MEDICAL MACHINE INTELLIGENCE: DATA-EFFICIENCY AND KNOWLEDGE-AWARENESS

by
Yuyin Zhou

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
October, 2020

Abstract

Traditional clinician diagnosis requires massive manual labor from experienced doctors, which is time-consuming and costly. Computer-aided systems are therefore proposed to reduce doctors' efforts by using machines to automatically make diagnosis and treatment recommendations. The recent success in deep learning has largely advanced the field of computer-aided diagnosis by offering an avenue to deliver automated medical image analysis. Despite such progress, there remain several challenges towards medical machine intelligence, such as unsatisfactory performance regarding challenging small targets, insufficient training data, high annotation cost, the lack of domain-specific knowledge, etc. These challenges cultivate the need for developing *data-efficient and knowledge-aware deep learning techniques which can generalize to different medical tasks without requiring intensive manual labeling efforts, and incorporate domain-specific knowledge in the learning process.*

In this thesis, we rethink the current progress of deep learning in medical image analysis, with a focus on the aforementioned challenges, and present different data-efficient and knowledge-aware deep learning approaches to address them accordingly. Firstly, we introduce coarse-to-fine mechanisms which use the prediction from the first (coarse) stage to shrink the input region for the second (fine) stage, to enhance the model performance especially for segmenting small challenging structures, such as the pancreas which occupies only a very small fraction (*e.g.*, $< 0.5\%$) of the entire CT volume. The method achieved the state-of-the-art result on the NIH pancreas segmentation dataset. Further extensions also demonstrated effectiveness for

segmenting neoplasms such as pancreatic cysts or multiple organs.

Secondly, we present a semi-supervised learning framework for medical image segmentation by leveraging both limited labeled data and abundant unlabeled data. Our learning method encourages the segmentation output to be consistent for the same input under different viewing conditions. More importantly, the outputs from different viewing directions are fused altogether to improve the quality of the target, which further enhances the overall performance. The comparison with fully-supervised methods on multi-organ segmentation confirms the effectiveness of this method.

Thirdly, we discuss how to incorporate knowledge priors for multi-organ segmentation. Noticing that the abdominal organ sizes exhibit similar distributions across different cohorts, we propose to explicitly incorporate anatomical priors on abdominal organ sizes, guiding the training process with domain-specific knowledge. The approach achieves 84.97% on the MICCAI 2015 challenge "Multi-Atlas Labeling Beyond the Cranial Vault", which significantly outperforms previous state-of-the-art even using fewer annotations.

Lastly, by rethinking how radiologists interpret medical images, we identify one limitation for existing deep-learning-based works on detecting pancreatic ductal adenocarcinoma is the lack of knowledge integration from multi-phase images. Thereby, we introduce a dual-path network where different paths are connected for multi-phase information exchange, and an additional loss is added for removing view divergence. By effectively incorporating multi-phase information, the presented method shows superior performance than prior arts on this matter.

Thesis Readers

Dr. Alan L. Yuille (Primary Advisor)
Bloomberg Distinguished Professor
Department of Computer
Johns Hopkins University

Dr. Vishal M. Patel
Assistant Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Dr. Wei Shen
Associate Professor
AI Institute
Shanghai Jiao Tong University

Dedicated to my family

and my friends.

Acknowledgements

First and foremost, I would like to thank my dear advisor Prof. Alan L. Yuille. It was my great pleasure and honor to work under his supervision. His insightful guidance, expertise, and wisdom have been truly instrumental to my Ph.D. career. Four years ago before I became a formal Ph.D. student, Alan introduced me to The Felix Project, which was funded by the Lustgarten Foundation, for automatically diagnosing the pancreas and pancreatic cancer in early stages. It was this very opportunity that lit up my research passion in medical image analysis, and enabled me to work with a group of remarkable scientists with various areas of expertise. Through all the ups and downs in my Ph.D. career, Alan has always been supporting me. I would have never got to where I am today without his support, encouragement, and continuous help.

Next, I want to express my sincere gratitude to Dr. Lingxi Xie, Dr. Wei Shen and Dr. Yan Wang for their meticulous guidance for various research works during my junior Ph.D. days. They are all excellent researchers who I look up to very much. In many of my projects, they have been wonderful collaborators who offer important suggestions about formulating research ideas, designing rigorous experimental validation, technical presentation and writing. Their envision in medical image analysis has surely influenced me deeply as an independent researcher. Besides, I also want to thank Prof. Alexander S. Szalay, Prof. Vishal M. Patel, Prof. Gregory D. Hager, Prof. Linda Chu, Dr. Elliot Fishman for participating in my GBO exam and providing invaluable suggestions regarding my research.

Additionally, I also want to thank all FELIX members, including Dr. Elliot Fishman, Dr. Seyoun Park, Dr. Linda C. Chu, Dr. Satomi Kawamoto, Dr. Daniel F. Fouladi, Dr. Shahab Shayesteh, Dr. Eva Zinreich, Dr. Karen M. Horton, Dr. Ralph H. Hruban, Dr. Kenneth W. Kinzler, and Dr. Bert Vogelstein, and many others. Their expertise has broadened my horizons and benefited my research career greatly. Among them, I want to especially thank Dr. Seyoun Park for giving me immeasurable guidance and invaluable advice in seeking research directions. Dr. Park's insightful research vision and strong engineering skills have always been very inspirational to me. In our collaboration, her comments were often crucial and beneficial to many of my research works. I also want to thank Prof. David Dreizin, for getting me involved in different medical research projects, and providing me funding for my last year's research.

I was fortunate to work as a research intern at Google twice to work on different medical image analysis problems. Those research experiences have been unique and insightful to me. As a junior student, I had my first research internship in Google Cloud AI, where I was privileged to get to know many great research scientists, including many renown scientists in computer vision: Dr. Mei Han, Dr. Liang-Chieh Chen, Dr. Wei Wei, Dr. Lu Jiang, Dr. Jia-Li Li, Prof. Fei-Fei Li, and many others. They have provided enormous help to me. Their ample experience, broad research scope and research philosophy have made me discover my own defects and reshaped me towards a more mature researcher. I also want to thank my other collaborators on the project, Dr. Zhe Li, Dr. Song Bai, Dr. Chong Wang and Dr. Xinlei Chen, for their time and devotion. I want to especially thank my host Dr. Mei Han, who has been so helpful and supportive to me from the beginning of my internship till present. Moreover, I also want to thank my hosts and collaborators during my second internship in Google Brain: Dr. Atilla Kiraly, Shahar Jamshy, Chace Lee, Dr. Wenxing Ye and Dr. Jie Yang. They have all been wonderful to me throughout my whole internship. In

addition, I would like to thank all people who have helped and supported me greatly in my academic career, including Prof. Lei Xing, Prof. Bennett Landman, Prof. Rama Chellappa, Prof. Ming-Hsuan Yang, Dr. Le Lu, and many others.

Being a member of CCVL is definitely one of the perks of my life. I want to thank all my colleagues of CCVL, including Qihang Yu, Yingda Xia, Fengze Liu, Jieneng Chen, Chen Wei, Zhuotun Zhu, Shuhao Fu, Yongyi Lu, who are also working on the FELIX project, Weichao Qiu, Chenxi Liu, Qi Chen, Zhishuai Zhang, Siyuan Qiao, Chenxu Luo, Huiyu Wang, Yingwei Li, Hongru Zhu, Jieru Mei, Yi Zhang, Yixiao Zhang, Zihao Xiao, Chenglin Yang, Yutong Bai, Qing Liu, Angtian Wang, Peng Tang, Jianyu Wang, Peng Wang, Fangting Xia, Xiaochen Lian, and many others. I also want to thank my boyfriend, Cihang Xie, for years of company and support.

Lastly, I would like to thank all the administrative staff in the computer science department. I want to deeply appreciate their hard work, which has made our lives much easier. In particular, I would like to thank Lilian Oonyu for helping me organize various conference trips, Zachary Burwell for monitoring my Ph.D. progress, and Kim Franklin for helping with my GBO and defense.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Deep learning has achieved remarkable progress in many domains including computer vision, natural language processing, and speech recognition, thereby propelling us into the era of artificial intelligence (AI). As a fundamentally important problem which could impact human lives, the development of intelligent medical machine learning systems for healthcare has garnered great research attention. In the meantime, medical images, which constitute the most commonly encountered healthcare data, have become one of the most important sources of evidence for clinical analysis and medical intervention [1].

Therefore, how to use AI-based techniques to drive automated interpretation of medical images has been widely studied. The successful applications include image registration, anatomical/cell structures detection, tissue segmentation, computer-aided disease diagnosis or prognosis, and so on [2, 3]. However, as deep learning exploits hierarchical feature representations which are highly data-driven, there remain several critical challenges towards medical machine intelligence, to name a few:

- The detection of small targets (*e.g.*, the pancreas, neoplasms) from medical images can be notoriously difficult due to their low resolution and noisy boundaries, which makes it easily confused by the complex and variable background.

- Deep learning models generally rely on large, representative, and well-annotated datasets to achieve high performance. However, annotating medical images demands extensive clinical expertise and manual labor, making it difficult to acquire large-scale datasets with complete and high-quality labels [4].

- Knowledge priors can take many forms: shape models; statistics on sizes/spatial locations; boundaries and edge polarity shape models; topology specification; geometrical interaction and distance prior between different regions/labels; atlas or pre-known models [5]. However, deep networks make decisions solely by extracting hierarchical features based on local textures or patterns, which generally make them lack the ability of leveraging prior information.

- In real clinical practice, medical images can take various forms (*e.g.*, different modalities/phases) based on the imaging protocol. The deep-learning-based techniques are expected to process different image forms. This can be important especially for cancer detection problems which are critical to give patients the best chance of recovery and survival.

These challenges have affected the applicability of such models being deployed in safety-critical medical scenarios. In the face of these challenges, we aim to develop *data-efficient and knowledge-aware deep learning techniques which can generalize to different medical tasks without requiring intensive manual labeling efforts, and incorporate domain-specific knowledge in the learning process.*

In this thesis, we will present our efforts towards medical machine intelligence by elaborating how to design data-efficient and knowledge-aware deep learning approaches from different aspects, including new training approaches, semi-supervised learning strategies, formulations of knowledge priors, and optimization methods, to enable effective learning from limited training data and incomplete labels.

1.2 Motivation

Medical image analysis is an essential requisite for many clinical applications such as computer-aided diagnosis, computer-aided surgery and radiation therapy. And it also plays a key role in modern healthcare diagnostics and procedures. With the advance of deep learning, how to harness powerful deep networks to solve medical image analysis problems has become an emerging topic. However, this can be quite challenging as medical images exhibit different traits from natural images as aforementioned in Section 1.1. These challenges inhibit the direct deployment of existing state-of-the-art networks in real-world clinical environments and motivate us to design methods which are tailored to medical data.

In terms of medical image segmentation, the current status is that for many abdominal organs (*e.g.*, liver, heart or kidneys), state-of-the-art performances in terms of Dice have already been far beyond 90%. Similar performances have also been observed for some targets outside the abdominal region (such as the brain) by directly applying deep learning approaches [6]. However, the segmentation of the pancreas report lower performances [7, 8]. In terms of this phenomenon, we have conducted some diagnosis experiments on the publicly available NIH pancreas segmentation dataset [7] consisting of 82 healthy cases. And we find that simply training and testing on the whole volume based on a state-of-the-art network only yields an average segmentation performance of 75.7% in terms of the average Dice score. For comparison, if we train/test with the same network but only inside the region-of-interest, in other words, we crop the pancreatic region from the CT image based on the groundtruth annotation and exclusively train/test on this region, we can then obtain a performance boost of around 8%. In the meantime, this huge performance gap doesn't occur for large organs such as the liver. Based on this observation, we identify that for small targets, deep networks are easily confused by the complex and variable background which

occupies a large fraction of the input volume. Therefore, to deal with small target segmentation, we design a multi-stage coarse-to-fine segmentation strategy where the first stage is used to extract the attentive region, based on which the second stage can then perform segmentation more accurately by effectively reducing the complex background and enhancing the salient features. The coarse stage and the fine stage can be either trained individually in a step-wise manner, or learned in an end-to-end fashion by cascading the two stages. Beyond single-target segmentation, we find that the multi-stage framework can also be well adapted to multi-organ segmentation.

Compared with natural image processing tasks, it is difficult and expensive to obtain large-scale datasets since collecting medical data is a complex and expensive procedure that requires the collaboration of researchers and radiologists [9]. By contrast, it is generally much easier to acquire unlabeled or partially-labeled data, which in turn motivates us to design semi/weakly-supervised approaches for *data-efficient* medical image analysis. By leveraging the power of additional unlabeled or partially-labeled data, our goal is to enhance model performance via maximizing the data utilization. Nevertheless, deep-learning-based semi-supervised learning in the medical domain has not drawn enough attention. One popular strategy is self-training [10, 11], which propagates labels from the labeled to the unlabeled data, and then using the larger, newly labeled set for training. However, in this approach, the error in the prediction (pseudo-label) can be reinforced during the training. To alleviate this negative effect, we exploit the fact that CT scans are high-resolution three-dimensional volumes which can be represented by multiple planes, *i.e.*, the axial, coronal, and sagittal planes. Inspired by this multi-view property, we use the co-training [12] paradigm to generate more accurate and robust pseudo-labels by utilizing the agreement among different learners.

In addition to *data-efficiency*, we also rethink existing deep-learning-based strategies in terms of *knowledge-awareness*. The inclusion of knowledge priors has been proved

useful for image segmentation by reducing the negative effects induced by noise, low contrast and objects' complexity [5]. However, how to explicitly embed priors in neural networks remain understudied. Therefore, we further discuss how to make deep neural networks aware of such knowledge priors, so as to approach the real clinical expertise. In the application of multi-organ segmentation, we sample various organs in the abdomen across different patients and datasets, and observe consistent anatomical similarities despite different imaging characteristics due to different scanners, image acquisition protocols or different patient populations. In particular, we exploit the fact that the size distributions of organs are similar across different patients, and formulate the statistics of size distributions as an explicit prior. Under this formulation, the prior can be easily embedded in the learning process by adding an additional objective to constrain the size distributions across training samples.

For medical images in various forms (*e.g.*, different modalities/phases), the identification of disease patterns usually requires combining multiple types of information. For instance, the texture changes of pancreatic ductal adenocarcinoma can be quite subtle in single-phase images. Consequently, in the real clinical practice, radiologists are recommended to interpret multi-phase information for providing more accurate diagnostics. Therefore, for the detection of pancreatic ductal adenocarcinoma, we design a multi-phase learning framework which incorporates multi-phase information by enabling the fusion of intermediate features from different imaging phases.

1.3 Thesis Outline

The organization of this thesis is outlined as follows:

In chapter 2, we provide an overview of the related techniques for medical image analysis.

In chapter 3, we present the coarse-to-fine approaches for pancreas segmentation [13–

15].

In chapter 4, we extend the coarse-to-fine approaches for pancreatic cyst segmentation by further introducing deep supervision [15–17].

In chapter 5, we adjust the coarse-to-fine framework for abdominal multi-organ segmentation by proposing organ-attention networks and statistical fusion [18].

In chapter 6, we propose deep multi-planar co-training for semi-supervised 3D abdominal multi-organ segmentation [19].

In chapter 7, we design a prior-aware neural network for partially-supervised multi-organ segmentation [20].

In chapter 8, we develop a multi-phase learning algorithm for Pancreatic Ductal Adenocarcinoma segmentation by proposing hyper-pairing network [21].

In chapter 9, we conclude the thesis and discuss potential future research directions.

1.4 Relevant Publications

The following publications constitute, or provide contexts and backgrounds for the ideas in this dissertation (* indicates equal contribution):

1. **Yuyin Zhou**, Lingxi Xie, Wei Shen, Yan Wang, Elliot Fishman, Alan Yuille. A Fixed-Point Model for Pancreas Segmentation in Abdominal CT Scans, *MICCAI 2017*

2. Qihang Yu, Lingxi Xie, Yan Wang, **Yuyin Zhou**, Elliot Fishman, Alan Yuille. Recurrent Saliency Transformation Network: Incorporating Multi-Stage Visual Cues for Small Organ Segmentation, *CVPR 2018*

3. **Yuyin Zhou**, Lingxi Xie, Elliot Fishman, Alan Yuille. Deep Supervision for Pancreatic Cyst Segmentation in Abdominal CT Scans, *MICCAI 2017*

4. Lingxi Xie, Qihang Yu, **Yuyin Zhou**, Yan Wang, Elliot Fishman, Alan Yuille. Recurrent Saliency Transformation Network for Tiny Target Segmentation in Abdominal CT Scans, *IEEE transactions on medical imaging 2019*

5. **Yuyin Zhou***, Qihang Yu*, Yan Wang, Lingxi Xie, Wei Shen, Elliot Fishman and Alan Yuille. 2D-Based Coarse-to-Fine Approaches for Small Target Segmentation in Abdominal CT Scans, in Deep Learning and Convolutional Neural Networks for Medical Image and Clinical Informatics, *Advances in Computer Vision and Pattern Recognition, Springer, 2019*

6. Yan Wang*, **Yuyin Zhou***, Wei Shen, Seyoun Park, Elliot Fishman, Alan Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion, *Medical Image Analysis 2019*

7. **Yuyin Zhou**, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot Fishman, Alan Yuille. SemiSupervised 3D Multi-Organ Segmentation via Deep Multi-Planar Co-Training, *WACV 2019*

8. **Yuyin Zhou***, Zhe Li*, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, Alan Yuille. Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, *ICCV 2019*

9. **Yuyin Zhou**, Yingwei Li, Zhishuai Zhang, Yan Wang, Alan Yuille, Seyoun Park. HyperPairing Network for Multi-Phase Pancreatic Ductal Adenocarcinoma Segmentation, *MICCAI 2019*

10. **Yuyin Zhou**, David Dreizin, Yingwei Li, Zhishuai Zhang, Yan Wang, Alan Yuille. MultiScale Attentional Network for Multi-Focal Segmentation of Active Bleed after Pelvic Fractures, *MLMI 2019*

11. Zhishuai Zhang, **Yuyin Zhou**, Wei Shen, Elliot Fishman, Alan Yuille. Lesion Detection by Efficiently Bridging 3D Context, *MLMI 2019*

12. Fengze Liu, **Yuyin Zhou**, Elliot Fishman, Alan Yuille. FusionNet: Incorporating Shape and Texture for Abnormality Detection in 3D Abdominal CT Scans, *MLMI 2019*

13. Yan Wang, **Yuyin Zhou**, Peng Tang, Wei Shen, Elliot Fishman, Alan Yuille. Training Multi-organ Segmentation Networks with Sample Selection by Relaxed Upper Confident Bound, *MICCAI 2018*

14. Yingwei Li*, Zhuotun Zhu*, **Yuyin Zhou**, Yingda Xia, Wei Shen, Elliot Fishman, and Alan Yuille. Volumetric Medical Image Segmentation: A 3D Deep Coarse-to-fine Framework and Its Adversarial Examples, in Deep Learning and Convolutional Neural Networks for Medical Image Computing, *Advances in Computer Vision and Pattern Recognition, Springer, 2019*

15. Yingda Xia*, Qihang Yu*, Wei Shen, **Yuyin Zhou**, Elliot Fishman, Alan Yuille. Detecting Pancreatic Adenocarcinoma in Multi-phase CT Scans via Alignment Ensemble, *MICCAI 2020*

16. Shuhao Fu, Yongyi Lu, Yan Wang, **Yuyin Zhou**, Wei Shen, Elliot Fishman, Alan Yuille. Domain Adaptive Relational Reasoning for 3D Multi-Organ Segmentation, *MICCAI 2020*

17. Yan Wang, Xu Wei, Fengze Liu, Jieneng Chen, **Yuyin Zhou**, Wei Shen, Elliot Fishman, Alan Yuille. Deep Distance Transform for Tubular Structure Segmentation in CT Scans, *CVPR 2020*

18. Cihang Xie, Jianyu Wang, Zhishuai Zhang,**Yuyin Zhou**, Lingxi Xie, Alan Yuille. Adversarial Examples for Semantic Segmentation and Object Detection, *ICCV 2017*

19. Cihang Xie, Zhishuai Zhang,**Yuyin Zhou**, Song Bai, Jianyu Wang, Zhou Ren,

Alan Yuille. Improving Transferability of Adversarial Examples with Input Diversity, *CVPR 2019*

20. Yingwei Li, Xiaojie Jin, Jieru Mei, Xiaochen Lian, Linjie Yang, Cihang Xie, Qihang Yu, Yuyin Zhou, Song Bai, Alan Yuille. Neural Architecture Search for Lightweight Non-Local Networks, *CVPR 2020*

21. Lifeng Huang, Chengying Gao, **Yuyin Zhou**, Cihang Xie, Alan Yuille, Changqing Zou, Ning Liu. UPC: Learning Universal Physical Camouflage Attacks on Object Detectors, *CVPR 2020*

22. Yingwei Li, Song Bai, **Yuyin Zhou**, Cihang Xie, Zhishuai Zhang, Alan Yuille. Learning Transferable Adversarial Examples via Ghost Networks, *AAAI 2020*

23. Song Bai, Yingwei Li, **Yuyin Zhou**, Qizhu Li, Philip HS Torr. Adversarial Metric Attack and Defense for Person Re-identification, *IEEE Transactions on Pattern Analysis and Machine Intelligence 2020*

Below lists other publications and researches I involved in during my Ph.D. study:

1. David Dreizin, **Yuyin Zhou**, Tina Chen, Guang Li, Alan Yuille, Ashley McLenithan, Jonathan Morrison. Deep learning-based quantitative visualization and measurement of extraperitoneal hematoma volumes in patients with pelvic fractures, *Journal of Trauma and Acute Care Surgery 2020*

2. David Dreizin, **Yuyin Zhou**, Yixiao Zhang, Nikki Tirada, Alan Yuille. Performance of a Deep Learning Algorithm for Automated Segmentation and Quantification of Traumatic Pelvic Hematomas on CT, *Journal of Digital Imaging 2019*

3. Linda C Chu, Seyoun Park, Satomi Kawamoto, Yan Wang, **Yuyin Zhou**, Wei Shen, Zhuotun Zhu, Yingda Xia, Lingxi Xie, Fengze Liu, Qihang Yu, Daniel

F. Fouladi, Shahab Shayesteh, Eva Zinreich, Jefferson S. Graves, Karen M. Horton, Alan Yuille, Ralph H. Hruban, Kenneth W. Kinzler, Bert Vogelstein, Elliot Fishman. Application of Deep Learning to Pancreatic Cancer Detection: Lessons Learned From Our Initial Experience, *Journal of the American College of Radiology 2019*

Chapter 2

Related Works

2.1 Deep Learning for Medical Image Analysis

Computer-aided diagnosis (CAD) is an important technique which can assist human doctors in many clinical scenarios. An important prerequisite of CAD is medical imaging analysis. As a popular and cheap way of medical imaging, contrast-enhanced computed tomography (CECT) produces detailed images of internal organs, bones, soft tissues and blood vessels. It is of great value to automatically segment organs and/or soft tissues from these CT volumes for further diagnosis [16, 22–24]. To capture specific properties of different organs, researchers often design individualized algorithms for each of them. Typical examples include the liver [25, 26], the *spleen* [27], the *kidneys* [28, 29], the *lungs* [30], the *pancreas* [31, 32], *etc.* Small organs (*e.g.*, the *pancreas*) are often more difficult to segment, partly due to their low contrast and large anatomical variability in size and (most often irregular) shape, as well as the complicated and unpredictable background contents. In particular, the internal neoplasms such as cysts [33] and tumors [34] can further change the anatomical property of the pancreas, making it even more difficult to recognize both targets.

Compared to previous works which used conventional approaches for segmentation, the progress of deep learning brought more powerful and efficient solutions. In particular, convolutional neural networks (CNNs) have been widely applied to a

wide range of vision tasks, such as image classification [35–37], object detection [38–40], and semantic segmentation [41, 42]. Recurrent neural networks, as a related class of networks, were first designed to process sequential data [43–45], and later generalized to image classification [46] and scene labeling [47] tasks. In the area of medical imaging analysis, in particular organ segmentation, different variants of fully convolutional networks (FCNs) [41] have been shown to significantly outperform conventional approaches, *e.g.*, segmenting the *liver* [48], the *lung* [49], or the *pancreas* [8, 50]. Additionally, researchers have also investigated the more challenging scenarios, such as liver lesion/tumor segmentation [51–53], brain lesion/tumor segmentation [6, 24], and general lesion detection [54–57]. Unlike these works, we aim to segment pancreatic cystic neoplasms and pancreatic tumors in this thesis, which has been rarely studied before.

The aforementioned studies are all focusing on single-organ segmentation. For multi-organ segmentation, atlas-based approaches were adopted for many applications [58–64]. The frameworks are usually problematic because 1) they are not able to capture the large inter-subject variations of abdominal regions and 2) computational time is tightly dependent on the number of atlases. Recently, learning-based approaches with relatively large datasets have been introduced for multi-organ segmentation [65–73]. Compared with multi-atlas-based approaches, CNNs based methods are generally more efficient and accurate. Recently deep-learning-based multi-organ segmentation has also been approached based on 3D FCNs [65, 74]. Later Roth *et al.* [66, 75] proposed to either use a hierarchical approach or to integrate multi-scale and varying context information for enhancing multi-organ segmentation.

We note that medical images differ from natural images in that data appear in a volumetric form. To deal with these data, researchers either slice a 3D volume into 2D slices [18, 74, 76], or train a 3D network directly [6, 77–79] for the applications of single- and multi- organ segmentation. In the latter case, 3D CNNs usually adopt

the sliding-window strategy to avoid the *out of memory* problem, leading to high time complexity induced by patch-based training and testing. Compared with 3D CNNs, 2D CNNs based algorithms can be directly end-to-end trained using 2D deep networks, which is less time-consuming. The trade-off between 2D and 3D approaches is discussed in [80].

2.2 Coarse-to-fine Approaches

By comparison to the entire CT volume, the organs and neoplasm (*e.g.*, pancreatic cyst) considered in this thesis often occupy a relatively small area. As deep segmentation networks such as FCN [41] are less accurate in depicting small targets, researchers proposed two types of ideas to improve detection and/or segmentation performance. The first type involved rescaling the image so that the target becomes comparable to the training samples [81], and the second one considered to focus on a sub-region of the image for each target to obtain higher accuracy in detection [82]. The coarse-to-fine idea was also well studied in the computer vision area for saliency detection [83] or semantic segmentation [84, 85]. In this thesis, we present coarse-to-fine frameworks for different medical image segmentation applications. The core idea is to first use the coarse stage to extract attentive regions, which is then fed to the fine stage to make dense predictions. The two stages can be trained either in a step-wise manner or in an end-to-end fashion.

Our method also belongs to attention-based methods. The attention mechanism has been successfully applied to various fields. Wang *et al.* [86] propose to model long-range relationships and design a non-local operator accordingly. Another type of attention is known as channel-wise attention [87, 88], which aims to model the relationships between different channels. In the field of medical image analysis, these attention modules are also widely used for different applications [89–92]. Different from these attention-based methods, in our approach we use a multi-stage framework

where the first stage explicitly extracts the attention, which is then used for facilitating the segmentation in the second stage.

