.

[69] S.-Z. Lai, W.-F. Li, L. Chen, W. Luo, Y.-Y. Chen, L.-Z. Liu, Y. Sun, A.-H. Lin, M.-Z. Liu, and J. Ma. How does intensity-modulated radiotherapy versus conventional two-dimensional radiotherapy influence the treatment results in nasopharyngeal carcinoma patients? *International Journal of Radiation Oncology* Biology* Physics*, 80(3):661–668, 2011.

[70] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

[71] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[72] B. Lassen, E. M. van Rikxoort, M. Schmidt, S. Kerkstra, B. van Ginneken, and J.-M. Kuhnigk. Automatic segmentation of the pulmonary lobes from chest ct scans based on fissures, vessels, and bronchi. *IEEE transactions on medical imaging*, 32(2):210–222, 2013.

[73] D. S. Lee, Y. S. Kim, J. S. Cheon, J. H. Song, S. H. Son, J. S. Jang, Y. N. Kang, J. H. Kang, S. L. Jung, I. R. Yoo, et al. Long-term outcome and toxicity of hypofractionated stereotactic body radiotherapy as a boost treatment for head and neck cancer: the importance of boost volume assessment. *Radiation Oncology*, 7(1):85, 2012.

[74] P. Levendag, M. Hoogeman, D. Teguh, T. Wolf, L. Hibbard, O. Wijers, B. Heijmen, P. Nowak, E. Vasquez-Osorio, and X. Han. Atlas based auto-segmentation of ct images: Clinical evaluation of using auto-contouring in high-dose, high-precision radiotherapy of cancer in the head and neck. *International Journal of Radiation Oncology• Biology• Physics*, 72(1):S401, 2008.

[75] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[76] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. *IEEE transactions on neural networks and learning systems*, 30(11):3484–3495, 2019.

[77] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, Oct 2017.

[78] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[79] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[80] S. Liu, D. Xu, S. K. Zhou, O. Pauly, S. Grbic, T. Mertelmeier, J. Wicklein, A. Jerebko, W. Cai, and D. Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 851–858. Springer, 2018.

[81] W.-S. Liu, C.-H. Hsin, Y.-H. Chou, J.-T. Liu, M.-F. Wu, S.-W. Tseng, J.-K. Lee, H.-C. Tseng, T.-H. Wang, M.-C. Su, et al. Long-term results of intensity-modulated radiotherapy concomitant with chemotherapy for hypopharyngeal carcinoma aimed at laryngeal preservation. *BMC cancer*, 10(1):102, 2010.

[82] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[83] J. Ma, F. Lin, S. Wesarg, and M. Erdt. A novel bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 480–487. Springer, 2018.

[84] M. Machtay, J. Moughan, A. Trotti, A. S. Garden, R. S. Weber, J. S. Cooper, A. Forastiere, and K. K. Ang. Factors associated with severe late toxicity after concurrent chemoradiation for locally advanced head and neck cancer: an rtog analysis. *Journal of Clinical Oncology*, 26(21):3582, 2008.

[85] K. Men, H. Geng, C. Cheng, H. Zhong, M. Huang, Y. Fan, J. P. Plastaras, A. Lin, and Y. Xiao. More accurate and efficient segmentation of organs-at-risk in radiotherapy with convolutional neural networks cascades. *Medical physics*, 2019.

[86] F. Milletari, N. Navab, and S. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR*, abs/1606.04797, 2016.

[87] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[88] D. Močnik, B. Ibragimov, L. Xing, P. Strojan, B. Likar, F. Pernuš, and T. Vrtovec. Segmentation of parotid glands from registered ct and mr images. *Physica Medica*, 52:33–41, 2018.

[89] A. K. Mondal, J. Dolz, and C. Desrosiers. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv preprint arXiv:1810.12241*, 2018.

[90] K. Murphy et al. A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification. *Medical Image Analysis*, 13(5):757–770, 2009.