2.3 Semi-/Weakly- Supervised Medical Image Segmentation

Currently, the most successful deep learning techniques for semantic segmentation stem from a common forerunner, *i.e.*, FCN [41]. Based on FCN, many recent advanced techniques have been proposed, such as DeepLab [93–95], PSPNet [96], RefineNet [85], *etc.* However, these methods are based on supervised learning, hence requiring a sufficient number of labeled training data to train. To cope with scenarios where supervision is limited, researchers begin to investigate the weakly-supervised setting [97–99], *e.g.*, only bounding-boxes or image-level labels are available, and the semi-supervised setting [97, 100], *i.e.*, unlabeled data are used to enlarge the training set. Papandreou *et al.* [97] propose EM-Adapt where the pseudo-labels of the unknown pixels are estimated in the expectation step and standard SGD is performed in the maximization step. Souly *et al.* [100] demonstrate the usefulness of generative adversarial networks for semi-supervised segmentation.

In the medical imaging domain, it becomes even more intractable to acquire sufficient labeled data due to the difficulty of annotation, as the annotation has to be done by experts. Although fully-supervised methods (*e.g.*, UNet [101], VoxResNet [102], DeepMedic [6], 3D-DSN [48], HNN [8]) have achieved remarkable performance improvement in tasks such as brain MR segmentation, abdominal single-organ segmentation and multi-organ segmentation, semi- or weakly-supervised learning is still a far more realistic solution. For these no-so-supervised settings [103], the most commonly used techniques include graph-based methods [104, 105], self-training [106, 107], co-training [12], multi-view learning [108], *etc.*

Graph-based semi-supervised methods define a graph where the nodes are labeled

and unlabeled examples in the dataset, and edges reflect the similarity of examples. These methods have been widely adopted in non-deep-learning based semi-supervised learning algorithms in the biomedical imaging domain [109–111].

In self-training, the classifier is iteratively re-trained using the training set augmented by adding the unlabeled data with their own predictions. The procedure is repeated until some convergence criteria are satisfied. In such case, one can imagine that a classification mistake can reinforce itself. Co-training [12] assumes that (1) features can be split into two independent sets and (2) each sub-feature set is sufficient to train a good classifier. During the learning process, each classifier is retrained with the additional training examples given by the other classifier. Co-training utilizes multiple sets of independent features which describe the same data, and therefore tends to yield more accurate and robust results than self-training [112]. Multi-view learning [108], in general, defines learning paradigms that utilize the agreement among different learners. Co-training is one of the earliest schemes for multi-view learning.

Built upon deep learning, self-training and co-training have witnessed good performances for different computer vision applications [11, 106, 107, 113]. In the medical imaging domain, similar attempts have been proved useful for semi-/weakly- supervised medical image segmentation. For instance, Bai *et al.* [10] propose an EM-based iterative method, where a CNN is alternately trained on labeled and post-processed unlabeled sets. In [114], supervised and unsupervised adversarial costs are involved to address semi-supervised gland segmentation. To make the learned models more robust, consistency-based methods [115–117] and uncertainty-driven approaches [118, 119] are proposed for different medical image classification and segmentation tasks. DeepCut [120] shows that bounding-boxes can be utilized by performing an iterative optimization scheme like [97] to benefit medical image segmentation. Kervadec *et al.* [121] further propose a constrained CNN method which suggests that weak annotations such as dots, scribbles can be also utilized for enhancing prostate segmentation.

2.4 Knowledge Priors in Deep Learning

The methods above are generally based on CNNs, which make them fail to capture the anatomical priors [122]. The inclusion of priors in medical imaging could potentially have much more impact compared with their usage in natural images since anatomical objects in medical images are naturally more constrained in terms of shape, location, size, *etc.* Earlier works suggest to employ priors through adjacency [123] and boundary [8, 124] conditions. Another popular strategy to explicitly employ prior structure for biomedical image segmentation is to use a conditional random field as a post-processing step [8, 24, 94, 125]. In [126, 127], shape priors are incorporated in neural networks by encouraging the computed segmentation to be similar to both the learned shape and the ground-truth. These approaches add priors simply by correcting segmentations produced by standard CNNs. Different from these studies, the recent study [122] demonstrates that priors can be learned by a generative model instead. But this method can incur additional computational overhead. Kervadec *et al.* [121] propose that directly imposing inequality constraints on sizes is also an effective way of incorporating anatomical priors. In this thesis, we propose to learn from partial annotations by embedding the abdominal region statistics in the training objective, which requires no additional training budgets.

Chapter 3

Coarse-to-Fine Approaches for Pancreas Segmentation

Deep neural networks have been widely adopted for automatic organ segmentation from abdominal CT scans. However, the segmentation accuracy of small organs (*e.g.*, the pancreas) is sometimes below satisfaction, arguably because deep networks are easily disrupted by the complex and variable background regions which occupy a large fraction of the input volume. In this chapter, we propose two coarse-to-fine mechanisms which use the prediction from the first (coarse) stage to shrink the input region for the second (fine) stage. More specifically, the two stages in the first method are trained individually in a step-wise manner, so that the entire input region and the region cropped according to the bounding box are treated separately. While the second method inserts a saliency transformation module between the two stages so that the segmentation probability map from the previous iteration can be repeatedly converted as spatial weights to the current iteration. In training, it allows joint optimization over both stages. In testing, it propagates multi-stage visual information throughout iterations to improve the segmentation accuracy. Experiments are performed on several CT datasets, including the NIH pancreas dataset and the JHMI multi-organ dataset, which confirms the effectiveness of our approach.

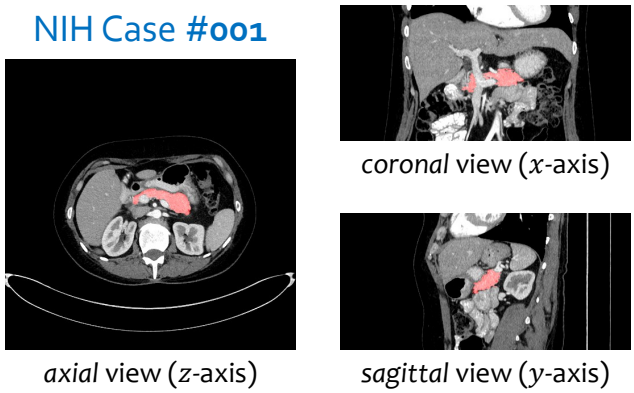


Figure 3-1. A typical example from the NIH *pancreas* segmentation dataset [7] (best viewed in color). We highlight the *pancreas* in red seen from three different viewpoints. It is a relatively small organ compared to the entire abdominal CT volume.

3.1 Introduction

This chapter focuses on small organ (*e.g.*, the pancreas) segmentation from abdominal CT scans, which is an important prerequisite for enabling computers to assist human doctors for clinical purposes. This problem falls into the research area named *medical imaging analysis*. Recently, great progress has been brought to this field by the fast development of deep learning, especially convolutional neural networks [35, 41]. Many conventional methods, such as the graph-based segmentation approaches [29] or those based on handcrafted local features [32], have been replaced by deep segmentation networks, which typically produce a higher segmentation accuracy [7, 18, 19, 101, 128].

Segmenting small structures (*e.g.*, the pancreas or neoplasms) from a CT scan is often challenging. As the target generally occupies a *small part* of the input data (*e.g.*, less than 1.5% in a 2D image, see Figure 3-1), deep segmentation networks such as FCN [41] and DeepLab [42] can be easily confused by the background region, which may contain complicated and variable contents. This motivates us to propose *coarse-to-fine* approaches, in which the coarse stage provides a rough localization, based on which the fine stage then performs accurate segmentation.

We propose two coarse-to-fine approaches in this chapter. In the first approach, we use the predicted segmentation mask to shrink the input region. With a relatively smaller input region (*e.g.*, a bounding box defined by the mask), it is straightforward to achieve more accurate segmentation. At the training stage, we fix the input region generated from the ground-truth annotation, and train two deep segmentation networks, *i.e.*, a coarse-scaled one and a fine-scaled one, to deal with the entire input region and the region cropped according to the bounding box, respectively. At the testing stage, the network parameters remain unchanged, and the coarse-scaled network was first used to obtain the rough position of the small target. Then the fine-scaled network was executed several times and the segmentation mask was updated iteratively until convergence. The iterative process can be formulated as a fixed-point model [129].

In spite of the state-of-the-art performance achieved for pancreas segmentation, this method suffers from *inconsistency* between its training and testing flowcharts, As in our first approach, the training phase dealt with coarse and fine stages individually and did not minimize a global energy function, but the testing phase assumed that these two stages can cooperate with each other in an iterative process. From another perspective, this also makes it difficult for multi-stage visual cues to be incorporated for enhancing the segmentation performance, *e.g.*, the previous segmentation mask which carries rich information is discarded except for the bounding box. In order to embed *consistency* between training and testing flowcharts, which is to say, we aim to minimize a global energy function in coarse and fine stages simultaneously during the training phase. To this end, we propose a Recurrent Saliency Transformation Network (RSTN) in our second approach. The chief innovation is to relate the coarse and fine stages with a saliency transformation module, which repeatedly transforms the segmentation probability map from previous iterations as spatial priors for the current iteration. This brings us two-fold advantages over the first method. First, in the

training phase, the coarse-scaled and fine-scaled networks are optimized jointly, so that the segmentation ability of each of them gets improved. Second, in the testing phase, the segmentation mask of each iteration is preserved and propagated throughout iterations, enabling multi-stage visual cues to be incorporated towards more accurate segmentation.

We show the superiority of our approaches on the NIH pancreas segmentation dataset [7] and the JHMI multi-organ dataset, which guarantees its efficiency and reliability in real clinical applications.

This chapter summarizes our previous works [13, 14]. The remainder of this chapter is organized as follows. Section 3.2 describes the proposed step-wise coarse-to-fine approach, and Section 3.3 presents our proposed end-to-end coarse-to-fine approach. After experiments are shown in Sections 3.4 and 3.5, we draw our conclusions in Section 3.6.

3.2 A Step-Wise Coarse-to-Fine Approach for Medical Image Segmentation

We investigate the problem of segmenting an organ from abdominal CT scans. Let a CT image be a 3D volume $\mathbf{X}$ of size $W \times H \times L$ which is annotated with a binary ground-truth segmentation $\mathbf{Y}$ where $y_i = 1$ indicates a foreground voxel. The goal of our work is to produce a binary output volume $\mathbf{Z}$ of the same dimension. Denote $\mathcal{Y}$ and $\mathcal{Z}$ as the set of foreground voxels in the ground-truth and prediction, *i.e.*, $\mathcal{Y} = \{i \mid y_i = 1\}$ and $\mathcal{Z} = \{i \mid z_i = 1\}$. The accuracy of segmentation is evaluated by the Dice-Sørensen coefficient (DSC): $\mathrm{DSC}(\mathcal{Y}, \mathcal{Z}) = \frac{2 \times |\mathcal{Y} \cap \mathcal{Z}|}{|\mathcal{Y}| + |\mathcal{Z}|}$. This metric falls in the range of $[0, 1]$ with 1 implying perfect segmentation.

3.2.1 Deep Segmentation Networks

Consider a segmentation model $\mathbb{M} : \mathbf{Z} = \mathbf{f}(\mathbf{X}; \mathbf{\Theta})$, where $\mathbf{\Theta}$ denotes the model parameters, and the loss function is written as $\mathcal{L}(\mathbf{Z}, \mathbf{Y})$. In the context of a deep segmentation network, we optimize $\mathcal{L}$ with respect to the network weights $\mathbf{\Theta}$ by gradient back-propagation. As the foreground region is often very small, we follow [78] to design a DSC-loss layer to prevent the model from being heavily biased towards the background class. We slightly modify the DSC of two voxel sets $\mathcal{A}$ and $\mathcal{B}$, $\text{DSC}(\mathcal{A}, \mathcal{B}) = \frac{2\times|\mathcal{A}\cap\mathcal{B}|}{|\mathcal{A}|+|\mathcal{B}|}$, into a loss function between the ground-truth mask $\mathbf{Y}$ and the predicted mask $\mathbf{Z}$, *i.e.*, $\mathcal{L}(\mathbf{Z}, \mathbf{Y}) = 1 - \frac{2\times\sum_i z_i y_i}{\sum_i z_i + \sum_i y_i}$. Note that this is a "soft" definition of DSC, and it is equivalent to the original form if all z_i's are either 0 or 1. The gradient computation is straightforward: $\frac{\partial\mathcal{L}(\mathbf{Z},\mathbf{Y})}{\partial z_j} = -2 \times \frac{y_j\left(\sum_i z_i + \sum_i y_i\right) - \sum_i z_i y_i}{\left(\sum_i z_i + \sum_i y_i\right)^2}$.

We train 2D deep networks for 3D segmentation[1]. Each 3D volume $\mathbf{X}$ is sliced along three axes, the *coronal*, *sagittal* and *axial* views, and these 2D slices are denoted by $\mathbf{X}_{\text{C},w}$ $(w = 1, 2, \ldots, W)$, $\mathbf{X}_{\text{S},h}$ $(h = 1, 2, \ldots, H)$ and $\mathbf{X}_{\text{A},l}$ $(l = 1, 2, \ldots, L)$, where the subscripts C, S and A stand for *coronal*, *sagittal* and *axial*, respectively. On each axis, an individual 2D-FCN [41] on a 16-layer VGGNet [36] is trained. In other words, we train three 2D-FCN models $\mathbb{M}_{\text{C}}$, $\mathbb{M}_{\text{S}}$ and $\mathbb{M}_{\text{A}}$ to perform segmentation through three views individually (images from three views are quite different). In testing, the segmentation results from three views are fused via majority voting. Both multi-slice segmentation (3 neighboring slices are combined as a basic unit in training and testing) and multi-axis fusion (majority voting over three axes) are performed to incorporate pseudo-3D information into segmentation.

3.2.2 Fixed-Point Optimization

The organs and neoplasms investigated in this chapter (*e.g.*, the *pancreas*) are relatively small. In each 2D slice, the fraction of the foreground pixels is often smaller than

[1]Please see Section 3.4.3 for the comparison to 3D networks.

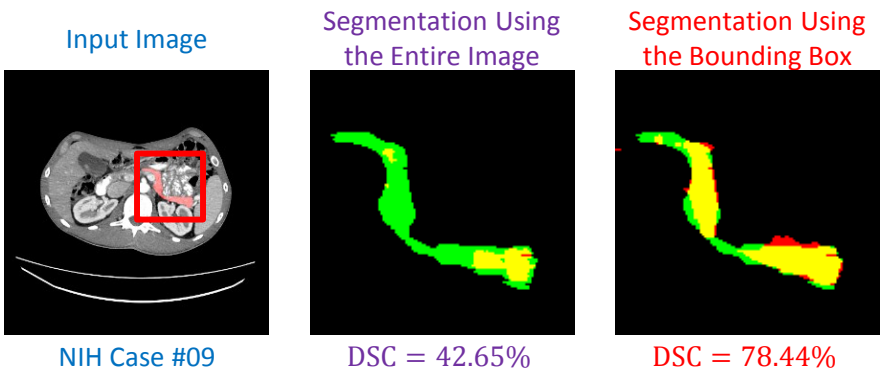


Figure 3-2. Segmentation results with different input regions (best viewed in color), either using the entire image or the bounding box (the red frame). Red, green and yellow indicate the prediction, ground-truth and overlapped pixels, respectively.

1.5%. It was observed [7] that deep segmentation networks such as FCN [41] produce less satisfying results when detecting small organs, arguably because the network is easily disrupted by the varying contents in the background regions. Much more accurate segmentation can be obtained by using a smaller input region around the region-of-interest. A typical example is shown in Figure 3-2.

This inspires us to make use of the predicted segmentation mask to shrink the input region. We introduce a transformation function $r(\mathbf{X}, \mathbf{Z}^\star)$ which generates the input region given the current segmentation $\mathbf{Z}^\star$. We rewrite the model as $\mathbf{Z} = \mathbf{f}(r(\mathbf{X}, \mathbf{Z}^\star)\,; \mathbf{\Theta})$, and the loss function is $\mathcal{L}(\mathbf{f}(r(\mathbf{X}, \mathbf{Z}^\star)\,; \mathbf{\Theta})\,, \mathbf{Y})$. Note that the segmentation mask ($\mathbf{Z}$ or $\mathbf{Z}^\star$) appears in both the input and output of $\mathbf{Z} = \mathbf{f}(r(\mathbf{X}, \mathbf{Z}^\star)\,; \mathbf{\Theta})$. This is a fixed-point model, and we apply the approach described in [129] for optimization, *i.e.*, finding a steady-state solution for $\mathbf{Z}$.

In training, the ground-truth annotation $\mathbf{Y}$ is used as the input mask $\mathbf{Z}^\star$. We train two sets of models (each set contains three models for different views) to deal with different input sizes. The *coarse-scaled* models are trained on those slices on which the pancreas occupies at least 100 pixels (approximately 25mm^2 in a 2D slice, our approach is not sensitive to this parameter) so as to prevent the model from being

heavily impacted by the background. For the *fine-scaled* models, we crop each slice according to the minimal 2D box covering the pancreas, add a frame around it, and fill it up with the original image data. The top, bottom, left and right margins of the frame are random integers sampled from $\{0, 1, \ldots, 60\}$. This strategy, known as data augmentation, helps to regularize the network and prevent over-fitting.

We initialize both networks using the FCN-8s model [41] pre-trained on the PascalVOC image segmentation task. The coarse-scaled model is fine-tuned with a learning rate of 10^{-5} for 80,000 iterations, and the fine-scaled model undergoes 60,000 iterations with a learning rate of 10^{-4}. Each mini-batch contains one training sample (a 2D image sliced from a 3D volume).

In testing, we use an iterative process to find a steady-state solution for $\mathbf{Z} = \mathbf{f}(r(\mathbf{X}, \mathbf{Z}^{\star}); \boldsymbol{\Theta})$. At the beginning, $\mathbf{Z}^{\star}$ is initialized as the entire 3D volume, and we compute the *coarse* segmentation $\mathbf{Z}^{(0)}$ using the *coarse-scaled* models. In each of the following T iterations, we slice the predicted mask $\mathbf{Z}^{(t-1)}$, find the smallest 2D box to cover all predicted foreground pixels in each slice, add a 30-pixel-wide frame around it (this is the mean value of the random distribution used in training), and use the *fine-scaled* models to compute $\mathbf{Z}^{(t)}$. The iteration terminates when a fixed number of iterations T is reached, or the similarity between successive segmentation results ($\mathbf{Z}^{(t-1)}$ and $\mathbf{Z}^{(t)}$) is larger than a given threshold R. The similarity is defined as the inter-iteration DSC, namely $d^{(t)} = \mathrm{DSC}\left(\mathbf{Z}^{(t-1)}, \mathbf{Z}^{(t)}\right) = \frac{2 \times \sum_i z_i^{(t-1)} z_i^{(t)}}{\sum_i z_i^{(t-1)} + \sum_i z_i^{(t)}}$. The testing stage is illustrated in Figure 3-3 and described in Algorithm 1.

3.3 An End-to-End Coarse-to-Fine Approach for Medical Image Segmentation

The step-wise coarse-to-fine approach is delicately designed for tiny target segmentation, but lacks global optimization of both the coarse and fine networks in the training stage. This motivates us to connect these two networks with a saliency transformation

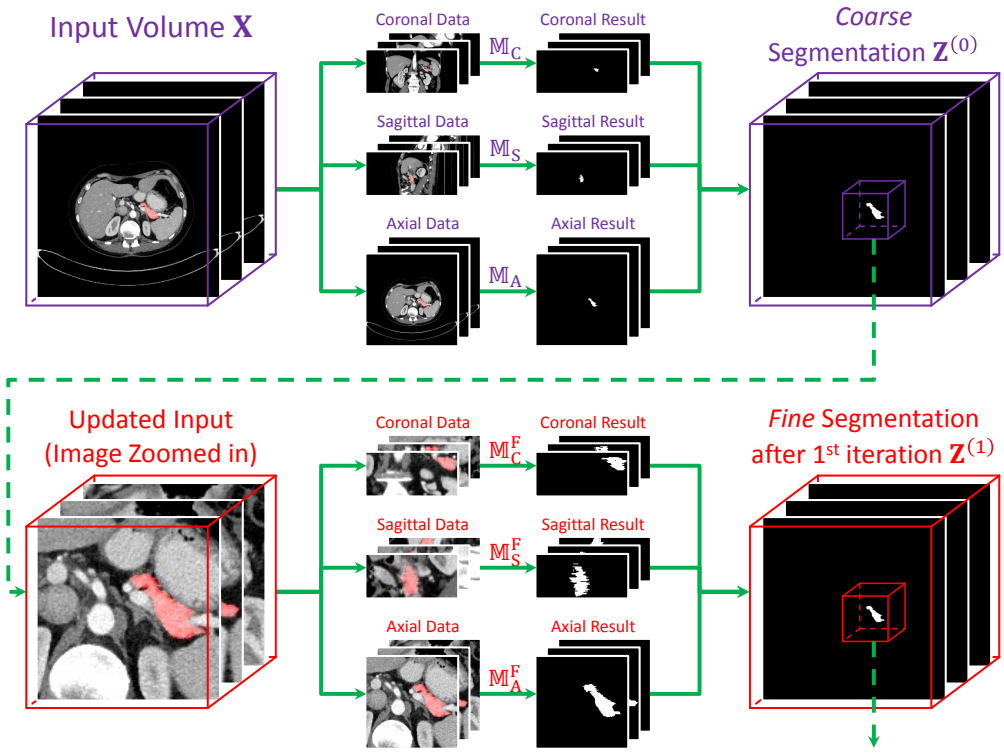


Figure 3-3. Illustration of the testing process (best viewed in color). Only one iteration is shown here. In practice, there are at most 10 iterations.

module, which leads to our end-to-end coarse-to-fine approach.

3.3.1 Recurrent Saliency Transformation Network

Following the step-wise coarse-to-fine approach, we also train an individual model for each of the three viewpoints. Without loss of generality, we consider a 2D slice along the *axial* view, denoted by $\mathbf{X}_{\mathrm{A},l}$. Our goal is to infer a binary segmentation mask $\mathbf{Z}_{\mathrm{A},l}$, which is achieved by first computing a *probability map* $\mathbf{P}_{\mathrm{A},l} = \mathbf{f}[\mathbf{X}_{\mathrm{A},l}; \boldsymbol{\theta}]$, where $\mathbf{f}[\cdot; \boldsymbol{\theta}]$ is a deep segmentation network with $\boldsymbol{\theta}$ being network parameters, and then binarizing $\mathbf{P}_{\mathrm{A},l}$ into $\mathbf{Z}_{\mathrm{A},l}$ using a fixed threshold of 0.5, *i.e.*, $\mathbf{Z}_{\mathrm{A},l} = \mathbb{I}[\mathbf{P}_{\mathrm{A},l} \geqslant 0.5]$.

In order to assist segmentation with the probability map, we introduce $\mathbf{P}_{\mathrm{A},l}$ as a latent variable. We introduce a *saliency transformation* module, which takes the probability map to generate an updated input image, *i.e.*, $\mathbf{I}_{\mathrm{A},l} = \mathbf{X}_{\mathrm{A},l} \odot \mathbf{g}(\mathbf{P}_{\mathrm{A},l}; \boldsymbol{\eta})$, and

Algorithm 1 Fixed-Point Model for Segmentation

Input: the testing volume $\mathbf{X}$, coarse-scaled models $\mathbb{M}_{\mathrm{C}}$, $\mathbb{M}_{\mathrm{S}}$ and $\mathbb{M}_{\mathrm{A}}$, fine-scaled models $\mathbb{M}_{\mathrm{C}}^{\mathrm{F}}$, $\mathbb{M}_{\mathrm{S}}^{\mathrm{F}}$ and $\mathbb{M}_{\mathrm{A}}^{\mathrm{F}}$, threshold R, maximal rounds in iteration T.
Initialization: using $\mathbb{M}_{\mathrm{C}}$, $\mathbb{M}_{\mathrm{S}}$ and $\mathbb{M}_{\mathrm{A}}$ to generate $\mathbf{Z}^{(0)}$ from $\mathbf{X}$;
for $t = 1, 2, \ldots, T$ **do**
  Using $\mathbb{M}_{\mathrm{C}}^{\mathrm{F}}$, $\mathbb{M}_{\mathrm{S}}^{\mathrm{F}}$ and $\mathbb{M}_{\mathrm{A}}^{\mathrm{F}}$ to generate $\mathbf{Z}^{(t)}$ from $\mathbf{Z}^{(t-1)}$;
  if $\mathrm{DSC}\Big(\mathbf{Z}^{(t-1)}, \mathbf{Z}^{(t)}\Big) \geqslant R$ **then**
    break;
  end if
end for
Output: the final segmentation $\mathbf{Z}^{\star} = \mathbf{Z}^{(t)}$.

uses the updated input $\mathbf{I}_{\mathrm{A},l}$ to replace $\mathbf{X}_{\mathrm{A},l}$. Here $\mathbf{g}[\cdot;\boldsymbol{\eta}]$ is the transformation function with parameters $\boldsymbol{\eta}$, and $\odot$ denotes element-wise product, *i.e.*, the transformation function adds spatial weights to the original input image. Thus, the segmentation process becomes:

$$\mathbf{P}_{\mathrm{A},l} = \mathbf{f}[\mathbf{X}_{\mathrm{A},l} \odot \mathbf{g}(\mathbf{P}_{\mathrm{A},l};\boldsymbol{\eta})\,;\boldsymbol{\theta}]. \tag{3.1}$$

This is a recurrent neural network. Note that the saliency transformation function $\mathbf{g}[\cdot,\boldsymbol{\eta}]$ needs to be differentiable so that the entire recurrent network can be optimized in an end-to-end manner. As $\mathbf{X}_{\mathrm{A},l}$ and $\mathbf{P}_{\mathrm{A},l}$ share the same spatial dimensionality, we set $\mathbf{g}[\cdot,\boldsymbol{\eta}]$ to be a *size-preserved* convolution, which allows the weight added to each pixel to be determined by the segmentation probabilities in a small neighborhood around it. As we will show in the experimental section (see Figure 3-7), the learned convolutional kernels are able to extract complementary information to help the next iteration.

To optimize Eqn. (3.1), we unfold the recurrent network into a plain form (see Figure 3-4). Given an input image $\mathbf{X}_{\mathrm{A},l}$ and an integer T which is the maximal number of iterations, we update $\mathbf{I}_{\mathrm{A},l}^{(t)}$ and $\mathbf{P}_{\mathrm{A},l}^{(t)}$, $t = 0, 1, \ldots, T$:

$$\mathbf{I}_{\mathrm{A},l}^{(t)} = \mathbf{X}_{\mathrm{A},l} \odot \mathbf{g}\Big(\mathbf{P}_{\mathrm{A},l}^{(t-1)};\boldsymbol{\eta}\Big), \tag{3.2}$$

$$\mathbf{P}_{\mathrm{A},l}^{(t)} = \mathbf{f}\Big[\mathbf{I}_{\mathrm{A},l}^{(t)};\boldsymbol{\theta}\Big]. \tag{3.3}$$

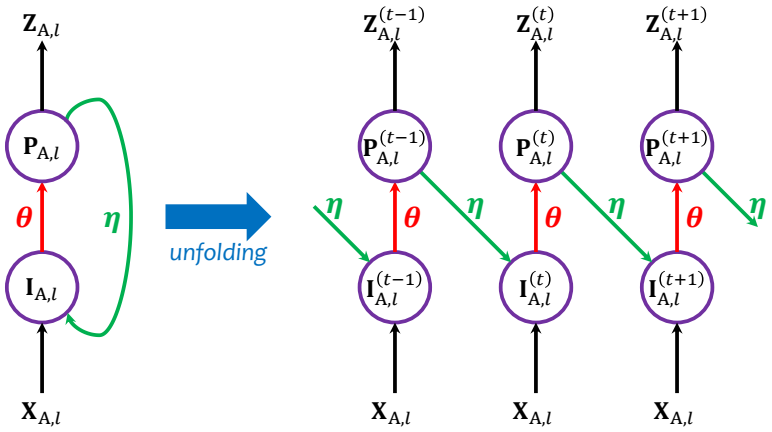


Figure 3-4. We formulate our approach into a recurrent network, and unfold it for optimization and inference.

Note that the original input image $\mathbf{X}_{\mathrm{A},l}$ does not change, and the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are shared by all iterations. At $t = 0$, we directly set $\mathbf{I}_{\mathrm{A},l}^{(0)} = \mathbf{X}_{\mathrm{A},l}$.

When segmentation masks $\mathbf{P}_{\mathrm{A},l}^{(t)}$ $(t = 0, 1, \ldots, T-1)$ are available for reference, deep networks benefit considerably from a shrunk input region especially when the target organ is very small. Thus, we define a *cropping* function $\mathrm{Crop}\left[\cdot; \mathbf{P}_{\mathrm{A},l}^{(t)}\right]$, which takes $\mathbf{P}_{\mathrm{A},l}^{(t)}$ as the *reference map*, binarizes it into $\mathbf{Z}_{\mathrm{A},l}^{(t)} = \mathbb{I}\left[\mathbf{P}_{\mathrm{A},l}^{(t)} \geqslant 0.5\right]$, finds the minimal rectangle covering all the activated pixels, and adds a K-pixel-wide margin (padding) around it. We fix K to be 20; our algorithm is not sensitive to this parameter.

Finally note that $\mathbf{I}_{\mathrm{A},l}^{(0)}$, the original input (the entire 2D slice), is much larger than the cropped inputs $\mathbf{I}_{\mathrm{A},l}^{(t)}$ for $t > 0$. We train two FCN's to deal with such a major difference in input data. The first one is named the *coarse-scaled* segmentation network, which is used *only* in the first iteration. The second one, the *fine-scaled* segmentation network, takes the charge of all the remaining iterations. We denote their parameters by $\boldsymbol{\theta}^{\mathrm{C}}$ and $\boldsymbol{\theta}^{\mathrm{F}}$, respectively. These two FCN's are optimized jointly.

We compute a DSC loss term on each probability map $\mathbf{P}_{\mathrm{A},l}^{(t)}$, $t = 0, 1, \ldots, T$, and denote it by $\mathcal{L}\left\{\mathbf{Y}_{\mathrm{A},l}, \mathbf{P}_{\mathrm{A},l}^{(t)}\right\}$. Here, $\mathbf{Y}_{\mathrm{A},l}$ is the ground-truth segmentation mask, and $\mathcal{L}\{\mathbf{Y}, \mathbf{P}\} = 1 - \frac{2 \times \sum_i Y_i P_i}{\sum_i Y_i + P_i}$ is based on the *soft* version of DSC [78]. Our goal is to

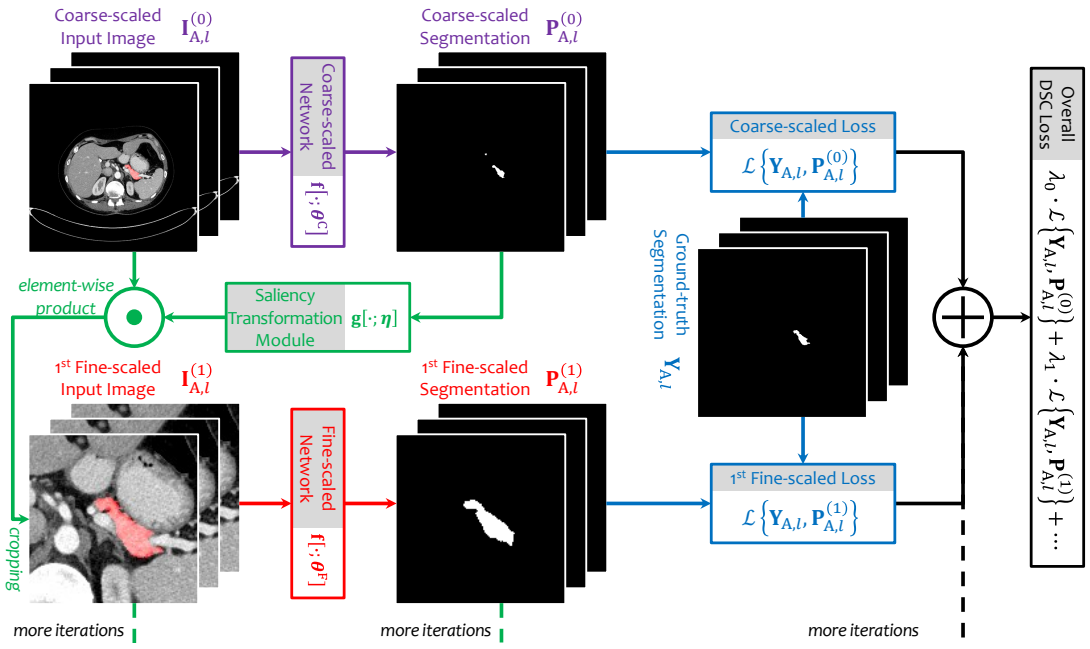


Figure 3-5. Illustration of the training process (best viewed in color). We display an input image along the *axial* view which contains 3 neighboring slices. To save space, we only plot the coarse stage and the first iteration in the fine stage.

minimize the overall loss:

$$\mathcal{L} = \sum_{t=0}^{T} \lambda_t \cdot \mathcal{L}\left\{\mathbf{Y}_{\mathrm{A},l}^{(t)}, \mathbf{Z}_{\mathrm{A},l}^{(t)}\right\}. \tag{3.4}$$

This leads to joint optimization over all iterations, which involves network parameters $\boldsymbol{\theta}^{\mathrm{C}}$, $\boldsymbol{\theta}^{\mathrm{F}}$, and transformation parameters $\boldsymbol{\eta}$. $\{\lambda_t\}_{t=0}^{T}$ controls the tradeoff among all loss terms. We set $2\lambda_0 = \lambda_1 = \ldots = \lambda_T = 2/(2T+1)$ so as to encourage accurate fine-scaled segmentation.

3.3.2 Training and Testing

The training phase is aimed at minimizing the loss function $\mathcal{L}$, defined in Eqn. (3.4), which is differentiable with respect to all parameters. In the early training stages, the coarse-scaled network cannot generate reasonable probability maps. To prevent the fine-scaled network from being confused by inaccurate input regions, we use the ground-truth mask $\mathbf{Y}_{\mathrm{A},l}$ as the reference map. After a sufficient number of

Algorithm 2 The Testing Phase for RSTN

Require: input volume $\mathbf{X}$, viewpoint $\mathcal{V} = \{\mathrm{C}, \mathrm{S}, \mathrm{A}\}$
Require: parameters $\boldsymbol{\theta}_{\mathrm{v}}^{\mathrm{C}}$
Require: $\boldsymbol{\theta}_{\mathrm{v}}^{\mathrm{F}}$ and $\boldsymbol{\eta}_{\mathrm{v}}$, $\mathrm{v} \in \mathcal{V}$;
Require: max number of iterations T, threshold thr;
$t \leftarrow 0$, $\mathbf{I}_{\mathrm{v}}^{(0)}$
$\leftarrow \mathbf{X}$, $\mathrm{v} \in \mathcal{V}$;
$\mathbf{P}_{\mathrm{v},l}^{(0)} \leftarrow \mathbf{f}\Big[\mathbf{I}_{\mathrm{v},l}^{(0)}; \boldsymbol{\theta}_{\mathrm{v}}^{\mathrm{C}}\Big]$, $\mathrm{v} \in \mathcal{V}$, $\forall l$;
$\mathbf{P}^{(0)} = \frac{\mathbf{P}_{\mathrm{C}}^{(0)} + \mathbf{P}_{\mathrm{S}}^{(0)} + \mathbf{P}_{\mathrm{A}}^{(0)}}{3}$, $\mathbf{Z}^{(0)} = \mathbb{I}\Big[\mathbf{P}^{(0)} \geqslant 0.5\Big]$;
repeat
  $t \leftarrow t + 1$;
  $\mathbf{I}_{\mathrm{v},l}^{(t)} \leftarrow \mathbf{X}_{\mathrm{v},l} \odot \mathbf{g}\Big(\mathbf{P}_{\mathrm{v},l}^{(t-1)}; \boldsymbol{\eta}\Big)$, $\mathrm{v} \in \mathcal{V}$, $\forall l$;
  $\mathbf{P}_{\mathrm{v},l}^{(t)} \leftarrow \mathbf{f}\Big[\mathrm{Crop}\Big[\mathbf{I}_{\mathrm{v},l}^{(t)}; \mathbf{P}_{\mathrm{v},l}^{(t-1)}\Big]; \boldsymbol{\theta}_{\mathrm{v}}^{\mathrm{F}}\Big]$, $\mathrm{v} \in \mathcal{V}$, $\forall l$;
  $\mathbf{P}^{(t)} = \frac{\mathbf{P}_{\mathrm{C}}^{(t)} + \mathbf{P}_{\mathrm{S}}^{(t)} + \mathbf{P}_{\mathrm{A}}^{(t)}}{3}$, $\mathbf{Z}^{(t)} = \mathbb{I}\Big[\mathbf{P}^{(t)} \geqslant 0.5\Big]$;
until $t = T$ **or** $\mathrm{DSC}\Big\{\mathbf{Z}^{(t-1)}, \mathbf{Z}^{(t)}\Big\} \geqslant \mathrm{thr}$
return $\mathbf{Z} \leftarrow \mathbf{Z}^{(t)}$.

training, we resume using $\mathbf{P}_{\mathrm{A},l}^{(t)}$ instead of $\mathbf{Y}_{\mathrm{A},l}$. In Section 3.4.3, we will see that this "fine-tuning" strategy improves segmentation accuracy considerably.

Due to the limitation in GPU memory, in each mini-batch containing one training sample, we set T to be the maximal integer (not larger than 5) so that we can fit the entire framework into the GPU memory. The overall framework is illustrated in Figure 3-5. As a side note, we find that setting $T \equiv 1$ also produces high accuracy, suggesting that major improvement is brought by joint optimization.

The testing phase follows the flowchart described in Algorithm 2. There are two minor differences from the training phase. First, as the ground-truth segmentation mask $\mathbf{Y}_{\mathrm{A},l}$ is not available, the probability map $\mathbf{P}_{\mathrm{A},l}^{(t)}$ is always taken as the reference map for image cropping. Second, the number of iterations is no longer limited by the GPU memory, as the intermediate outputs can be discarded on the way. In practice, we terminate our algorithm when the similarity of two consecutive predictions, measured by $\mathrm{DSC}\Big\{\mathbf{Z}^{(t-1)}, \mathbf{Z}^{(t)}\Big\} = \frac{2 \times \sum_i Z_i^{(t-1)} Z_i^{(t)}}{\sum_i Z_i^{(t-1)} + Z_i^{(t)}}$, reaches a threshold thr, or a fixed number (T) of iterations are executed. We will discuss these parameters in Section 3.4.3.

3.4 Pancreas Segmentation Experiments

3.4.1 Dataset and Evaluation

We evaluate our approach on the NIH *pancreas* segmentation dataset [7], which contains 82 contrast-enhanced abdominal CT volumes. The resolution of each scan is $512 \times 512 \times L$, where $L \in [181, 466]$ is the number of slices along the long axis of the body. The distance between neighboring voxels ranges from 0.5mm to 1.0mm.

Following the standard cross-validation strategy, we split the dataset into 4 fixed folds, each of which contains approximately the same number of samples. We apply cross validation, *i.e.*, training the models on 3 out of 4 subsets and testing them on the remaining one. We measure the segmentation accuracy by computing the Dice-Sørensen coefficient (DSC) for each sample, and report the average and standard deviation over all 82 cases.

3.4.2 Evaluation of the Step-Wise Coarse-to-Fine Approach

We initialize both networks using the FCN-8s model [41] pre-trained on the PascalVOC image segmentation task. The coarse-scaled model is fine-tuned with a learning rate of 10^{-5} for 80,000 iterations, and the fine-scaled model undergoes 60,000 iterations with a learning rate of 10^{-4}. Each mini-batch contains one training sample (a 2D image sliced from a 3D volume).

We first evaluate the baseline (coarse-scaled) approach. Using the coarse-scaled models trained from three different views (*i.e.*, $\mathbb{M}_{\mathrm{C}}$, $\mathbb{M}_{\mathrm{S}}$ and $\mathbb{M}_{\mathrm{A}}$), we obtain $66.88\% \pm 11.08\%$, $71.41\% \pm 11.12\%$ and $73.08\% \pm 9.60\%$ average DSC, respectively. Fusing these three models via majority voting yields $75.74\% \pm 10.47\%$, suggesting that complementary information is captured by different views. This is used as the starting point $\mathbf{Z}^{(0)}$ for the later iterations.

To apply the fixed-point model for segmentation, we first compute $d^{(t)}$ to observe

Method	Mean DSC	# Iterations	Max DSC	Min DSC
Roth *et al.*, MICCAI'15 [7]	71.42 ± 10.11	–	86.29	23.99
Roth *et al.*, MICCAI'16 [8]	78.01 ± 8.20	–	88.65	34.11
Coarse Segmentation	75.74 ± 10.47	–	88.12	39.99
After 1 Iteration	82.16 ± 6.29	1	**90.85**	54.39
After 2 Iterations	82.13 ± 6.30	2	90.77	57.05
After 3 Iterations	82.09 ± 6.17	3	90.78	58.39
After 5 Iterations	82.11 ± 6.09	5	90.75	62.40
After 10 Iterations	82.25 ± 5.73	10	90.76	61.73
After $d_t > 0.90$	82.13 ± 6.35	1.83 ± 0.47	**90.85**	54.39
After $d_t > 0.95$	$\mathbf{82.37} \pm 5.68$	2.89 ± 1.75	**90.85**	**62.43**
After $d_t > 0.99$	82.28 ± 5.72	9.87 ± 0.73	90.77	61.94
Best among All Iterations	82.65 ± 5.47	3.49 ± 2.92	90.85	63.02
Oracle Bounding Box	83.18 ± 4.81	–	91.03	65.10

Table 3-I. Segmentation accuracy (measured by DSC, %) reported by different approaches. We start from the initial (coarse) segmentation $\mathbf{Z}^{(0)}$, and explore different terminating conditions, including a fixed number of iterations and a fixed threshold of inter-iteration DSC. The last two lines show two upper-bounds of our approach, *i.e.*, "Best of All Iterations" means that we choose the highest DSC value over 10 iterations, and "Oracle Bounding Box" corresponds to using the ground-truth segmentation to generate the bounding box in testing. We also compare our results with the state-of-the-art [7, 8], demonstrating our advantage over all statistics.

the convergence of the iterations. After 10 iterations, the average $d^{(t)}$ value over all samples is 0.9767, the median is 0.9794, and the minimum is 0.9362. These numbers indicate that the iteration process is generally stable.

Now, we investigate the fixed-point model using the threshold $R = 0.95$ and the maximal number of iterations $T = 10$. The average DSC is boosted by 6.63%, which is impressive given the relatively high baseline (75.74%). This verifies our hypothesis, *i.e.*, a fine-scaled model depicts a small organ more accurately.

We also summarize the results generated by different terminating conditions in Table 3-I. We find that performing merely 1 iteration is enough to significantly boost the segmentation accuracy (+6.42%). However, more iterations help to improve the accuracy of the worst case, as for some challenging cases (*e.g.*, Case #09, see

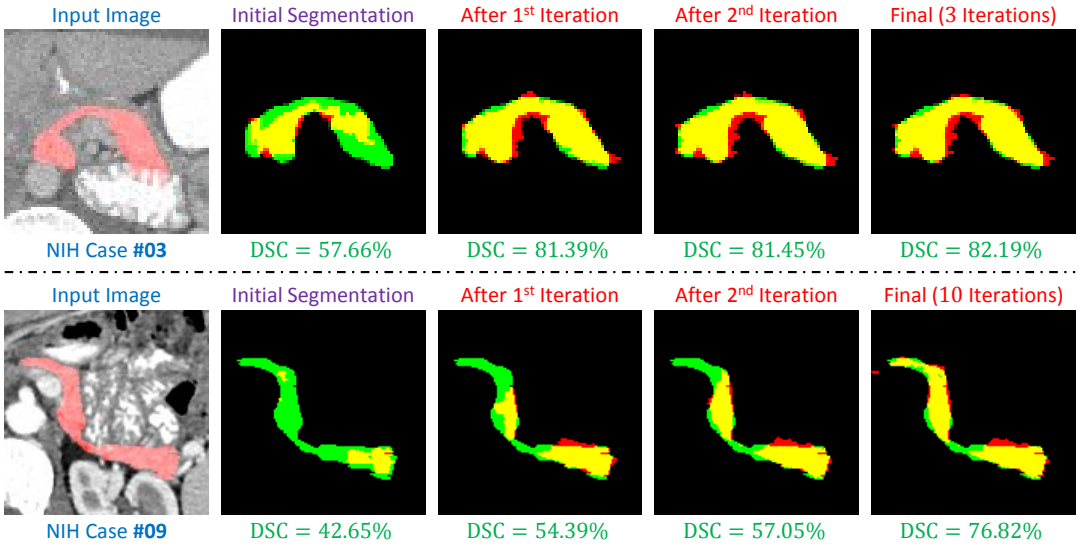


Figure 3-6. Examples of segmentation results throughout the iteration process (best viewed in color). We only show a small region covering the pancreas in the axial view. The terminating condition is $d^{(t)} \geqslant 0.95$. Red, green and yellow indicate the prediction, ground-truth and overlapped regions, respectively.

Figure 3-6), the missing parts in coarse segmentation are recovered gradually. The best average accuracy comes from setting $R = 0.95$. Using a larger threshold (*e.g.*, 0.99) does not produce accuracy gain, but requires more iterations and, consequently, more computation at the testing stage. In average, it takes less than 3 iterations to reach the threshold 0.95. On a modern GPU, we need about 3 minutes on each testing sample, comparable to recent work [8], but we report much higher segmentation accuracy (82.37% vs. 78.01%).

As a diagnostic experiment, we use the ground-truth (oracle) bounding box of each testing case to generate the input volume. This results in a 83.18% average accuracy (no iteration is needed in this case). By comparison, we report a comparable 82.37% average accuracy, indicating that our approach has almost reached the upper-bound of the current deep segmentation network.

We also compare our segmentation results with the state-of-the-art approaches. Using DSC as the evaluation metric, our approach outperforms the recent published

Model	Average	Max	Min
3×3 kernels in saliency transformation	83.47 ± 5.78	90.63	57.85
1×1 kernels in saliency transformation	82.85 ± 6.68	90.40	53.44
5×5 kernels in saliency transformation	83.64 ± 5.29	90.35	66.35
Two-layer saliency transformation (3×3 kernels)	83.93 ± 5.43	90.52	64.78
Fine-tuning with noisy data (3×3 kernels)	83.99 ± 5.09	90.57	65.05

Table 3-II. Accuracy (DSC, %) comparison of different settings of our approach. Please see the texts in Section 3.4.3 for detailed descriptions of these variants.

work [8] significantly. The average accuracy over 82 samples increases remarkably from 78.01% to 82.37%, and the standard deviation decreases from 8.20% to 5.68%, implying that our approach are more stable. We also implement a recently published coarse-to-fine approach [130], and get a 77.89% average accuracy. In particular, [8] reported 34.11% for the worst case (some previous works [31, 32] reported even lower numbers), and this number is boosted considerably to 62.43% by our approach. We point out that these improvements are mainly due to the fine-tuning iterations. Without it, the average accuracy is 75.74%, and the accuracy on the worst case is merely 39.99%. Figure 3-6 shows examples of how the segmentation quality is improved in two challenging cases.

3.4.3 Evaluation of the End-to-End Coarse-to-Fine Approach

Different Settings. We initialize the up-sampling layers in FCN-8s model [41] pre-trained on PascalVOC [131] with random weights, set the learning rate to be 10^{-4} and run 80,000 iterations. Different options are evaluated, including using different kernel sizes in saliency transformation, and whether to fine-tune the models using the predicted segmentations as reference maps (see the description in Section 3.3.2). Quantitative results are summarized in Table 3-II.

As the saliency transformation module is implemented by a size-preserved convolution (see Section 3.3.1), the size of convolutional kernels determines the range

that a pixel can use to judge its saliency. In general, a larger kernel size improves segmentation accuracy (3×3 works significantly better than 1×1), but we observe the marginal effect: the improvement of 5×5 over 3×3 is limited. As we use 7×7 kernels, the segmentation accuracy is slightly lower than that of 5×5. This may be caused by the larger number of parameters introduced to this module. Another way of increasing the receptive field size is to use two convolutional layers with 3×3 kernels. This strategy, while containing a smaller number of parameters, works even better than using one 5×5 layer. But, we do not add more layers, as the performance saturates while computational costs increase.

As described in Section 3.3.2, we fine-tune these models with images cropped from the coarse-scaled segmentation mask. This is to adjust the models to the testing phase, in which the ground-truth mask is unknown, so that the fine-scaled segmentation needs to start with, and be able to revise the coarse-scaled segmentation mask. We use a smaller learning rate (10^{-6}) and run another 40,000 iterations. This strategy not only reports 0.52% overall accuracy gain, but also alleviates over-fitting.

In summary, all these variants produce higher accuracy than our step-wise coarse-to-fine approach (82.37%), which verifies that the major contribution of our end-to-end approach comes from our recurrent framework which enables joint optimization. In the later experiments, we inherit the best variant learned from this section, including in a large-scale multi-organ dataset (see Section 3.5). That is to say, we use two 3×3 convolutional layers for saliency transformation, and fine-tune the models with coarse-scaled segmentation. This setting produces an average accuracy of 84.50%, as shown in Table 3-III.

Performance Comparison. We show that our end-to-end coarse-to-fine approach works better than the step-wise coarse-to-fine approach. As shown in Table 3-III, the average improvement over 82 cases is $2.13 \pm 2.67\%$. In addition, the student's t-test

Approach	Average	Max	Min
Roth *et al.* [7]	71.42 ± 10.11	86.29	23.99
Roth *et al.* [8]	78.01 ± 8.20	88.65	34.11
Zhang *et al.* [130]	77.89 ± 8.52	89.17	43.67
Roth *et al.* [8]	81.27 ± 6.27	88.96	50.69
Cai *et al.* [50]	82.4 ± 6.7	90.1	60.0
Our Step-Wise Approach	82.37 ± 5.68	90.85	62.43
Our End-to-End Approach	$\mathbf{84.50} \pm 4.97$	**91.02**	**62.81**

Table 3-III. Accuracy (DSC, %) comparison between our approach and the state-of-the-arts on the NIH *pancreas* segmentation dataset [7].

suggests statistical significance ($p = 3.62 \times 10^{-8}$) of our improvement. A case-by-case study reveals that our end-to-end approach reports higher accuracies on 67 out of 82 cases, with the largest advantage being +17.60% and the largest deficit being merely −3.85%. We analyze the sources of improvement in Section 3.4.4.

We briefly discuss the advantages and disadvantages of using 3D networks. 3D networks capture richer contextual information, but also require training more parameters. Our 2D approach makes use of 3D contexts more efficiently. At the end of each iteration, predictions from three views are fused, and thus the saliency transformation module carries this information to the next iteration. We implement VNet [78], and obtain an average accuracy of 83.18% with a 3D *ground-truth* bounding box provided for each case. Without the ground-truth, a sliding-window process is required which is really slow – an average of 5 minutes on a Titan-X Pascal GPU. In comparison, our end-to-end approach needs 1.3 minutes, slower than our step-wise approach (0.9 minutes), but faster than other 2D approaches [7, 8] (2–3 minutes).

3.4.4 Diagnosis

Joint Optimization and Multi-Stage Cues. Our end-to-end approach enables joint training, which improves both the coarse and fine stages individually. We denote the two networks trained by our step-wise approach by $\mathbb{I}^{\mathrm{C}}$ and $\mathbb{I}^{\mathrm{F}}$, and similarly, those

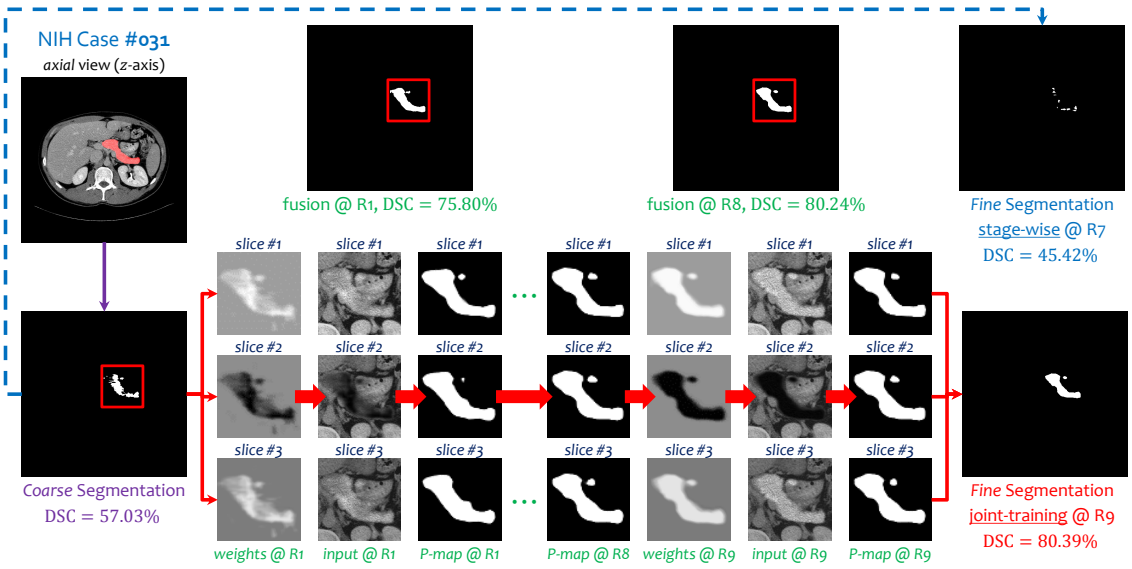


Figure 3-7. Visualization of how recurrent saliency transformation works in coarse-to-fine segmentation (best viewed in color). Segmentation accuracy is largely improved by making use of the probability map from the previous iteration to help the current iteration. Note that the three weight maps capture different visual cues, with two of them focused on the foreground region, and the remaining one focused on the background region.

trained in our approach by $\mathbb{J}^{\mathrm{C}}$ and $\mathbb{J}^{\mathrm{F}}$, respectively. In the coarse stage, $\mathbb{I}^{\mathrm{C}}$ reports 75.74% and $\mathbb{J}^{\mathrm{C}}$ reports 78.23%. In the fine stage, applying $\mathbb{J}^{\mathrm{F}}$ on top of the output of $\mathbb{I}^{\mathrm{C}}$ gets 83.80%, which is considerably higher than 82.37% ($\mathbb{I}^{\mathrm{F}}$ on top of $\mathbb{I}^{\mathrm{C}}$) but lower than 84.50% ($\mathbb{J}^{\mathrm{F}}$ on top of $\mathbb{J}^{\mathrm{C}}$). Therefore, we conclude that both the coarse-scaled and fine-scaled networks benefit from joint optimization. A stronger coarse stage provides a better starting point, and a stronger fine stage improves the upper-bound.