[91] P. F. Nguyen-Tan, Q. Zhang, K. K. Ang, R. S. Weber, D. I. Rosenthal, D. Soulieres, H. Kim, C. Silverman, A. Raben, T. J. Galloway, et al. Randomized phase iii trial to test accelerated versus standard fractionation in combination with concurrent cisplatin for head and neck carcinomas in the radiation therapy oncology group 0129 trial: long-term report of efficacy and toxicity. *Journal of Clinical Oncology*, 32(34):3858, 2014.

[92] S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-Paredes, A. Karthikesalingam, C. Chu, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

[93] J. J. Nirschl, A. Janowczyk, E. G. Peyster, R. Frank, K. B. Margulies, M. D. Feldman, and A. Madabhushi. Deep learning tissue segmentation in cardiac histopathology images. In *Deep learning for medical image analysis*, pages 179–195. Elsevier, 2017.

[94] 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.

[95] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *European Conference on Computer Vision*, pages 762–780. Springer, 2020.

[96] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan, and D. Rueckert. Data efficient unsupervised domain adaptation for cross-modality image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 669–677. Springer, 2019.

[97] G. V. Pednekar, J. K. Udupa, D. J. McLaughlin, X. Wu, Y. Tong, C. B. Simone, J. Camaratta, and D. A. Torigian. Image quality and segmentation. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 105762N. International Society for Optics and Photonics, 2018.

[98] G. Peng, T. Wang, K.-y. Yang, S. Zhang, T. Zhang, Q. Li, J. Han, and G. Wu. A prospective, randomized study comparing outcomes and toxicities of intensity-modulated radiotherapy vs. conventional two-dimensional radiotherapy for the treatment of nasopharyngeal carcinoma. *Radiotherapy and oncology*, 104(3):286–293, 2012.

[99] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2020.

[100] J.-M. Prades, B. Lallemant, R. Garrel, E. Reyt, C. Righini, T. Schmitt, N. Remini, L. Saban-Roche, A. P. Timoshenko, B. Trombert, et al. Randomized phase iii trial comparing induction chemotherapy followed by radiotherapy to concomitant chemoradiotherapy for laryngeal preservation in t3m0 pyriform sinus carcinoma. *Acta oto-laryngologica*, 130(1):150–155, 2010.

[101] A. A. Qazi, V. Pekar, J. Kim, J. Xie, S. L. Breen, and D. A. Jaffray. Auto-segmentation of normal and target structures in head and neck ct images: A feature-driven model-based approach. *Medical physics*, 38(11):6160–6170, 2011.

[102] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, and S. Levine. Conditional networks for few-shot semantic segmentation. 2018.

[103] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[104] P. F. Raudaschl, P. Zaffino, G. C. Sharp, M. F. Spadea, A. Chen, B. M. Dawant, T. Albrecht, T. Gass, C. Langguth, M. Lüthi, et al. Evaluation of segmentation methods on head and neck ct: auto-segmentation challenge 2015. *Medical physics*, 44(5):2020–2036, 2017.

[105] S. Ren et al. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[106] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[107] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[108] X. Ren, L. Xiang, D. Nie, Y. Shao, H. Zhang, D. Shen, and Q. Wang. Interleaved 3d-cnn s for joint segmentation of small-volume structures in head and neck ct images. *Medical physics*, 45(5):2063–2075, 2018.

[109] O. Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[110] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[111] 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*, 45:94–107, 2018.

[112] C. Rother, V. Kolmogorov, and A. Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[113] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger. 'squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020.

[114] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing. Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images. *IEEE transactions on medical imaging*, 39(5):1316–1325, 2019.

[115] A. Setio, F. Ciompi, et al. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE Trans. on Med. Imaging*, 35(5):1160–1169, 2016.

[116] 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*, 35(5):1160–1169, 2016.

[117] A. A. A. Setio et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *CoRR*, abs/1612.08012, 2016.

[118] 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*, 42:1–13, 2017.