In Figure 3-7, we visualize how the recurrent network assists segmentation by incorporating multi-stage visual cues. It is interesting to see that in saliency transformation, different channels deliver complementary information, *i.e.*, two of them focus on the target organ, and the remaining one adds most weights to the background region. Similar phenomena happen in the models trained in different viewpoints and different folds. This reveals that, except for foreground, background and boundary also contribute to visual recognition [132].

Convergence. We study convergence, which is a very important criterion to judge the reliability of our end-to-end approach. We choose the best model reporting an average accuracy of 84.50%, and record the inter-iteration DSC throughout the testing process: $d^{(t)} = \text{DSC}\Big\{\mathbf{Z}^{(t-1)}, \mathbf{Z}^{(t)}\Big\} = \frac{2\times\sum_i Z_i^{(t-1)} Z_i^{(t)}}{\sum_i Z_i^{(t-1)} + Z_i^{(t)}}$.

After 1, 2, 3, 5 and 10 iterations, these numbers are 0.9037, 0.9677, 0.9814, 0.9908 and 0.9964 for our approach, and 0.8286, 0.9477, 0.9661, 0.9743 and 0.9774 for our step-wise approach, respectively. Each number reported by our end-to-end approach is considerably higher than that by the step-wise approach. The better convergence property provides us with the opportunity to set a more strict terminating condition, *e.g.*, using thr = 0.99 rather than thr = 0.95.

When the threshold is increased from 0.95 to 0.99 in our end-to-end approach, 80 out of 82 cases converge (in an average of 5.22 iterations), and the average accuracy is improved from 83.93% to 84.50%. On a Titan-X Pascal GPU, one iteration takes 0.2 minutes, so using thr = 0.99 requires an average of 1.3 minutes in each testing case.

The Over-Fitting Issue. Finally, we investigate the over-fitting issue of our end-to-end approach by making use of *oracle* information in the testing process. We use the ground-truth bounding box *on each slice*, which is used to crop the input region in *every* iteration. Note that annotating a bounding box in each slice is expensive and thus not applicable in real-world clinical applications. This experiment is aimed at exploring the upper-bound of our segmentation networks under perfect localization.

With oracle information provided, our best model reports 86.37%, which is considerably higher than the number (84.50%) without using oracle information. If we do not fine-tune the networks using coarse-scaled segmentation (see Table 3-II), the above numbers are 86.26% and 83.68%, respectively. This is to say, fine-tuning prevents our model from relying on the ground-truth mask. It not only improves the average accuracy, but also alleviates over-fitting (the disadvantage of our model against that

Organ	Stepwise-**C**	Stepwise-**F**	End2end-**C**	End2end-**F**
adrenal g.	57.38	61.65	60.70	**63.76**
duodenum	67.42	69.39	71.40	**73.42**
gallbladder	82.57	♯82.12	87.08	**87.10**
inferior v.c.	71.77	♯71.15	79.12	**79.69**
kidney l.	92.56	92.78	96.08	**96.21**
kidney r.	94.98	95.39	95.80	**95.97**
pancreas	83.68	85.79	86.09	**87.60**

Table 3-IV. Comparison of coarse-scaled (**C**) and fine-scaled (**F**) segmentation by our step-wise approach and end-to-end approach on our JHMI multi-organ dataset. A fine-scaled accuracy is indicated by ♯ if it is lower than the coarse-scaled one. The *pancreas* segmentation accuracies are higher than those in Table 3-III, due to the increased number of training samples and the higher resolution in CT scans.

with oracle information is decreased by 0.67%).

3.5 JHMI Multi-Organ Segmentation Experiments

To verify that our approach can be applied to other organs, the radiologists in our team collect a large dataset which contains 200 CT scans, 11 abdominal organs and 5 blood vessels. This corpus took 4 full-time radiologists around 3 months to annotate. To the best of our knowledge, this dataset is larger and contains more organs than any public datasets. We choose 5 most challenging targets including the *pancreas* and a blood vessel, as well as two *kidneys* which are relatively easier. Other easy organs such as the *liver* are not included. To the best of our knowledge, some of these organs were never investigated before, but they are important in diagnosing pancreatic diseases and detecting the pancreatic cancer at an early stage. We randomly partition the dataset into 4 folds for cross validation. Each organ is trained and tested individually. When a pixel is predicted as more than one organ, we choose the one with the largest confidence score.

Results of our two approaches are summarized in Table 3-IV. Our end-to-end approach performs generally better than the step-wise approach. It reports a 4.29%

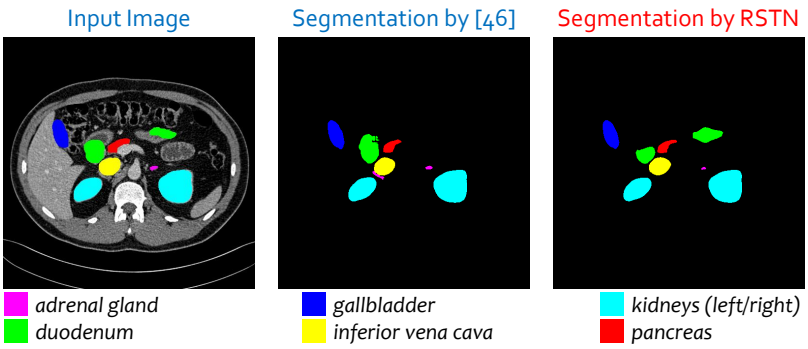


Figure 3-8. Multi-organ segmentation in the *axial* view (best viewed in color). Organs are marked in different colors (input image is shown with the ground-truth annotation).

average improvement over 5 challenging organs (the *kidneys* excluded). An example is displayed in Figure 3-8.

3.6 Summary

This work is motivated by the difficulty of small target segmentation, which is required to focus on a local input region. Two coarse-to-fine approaches are proposed, namely, step-wise coarse-to-fine and end-to-end coarse-to-fine. The step-wise algorithm is formulated as a fixed-point model taking the segmentation mask as both the input and the output, whereas the end-to-end algorithm jointly optimizes over the two stages, and generally achieves better results compared with the step-wise one.

Our approaches are applied to different datasets for pancreas segmentation and multi-organ segmentation, and outperforms the baseline approach as well as previous state-of-the-arts significantly. Confirmed by the radiologists in our team, these segmentation results are helpful to computer-assisted clinical diagnoses.

Chapter 4

Deep Supervision for Pancreatic Cyst Segmentation

The automatic segmentation of an organ and its cystic region is a prerequisite of computer-aided diagnosis. In this chapter, we focus on pancreatic cyst segmentation in an abdominal CT scan. This task is important and very useful in clinical practice yet challenging due to the low contrast in boundary, the variability in location, shape and the different stages of pancreatic cancer. Inspired by the high relevance between the location of a pancreas and its cystic region, we introduce extra deep supervision into the segmentation network, so that cyst segmentation can be improved with the help of relatively easier pancreas segmentation. Under a reasonable transformation function, our approach can be factorized into two stages, and each stage can be efficiently optimized via gradient back-propagation throughout the deep networks. We collect a new dataset with 131 pathological samples, which, to the best of our knowledge, is the largest set for pancreatic cyst segmentation. Without human assistance, our approach consistently outperforms results achieved without deep supervision.

4.1 Introduction

In 2012, pancreatic cancers of all types were the 7th most common cause of cancer deaths, resulting in 330,000 deaths globally. By the time of diagnosis, pancreatic

cancer has often spread to other parts of the body. Therefore, it is very important to use medical imaging analysis to assist identifying malignant cysts in the early stages of pancreatic cancer to increase the survival chance of a patient [133]. The emerge of deep learning has largely advanced the field of computer-aided diagnosis (CAD). With the help of the state-of-the-art deep convolutional neural networks [35, 36], such as the fully-convolutional networks (FCN) [41] for semantic segmentation, researchers have achieved accurate segmentation on many abdominal organs. There are often different frameworks for segmenting different structures [7, 134]. Meanwhile, it is of great interest to find the lesion area in a structure [23, 24, 51], which, frequently, is even more challenging due to the high variability in the shape, texture, size, *etc.*

This chapter focuses on segmenting pancreatic cyst from abdominal CT images. The pancreas is one of the abdominal organs that are very difficult to be segmented even in the healthy cases [7, 8, 13], mainly due to the low contrast in the boundary and the high variability in its geometric properties. In the pathological cases, the difference in the pancreatic cancer stage also impacts both the morphology of the pancreas and the cyst [135, 136]. Despite the importance of pancreatic cyst segmentation, this topic is less studied: some of the existing methods are based on old-fashioned models [137], and a state-of-the-art approach [133] requires a bounding box of the cyst to be annotated beforehand, as well as a lot of interactive operations throughout the segmentation process to annotate some voxels on or off the target. These requirements are often unpractical when the user is not well knowledgeable in medicine. To the best of our knowledge, our method is the first to produce reasonable pancreatic cyst segmentation **without human assistance** on the testing stage.

Intuitively, the pancreatic cyst is often closely related to the pancreas, and thus segmenting the pancreas (relatively easier) may assist the localization and segmentation of the cyst. To this end, we introduce deep supervision [138] into the original segmentation network, leading to a joint objective function taking both the pancreas

and the cyst into consideration. Using a reasonable transformation function, the optimization process can be factorized into two stages, in which we first find the pancreas, and then localize and segment the cyst based on the predicted pancreas mask. Our approach works efficiently based on the coarse-to-fine approaches [13, 14] introduced in Chapter 3 for pancreas segmentation. We perform experiments on a newly collected dataset with 131 pathological samples from CT scan. Without human assistance, our approach consistently outperforms results achieved without deep supervision.

4.2 Approach

4.2.1 Formulation

Let the 3D CT-scanned volume $\mathbf{X}$ annotated with ground-truth pancreas segmentation $\mathbf{P}^\star$ and cyst segmentation $\mathbf{C}^\star$, and both of them are of the same dimensionality as $\mathbf{X}$. $P_i^\star = 1$ and $C_i^\star = 1$ indicate a foreground voxel of pancreas and cyst, respectively. Denote a cyst segmentation model as $\mathbb{M} : \mathbf{C} = \mathbf{f}(\mathbf{X}; \mathbf{\Theta})$, where $\mathbf{\Theta}$ denotes the model parameters. The loss function can be written as $\mathcal{L}(\mathbf{C}, \mathbf{C}^\star)$. In a regular deep neural network such as our baseline, the fully-convolutional network (FCN) [41], we optimize $\mathcal{L}$ with respect to the network weights $\mathbf{\Theta}$ via gradient back-propagation. To deal with small targets, we also follow [78] to compute the DSC loss function: $\mathcal{L}(\mathbf{C}, \mathbf{C}^\star) = \frac{2\times\sum_i C_i C_i^\star}{\sum_i C_i + \sum_i C_i^\star}$. The gradient $\frac{\partial \mathcal{L}(\mathbf{C}, \mathbf{C}^\star)}{\partial \mathbf{C}}$ can be easily computed.

The pancreas is a small organ, and the pancreatic cyst is even smaller. In our newly collected dataset, the fraction of the cyst, relative to the entire volume, is often much smaller than 0.1%. In a very challenging case, the cyst only occupies 0.0015% of the volume, or around 1.5% of the pancreas. This largely increases the difficulty of segmentation or even localization. Figure 4-1 shows a representative example where cyst segmentation fails completely when we take the entire 2D slice as the input.

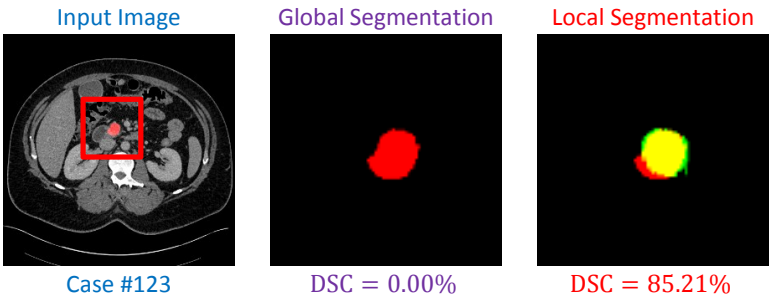


Figure 4-1. A relatively difficult case in pancreatic cyst segmentation and the results produced by different input regions, namely using the entire image and the region around the ground-truth pancreas mask (best viewed in color). The cystic, predicted and overlapping regions are marked by red, green and yellow, respectively. For better visualization, the right two figures are zoomed in *w.r.t.* the red frame.

To deal with this problem, we note that the location of the pancreatic cyst is highly relevant to the pancreas. Denote the set of voxels of the pancreas as $\mathcal{P}^\star = \{i \mid P_i^\star = 1\}$, and similarly, the set of cyst voxels as $\mathcal{C}^\star = \{i \mid C_i^\star = 1\}$. Frequently, a large fraction of $\mathcal{C}^\star$ falls within $\mathcal{P}^\star$ (*e.g.*, $|\mathcal{P}^\star \cap \mathcal{C}^\star| \,/\, |\mathcal{C}^\star| > 95\%$ in 121 out of 131 cases in our dataset). Starting from the pancreas mask increases the chance of accurately segmenting the cyst. Figure 4-1 shows an example of using the ground-truth pancreas mask to recover the failure case of cyst segmentation.

This inspires us to perform cyst segmentation based on the pancreas region, which is relatively easy to detect. To this end, we introduce the pancreas mask $\mathbf{P}$ as an explicit variable of our approach, and append another term to the loss function to jointly optimize both pancreas and cyst segmentation networks. Mathematically, let the pancreas segmentation model be $\mathbb{M}_\text{P} : \mathbf{P} = \mathbf{f}_\text{P}(\mathbf{X}; \mathbf{\Theta}_\text{P})$, and the corresponding loss term be $\mathcal{L}_\text{P}(\mathbf{P}, \mathbf{P}^\star)$. Based on $\mathbf{P}$, we create a smaller input region by applying a transformation $\mathbf{X}' = \sigma[\mathbf{X}, \mathbf{P}]$, and feed $\mathbf{X}'$ to the next stage. Thus, the cyst segmentation model can be written as $\mathbb{M}_\text{C} : \mathbf{C} = \mathbf{f}_\text{C}(\mathbf{X}'; \mathbf{\Theta}_\text{C})$, and we have the corresponding loss them $\mathcal{L}_\text{C}(\mathbf{C}, \mathbf{C}^\star)$. To optimize both $\mathbf{\Theta}_\text{P}$ and $\mathbf{\Theta}_\text{C}$, we consider the

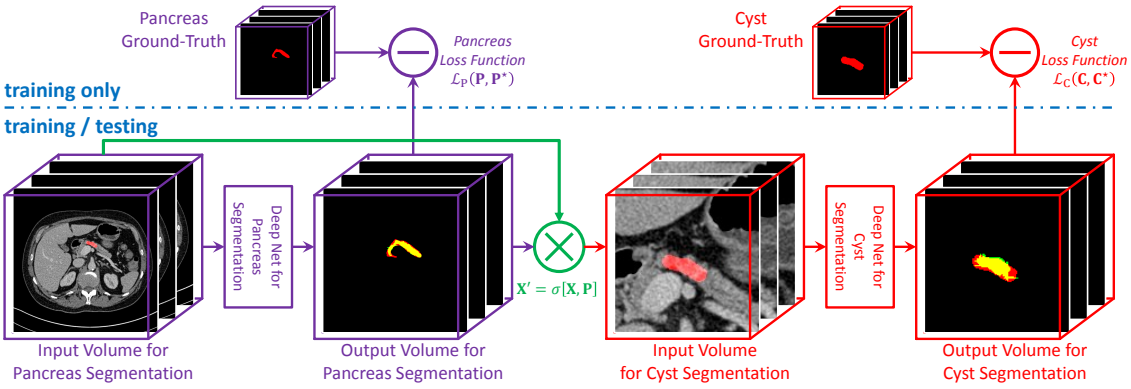


Figure 4-2. The framework of our approach (best viewed in color). Two deep segmentation networks are stacked, and two loss functions are computed. The predicted pancreas mask is used in transforming the input image for cyst segmentation.

following loss function:

$$\mathcal{L}(\mathbf{P}, \mathbf{P}^\star, \mathbf{C}, \mathbf{C}^\star) = \lambda \mathcal{L}_{\mathrm{P}}(\mathbf{P}, \mathbf{P}^\star) + (1 - \lambda)\, \mathcal{L}_{\mathrm{C}}(\mathbf{C}, \mathbf{C}^\star), \tag{4.1}$$

where λ is the balancing parameter defining the weight between either terms.

4.2.2 Optimization

We use gradient descent for optimization, which involves computing the gradients over $\boldsymbol{\Theta}_{\mathrm{P}}$ and $\boldsymbol{\Theta}_{\mathrm{C}}$. Among these, $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Theta}_{\mathrm{C}}} = \frac{\partial \mathcal{L}_{\mathrm{C}}}{\partial \boldsymbol{\Theta}_{\mathrm{C}}}$, and thus we can compute it via standard back-propagation in a deep neural network. On the other hand, $\boldsymbol{\Theta}_{\mathrm{P}}$ is involved in both loss terms, and applying the chain rule yields:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Theta}_{\mathrm{P}}} = \frac{\partial \mathcal{L}_{\mathrm{P}}}{\partial \boldsymbol{\Theta}_{\mathrm{P}}} + \frac{\partial \mathcal{L}_{\mathrm{C}}}{\partial \mathbf{X}'} \cdot \frac{\partial \mathbf{X}'}{\partial \mathbf{P}} \cdot \frac{\partial \mathbf{P}}{\partial \boldsymbol{\Theta}_{\mathrm{P}}}. \tag{4.2}$$

The second term on the right-hand side depends on the definition of $\mathbf{X}' = \sigma[\mathbf{X}, \mathbf{P}]$. In practice, we define a simple transformation to simplify the computation. The intensity value (directly related to the Hounsfield units in CT scan) of each voxel is either preserved or set as 0, and the criterion is whether there exists a nearby voxel which is likely to fall within the pancreas region:

$$X'_i = X_i \times \mathbb{I}\{\exists j \mid P_j > 0.5 \wedge |i - j| < t\}, \tag{4.3}$$

where t is the threshold which is the farthest distance from a cyst voxel to the pancreas volume. We set $t = 15$ in practice, and our approach is not sensitive to this parameter. With this formulation, *i.e.*, $\frac{\partial X'_i}{\partial P_j} = 0$ almost everywhere. Thus, we have $\frac{\partial \mathbf{X}'}{\partial \mathbf{P}} = \mathbf{0}$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{\Theta}_{\mathrm{P}}} = \frac{\partial \mathcal{L}_{\mathrm{P}}}{\partial \mathbf{\Theta}_{\mathrm{P}}}$. This allows us to factorize the optimization into two stages in both training and testing. Since $\frac{\partial \mathcal{L}}{\partial \mathbf{\Theta}_{\mathrm{P}}}$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{\Theta}_{\mathrm{C}}}$ are individually optimized, the balancing parameter λ in Eqn. (4.1) can be ignored. The overall framework is illustrated in Figure 4-2. In training, we directly set $\mathbf{X}' = \sigma[\mathbf{X}, \mathbf{P}^{\star}]$, so that the cyst segmentation model $\mathbb{M}_{\mathrm{C}}$ receives more reliable supervision. In testing, starting from $\mathbf{X}$, we compute $\mathbf{P}$, $\mathbf{X}'$ and $\mathbf{C}$ orderly. Dealing with two stages individually reduces the computational overheads. It is also possible to formulate the second stage as multi-label segmentation.

4.3 Experiments

4.3.1 Dataset and Evaluation

We evaluate our approach on a cyst dataset collected by the radiologists in our team. This dataset contains 131 contrast-enhanced abdominal CT volumes, and each of them is manually labeled with both pancreas and pancreatic cyst masks. The resolution of each CT scan is $512 \times 512 \times L$, where $L \in [358, 1121]$ is the number of sampling slices along the long axis of the body. The slice thickness varies from 0.5mm–1.0mm. We split the dataset into 4 fixed folds, and each of them contains approximately the same number of samples. We apply cross validation, *i.e.*, training our approach on 3 out of 4 folds and testing it on the remaining one. The same as before, we measure the segmentation accuracy by computing the Dice-Sørensen Coefficient (DSC) for each 3D volume. We report the average DSC score together with other statistics over all 131 testing cases from 4 testing folds.

4.3.2 Implementation Details

We follow Section 4.2 to use a multi-stage approach, which first finds the regular organ (pancreas), and then locates the neoplasm (cyst) by referring to that organ. Based on the coarse-to-fine approaches [13, 14] introduced in Chapter 3 for pancreas segmentation, we hereby introduce the following two different implementations for pancreatic cyst segmentation. First, based on the step-wise coarse-to-fine framework, we adopt a four-stage strategy, *i.e.*, coarse-scaled and fine-scaled pancreas segmentation, as well as coarse-scaled and fine-scaled cyst segmentation are performed sequentially in a step-wise manner. Second, this can be also implemented by two RSTN (*i.e.*, end-to-end coarse-to-fine) modules, where the first RSTN segments the pancreas given the CT images while the second segments the pancreatic cyst given the pancreas-cropped region as the input.

4.3.3 Results and Discussion

We report both pancreas and cyst segmentation results in Table 4-I, where we summarize the results of pancreas segmentation, pancreatic cyst segmentation without pancreas supervision (*i.e.*, two-stage coarse-to-fine approach, w/o deep supervision), and pancreatic cyst segmentation with pancreas supervision (*i.e.*, four-stage strategy, w/ deep supervision). It is interesting to see that without deep supervision, our two approaches perform comparably with each other, but with deep supervision, the end-to-end approach works better than the step-wise one. This is because a much better pancreas segmentation result (*i.e.*, 83.81% compared with 79.32%) provides more accurate contextual information for cyst segmentation. In addition, our approaches yield even better results by adopting a stronger backbone, *e.g.*, under the setting of Step-Wise, w/ Deep Supervision, when we employ DeepLab [42] as the backbone network in the coarse-stage for pancreas segmentation, we can even achieve $69.38 \pm 27.60\%$ in DSC for cyst segmentation.

Three representative cases are shown in Figure 4-3. In the first case, both the pancreas and the cyst can be segmented accurately from the original CT scan. In the second case, however, the cyst is small in volume and less discriminative in contrast, and thus an accurate segmentation is only possible when we roughly localize the pancreas and shrink the input image size accordingly. The accuracy gain of our approach mainly owes to the accuracy gain of this type of cases. The third case shows a failure example of our approach, in which an inaccurate pancreas segmentation leads to a complete missing in cyst detection. Note that the baseline approach reports a 59.93% DSC in this case, and, if the oracle pancreas bounding box is provided, we can still achieve a DSC of 77.56%. This inspires us that cyst segmentation can sometimes help pancreas segmentation, and this topic is left for future research.

To the best of our knowledge, pancreatic cyst segmentation has been little studied previously. A competitor is [33] published in 2016, which combines random walk and region growth for segmentation. However, it requires the user to annotate the region-of-interest (ROI) beforehand, and provide interactive annotations on foreground/background voxels throughout the segmentation process. In comparison, our approaches can be widely applied to automatic diagnosis, especially for the common users without professional knowledge in medicine. Our studies on the pancreas and pancreatic cyst segmentation are summarized in [15].

4.4 Summary

This chapter presents the first system for pancreatic cyst segmentation which can work without human assistance on the testing stage. Motivated by the high relevance of a cystic pancreas and a pancreatic cyst, we formulate pancreas segmentation as an explicit variable in the formulation, and introduce deep supervision to assist the network training process. The joint optimization can be factorized into two stages, making our approach very easy to implement. We collect a dataset with 131

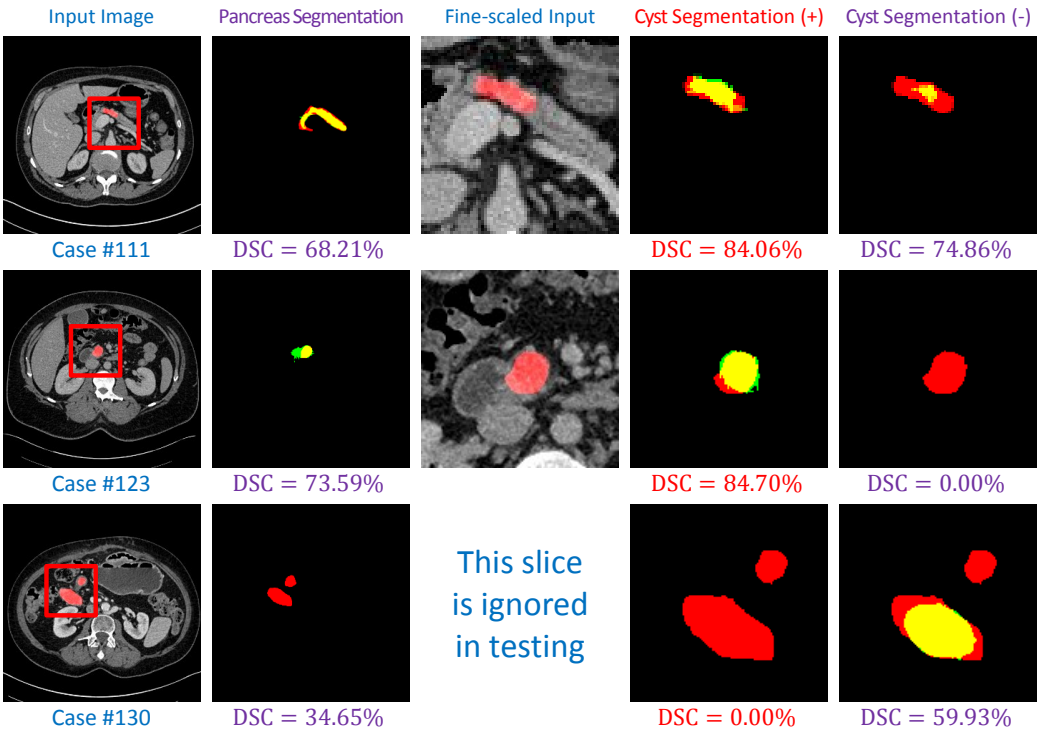


Figure 4-3. Sample pancreas and pancreatic cyst segmentation results (best viewed in color). From left to right: input image (in which pancreas and cyst are marked in red and green, respectively), pancreas segmentation result, and cyst segmentation results when we apply deep supervision (denoted by +) or not (-). The figures in the right three columns are zoomed in *w.r.t.* the red frames. In the last example, pancreas segmentation fails in this slice, resulting in a complete failure in cyst segmentation.

Target	Method	Average	Max	Min
pancreas	step-wise	79.23 ± 9.72	93.82	**69.54**
pancreas	end-to-end	$\mathbf{83.81} \pm 10.51$	**94.34**	20.77
cyst	step-wise, w/o Deep Supervision	60.46 ± 31.37	95.67	0.00
cyst	end-to-end, w/o Deep supervision	$\mathbf{60.73} \pm 32.46$	**96.50**	0.00
cyst	step-wise, w/ Deep Supervision	63.44 ± 27.71	95.55	0.00
cyst	end-to-end, w/ Deep Supervision	$\mathbf{67.19} \pm 27.91$	**96.05**	0.00

Table 4-I. Accuracy (DSC, %) comparison on different targets (*pancreas* or *cyst*) and different approaches. For *cyst* segmentation, w/o Deep Supervision means directly apply our coarse-to-fine approaches on cyst segmentation, given the whole CT image, while w/ Deep Supervision means segmenting the *pancreas* first, and then segmenting the *cyst* in the input image cropped by the *pancreas* region.

pathological cases. Based on the coarse-to-fine segmentation algorithms, our approach produces reasonable cyst segmentation results. It is worth emphasizing that our

approach does not require any extra human annotations on the testing stage, which is especially practical in assisting common patients in cheap and periodic clinical applications.

This work teaches us that a lesion can be detected more effectively by considering its highly related organ(s). This knowledge, being simple and straightforward, is useful in future work, especially for the pathological organ or lesion segmentation.

Chapter 5

Abdominal Multi-Organ Segmentation with Organ-Attention Networks and Statistical Fusion

Accurate and robust segmentation of abdominal organs on CT is essential for many clinical applications such as computer-aided diagnosis and computer-aided surgery. But this task is challenging due to the weak boundaries of organs, the complexity of the background, and the variable sizes of different organs. To address these challenges, we introduce a novel framework for multi-organ segmentation of abdominal regions by using organ-attention networks with reverse connections (OAN-RCs) which are applied to 2D views, of the 3D CT volume, and output estimates which are combined by statistical fusion exploiting structural similarity. More specifically, OAN is a two-stage deep convolutional network, where deep network features from the first stage are combined with the original image, in a second stage, to reduce the complex background and enhance the discriminative information for the target organs. Intuitively, OAN reduces the effect of the complex background by focusing attention so that each organ only needs to be discriminated from its local background. RCs are added to the first stage to give the lower layers more semantic information thereby enabling them to adapt to the sizes of different organs. Our networks are trained on 2D views (slices)

enabling us to use holistic information and allowing efficient computation (compared to using 3D patches). To compensate for the limited cross-sectional information of the original 3D volumetric CT, *e.g.*, the connectivity between neighbor slices, multi-sectional images are reconstructed from the three different 2D view directions. Then we combine the segmentation results from the different views using statistical fusion, with a novel term relating the structural similarity of the 2D views to the original 3D structure. To train the network and evaluate results, 13 structures were manually annotated by four human raters and confirmed by a senior expert on 236 normal cases. We tested our algorithm by 4-fold cross-validation and computed Dice-Sørensen similarity coefficients (DSC) and surface distances for evaluating our estimates of the 13 structures. Our experiments show that the proposed approach gives strong results and outperforms 2D- and 3D-patch based state-of-the-art methods in terms of DSC and mean surface distances.

5.1 Introduction

Segmentation of the internal structures, like body organs, in medical images is an essential task for many clinical applications such as computer-aided diagnosis (CAD), computer-aided surgery (CAS) and radiation therapy (RT). However, despite intensive studies of automatic or semi-automatic segmentation methods, there remain challenges which need to be overcome before these methods can be applied to clinical environments. In particular, detailed abdominal organ segmentation on CT is a challenging task both for manual human annotation and for automatic segmentation algorithms for various reasons including the morphological complexity of the structures, the large variations between inter- and intra-subjects, and image characteristics such as low contrast of soft tissues.

Early studies of abdominal organ segmentation focused on specific single organs, for example relatively large isolated structures such as the liver [26, 139, 140] or critical

structures such as blood vessels [141, 142]. However, most of the algorithms were based on specific features of the target organ, and so extensibility to the simultaneous segmentation of multiple organs was limited. For multi-organ segmentation, atlas-based approaches were adopted for many applications [58–64]. The general framework of atlas-based segmentation is to deformably register selected atlas images with segmented structures to the target image. Critical issues for this approach, which affect performance accuracy, include proper atlas selection, accurate deformable image registration, and label fusion. In particular, for the abdominal region, inter-subject variations are relatively large compared with other parts of the body (*e.g.*, the brain) so the segmentation results are dependent on deformable registration between inter-subjects from the limited set of atlases, which is a challenging problem that critically affects the final accuracies. In addition, the computational time is strongly dependent on the number of atlases. Therefore, the selection of the proper number and types of atlases is a critical factor for both accuracy and efficiency.

Recently, learning-based approaches exploiting large datasets have been applied to the segmentation of medical images [6, 24, 48, 78, 102, 120, 143–146]. In particular, deep convolutional neural networks (CNN) have been very successful [6, 8, 24, 48, 78, 102, 120, 143, 145]. Targets include regions in the brain [6, 24, 102], chest [145], and abdomen [7, 8, 48]. The performance results of CNNs for organs (and even tumors) reach, or outperform, alternative state-of-the-art methods. Unlike multi-atlas-based approaches, deep networks do not require selecting a specific atlas or require deformable registration from training sets to a target image. In this study, we apply deep network approaches to abdominal organ segmentation.

Most studies based on deep networks, however, focused on a single structure segmentation, particularly for abdominal regions, and there are few studies of multi-organ segmentation partly due to technical challenges discussed later. We note that fully convolutional networks (FCNs) [41] have been generally accepted for organ

segmentation on CT scans [8, 13, 143] partly because they give state-of-the-art performance for semantic segmentation of natural images [41, 102]. But there are three major characteristics of abdominal CT which we must address in order to obtain strong performance on multi-organ segmentation.

Firstly, many abdominal organs have weak boundaries between spatially adjacent structures on CT, *e.g.*, between the head of the pancreas and the duodenum. In addition, the entire CT volume includes a large variety of different complex structures. Morphological and topological complexity includes anatomically connected structures such as the gastrointestinal (GI) track (stomach, duodenum, small bowel and colon) and vascular structures. The correct anatomical borders between connected structures may not be always visible in CT, especially in sectional images (*i.e.*, 2D slices), and may be indicated only by subtle texture and shape change, which causes uncertainty even for human experts. This makes it hard for deep networks to distinguish the target organs from the complex background.

Secondly, there are large variations in the relative sizes of different target organs, *e.g.*, the liver compared to the gallbladder. This causes problems when applying deep networks to multi-organ segmentation because lower layers typically lack semantic information when segmenting small structures. The same problem has been observed in semantic segmentation of natural images where the segmentation performance on small regions is typically much worse than on large regions, motivating the need to introduce mechanisms which attend to the scale [147].

Thirdly, although CT scans are high-resolution three-dimensional volumes, most current deep network methods were designed for 2D images. To overcome the limitations of using 2D CNNs for 3D images, [145] used multiple 2D patches reconstructed from 9 different directions around the target region for the task of pulmonary nodule detection. [63] used 2D axial, coronal, and sagittal slices for pancreas detection at the coarse level and also for segmentation at the finer level. More recently, there are

studies which use 3D deep networks [6, 66, 78, 143]. These, however, are not networks that act on the entire 3D CT volume but instead are local patch-based approaches (due to complex challenges of 3D deep networks discussed later in this paragraph). To address the problems caused by restricting to image patches, [75] and [6] used a hierarchical approach with multi-resolutions, which reduces the dimension of the whole volume for initial detection and focuses on smaller regions at the finer resolution. But this strategy is best suited to a single target structure. [66] applied a bigger patch size to deal with the whole dense pancreatic volume, but this was also for single pancreas segmentation and hard to extend to the whole abdominal region. In general, 3D deep networks face far greater complex challenges than 2D deep networks. Both approaches rely heavily on graphics processing units (GPUs) but these GPUs have limited memory size which makes it difficult when dealing with full 3D CT volumes compared to 2D CT slices (which require much less memory). In addition, 3D deep networks typically require many more parameters than 2D deep networks and hence require much more training data, unless they are restricted to patches. But there is limited training data for abdominal CT images, because annotating them is challenging and requires expert human radiologists, which makes it particularly difficult to apply 3D deep networks to abdominal multi-organ segmentation. We have, however, implemented a 3D patch based approach for comparison.

To deal with the technical difficulties for abdominal multi-organ segmentation on CT, we introduce a novel framework of an organ-attention 2D deep networks with reverse connections (OAN-RC) followed by statistical fusion to combine the information from the three different views exploiting structural similarity using local isotropic 3D patches. OAN is a two-stage deep network, which computes an organ-attention map (OAM) from typical probability map of labels for input images in the first stage and combines OAM to the original input image for the second stage. This two-stage strategy effectively reduces the complexity of the background while enhancing the

discriminative information of target structures (by concentrating attention close to the target structures). By training OAM with additional deep network, uncertainties and errors from the first stage are adjusted and the fidelity of the final probability map is improved. In this procedure, we apply reverse connections [148] to the first stage so that we can localize organ information at different scales by assisting the lower layers with semantic information.

More specifically, we apply OAN-RC to each sectional slice, which is an extreme form of anisotropic local patches but include the whole semantic (*i.e.*, volume) information from one viewing direction. This yields segmentation information from separate sets of multi-sectional images (axial, coronal, and sagittal planes in this study similarly to most of medical image platforms for 2D visualization). We statistically fuse the three sources of information using local isotropic 3D patches based on direction-dependent local structural similarity. The basic fusion framework uses expectation-maximization (EM) similar to [59, 149]. But, unlike typical statistical fusion methods used for atlas-based segmentation, the input volumes and the target volumes for segmentation in our problem are the same. But different structures and texture patterns, from different viewing directions, will often generate nonidentical segmentations in 3D. Our strategy is to exploit structural similarity by computing a direction-dependent local property at each voxel. This models the structural similarity from the 2D images to the original 3D structure (in the 3D volume) by local weights. This structural statistical fusion improves our overall performance by combining the information from the three different views in a principled manner and also imposing local structure.

Figure 5-1 describes the graphical concept of our framework. Our proposed algorithm was tested on 236 abdominal CT scans of normal cases collected as a part of FELIX project for pancreatic cancer research [150]. By experiments, our method showed robust and high fidelities to the ground-truth for all target structures with

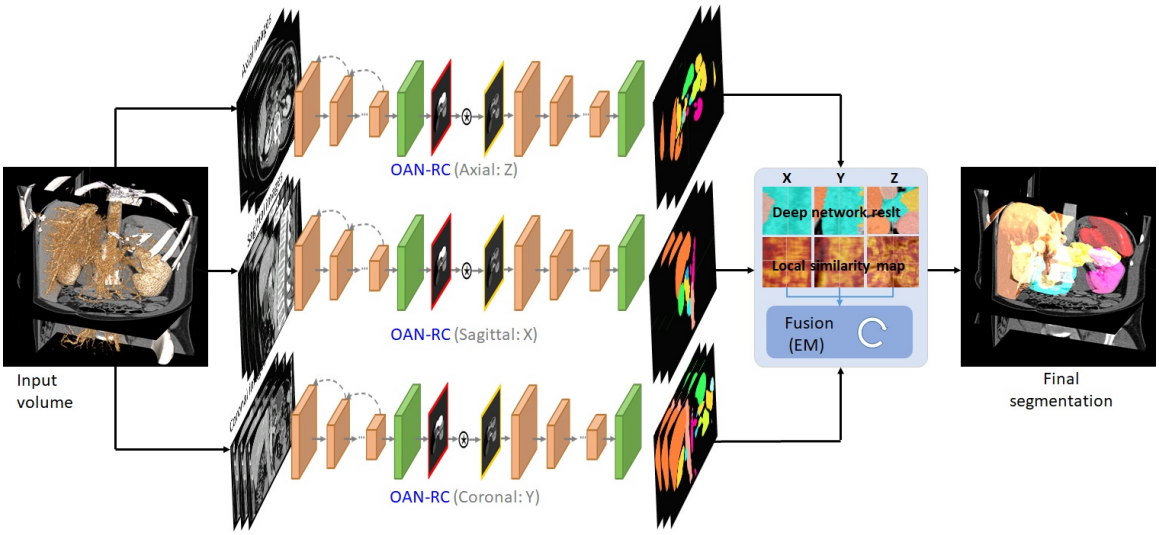


Figure 5-1. The overall framework for multi-organ segmentation.

smooth boundaries. It outperformed 3D patch-based algorithms as well as 2D-based in terms of DICE-similarity coefficient and average surface distance with memory and computational efficiency.

5.2 Organ-Attention Networks with Reverse Connections

Given a 3D volume of interest (VOI) of a scanned CT image $V \subset \mathbb{R}^3$, our goal is to find the label of each voxel $v \in V$. The target structures (*i.e.*, the labeled structures) are restricted to be organs which do not overlap with each other, so every voxel v should be assigned to a label in a finite set $\mathcal{L}$. In this section we introduce our proposed organ-attention networks with Reverse connections (annotated as OAN-RC) which is run separately on three different views, and then in the next section we describe our novel structural similarity statistical fusion method which combines the segmentation results obtained from the OAN-RCs on the three different views.

5.2.1 Two-stage Organ Attention Network

We first introduce the OAN, which is composed of two jointly optimized stages. The first stage (stage-I) transforms the organ segmentation probability map to provide spatial attention to the second stage (stage-II), so that the segmentation network trained in stage-II is more discriminative for segmenting organs (because it only has to deal with local context). To assist the lower layers in stage-I with more semantic information, we employ reverse connections (Section 5.2.2), which pass semantic information down from high layers to low layers. The OAN is trained in an end-to-end fashion to enhance the learning ability of all stages.

The input images to our OAN are reconstructed 2D slices from axial, sagittal and coronal directions. Based on the normal vector directions of the sagittal (X), coronal (Y) and axial (Z) planes, we denote the 2D images by $\mathbf{I}_i^X$, $\mathbf{I}_j^Y$ and $\mathbf{I}_k^Z$ respectively, where $i = 1, \ldots, n_x,\ j = 1, \ldots, n_y,\ k = 1, \ldots, n_z$ and $n_x,\ n_y,\ n_z$ are the numbers of slices for the three directions, respectively, and $\bigcup_i \mathbf{I}_i^X = \bigcup_j \mathbf{I}_j^Y = \bigcup_k \mathbf{I}_k^Z = V$. Following the work of [13], we train an individual OAN for each direction.

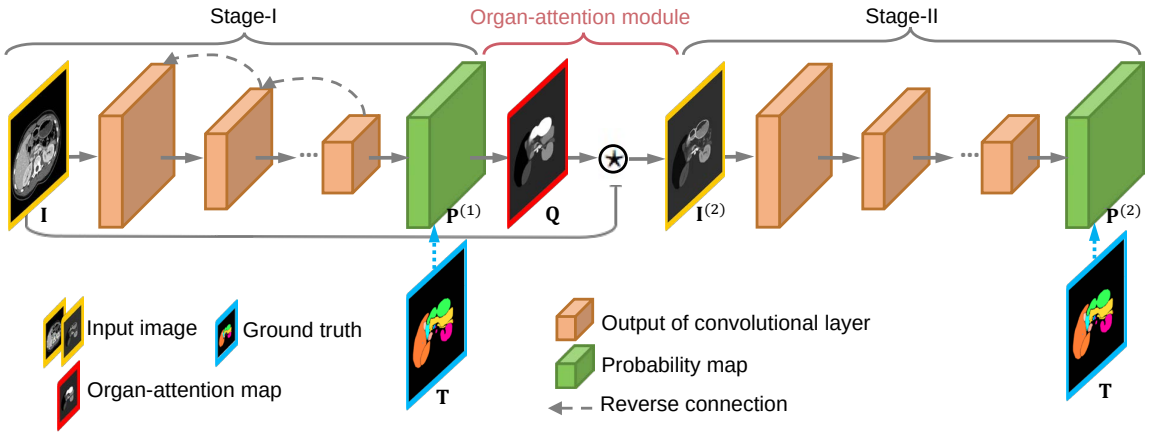


Figure 5-2. The architecture of our two-stage organ-attention network with reverse connections. The organ-attention network (OAN) is composed of two jointly optimized stages, where the first stage (stage-I) transforms the organ segmentation probability map by spatial attention to the second stage (stage-II). Hence the organ segmentation map generated in the organ-attention module guides the latter computation. The reverse connections, described in Section 5.2.2, modify the first stage of OAN as shown by dashed lines.

Figure 5-2 illustrates our organ-attention-network architecture. The network consists of two stages, where each stage is a segmentation network. For notational simplicity, we denote an input 2D slice by $\mathbf{I} \subset \mathbb{R}^{H \times W}$ and its corresponding label map by $\mathbf{T} = \{t_i\}_{i=1,...,H \times W}$. Stage-I outputs a probability map $\mathbf{P}^{(1)} = f(\mathbf{I}; \mathbf{\Theta}^{(1)}) \subset \mathbb{R}^{H \times W \times |\mathcal{L}|}$ for each label at every pixel, where the probability density function $f(\cdot; \mathbf{\Theta}^{(1)})$ is a segmentation network parameterized by $\mathbf{\Theta}^{(1)}$. We use FCN [41] with reverse connections, which is explained in Section 5.2.2, as $\mathbf{\Theta}^{(1)}$. FCN is the backbone network throughout the chapter. Each element $p_{i,l}^{(1)} \in \mathbf{P}^{(1)}$ is the probability that the i-th pixel in the input slice belongs to label l, where $l = 0$ is the background, and $l = 1, ..., |\mathcal{L}|$ are target organs. We define $p_{i,l}^{(1)} = \sigma(a_{i,l}^{(1)}) = \frac{\exp(a_{i,l}^{(1)})}{\sum_{t=0}^{|\mathcal{L}|} \exp(a_{i,t}^{(1)})}$, where $a_{i,l}^{(1)}$ is the activation value of the i-th pixel on the l-th channel dimension. Let $\mathbf{A}^{(1)} = \{a_{i,l}^{(1)}\}_{i=1,...,H \times W, l=0,...,|\mathcal{L}|}$ be the activation map. The objective function to minimize for $\mathbf{\Theta}^{(1)}$ is given by

$$\mathcal{J}^{(1)}(\mathbf{\Theta}^{(1)}) = -\frac{1}{H \times W} \left[\sum_{i=1}^{H \times W} \sum_{j=0}^{|\mathcal{L}|} \mathbf{1}\left(t_i = l\right) \log p_{i,l}^{(1)} \right], \tag{5.1}$$

where $\mathbf{1}(\cdot)$ is an indicator function.

Using a preliminary organ segmentation map to guide the computation of a better organ segmentation can be thought as employing an attentional mechanism. Towards this end, we propose an organ-attention module by

$$\mathbf{Q} = \mathbf{W} * \mathbf{P}^{(1)} + \mathbf{b}, \tag{5.2}$$

where $*$ denotes the convolution operator, $\mathbf{W}$ indicates the convolutional filters, and $\mathbf{b}$ is the bias. Eqn. (5.2) embeds cross-organ information into a single organ-attention map, $\mathbf{Q}$, which learns discriminative spatial attention for different organs automatically. By combining $\mathbf{Q}$ with the original input $\mathbf{I}$, we get an image which emphasizes each organ by

$$\mathbf{I}^{(2)} = \mathbf{I} \star \mathbf{Q}, \tag{5.3}$$

where $\star$ is the element-wise product operator. We apply $\mathbf{I}^{(2)}$ to the input of stage-II, and the probability of stage-II then becomes $\mathbf{P}^{(2)} = f(\mathbf{I}^{(2)}; \mathbf{\Theta}^{(2)})$.

In order to drive stage-II to focus on organ regions without needing to deal with complicated non-local background, we define a selection function, $\mathbf{1}(\mathbf{P}_0^{(1)} \leqslant \rho)$ where $\mathbf{P}_0^{(1)} = \{p_{i,0}^{(1)}\}_{i=1,\ldots,H\times W}$ is the probability map provided by stage-I. In stage-II, we only accept the region if $p_{i,0}^{(1)} > \rho$ and do not back-propagate it to stage-I. The loss function for stage-II is formulated as

$$\mathcal{J}^{(2)}(\mathbf{\Theta}^{(2)}, \mathbf{W}, \mathbf{b}) = -\frac{1}{H \times W}\left[\sum_{i=1}^{H\times W}\sum_{j=0}^{|\mathcal{L}|}\mathbf{1}\left(p_{i,0}^{(1)} \leqslant \rho\right) \cdot \mathbf{1}\left(t_i = l\right)\log p_{i,l}^{(2)}\right]. \tag{5.4}$$

To jointly optimize stage-I and stage-II, we define a loss function aiming at estimating parameters $\mathbf{\Theta}^{(1)}$, $\mathbf{\Theta}^{(2)}$, $\mathbf{W}$, and $\mathbf{b}$ by optimizing

$$\mathcal{J} = h^{(1)}\mathcal{J}^{(1)}(\mathbf{\Theta}^{(1)}) + h^{(2)}\mathcal{J}^{(2)}(\mathbf{\Theta}^{(2)}, \mathbf{W}, \mathbf{b}), \tag{5.5}$$

where $h^{(1)}$ and $h^{(2)}$ are the fusion weights.

5.2.2 Reverse Connections

FCNs [41] have shown good segmentation results in recent studies, especially for single organ segmentation. However, for multi-organ segmentation, lower layers typically lack semantic information, which may lead to inaccurate segmentation particularly for smaller structures. Therefore, we propose reverse connections which feed coarse-scale (high) layer information backward to fine-scale (low) layer for semantic segmentation of multi-scale structures, inspired by [148]. This enables us to connect abstract high-level semantic information to the more detailed lower layers so that all the target organs have similar levels of details and abstract information at the same layer. The reverse connections framework for stage-I is shown in Figure 5-3. Figure 5-4 illustrates a reverse connection block. Let $\mathbf{R}_n$ denote the reverse connection map of the n-th convolutional layer in the backbone network, *i.e.*, FCN in this study, where $\mathbf{C}_n$ is the output of the n-th convolutional layer. A convolutional layer (with 512 channels by 3×3 kernels) is added after $\mathbf{C}_n$, and a deconvolutional layer (with 512 channels by 4×4 kernels) is applied after $\mathbf{R}_{n+1}$. $\mathbf{R}_n$ is then obtained via an element-wise

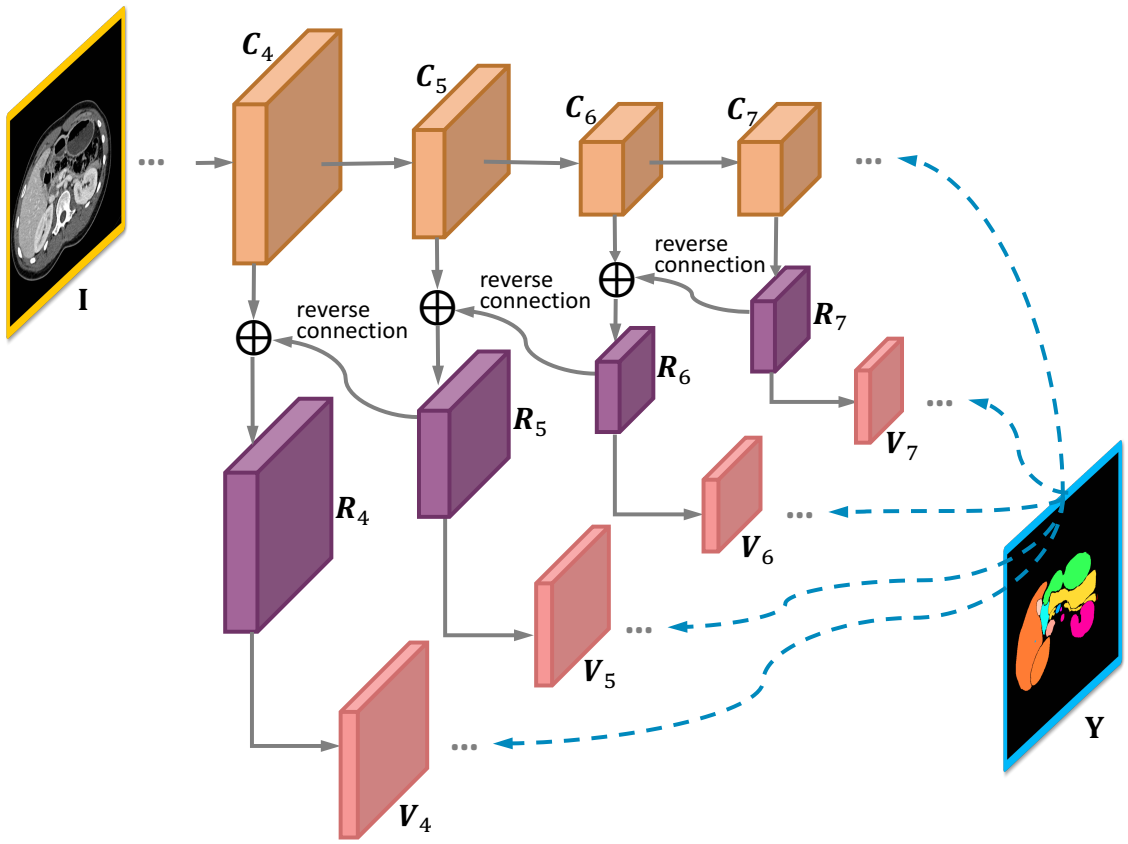


Figure 5-3. The reverse connections architecture of OAN stage-I. The network has reverse connections to the output of convolutional layers. In the training step, both backbone network and reverse connection side-outputs are supervised by the ground-truth. Finally, all reverse connection side-outputs and the output of backbone network are fused and made to approach ground-truth.

summation of these two maps. $\mathbf{R}_7$ is the output of a convolutional layer (with 512 channels by 2×2 kernels) grafted onto $\mathbf{C}_7$. Let $\mathbf{w}^n$ denote the corresponding weights for obtaining $\mathbf{R}_n$. Following [148], we add reverse connections from $\mathbf{C}_4$ to $\mathbf{C}_7$.