[119] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[120] D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[121] A. Shrivastava et al. Training region-based object detectors with online hard example mining. *CoRR*, abs/1604.03540, 2016.

[122] M. Siam, B. N. Oreshkin, and M. Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[123] R. L. Siegel et al. Cancer statistics, 2015. *CA:a cancer journal for clinicians*, 65(1):5–29, 2015.

[124] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5–29, 2015.

[125] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[126] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[127] R. Sims, A. Isambert, V. Grégoire, F. Bidault, L. Fresco, J. Sage, J. Mills, J. Bourhis, D. Lefkopoulos, O. Commowick, et al. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. *Radiotherapy and Oncology*, 93(3):474–478, 2009.

[128] J. Snell, K. Swersky, and R. S. Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[129] L. Sun, W. Ma, X. Ding, Y. Huang, D. Liang, and J. Paisley. A 3d spatially weighted network for segmentation of brain tissue from mri. *IEEE transactions on medical imaging*, 39(4):898–909, 2019.

[130] M. Sun Yoon, W.-K. Chung, S.-J. Ahn, T.-K. Nam, B.-S. Nah, J.-Y. Song, S. Cheol Lim, and J. Kyoo Lee. Concurrent chemoradiotherapy with cisplatin and fluorouracil for locally advanced hypopharyngeal carcinoma. *Acta oto-laryngologica*, 128(5):590–596, 2008.

[131] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[132] J. Sykes. Reflections on the current status of commercial automated segmentation systems in clinical practice. *Journal of medical radiation sciences*, 61(3):131–134, 2014.

[133] C. Tam, X. Yang, S. Tian, X. Jiang, J. Beitler, and S. Li. Automated delineation of organs-at-risk in head and neck ct images using multi-output support vector regression. In *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10578, page 1057824. International Society for Optics and Photonics, 2018.

[134] H. Tang, X. Chen, Y. Liu, Z. Lu, J. You, M. Yang, S. Yao, G. Zhao, Y. Xu, T. Chen, et al. Clinically applicable deep learning framework for organs at risk delineation in ct images. *Nature Machine Intelligence*, pages 1–12, 2019.

[135] H. Tang, D. R. Kim, and X. Xie. Automated pulmonary nodule detection using 3d deep convolutional neural networks. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 523–526. IEEE, 2018.

[136] H. Tang, X. Liu, K. Han, X. Xie, X. Chen, H. Qian, Y. Liu, S. Sun, and N. Bai. Spatial context-aware self-attention model for multi-organ segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 939–949, 2021.

[137] H. Tang, X. Liu, and X. Xie. An end-to-end framework for integrated pulmonary nodule detection and false positive reduction. In *Biomedical Imaging (ISBI 2019), 2019 IEEE 16th International Symposium on*. IEEE, 2019.

[138] H. Tang, C. Zhang, and X. Xie. Automatic pulmonary lobe segmentation using deep learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1225–1228. IEEE, 2019.

[139] H. Tang, C. Zhang, and X. Xie. Nodulenet: Decoupled false positive reductionfor pulmonary nodule detection and segmentation. *arXiv preprint arXiv:1907.11320*, 2019.

[140] Z. Teed and J. Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020.

[141] D. N. Teguh, P. C. Levendag, P. W. Voet, A. Al-Mamgani, X. Han, T. K. Wolf, L. S. Hibbard, P. Nowak, H. Akhiat, M. L. Dirkx, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *International Journal of Radiation Oncology* Biology* Physics*, 81(4):950–957, 2011.

[142] J.-P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[143] D. Thomson, C. Boylan, T. Liptrot, A. Aitkenhead, L. Lee, B. Yap, A. Sykes, C. Rowbottom, and N. Slevin. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiation Oncology*, 9(1):173, 2014.

[144] N. Tong, S. Gou, S. Yang, D. Ruan, and K. Sheng. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Medical physics*, 45(10):4558–4567, 2018.