With these learnable reverse connections, the semantic information of the lower layers can be enriched. In order to drive learned reverse connection maps to produce segmentation results approaching the ground-truth, we make each reverse connection map associate with a classifier. As the side-output layers proposed in [148] are designed for detection purposes, they are not suitable for our task. Instead we follow the side-outputs used in [151]. More specifically, a convolutional layer (with $|\mathcal{L}|$ channels by 1×1 kernels) is added on top of $\mathbf{R}_n$, whose output is denoted as $\mathbf{V}_n$, and followed

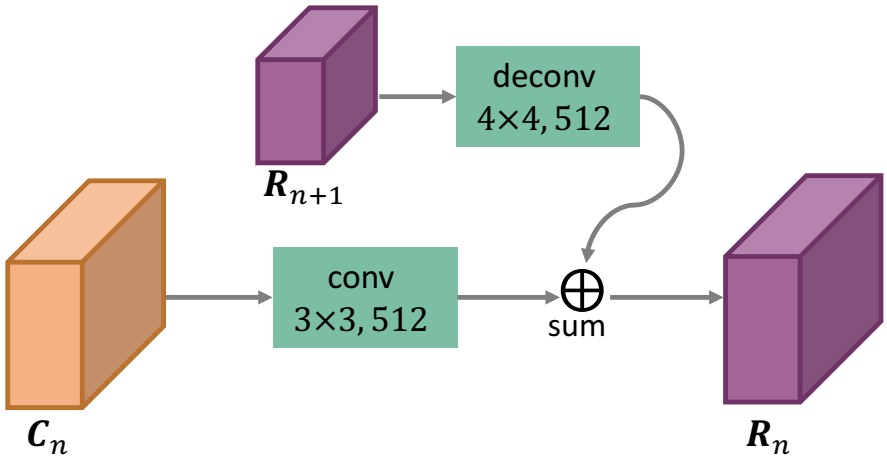


Figure 5-4. A reverse connection block.

by a deconvolutional layer (with $|\mathcal{L}|$ channels). We denote the weights of the n-th side-output layer by $\boldsymbol{\theta}^n$. The loss function for side-output layers $\mathcal{J}^{(s,1)}$ is defined as

$$\mathcal{J}^{(s,1)}(\mathbf{\Theta}^{(1)}, \mathbf{w}, \boldsymbol{\theta}) = \sum_{n=4}^{7} h_n^{(s,1)} \ell_n^{(s,1)} \left(\mathbf{\Theta}^{(1)}, \mathbf{w}^n, \boldsymbol{\theta}^n\right), \tag{5.6}$$

where $\ell_n^{(s,1)} = -\frac{1}{H\times W}\left[\sum_{i=1}^{H\times W}\sum_{j=0}^{|\mathcal{L}|} \mathbf{1}\left(t_i = l\right) \log p_{i,l}^{(s,1)}\right]$ and $p_{i,l}^{(s,1)}$ is the probability output of the n-th side-output layer.

In order to combine the learned reverse connection maps of fine layers and coarse layers, we add up the predictions (*i.e.*, $\mathbf{V}_n$) of the reverse connection maps from high layer to low layer gradually. First, $\mathbf{V}_6$ is fused with a $2\times$ upsampling of $\mathbf{V}_7$ by an element-wisely addition. Then we follow the same strategy and gradually merge $\mathbf{V}_5$ and $\mathbf{V}_4$, as shown in Figure 5-5. To obtain a fused activation map $\mathbf{A}^{(f,1)} = \{a_{i,l}^{(f,1)}\}_{i=1,\ldots,H\times W, l=0,\ldots,|\mathcal{L}|}$ from the activation map of both side-outputs (*i.e.*, $\mathbf{A}^{(r,1)}$) and convolutional layers in the backbone network (*i.e.*, $\mathbf{A}^{(b,1)}$), a scale function is adopted followed by an element-wise addition by

$$\mathbf{A}_l^{(f,1)} = h_l^{(r,1)} \mathbf{A}_l^{(r,1)} + h_l^{(b,1)} \mathbf{A}_l^{(b,1)}, \qquad l = 0, \ldots, |\mathcal{L}| \tag{5.7}$$

where $\mathbf{A}_l$ indicates the l-th channel of the activation map. $h_l^{(r,1)}$ and $h_l^{(b,1)}$ are fusion weights. Then the fused probability map, $\mathbf{P}^{(f,1)} = \{p_{i,l}^{(f,1)}\}_{i=1,\ldots,H\times W, l=0,\ldots,|\mathcal{L}|}$, can be

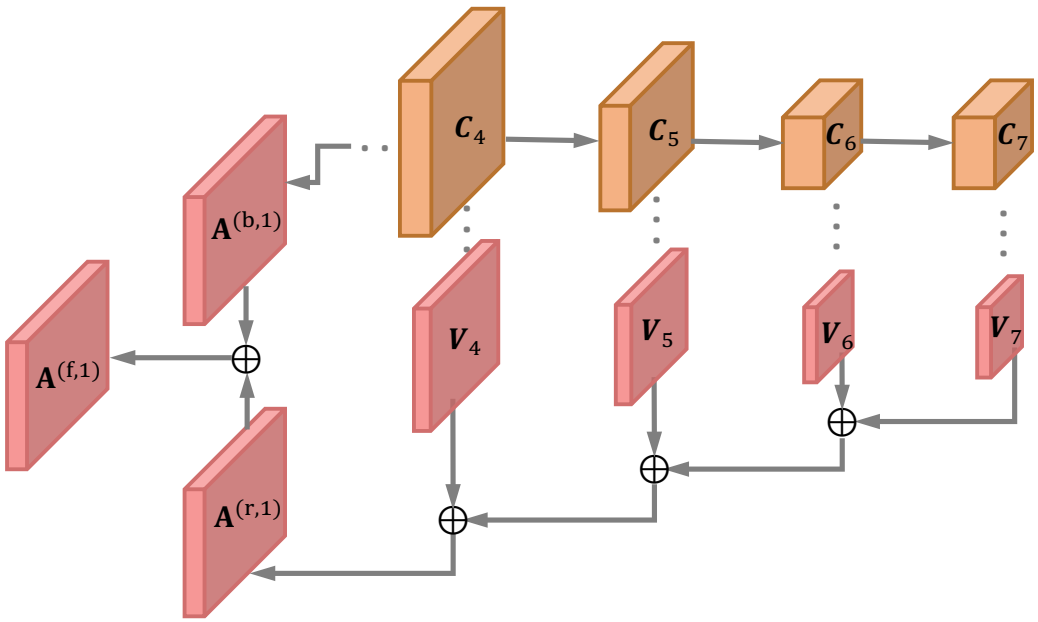


Figure 5-5. Feature fusion strategy. A deep-to-shallow refinement is adopted for multi-scale side-output features. The final activation map ($\mathbf{A}^{(f,1)}$) for stage-I is an element-wise addition of the side-output activation map ($\mathbf{A}^{(r,1)}$) and the backbone network activation map ($\mathbf{A}^{(b,1)}$).

obtained by $p_{i,l}^{(f,1)} = \sigma(a_{i,l}^{(f,1)})$. The final objective function for stage-I is defined by

$$\begin{aligned} &\mathcal{J}^{(1)}(\mathbf{\Theta}^{(1)}, \mathbf{w}, \boldsymbol{\theta}) = h^{(b,1)}\mathcal{J}^{(b,1)}(\mathbf{\Theta}^{(1)}) \\ &+ h^{(s,1)}\mathcal{J}^{(s,1)}(\mathbf{\Theta}^{(1)}, \mathbf{w}, \boldsymbol{\theta}) + h^{(f,1)}\mathcal{J}^{(f,1)}(\mathbf{\Theta}^{(1)}, \mathbf{w}, \boldsymbol{\theta}), \end{aligned} \tag{5.8}$$

where $h^{(b,1)}$, $h^{(s,1)}$ and $h^{(f,1)}$ are fusion weights, and

$$\mathcal{J}^{(f,1)}(\mathbf{\Theta}^{(1)}, \mathbf{w}, \boldsymbol{\theta}) = -\frac{1}{H \times W}\left[\sum_{i=1}^{H\times W}\sum_{j=0}^{|\mathcal{L}|}\mathbf{1}\left(t_i = l\right)\log p_{i,l}^{(f,1)}\right]. \tag{5.9}$$

Note that in our full system with the two-stage organ-attention network and reverse connections, all the parameters are optimized simultaneously by standard back-propagation

$$\begin{aligned} &(\hat{\mathbf{\Theta}}^{(1)}, \hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{\Theta}}^{(2)}, \hat{\mathbf{W}}, \hat{\mathbf{b}}) \\ &= \arg\min\{\mathcal{J}^{(1)}(\mathbf{\Theta}^{(1)}, \mathbf{w}, \boldsymbol{\theta}) + h^{(2)}\mathcal{J}^{(2)}(\mathbf{\Theta}^{(2)}, \mathbf{W}, \mathbf{b})\}. \end{aligned} \tag{5.10}$$

5.2.3 Testing Phase

In the testing stage, given a slice $\mathbf{I}$, we obtain the stage-I and stage-II probability map by

$$\begin{aligned} \mathbf{P}^{(1)} &= f(\mathbf{I}; \hat{\boldsymbol{\Theta}}^{(1)}, \hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}) \\ \mathbf{P}^{(2)} &= f(\mathbf{I}; \hat{\boldsymbol{\Theta}}^{(2)}, \hat{\mathbf{W}}, \hat{\mathbf{b}}), \end{aligned} \tag{5.11}$$

where $f(\cdot, \cdot)$ is the network functions defined in Section 5.2.1. A fused probability map of $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ is then given by

$$\mathbf{P} = \mathbf{P}^{(1)} \circ \mathbf{1}(\mathbf{P}_0^{(1)} > \rho) + \mathbf{P}^{(2)} \circ \mathbf{1}(\mathbf{P}_0^{(1)} \leqslant \rho). \tag{5.12}$$

The final label map $\mathbf{S} = \{s_i\}_{i=1,\ldots,H\times W}$ is determined by $s_i = \arg\min_{l\in\mathcal{L}} p_{i,l}$.

5.3 Statistical Label Fusion Based on Local Structural Similarity

As described in Section 5.1, our OAN-RC is based on 2D images which is an extreme case of 3D anisotropic patches. In this section, we propose to fuse anisotropic information obtained from different viewing directions using isotropic 3D local patches to estimate the final segmentation. Let us denote the segmentation results by $\mathbf{S}^j, (j = 1, \ldots, M = 3)$, which are obtained as described in Section 5.2.3 from the axial (Z), sagittal (X), and coronal (Y) OAN-RCs. Depending on the viewing directions, sectional images contain different structures and may have different texture patterns in the same organs. These differences can cause nonidentical segmentations by the deep network as shown in Figure 5-6 in 3D. In addition, there is no guarantee of connectivity between neighbor slices by independent use of slices for training and testing. Possible naïve approaches for determining the final segmentation in 3D from the OAN-RC results can be boolean operations such as union or intersection. Majority voting (MV) is another candidate for efficient fusion, however, theses approaches assume the same

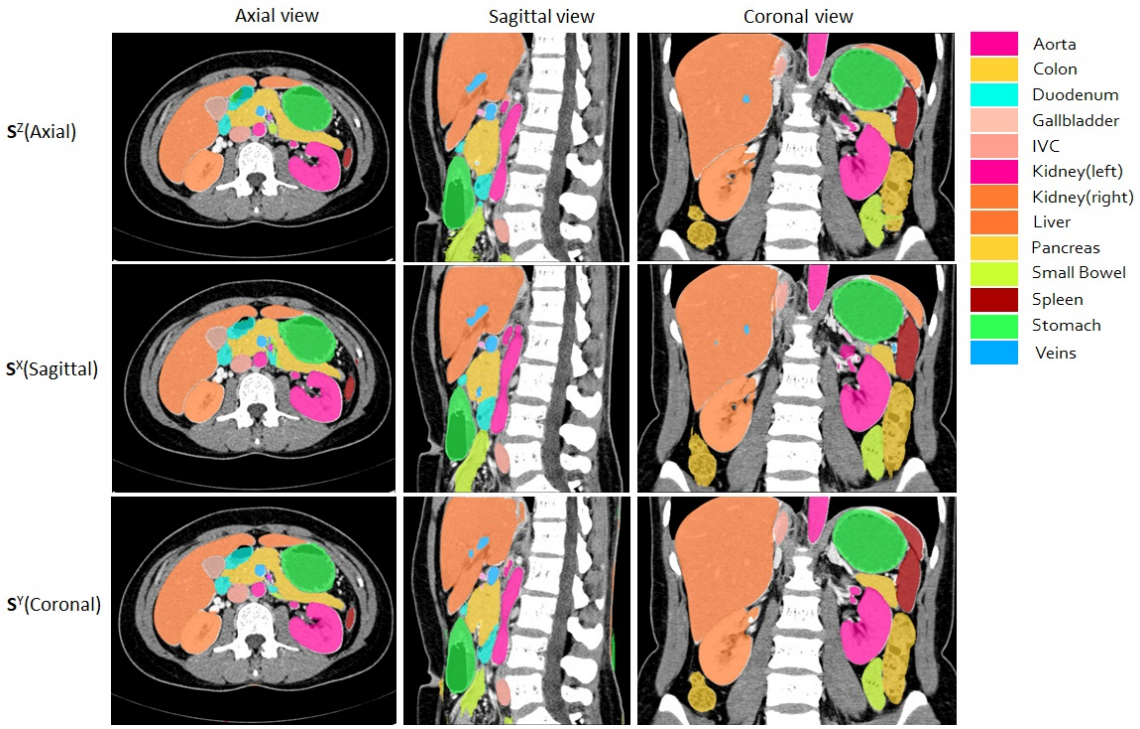


Figure 5-6. An example of multi-planar reconstruction view of OAN-RC estimations

global weights of OAN-RC results. From the observations that the performance level of segmentation, *e.g.*, sensitivity, can be different from viewing directions for each organ, we set the performance level to be an unknown variable when computing the probability of labeling. This concept is similar to the label fusion algorithms using expectation-maximization (EM) framework such as STAPLE (simultaneous truth and performance level estimation) and its extensions [59, 149, 152].

Let us denote the true label of the V by $\mathbf{T}$, which is unknown, and the unknown performance level parameter of segmentation by θ. The segmentations from the deep networks $\mathbf{S} = \{\mathbf{S}^j | j = 1, ..., M\}$ are observed values. Under this condition, the basic EM framework is performed by following two steps in an iterative manner: 1) to compute $Q^0(\theta|\theta^{(k)}) = E_{\mathbf{T}}\left[\ln L(\theta|\mathbf{S}, \mathbf{T})|\mathbf{S}, \theta^{(k)}\right]$ which is the expected value of the log likelihood, $\ln L(\theta|\mathbf{S}, \mathbf{T}) = \ln P(\mathbf{S}, \mathbf{T}|\theta)$, under the current estimate of the parameters $\theta^{(k)}$ at k^{th} iteration, and 2) to find the parameter $\theta^{(k+1)}$ which maximizes $Q^0(\theta|\theta^{(k)})$.

The maximization step can be written as

$$\begin{aligned}\theta^{(k+1)} &= \arg\max_{\theta} E_{\mathbf{T}}\left[\ln P(\mathbf{S},\mathbf{T}|\theta)|\mathbf{S},\theta^{(k)}\right] \\ &= \arg\max_{\theta} E_{\mathbf{T}}\left[\ln P(\mathbf{S}|\mathbf{T},\theta)P(\mathbf{T})|\mathbf{S},\theta^{(k)}\right] \\ &= \arg\max_{\theta} \sum_{\mathbf{T}} \ln\left\{P(\mathbf{S}|\mathbf{T},\theta)P(\mathbf{T})\right\} P(\mathbf{T}|\mathbf{S},\theta^{(k)}) \\ &= \arg\max_{\theta} \sum_{\mathbf{T}} \left\{\ln P(\mathbf{S}|\mathbf{T},\theta) + \ln P(\mathbf{T})\right\} P(\mathbf{T}|\mathbf{S},\theta^{(k)}).\end{aligned} \tag{5.13}$$

By assuming independence between $\mathbf{T}$ and θ in our problem, the second term $\sum_{\mathbf{T}} \ln P(\mathbf{T})P(\mathbf{T}|\mathbf{S},\theta^{(k)})$ in Eqn. (5.13) becomes free of θ and the maximization step can be written as

$$\begin{aligned}\theta^{(k+1)} &= \arg\max_{\theta} \sum_{\mathbf{T}} \ln P(\mathbf{S}|\mathbf{T},\theta)P(\mathbf{T}|\mathbf{S},\theta^{(k)}) \\ &= \arg\max_{\theta} E_{\mathbf{T}}\left[\ln P(\mathbf{S}|\mathbf{T},\theta)|\mathbf{S},\theta^{(k)}\right].\end{aligned} \tag{5.14}$$

Therefore, we redefine $Q^0(\theta|\theta^{(k)})$ as $Q(\theta|\theta^{(k)}) = E_{\mathbf{T}}\left[\ln P(\mathbf{S}|\mathbf{T},\theta)|\mathbf{S},\theta^{(k)}\right]$.

The performance level parameter in this framework is a global property representing the overall confidence of deep network segmentation for the whole volume. However, it can also vary according to the voxel spatial locations via the local and neighbor structures as we use 2D slices for the initial segmentation. Therefore, we propose to combine local structural similarity shown from a specific viewing direction to the original 3D volume and the global performance level, conceptually similar to local weighted voting [153]. We compute the probability of correspondence between 2D images and the 3D volume by structural similarity (SSIM) [154] by

$$\begin{aligned}\alpha_i^j &= P\left(\ell_2(I_i^j)|\ell_3(V_i)\right) \equiv SSIM\left(\ell_2(I_i^j),\ell_3(V_i)\right) \\ &= \frac{\left(2\mu_{\ell_2(I_i^j)}\mu_{\ell_3(V_i)} + c_1\right)\left(2\sigma_{\ell_2(I_i^j)\ell_3(V_i)} + c_2\right)}{\left(\mu^2_{\ell_2(I_i^j)} + \mu^2_{\ell_3(V_i)} + c_1\right)\left(\sigma^2_{\ell_2(I_i^j)} + \sigma^2_{\ell_3(V_i)} + c_2\right)},\end{aligned} \tag{5.15}$$

where α_i^j is the SSIM from the j^{th} viewing direction at the i^{th} voxel. c_1 and c_2 are user-defined constants, and $\ell_2(I_i)$ and $\ell_3(V_i)$ represent local 2D and 3D patches centered at the i^{th} voxel, respectively. μ_ℓ and σ_ℓ are the average and standard deviation of the patch ℓ, respectively, and $\sigma_{\ell_2(I_i)\ell_3(V_i)}$ is the covariance of $\ell_2(I_i)$ and $\ell_3(V_i)$. Figure 5-7

shows an example of the structural similarity computed on different viewing directions as a color map.

Considering the local image properties, the expectation of log likelihood function in our problem becomes

$$\begin{aligned} Q\left(\theta|\theta^{(k)}\right) &= E\left[\ln P\left(\mathbf{S}, I|\mathbf{T}, V, \theta\right)|\mathbf{S}, I, V, \theta^{(k)}\right] \\ &= \sum_{\mathbf{T}} \ln P\left(\mathbf{S}, I|\mathbf{T}, V, \theta\right) P\left(\mathbf{T}|\mathbf{S}, I, V, \theta^{(k)}\right). \end{aligned} \tag{5.16}$$

The global underlying performance level parameters of the deep network segmentations is defined as

$$\theta_{js's} \equiv P\left(\mathbf{S}_i^j = s'|\mathbf{T}_i = s, \theta_{js's}^{(k)}\right), \tag{5.17}$$

where $\theta_{js's}$ is the probability of the voxel labeled as s' from the j^{th} deep network with the current estimated performance value $\theta_{js's}^{(k)}$, when the true label is s.

To make the problem simple, we assume conditional independence between labeling and the original volume intensities. The labeling probability with the target image intensity then becomes

$$\begin{aligned} & P\left(\mathbf{S}_i^j = s', \ell_2(I_i^j)|\mathbf{T}_i = s, \ell_3(V_i), \theta_{js's}^{(k)}\right) \\ &= P\left(\mathbf{S}_i^j = s'|\mathbf{T}_i = s, \theta_{js's}^{(k)}\right) P\left(\ell_2(I_i^j)|\ell_3(V_i)\right) \\ &= \theta_{js's}\alpha_i^j. \end{aligned} \tag{5.18}$$

5.3.1 E-step

In the expectation step (E-step), we estimate the probability of voxelwise labels. Let us denote the probability that the true label of i^{th} voxel is $s \in \mathcal{L}$ at the k^{th} iteration by $\omega_{si}^{(k)}$. When the deep network segmentations $\mathbf{S}$ and performance level parameters at the k^{th} iteration $\theta^{(k)}$ are given, $\omega_{si}^{(k)}$ can be then described as

$$P\left(\mathbf{T}_i = s|\mathbf{S}, I, V, \theta^{(k)}\right) \equiv \omega_{si}^{(k)}, \tag{5.19}$$

where $\theta \in \mathbb{R}^{N\times|\mathcal{L}|\times|\mathcal{L}|}$ is the vector of all $(\theta_{js's})^T$. From the independence between $\mathbf{S}^X$, $\mathbf{S}^Y$, and $\mathbf{S}^Z$, we apply Bayesian theorem to Eqn. (5.19).

$$\omega_{si}^{(k)} = \frac{P(\mathbf{T}_i = s)\prod_j P\left(\mathbf{S}_i^j = s', \ell_2(I_i^j)|\mathbf{T}_i = s, \ell_3(V_i), \theta_j^{(k)}\right)}{\sum_n P(\mathbf{T}_i = n)\prod_j P\left(\mathbf{S}_i^j = s', \ell_2(I_i^j)|\mathbf{T}_i = n, \ell_3(V_i), \theta_j^{(k)}\right)}, \tag{5.20}$$

where $P(T_i = s)$ is a *priori* of the i^{th} voxel. By applying Eqn. (5.18) to Eqn. (5.20), we then obtain the probability of voxelwise labeling as

$$\omega_{si}^{(k)} = \frac{P(\mathbf{T}_i = s)\prod_j \theta_{js's}^{(k)}\alpha_i^j}{\sum_n P(\mathbf{T}_i = n)\prod_j \theta_{js'n}^{(k)}\alpha_i^j}. \tag{5.21}$$

5.3.2 M-step

In the maximization step (*M-step*), the goal is to find the performance parameters, θ, which maximize Eqn. (5.16) with the current given parameters. Considering each $\mathbf{S}^j$ and θ_j independently, the expectation of log likelihood function in Eqn. (5.16) can be expressed with the estimated voxelwise probability in *E-step*. Then the performance parameter of each segmentation can be formulated to find the solution which maximizes the summation of voxelwise probability as

$$\theta_j^{(k+1)} = \arg\max_{\theta_j} Q\left(\theta_j|\theta_j^{(k)}\right) = \arg\max_{\theta_j}\sum_i Q_i\left(\theta_j|\theta_j^{(k)}\right), \tag{5.22}$$

where $Q_i = E[\ln P(\mathbf{S}_i, \ell_2(I_i)|\mathbf{T}_i, \ell_3(V_i), \theta^{(k)})|\mathbf{S}, I, V, \theta^{(k)}]$ at i^{th} voxel. By applying Eqn. (5.19) and Eqn. (5.18), Eqn. (5.22) becomes

$$\begin{aligned}
\theta_j^{(k+1)} &= \arg\max_{\theta_j}\sum_i\sum_s P(\mathbf{T}_i = s|\mathbf{S}, I, V, \theta^{(k)}) \times \ln P\left(\mathbf{S}_i^j, \ell_2(I_i^j)|\mathbf{T}_i = s, \ell_3(V_i), \theta_j^{(k)}\right) \\
&= \arg\max_{\theta_j}\sum_i\sum_s \omega_{si}^{(k)} \ln P\left(\mathbf{S}_i^j, \ell_2(I_i^j)|\mathbf{T}_i = s, \ell_3(V_i), \theta_j^{(k)}\right) \\
&= \arg\max_{\theta_j}\sum_{s'}\sum_{i:\mathbf{S}_i^j = s'}\sum_s \omega_{si}^{(k)} \times lnP\left(\mathbf{S}_i^j = s', \ell_2(I_i^j)|\mathbf{T}_i = s, \ell_3(V_i), \theta_j^{(k)}\right) \\
&= \arg\max_{\theta_j}\sum_{s'}\sum_{i:\mathbf{S}_i^j = s'}\sum_s \omega_{si}^{(k)} \ln \theta_{js's}\alpha_i^j.
\end{aligned} \tag{5.23}$$

From the definition of θ in Eqn. (5.17), the summation of probability mass function, $\sum_{s'}\theta_{js's}^{(k)}$, must be 1, and Eqn. (5.22) becomes a constrained optimization problem

which can be solved by introducing a Lagrange multiplier, λ. We then obtain the optimal solution by making the first gradient zero as

$$0 = \frac{\partial}{\partial \theta_{js's}} \left[Q\left(\theta_j | \theta_j^{(k)}\right) + \lambda \sum_{s'} \theta_{js's} \right]. \tag{5.24}$$

By applying the derivation of Q in Eqn. (5.16), Eqn. (5.22) and Eqn. (5.23), Eqn. (5.24) becomes

$$\begin{aligned} 0 &= \frac{\sum_{i:\mathbf{S}_i^j = s'} \omega_{si}^{(k)} \alpha_i^j}{\theta_{js's}} + \lambda \\ \theta_{js's}^{(k+1)} &= \frac{\sum_{i:\mathbf{S}_i^j = s'} \omega_{si}^{(k)} \alpha_i^j}{-\lambda}. \end{aligned} \tag{5.25}$$

By substituting the constraint of $\sum_{s'} \theta_{js's}^{(k)} = 1$, we can obtain the final optimal solution as

$$\theta_{js's}^{(k+1)} = \frac{\sum_{i:\mathbf{S}_i^j = s'} \alpha_i^j \omega_{si}^{(k)}}{\sum_i \omega_{si}^{(k)}}. \tag{5.26}$$

The two steps, Eqn. (5.21) and Eqn. (5.26), are then computed alternatively in the EM iterations until they converge. From the final values of Eqn. (5.21), the final segmentation can be computed by graph-based approaches such as [155].

5.3.3 Parallel computing using GPUs

The fusion step can be efficiently computed in a parallel way on a GPU. The local structural similarity α_i^j of i-th voxel in jth deep network and *priori* $P(T_i)$ can be computed for each voxel and saved as a pre-processing step. In the EM iterations, as shown in Eqn. (5.21), the probability can be computed and updated for each structure at each voxel. In our implementation, a GPU thread is logically allocated for each voxel. However, to reduce the used memory and computation cost, the target volume of interest (VOI) for each structure s is computed in an extended region as $\delta = 4$ voxels for each direction from $V(\bigcup_j \mathbf{S}^j = s)$ in our implementation. For parallel computing, one CPU thread is allocated to a structure and launches a kernel of one GPU to compute EM iteration for each structure.