[145] Y. Tong, J. K. Udupa, X. Wu, D. Odhner, G. Pednekar, C. B. Simone, D. McLaughlin, C. Apinorasethkul, G. Shammo, P. James, et al. Hierarchical model-based object localization for auto-contouring in head and neck radiation therapy planning. In *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10578, page 1057822. International Society for Optics and Photonics, 2018.

[146] I. R. S. Valente, P. C. Cortez, E. C. Neto, J. M. Soares, V. H. C. de Albuquerque, and J. M. R. Tavares. Automatic 3d pulmonary nodule detection in ct images: a survey. *Computer methods and programs in biomedicine*, 124:91–107, 2016.

[147] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. J. Aerts, N. Khaouam, P. F. Nguyen-Tan, C.-S. Wang, K. Sultanem, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific reports*, 7(1):10117, 2017.

[148] J. van der Veen, S. Willems, S. Deschuymer, D. Robben, W. Crijns, F. Maes, and S. Nuyts. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiotherapy and Oncology*, 138:68–74, 2019.

[149] B. Van Ginneken et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the anode09 study. *Medical Image Analysis*, 14(6):707–722, 2010.

[150] B. Van Ginneken, C. M. Schaefer-Prokop, and M. Prokop. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, 261(3):719–732, 2011.

[151] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[152] R. F. Verhaart, V. Fortunati, G. M. Verduijn, A. Lugt, T. Walsum, J. F. Veenland, and M. M. Paulides. The relevance of mri for patient modeling in head and neck hyperthermia treatment planning: A comparison of ct and ct-mri based tissue segmentation on simulated temperature. *Medical physics*, 41(12), 2014.

[153] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

[154] P. W. Voet, M. L. Dirkx, D. N. Teguh, M. S. Hoogeman, P. C. Levendag, and B. J. Heijmen. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? a dosimetric analysis. *Radiotherapy and Oncology*, 98(3):373–377, 2011.

[155] E. Vorontsov, A. Tang, C. Pal, and S. Kadoury. Liver lesion segmentation informed by joint liver segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1332–1335. IEEE, 2018.

[156] Q. D. Vu, S. Graham, T. Kurc, M. N. N. To, M. Shaban, T. Qaiser, N. A. Koohbanani, S. A. Khurram, J. Kalpathy-Cramer, T. Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 7:53, 2019.

[157] C. Wachinger, K. Fritscher, G. Sharp, and P. Golland. Contour-driven atlas-based segmentation. *IEEE transactions on medical imaging*, 34(12):2492–2505, 2015.

[158] G. V. Walker, M. Awan, R. Tao, E. J. Koay, N. S. Boehling, J. D. Grant, D. F. Sittig, G. B. Gunn, A. S. Garden, J. Phan, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiotherapy and Oncology*, 112(3):321–325, 2014.

[159] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019.

[160] S. Wang, M. Zhou, Z. Liu, Z. Liu, D. Gu, Y. Zang, D. Dong, O. Gevaert, and J. Tian. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical image analysis*, 40:172–183, 2017.

[161] W. Wang, Q. Song, R. Feng, T. Chen, J. Chen, D. Z. Chen, and J. Wu. A fully 3d cascaded framework for pancreas segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 207–211. IEEE, 2020.

[162] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[163] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical image analysis*, 55:88–102, 2019.

[164] Z. Wang, L. Wei, L. Wang, Y. Gao, W. Chen, and D. Shen. Hierarchical vertex regression-based segmentation of head and neck ct images for radiotherapy planning. *IEEE Transactions on Image Processing*, 27(2):923–937, 2018.

[165] B. Wu, Z. Zhou, J. Wang, and Y. Wang. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1109–1113. IEEE, 2018.

[166] X. Wu, J. K. Udupa, Y. Tong, D. Odhner, G. V. Pednekar, C. B. Simone, D. McLaughlin, C. Apinorasethkul, J. Lukens, D. Mihailidis, et al. Auto-contouring via automatic anatomy recognition of organs at risk in head and neck cancer on ct images. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 1057617. International Society for Optics and Photonics, 2018.