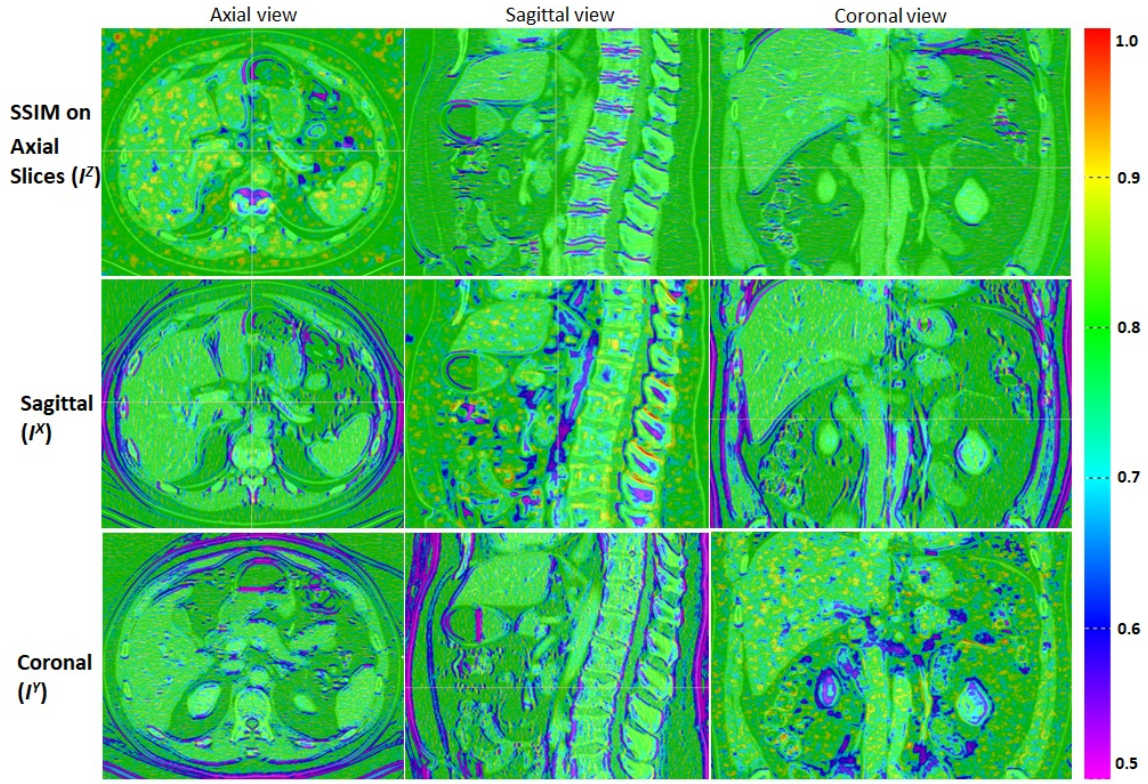


Figure 5-7. The local structural similarity map between 2D slices and the 3D volume. Each row is captured from the same similarity map computed on one viewing direction. Each column shows the captures images at the same location computed from different viewing directions.

5.4 Experimental Results

We evaluated our methods on 236 abdominal CT images of normal cases under an IRB (Institutional Review Board) approved protocol in Johns Hopkins Hospital as a part of the FELIX project for pancreatic cancer research [150]. CT images were obtained by Siemens Healthineers (Erlangen,Germany) SOMATOM Sensation and Definition CT scanners. CT scans are composed of $(319 - 1051)$ slices of (512×512) images, and have voxel spatial resolution of $([0.523 - 0.977] \times [0.523 - 0.977] \times 0.5)\, mm^3$. All CT scans are contrast enhanced images and obtained in the portal venous phase.

A total of 13 structures for each case were segmented by four human annotators/raters, one case by one person, and confirmed by an independent senior expert. The structures include the aorta, colon, duodenum, gallbladder, interior vena cava (IVC), kidney (left, right), liver, pancreas, small bowel, spleen, stomach, and large veins. Vascular structures were segmented only outside of the organs in order to make the structures exclusive to each other (*i.e.*, no overlaps).

As explained in Section 5.2, we used OAN-RCs for multi-organ segmentation whose backbone FCNs had been pre-trained by *PascalVOC* dataset [131]. From the possible variants of FCNs (*e.g.*, FCN-32s, FCN-16s, and FCN-8s), which depend on how they combine the fine detailed predictions [156], we selected FCN-8s in this study because it captures very fine details in the 3^{rd} and 4^{th} pooling layer, and keeps high-level semantic contextual information from the final layer. Our algorithm was implemented and tested on a workstation with Intel i7-6850K CPU, NVidia TITAN X (PASCAL) GPU. With 236 cases, the initial segmentations using OAN-RCs were tested by four-fold cross-validation. All the input images of OAN-RCs are 1.5 times enlarged by upsampling, which lead to improved performance in our experiments.

In the fusion step, the average probability of $\mathbf{S}^X, \mathbf{S}^Y, \mathbf{S}^Z$ are taken as a *priors* in Eqn. (5.21) and the initial performance levels $\theta^{(0)}_{js's}$ were computed by randomly

selecting 5 cases and by comparing them to the ground-truth. To compute the local patch-based structural similarity in Eqn. (5.15), patches of $(4.5 \times 4.5 \times 4.5)mm^3$ size cubes were used for 3D volume. Since CT voxels are not always isotropic and spatial resolutions can be different between scan volumes, we re-sampled the 3D patch with $0.5mm$ length cubic voxels so that the same size of $(9 \times 9 \times 9)$ 3D patches and (9×9) 2D patches from all directions can be used for all cases in our experiments.

The final segmentation results using OAN-RC with local structural similarity-based statistical fusion (LSSF) were compared with the 3D-patch based state-of-the-art approaches, 3D Unet [143] and hierarchical 3D FCN (HFCN) [75] as well as 2D-based FCN, OAN and OAN-RC with majority voting (MV). For a quantitative comparison, we computed the well-known Dice-Sørensen similarity coefficient (DSC) and the surface distances based on the manual annotations as ground-truth. For a structure s, DSC is computed as $\frac{2V(\mathbf{S}=s \bigcap \mathbf{T}=s)}{V(\mathbf{S}=s)+V(\mathbf{T}=s)}$ where $\mathbf{S}$ is the estimated segmentation and $\mathbf{T}$ is the ground-truth, *i.e.*, manual annotations in this study. The surface distance was computed from each vertex of the ground-truth and to the estimates of our algorithms. Figure 5-8 shows comparison results by box plots, while Table 5-I and Table 5-II represent the mean and standard deviations for all the 236 cases.

As shown in Figure 5-8, the basic OAN-RC outperforms other state-of-the-art approaches and our local structural similarity-based fusion improves the results even more. We note that although DSC shows the relative overall volume similarity, it does not quantify the boundary smoothness or the boundary noise of the results. But evaluating the surface distances, see below, shows that our method works effectively for both the whole volumes and the boundaries of the organs.

Table 5-I and Table 5-II represent the mean and standard deviations of performance measures for 13 critical organs. Similar to the box plots, they show that our OAN-RCs with statistical fusion improves the overall mean performance and also reduces the standard deviations significantly. Figure 5-9 shows an example generated by

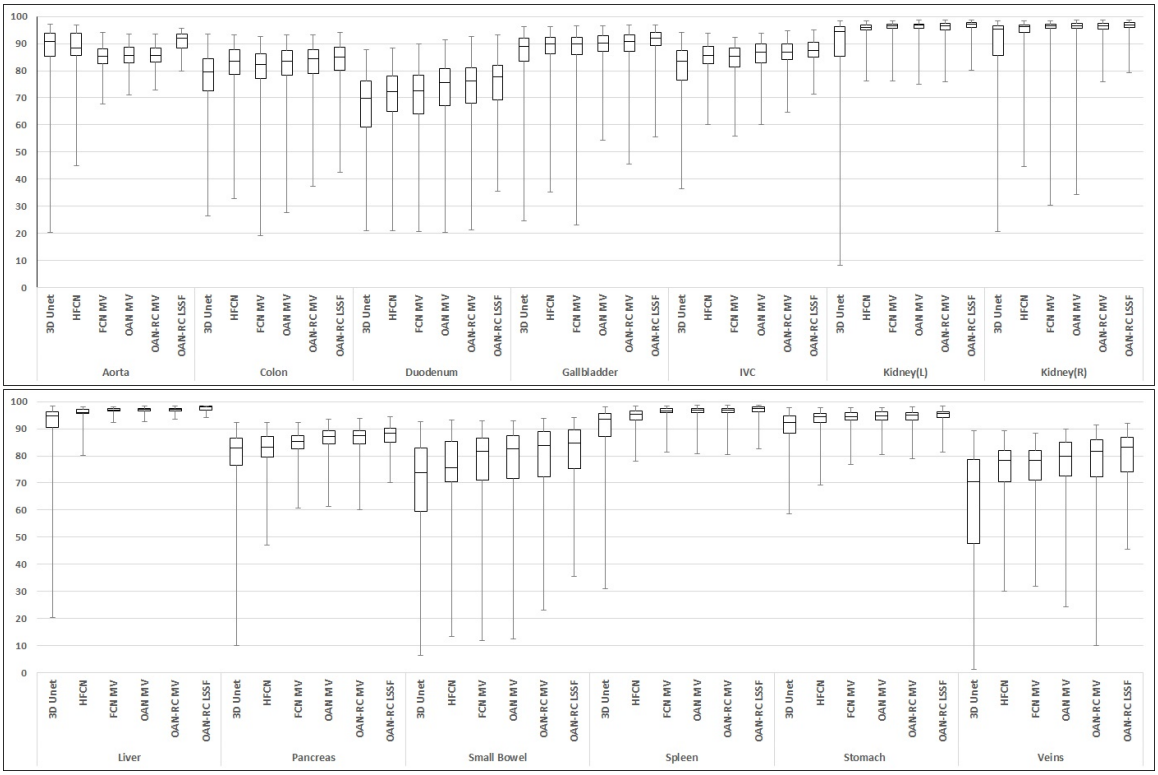


Figure 5-8. Box plots of the Dice-Sørensen similarity coefficients of 13 structures to compare performance. As in typical box plots, the box represents the first quartile, median, and the third quartile from the lower border, middle and the upper boarder, respectively, and the lower and the upper whiskers show the minimum and the maximum values. (LSSF: Local Similarity-based Statistical Fusion.)

Table 5-I. DICE-Sørensen similarity coefficient (DSC, %) of thirteen segmented organs (mean ± standard deviation of 236 cases).

Structure	3D U-net	HFCN	FCN MV	OAN MV	OAN-RC MV	OAN-RC LSSF
Aorta	87.0±12.3	88.3± 8.8	85.0±4.2	85.5± 4.2	85.3± 4.1	**91.8**± 3.5
Colon	77.0±11.0	79.3± 9.2	80.3± 9.1	81.5± 9.4	82.0± 8.8	**83.0**± 7.4
Duodenum	66.8±12.8	70.3± 10.4	70.2±11.3	72.6±11.4	73.4±11.1	**75.4**± 9.1
Gallbladder	85.4±10.3	87.9± 7.5	87.8± 8.3	88.9± 6.2	89.4± 6.1	**90.5**± 5.3
IVC	80.8±10.2	84.7± 5.9	84.0± 6.0	85.6± 5.8	86.0± 5.5	**87.0**± 4.2
Kidney(L)	83.9±22.4	95.2± 2.6	96.1± 2.0	96.2± 2.2	95.9± 2.3	**96.8**± 1.9
Kidney(R)	88.0±14.4	95.6± 4.5	95.8± 4.9	95.9± 4.9	96.0± 2.5	**98.4**± 2.1
Liver	91.4± 9.9	95.7± 1.8	96.8± 0.8	97.0± 0.9	97.0± 0.8	**98.0**± 0.7
Pancreas	79.3±11.7	81.4±10.8	84.3± 4.9	86.2± 4.5	86.6± 4.3	**87.8**± 3.1
Small bowel	69.9±17.3	71.1±15.0	76.9±14.0	78.0±13.8	79.0±13.4	**80.1**±10.2
Spleen	89.6±9.5	93.1± 2.1	96.3± 1.9	96.4± 1.9	96.4± 1.7	**97.1**± 1.5
Stomach	90.1± 7.2	93.2± 5.4	93.9± 3.2	94.2± 2.9	94.2± 3.0	**95.2**± 2.6
Veins	60.7±23.7	74.5±10.5	74.8±10.7	76.8±11.2	77.4±12.1	**80.7**± 9.3

Table 5-II. Average surface distances of thirteen segmented organs for all 236 cases (mean ± standard deviation of average surface distances in mm).

Structure	3D U-net	HFCN	FCN MV	OAN MV	OAN-RC MV	OAN-RC LSSF
Aorta	0.44 ±1.01	0.42±0.58	0.56±0.47	0.47±0.42	0.44±0.28	**0.39**±0.21
Colon	6.75±9.01	6,35±8.12	6.27±7.44	5.65±7.25	4.07±5.72	**3.59**±4.17
Duodenum	2.01±2.46	1.70±2.18	1.71±2.25	1.49±1.87	1.54±1.43	**1.36**±1.31
Gallbladder	1.31±0.76	1.21±0.50	1.22±0.52	1.12±0.50	1.05±0.41	**0.95**±0.37
IVC	1.57±1.53	1.15±1.05	1.26±1.08	1.16±1.38	1.12±1.24	**1.08**±1.03
Kidney(L)	0.77±1.04	0.41±0.42	0.36±0.47	0.34±0.47	0.30±0.33	**0.30**±0.30
Kidney(R)	1.39±2.01	1.03±1.68	1.05±1.74	0.74±1.32	0.54±1.09	**0.45**±0.89
Liver	1.89±3.21	1.60± 0	1.61±2.98	1.39±2.64	1.32±1.74	**1.23**±1.52
Pancreas	1.78±1.05	1.51±0.80	1.41±0.88	1.19±0.82	1.17±0.72	**1.05**±0.65
Small bowel	4.21±5.78	4.01±6.01	3.91±6.05	3.20±4.05	3.37±5.48	**3.01**±3.35
Spleen	0.98±0.56	0.59±0.37	0.60±0.36	0.56±0.40	0.47 ±0.27	**0.42**±0.25
Stomach	2.78±5.89	2.50±5.02	2.51±5.13	2.36±5.65	1.88±1.64	**1.68**±1.55
Veins	2.31±4.51	1.75±3.51	1.69±3.61	1.92±6.48	1.40±3.61	**1.21**±3.05

our proposed OAN-RC with LSSF, which is visually indistinguishable from manual segmentation for almost all target structures.

The OAN-RC training and testing can be computed in parallel for each view direction. In our experiments, the training took 40 hours for $120,000$ iterations for 177 training cases and the average testing time for each volume was 76.73 seconds. The fusion time depended on the volume of the target structure, and the average computation time for 13 organs was 6.87 seconds.

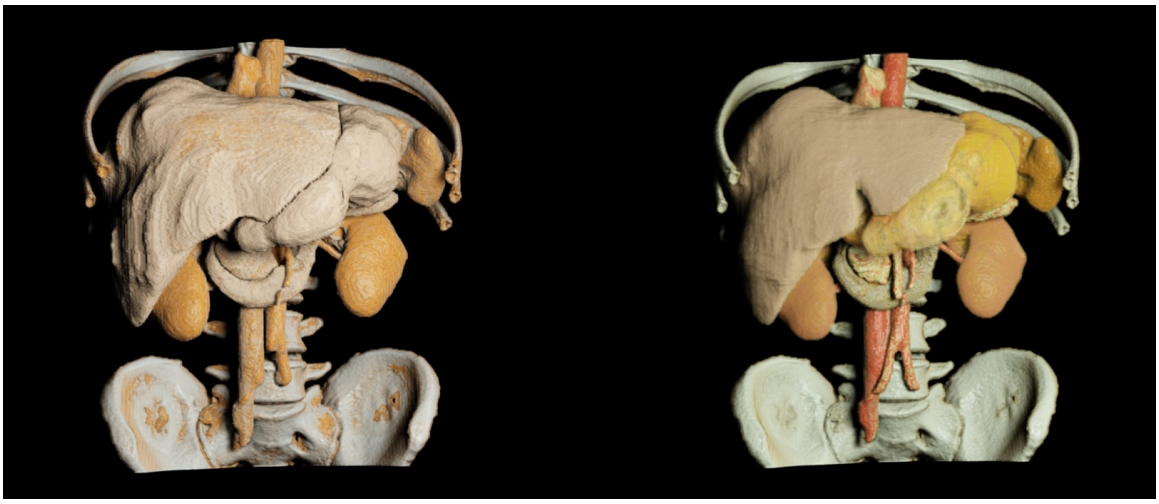


Figure 5-9. 3D photo-realistic rendering of the ground-truth (left) and the results from OAN-RC with statistical fusion (right). The aorta, duodenum, IVC, liver, kidneys, pancreas, duodenum, spleen, and stomach are rendered. The difference between our results and the ground-truth are almost visually indistinguishable. To differentiate adjacent organs and from manual segmentation, different color setting were applied to the our methods results.

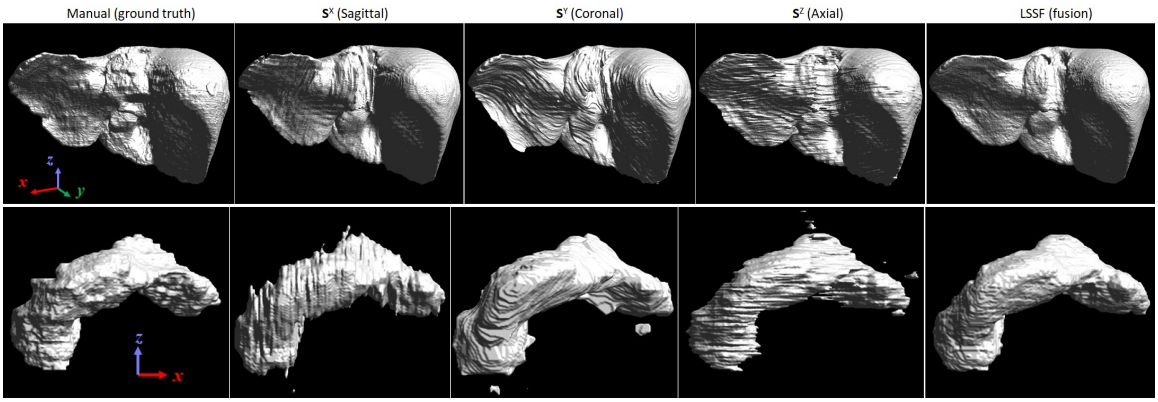
5.5 Discussion

Multi-organ segmentation using OAN-RCs alone, without the statistical fusion, gave similar or better performance compared with the state-of-the-art approaches summarized in [64]. In the specific case of the pancreas, state-of-the-art methods showed (mean $\pm$ standard deviations) segmentation accuracies as $74.4 \pm 20.2(\%)$ on 140 cases [157], $78.5 \pm 14.0(\%)$ on 150 cases [64], $78.0 \pm 8.2(\%)$ on 82 cases [8] and $75.74 \pm 10.47(\%)$ (on the whole slice) versus $82.4 \pm 5.7(\%)$ (reduced region of interest) on 82 cases [13] in terms of DSC. We cannot make a direct comparison because in these datasets CT images and manual segmentations (*i.e.*, annotation) for the ground-truth are different from each other. But our OAN-RCs segmentations on our larger dataset shows similar or better performances in terms of DSC. Among target organs, our performance on structures such as gallbladder and pancreas, whose sizes are relatively small and have particularly weak boundaries improves significantly from using basic FCNs or using OANs without reverse connections.

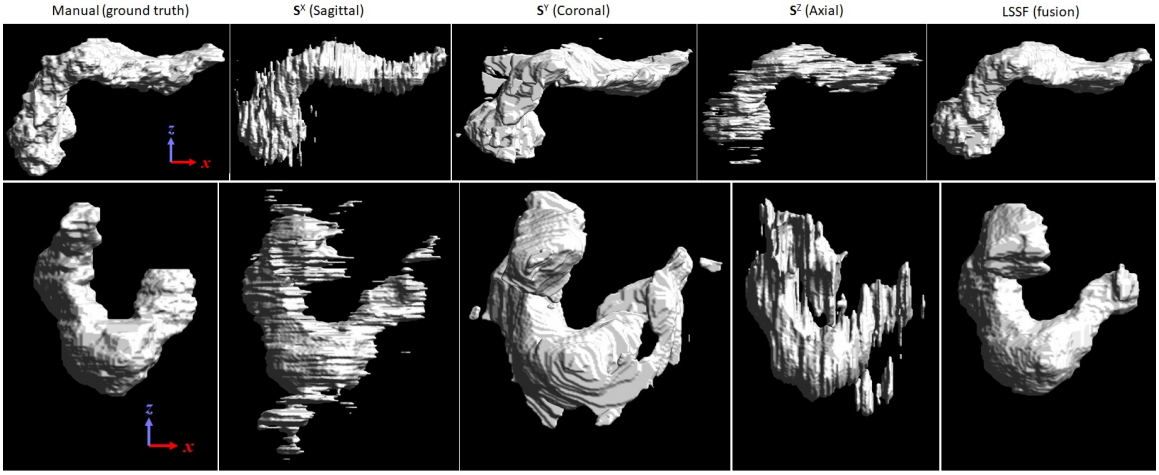
Moreover, as shown in section 5.4, our statistical fusion based on local structural similarity improves the overall segmentation accuracies in terms of both DSC and average surface distances. In particular, there are significant performance improvements for the minimum values as shown in Figure 5-8, which helps explain the robustness of the algorithm. The differences can be depicted more clearly by visualizing the 3D surfaces as shown in Figure 5-10 and Figure 5-11. The noise of the deep network segmentations is distributed over large regions, without much connectivity, and occasionally they show significantly different patterns. But our fusion step exploits structural similarity which outputs clean and smooth boundaries by effectively combining different information based on the local structure of the original 3D volume.

When applying our proposed method and interpreting the evaluation results, we must address several considerations. As shown in our experiments, our proposed algorithm also outperforms 3D patch based approaches. But 3D (isotropic) patch-based approaches have several issues which make it hard to apply to this problem. To make bigger patch size, they require more parameters and hence require more training data or, if this is not available, significant data augmentation (e.g, by scaling, rotation, and elastic deformation). In addition, there can be practical memory limitation on GPUs which restricts the expandable patch size. The limited patch size means that the deep networks receptive field sizes contains only limited local information which is problematic for multi-organ segmentation and the discontinuities between the patches also raises problems. It is possible that solutions to these three problems may make 3D patch based methods work better in the future. Unlike 3D approaches, the local structure-similarity used in our fusion method effectively combine the information from anisotropic patches to 3D at each voxel.

The ground-truth used in this study for training and evaluation was specified using manual annotations by human observers. It is well known that there can be significant inter-/intra-observer variations in manual segmentation. But, as explained before,



(a) Liver (upper) and Pancreas (lower)



(b) Pancreas (Upper) and Duodenum (lower)

Figure 5-10. Effects of local structural similarity-based statistical fusion (LSSF) for estimating 3D surfaces. From left to right, the manual segmentation (ground-truth), initial segmentations from OAN-RCs with X, Y, Z slices, and the results of our proposed algorithm with statistical fusion. (a) When $\mathbf{S}^X$, $\mathbf{S}^Y$, and $\mathbf{S}^Z$ show similar result, statistical fusion produces smoother and less-noisy boundaries. (b) Surface estimation examples when initial OAN-RCs give differing results. But our approach effectively fuses the information, exploiting the local structural similarity.

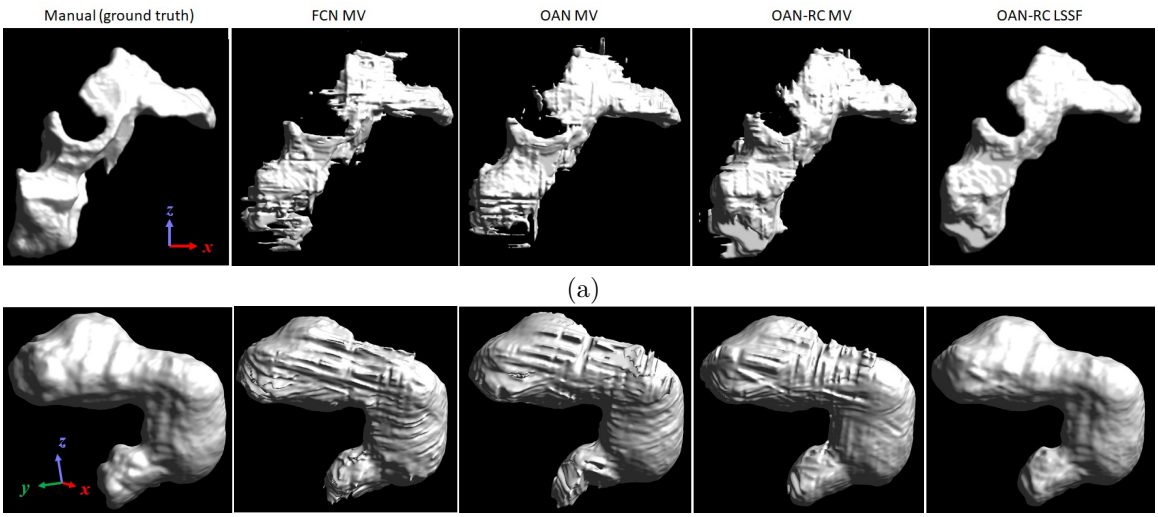


Figure 5-11. Examples of FCN, OAN, OAN-RC, and OAN-RC. The manual segmentation (ground-truth), FCN MV, OAN MV, OAN-RC MV, OAN-RC LSFF (from left to right). (a) Pancreas: DSC(%) and surface distances (mean± standard deviation in mm) to the ground-truth are 72.5 and 2.13 ± 1.74 (FCN MV), 77.2 and 1.90 ± 1.77 (OAN MV), 82.4 and 1.33 ± 1.31 (OAN-RC MV), and 85.5 and 0.71 ± 0.81 (OAN-RC LSSF), respectively. (b) Stomach: DSC(%) and surface distances (mean± standard deviation in mm) to the ground-truth are 92.5 and 2.44 ± 1.27 (FCN MV), 93.6 and 1.63 ± 1.14 (OAN MV), 94.9 and 2.25 ± 1.30 (OAN-RC MV), and 97.1 and 1.26 ± 0.88 (OAN-RC LSSF), respectively.

the ground-truth was created by four human observers and checked by experts in a visual way, and we randomly divided testing groups in our 4-fold cross-validation to avoid biased comparison. However, it is still possible that inaccuracies due to human variability may affect the evaluation as well as the training. This can be further intensively explored as separate experiments.