[167] X. Wu, J. K. Udupa, Y. Tong, D. Odhner, G. V. Pednekar, C. B. Simone II, D. McLaughlin, C. Apinorasethkul, O. Apinorasethkul, J. Lukens, et al. Aar-rt–a system for auto-contouring organs at risk on ct images for radiation therapy planning: Principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. *Medical image analysis*, 54:45–62, 2019.

[168] Y. Xia, L. Xie, F. Liu, Z. Zhu, E. K. Fishman, and A. L. Yuille. Bridging the gap between 2d and 3d organ segmentation with volumetric fusion net. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 445–453. Springer, 2018.

[169] S.-A. Yeh, Y. Tang, C.-C. Lui, Y.-J. Huang, and E.-Y. Huang. Treatment outcomes and late complications of 849 patients with nasopharyngeal carcinoma treated with radiotherapy alone. *International Journal of Radiation Oncology* Biology* Physics*, 62(3):672–679, 2005.

[170] H. Yu, S. Sun, H. Yu, X. Chen, H. Shi, T. S. Huang, and T. Chen. Foal: Fast online adaptive learning for cardiac motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4313–4323, 2020.

[171] 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*, pages 8280–8289, 2018.

[172] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9587–9595, 2019.

[173] T. Zhang, Y. Chi, E. Meldolesi, and D. Yan. Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*, 68(2):522–530, 2007.

[174] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50(9):3855–3865, 2020.

[175] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8553, 2019.

[176] N. Zhao, N. Tong, D. Ruan, and K. Sheng. Fully automated pancreas segmentation with two-stage 3d convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 201–209. Springer, 2019.

[177] W. Zhao, J. Yang, Y. Sun, C. Li, W. Wu, L. Jin, Z. Yang, B. Ni, P. Gao, P. Wang, et al. 3d deep learning from ct scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. *Cancer research*, 78(24):6881–6889, 2018.

[178] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan. A deep learning model integrating fcnns and crfs for brain tumor segmentation. *Medical image analysis*, 43:98–111, 2018.

[179] X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara, and H. Fujita. Three-dimensional ct image segmentation by combining 2d fully convolutional network with 3d majority voting. In *Deep Learning and Data Labeling for Medical Applications*, pages 111–120. Springer, 2016.

[180] Y. Zhou, Z. Li, S. Bai, C. Wang, X. Chen, M. Han, E. Fishman, and A. L. Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10672–10681, 2019.

[181] 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*, pages 693–701. Springer, 2017.

[182] W. Zhu, Y. Huang, H. Tang, Z. Qian, N. Du, W. Fan, and X. Xie. Anatomynet: Deep 3d squeeze-and-excitation u-nets for fast and fully automated whole-volume anatomical segmentation. *arXiv preprint arXiv:1808.05238*, 2018.

[183] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie. Anatomynet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics*, 46(2):576–589, 2019.

[184] W. Zhu, C. Liu, W. Fan, and X. Xie. Deeplung: 3d deep convolutional nets for automated pulmonary nodule detection and classification. *CoRR*, abs/1709.05538, 2017.

[185] W. Zhu, C. Liu, W. Fan, and X. Xie. Deeplung: 3d deep convolutional nets for automated pulmonary nodule detection and classification. *arXiv preprint arXiv:1709.05538*, 2017.

[186] W. Zhu, Y. S. Vang, Y. Huang, and X. Xie. Deepem: Deep 3d convnets with em for weakly supervised pulmonary nodule detection. *arXiv preprint arXiv:1805.05373*, 2018.

[187] Z. Zhu, Y. Xia, W. Shen, E. Fishman, and A. Yuille. A 3d coarse-to-fine framework for volumetric medical image segmentation. In *2018 International Conference on 3D Vision (3DV)*, pages 682–690. IEEE, 2018.