Another possible consideration when applying the proposed approach is the image quality which can affect both of manual annotations and deep network segmentation results. Various factors such as spatial resolution, level of artifacts and reconstruction kernels should be considered. The dataset used in this study has been collected between 2005 to 2009 in the same institute with control over the scanning parameters. As explained in Section 5.4, the CT protocol is the portal venous phase and the spatial resolution is almost isotropic. But different scanning parameters and artifacts may affect our algorithms performance when applied to other datasets.

The same issues about manual segmentations and image qualities can be raised in general segmentation and evaluations. Specifically for our proposed approach, especially in the fusion step, the way of computing *priori*, $P(T)$, used in Eqn. (5.21) can in practice affect the final segmentation. But considering that the deep network segmentation results from different viewing-directions are independently obtained, the mean can be accepted in general. However, if the deep network segmentations show clear tendencies towards over-estimation or under-estimation, then different types of models for *priors* may need to be used in order to improve the final result for practical applications.

One of the main advantages of our algorithm is the efficient computation time. The segmentation of 13 organs of the whole volume takes similar to or less than 1 minute with better performance reported than the state-of-the-art methods [64]. Hence our approach can be practically useful in clinical environments.

5.6 Summary

In this chapter, we proposed a novel framework for multi-organ segmentation using OAN-RCs with statistical fusion exploiting structural similarity. Our two-stage organ-attention network reduces uncertainties at weak boundaries, focuses attention on organ regions with simple context, and adjusts FCN error by training the combination of original images and OAMs. Reverse connections deliver abstract level semantic information to lower layers so that hidden layers can be assisted to contain more semantic information and give good results even for small organs. The results are improved by the statistical fusion, based on local structural similarity, which smooths our noise and removes biases leading to better overall segmentation performance in terms of DSC and surface distances. We showed that our performance is better than previous state-of-the-art algorithms. Our framework is not specific to any particular body region, but gives high quality and robust results for abdominal CTs, which are typically challenging regions due to their low contrast, large intra-/inter-variations, and different scales. In addition, the efficient computational time of our algorithm makes our approach practical for clinical environments such as CAD, CAS or RT.

Chapter 6

Semi-Supervised 3D Abdominal Multi-Organ Segmentation via Deep Multi-Planar Co-Training

In multi-organ segmentation of abdominal CT scans, most existing fully supervised deep learning algorithms require lots of voxel-wise annotations, which are usually difficult, expensive, and slow to obtain. In comparison, massive unlabeled 3D CT volumes are usually easily accessible. Current mainstream works to address semi-supervised biomedical image segmentation problem are mostly graph-based. By contrast, deep network based semi-supervised learning methods have not drawn much attention in this field. In this chapter, we propose Deep Multi-Planar Co-Training (DMPCT), whose contributions can be divided into two folds: 1) The deep model is learned in a co-training style which can mine consensus information from multiple planes like the sagittal, coronal, and axial planes; 2) Multi-planar fusion is applied to generate more reliable pseudo-labels, which alleviates the errors occurring in the pseudo-labels and thus can help to train better segmentation networks. Experiments are done on our newly collected large dataset with 100 unlabeled cases as well as 210 labeled cases where 16 anatomical structures are manually annotated by four radiologists and confirmed by a senior expert. The results suggest that DMPCT significantly outperforms the fully supervised method by more than 4% especially

when only a small set of annotations is used.

6.1 Introduction

Multi-organ segmentation of radiology images is a critical task which is essential to many clinical applications such as computer-aided diagnosis, computer-aided surgery, and radiation therapy. Compared with other internal human structures like brain or heart, segmenting abdominal organs appears to be much more challenging due to the low contrast and high variability of shape in CT images. In this chapter, we focus on the problem of multi-organ segmentation in abdominal regions, *e.g.*, liver, pancreas, kidney, *etc.*

Fully supervised approaches can usually achieve high accuracy with a large labeled training set which consists of pairs of radiology images as well as their corresponding pixel-wise label maps. However, it is quite time-consuming and costly to obtain such a large training set especially in the medical imaging domain due to the following reasons: 1) precise annotations of radiology images must be hand annotated by experienced radiologists and carefully checked by additional experts and 2) contouring organs or tissues in 3D volumes requires tedious manual input. By contrast, large unannotated datasets of CT images are much easier to obtain. Thereby our study mainly focuses on multi-organ segmentation in a semi-supervised fashion, *i.e.*, how to fully leverage unlabeled data to boost performance, so as to alleviate the need for such a large annotated training set.

In the biomedical imaging domain, traditional methods for semi-supervised learning usually adopt graph-based methods [109, 110] with a clustering assumption to segment pixels (voxels) into meaningful regions, *e.g.*, superpixels. These methods were studied for tissue or anatomical structures segmentation in 3D brain MR images, ultrasound images, *etc.* Other machine learning methods such as kernel-based large margin

algorithms [111] have been suggested for white matter hyperintensities segmentation. Although widely applied to biomedical imaging segmentation tasks in the past decade, the traditional methods cannot always produce a satisfactory result due to the lack of advanced techniques.

With the recent advance of deep learning and its applications [35, 158–160], fully convolutional networks (FCNs) [41] have been successfully applied to many biomedical segmentation tasks such as neuronal structures segmentation [45, 93, 101, 161], single organ segmentation [8, 13, 16], and multi-organ segmentation [18, 75, 128] in a fully supervised manner. Their impressive performances have shown that we are now equipped with much more powerful techniques than traditional methods. Nevertheless, network-based semi-supervised learning for biomedical image segmentation has not drawn enough attention. The current usage of deep learning for semi-supervised multi-organ segmentation in the biomedical imaging domain is to train an FCN on both labeled and unlabeled data, and alternately update automated segmentations (pseudo-labels) for unlabeled data and the network parameters [10]. However, if an error occurs in the initial pseudo-label of the unlabeled data, the error will be reinforced by the network during the following iterations. How to improve the quality of pseudo-labels for unlabeled data hence becomes a promising direction to alleviate this negative effect.

In our approach, we exploit the fact that CT scans are high-resolution three-dimensional volumes which can be represented by multiple planes, *i.e.*, the axial, coronal, and sagittal planes. Taking advantages of this multi-view property, we propose Deep Multi-Planar Co-Training (DMPCT), a systematic EM-like semi-supervised learning framework. DMPCT consists of a teacher model, a multi-planar fusion module, and a student model. While the teacher model is trained from multiple planes separately in a slice-by-slice manner with a few annotations, the key advantage of DMPCT is that it enjoys the additional benefit of continuously generating more

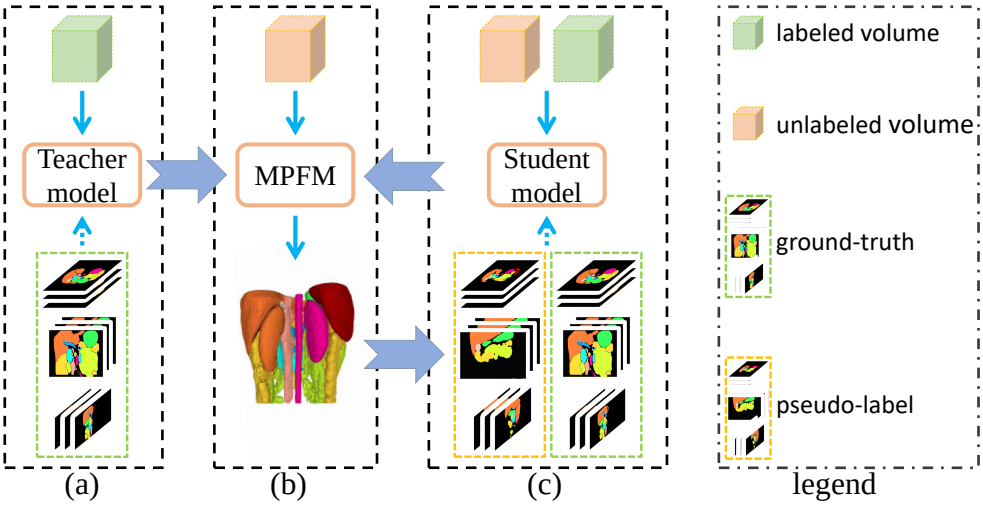


Figure 6-1. Illustration of the Deep Multi-Planar Co-Training (DMPCT) framework. (a) We first train a teacher model on the labeled dataset. (b) The trained model is the used to assign pseudo-labels to the unlabeled data using our multi-planar fusion module as demonstrated in Figure 6-2. (c) Finally, we train a student model over the union of both the labeled and the unlabeled data. Step (b) and (c) are performed in an iterative manner.

reliable pseudo-labels by the multi-planar fusion module, which can afterward help train the student model by making full usage of massive unlabeled data. As there are multiple segmentation networks corresponding to different planes in the teacher model and the student model, co-training [12, 113] is introduced so that these networks can be trained simultaneously in our unified framework and benefit from each other. We evaluate our algorithm on our newly collected large dataset and observe a significant improvement of 4.23% compared with the fully supervised method. At last, as DMPCT is a generic and flexible framework, it can be envisioned that better backbone models and fusion strategies can be easily plugged into our framework. Our unified system can be also practically useful for current clinical environments due to the efficiency in leveraging massive unlabeled data to boost segmentation performance.

6.2 Deep Multi-Planar Co-Training

We propose Deep Multi-Planar Co-Training (DMPCT), a semi-supervised multi-organ segmentation method which exploits multi-planar information to generate pseudo-labels for unlabeled 3D CT volumes. Assume that we are given a 3D CT volume dataset $\mathcal{S}$ containing K organs. This includes labeled volumes $\mathcal{S}_{\mathrm{L}} = \{(\mathbf{I}_m, \mathbf{Y}_m)\}_{m=1}^{l}$ and unlabeled volumes $\mathcal{S}_{\mathrm{U}} = \{\mathbf{I}_m\}_{m=l+1}^{M}$, where $\mathbf{I}_m$ and $\mathbf{Y}_m$ denote a 3D input volume and its corresponding ground-truth segmentation mask. l and $M-l$ are the numbers of labeled and unlabeled volumes, respectively. Typically $l \ll M$. As shown in Figure 6-1, DMPCT involves the following steps:

- **Step 1:** train a *teacher model* on the manually labeled data $\mathcal{S}_{\mathrm{L}}$ in the fully supervised setting (see Section 6.2.1).
- **Step 2:** the trained model is then used to assign pseudo-labels $\{\hat{\mathbf{Y}}_m\}_{m=l+1}^{M}$ to the unlabeled data $\mathcal{S}_{\mathrm{U}}$ by fusing the estimations from all planes (see Section 6.2.2).
- **Step 3:** train a *student model* on the union of the manually labeled data and automatically labeled data $\mathcal{S}_{\mathrm{L}} \cup \{(\mathbf{I}_m, \hat{\mathbf{Y}}_m)\}_{m=l+1}^{M}$ (see Section 6.2.3).
- **Step 4:** perform step 2 & 3 in an iterative manner.

6.2.1 Teacher Model

We train the teacher model on the labeled dataset $\mathcal{S}_{\mathrm{L}}$. By splitting each volume and its corresponding label mask from the sagittal (S), coronal (C), and axial (A) planes, we can get three sets of 2D slices, *i.e.*, $\mathcal{S}_{\mathrm{L}}^{V} = \{(\mathbf{I}_n^V, \mathbf{Y}_n^V)\}_{n=1}^{N_V}$, $V \in \{\mathrm{S}, \mathrm{C}, \mathrm{A}\}$, where N_V is the number of 2D slices obtained from plane V. We train a 2D-FCN model (we use [41] as our reference CNN model throughout this chapter) to perform segmentation from each plane individually.

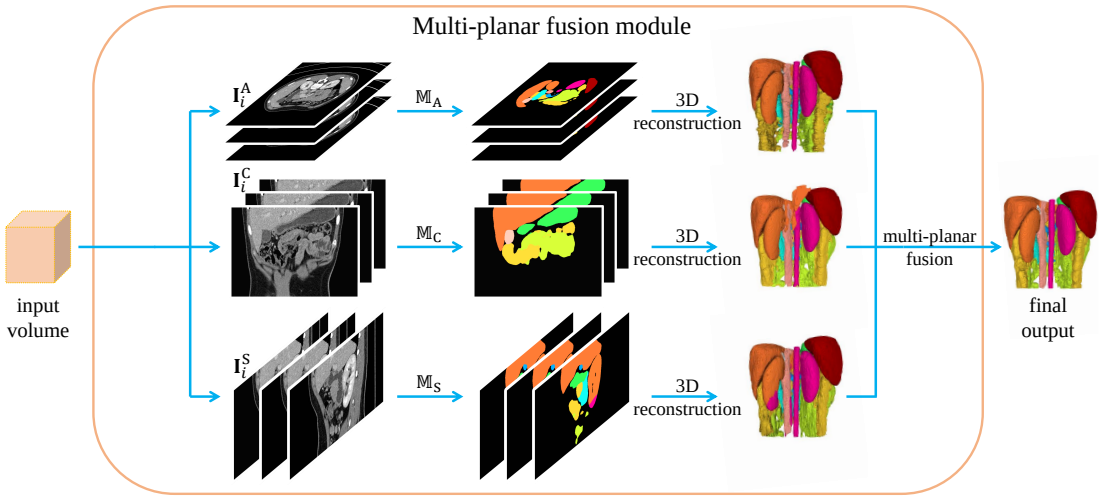


Figure 6-2. Illustration of the multi-planar fusion module, where the input 3D volume is first parsed into 3 sets of slices along the sagittal, coronal, and axial planes to be evaluated respectively. Then the final 3D estimation is obtained by fusing predictions from each individual plane.

Without loss of generality, let $\mathbf{I}^V \in \mathbb{R}^{W \times H}$ and $\mathbf{Y}^V = \{y_i^V\}_{i=1}^{W \times H}$ denote a 2D slice and its corresponding label mask in $\mathcal{S}_{\mathrm{L}}^{\mathrm{V}}$, where $y_i^V \in \{0, 1, \ldots, K\}$ is the organ label (0 means background) of the i-th pixel in $\mathbf{I}^V$. Consider a segmentation model $\mathbb{M}^V : \hat{\mathbf{Y}} = \mathbf{f}\left(\mathbf{I}^V; \theta\right)$, where θ denotes the model parameters and $\hat{\mathbf{Y}}$ denotes the prediction for $\mathbf{I}^V$. Our objective function is

$$\mathcal{L}(\mathbf{I}^V, \mathbf{Y}^V; \theta) = -\frac{1}{W \times H}\left[\sum_{i=1}^{W \times H}\sum_{k=0}^{K} \mathbb{1}(y_i^V = k) \log p_{i,k}^V\right], \tag{6.1}$$

where $p_{i,k}^V$ denotes the probability of the i-th pixel been classified as label k on 2D slice $\mathbf{I}^V$ and $\mathbb{1}(\cdot)$ is the indicator function. We train the teacher model by optimizing $\mathcal{L}$ *w.r.t.* θ by stochastic gradient descent.

6.2.2 Multi-Planar Fusion Module

Given a well-trained teacher model $\{\mathbb{M}^V | V \in \{\mathrm{S}, \mathrm{C}, \mathrm{A}\}\}$, our goal of the multi-planar fusion module is to generate the pseudo-labels $\{\hat{\mathbf{Y}}_m\}_{m=l+1}^{M}$ for the unlabeled data $\mathcal{S}_{\mathrm{U}}$.We first make predictions on the 2D slices from each plane and then reconstruct the 3D volume by stacking all slices back together. Several previous studies [162,

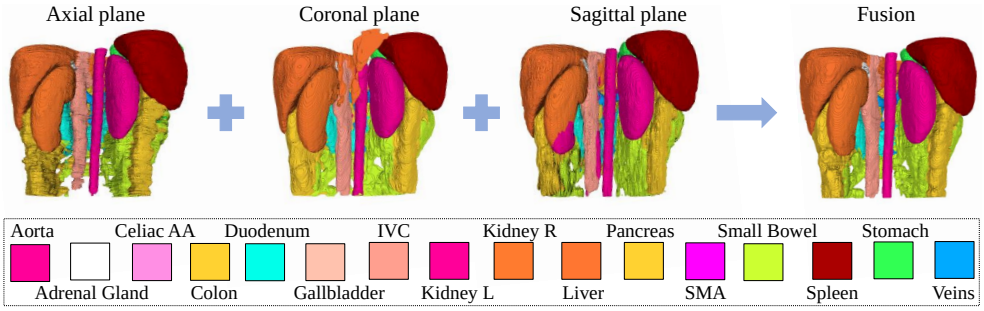


Figure 6-3. An example of 3D predictions reconstructed from the sagittal, coronal, and axial planes as well as their fusion output. Estimations from single planes are already reasonably well, whereas the single fusion outcome is superior to estimation from any single plane.

163] suggest that combining predictions from multiple views can often improve the accuracy and the robustness of the final decision since complementary information can be exploited from multiple views simultaneously. Thereby, the fused prediction from multiple planes is superior to any estimation of a single plane. The overall module is shown in Figure 6-2.

More specifically, majority voting is applied to fuse the hard estimations by seeking an agreement among different planes. If the predictions from all planes do not agree on a voxel, then we select the prediction for that voxel with the maximum confidence. As simple as this strategy might sound, this method has been shown to result in highly robust and efficient outcome in various previous studies [13, 162–164]. The final decision for the i-th voxel $y_i^\star$ of $\hat{\mathbf{Y}}_m$ is computed by:

$$y_i^\star = \begin{cases} y_i^V, & \text{if } \exists\ V,\ V' \in \{\mathrm{S},\mathrm{C},\mathrm{A}\},\ V \neq V' \mid y_i^V = y_i^{V'} \\ y_i^{V^\star}, & \text{otherwise} \end{cases}, \tag{6.2}$$

where $V^\star = \underset{V \in \{\mathrm{S},\mathrm{C},\mathrm{A}\}}{\arg\max} \max_j p_{i,j}^V$. $p_{i,j}^{\mathrm{S}}$, $p_{i,j}^{\mathrm{C}}$, and $p_{i,j}^{\mathrm{A}}$ denote the probabilities of the i-th pixel classified as label j from the sagittal, coronal, and axial planes, respectively. y_i^V denotes the hard estimation for the i-th pixel on plane V, *i.e.*, $y_i^V = \arg\max_j p_{i,j}^V$.

As shown in Figure 6-3, our multi-planar fusion module improves both over- and under-estimation by fusing aspects from different planes and therefore yields a much

Algorithm 3 Multi-planar co-training for multi-organ segmentation

Require: A set of labeled data $\mathcal{S}_{\mathrm{L}} = \{(\mathbf{I}_m, \mathbf{Y}_m)\}_{m=1}^{l}$ and unlabeled volumes $\mathcal{S}_{\mathrm{U}} = \{\mathbf{I}_m\}_{m=l+1}^{M}$.

Ensure: A trained multi-organ segmentation model $\{\mathbb{M}^{\mathrm{S}}, \mathbb{M}^{\mathrm{C}}, \mathbb{M}^{\mathrm{A}}\}$.

```
S ← S_L
for t = 1 to T do
    Parse S into S^S, S^C ,S^A;
    Train M^S, M^C, and M^A on S^S, S^C, and S^A respectively;
    Generate pseudo-class labels {Ŷ_m}_{m=l+1}^M for the unlabeled dataset S_U by Eqn. (6.2);
    Augment the training set by adding the self-labeled examples to S, i.e., S = S_L ∪ {(I_m, Ŷ_m)}_{m=l+1}^M.
end for
Parse S into S^S, S^C ,S^A.
Train M^S, M^C, and M^A on S^S, S^C, and S^A respectively.
```

better outcome. Note that other rules [18, 165] can also be easily adapted to this module. We do not focus on discussing the influence of the fusion module here, although intuitively a stronger fusion module should lead to a higher performance.

6.2.3 Student Model

After generating the pseudo-labels $\{\hat{\mathbf{Y}}_m\}_{m=l+1}^{M}$ for the unlabeled dataset $\mathcal{S}_{\mathrm{U}}$, the training set can be then enlarged by taking the union of both the labeled and the unlabeled dataset, *i.e.*, $\mathcal{S} = \mathcal{S}_{\mathrm{L}} \cup \{(\mathbf{I}_m, \hat{\mathbf{Y}}_m)\}_{m=l+1}^{M}$. The student model is trained on this augmented dataset $\mathcal{S}$ the same way we train the teacher model as described in Section 6.2.1. The overall training procedure is summarized in Algorithm 3. In the training stage, we first train a teacher model in a supervised manner and then use it to generate the pseudo-labels for the unlabeled dataset. Then we alternate the training of the student model and the pseudo-label generation procedures in an iterative manner to optimize the student model T times. In the testing stage, we follow the method in Section 6.2.2 to generate the final estimation using the T-th student model.

6.3 Experiments

6.3.1 Dataset and Evaluation

Our fully-labeled dataset includes 210 contrast-enhanced abdominal clinical CT images in the portal venous phase, in which we randomly choose 50/30/80 patients for training, validation, and testing, unless otherwise specified. A total of 16 structures (Aorta, Adrenal gland, Celiac AA, Colon, Duodenum, Gallbladder, Interior Vena Cava (IVC), Kidney (left, right), Liver, Pancreas, Superior Mesenteric Artery (SMA), Small bowel, Spleen, Stomach, Veins) for each case were segmented by four experienced radiologists, and confirmed by an independent senior expert. Our unlabeled dataset consists of 100 unlabeled cases acquired from a local hospital. To the best of our knowledge, this is the largest abdominal CT dataset with the most number of organs segmented. Each CT volume consists of $319 \sim 1051$ slices of 512×512 pixels, and have voxel spatial resolution of $([0.523 \sim 0.977] \times [0.523 \sim 0.977] \times 0.5)\text{mm}^3$. The metric we use is the Dice-Sørensen Coefficient (DSC), which measures the similarity between the prediction voxel set $\mathcal{Z}$ and the ground-truth set $\mathcal{Y}$, with the mathematical form of $\text{DSC}(\mathcal{Z}, \mathcal{Y}) = \frac{2\times|\mathcal{Z}\cap\mathcal{Y}|}{|\mathcal{Z}|+|\mathcal{Y}|}$. For each organ, we report an average DSC together with the standard deviation over all the testing cases.

6.3.2 Implementation Details

We set the learning rate to be 10^{-9}. The teacher model and the student model are trained for $80,000$ and $160,000$ iterations respectively. The validation set is used for tuning the hyper-parameters. Similar to [166], we use three windows of $[-125,\ 275]$, $[-160,\ 240]$, and $[-1000,\ 1000]$ Housefield Units as the three input channels respectively. The intensities of each slice are rescaled to [0.0, 1.0]. Similar to [13, 14, 18], we initialize the network parameters θ by using the FCN-8s model [41] pre-trained on the PascalVOC image segmentation dataset. The iteration number T in Algorithm 3 is set to 2, *i.e.*, $T = 2$, as the performance of the validation set gets saturated.

Table 6-I. The comparison of segmentation accuracy (DSC, %) by using 50 labeled data and varying the number of unlabeled data (*e.g.*, 50-0 indicates 50 labeled data and 0 unlabeled data). We report the mean and standard deviation over 80 cases. The *p*-values for testing significant difference between DMPCT (50-100) and FCN (50-0) are shown. Significant statistical improvement is shown in *italic* with $p < 0.05$. See Section 6.3.3 for definitions of FCN, SPSL, and DMPCT (Ours).

Organ Type	FCN	SPSL		DMPCT (Ours)		*p*-value
	50 - 0	50 - 50	50 - 100	50 - 50	50 - 100	
Aorta	89.14 ± 7.95	91.10 ± 5.52	90.76 ± 5.90	91.43 ± 4.88	91.54 ± 4.65	3.32×10^{-5}
Adrenal gland	26.45 ± 12.1	29.92 ± 14.7	26.93 ± 15.6	30.58 ± 12.7	35.48 ± 11.8	1.98×10^{-15}
Celiac AA	35.01 ± 19.7	37.27 ± 19.0	39.78 ± 18.4	36.25 ± 20.5	40.50 ± 18.9	1.00×10^{-5}
Colon	71.81 ± 14.9	78.28 ± 13.0	79.58 ± 12.9	79.61 ± 12.3	80.53 ± 11.6	7.69×10^{-12}
Duodenum	54.89 ± 15.5	57.77 ± 17.3	62.22 ± 14.8	66.95 ± 12.6	64.78 ± 13.8	1.95×10^{-19}
Gallbladder	86.53 ± 6.21	87.87 ± 5.45	88.02 ± 5.83	88.45 ± 5.07	87.77 ± 6.29	0.002
IVC	77.67 ± 9.49	81.28 ± 8.87	82.63 ± 7.31	83.49 ± 6.94	83.43 ± 7.02	9.30×10^{-14}
Kidney (L)	95.12 ± 5.01	95.59 ± 4.97	95.88 ± 3.68	95.82 ± 3.60	96.09 ± 3.42	3.69×10^{-6}
Kidney (R)	95.69 ± 2.36	95.77 ± 4.93	96.14 ± 2.94	96.17 ± 2.75	96.26 ± 2.29	1.74×10^{-7}
Liver	95.45 ± 2.41	96.06 ± 0.99	96.07 ± 1.03	96.11 ± 0.97	96.15 ± 0.92	0.005
Pancreas	76.49 ± 11.6	80.12 ± 7.52	80.93 ± 6.84	81.46 ± 6.32	82.03 ± 6.16	2.97×10^{-8}
SMA	52.26 ± 17.1	51.81 ± 18.2	51.94 ± 17.1	49.40 ± 19.2	52.70 ± 17.7	0.667
Small bowel	71.13 ± 13.1	78.93 ± 12.6	79.97 ± 12.8	79.49 ± 12.1	79.25 ± 12.6	2.53×10^{-22}
Spleen	94.81 ± 2.64	95.46 ± 2.09	95.58 ± 1.90	95.73 ± 2.03	95.98 ± 1.59	1.83×10^{-10}
Stomach	91.38 ± 3.94	92.62 ± 3.71	92.92 ± 3.65	93.33 ± 3.47	93.42 ± 3.21	3.30×10^{-23}
Veins	64.75 ± 15.4	70.43 ± 14.3	69.66 ± 14.6	69.82 ± 14.5	70.23 ± 14.4	4.16×10^{-15}
Mean	73.71 ± 9.97	76.32 ± 9.58	76.87 ± 9.08	77.20 ± 8.75	$\mathbf{77.94} \pm 8.51$	4.74×10^{-90}

6.3.3 Comparison with the Baseline

We show that our proposed DMPCT works better than other methods: 1) fully supervised learning method [41] (denoted as FCN), and 2) single planar based semi-supervised learning approach [10] (denoted as SPSL). Both 1) and 2) are applied on each individual plane separately, and then the final result is obtained via multi-planar fusion (see Sec 6.2.2). As shown in Table 6-I, with 50 labeled data, by varying the number of unlabeled data from 0 to 100, the average DSC of DMPCT increases from 73.71% to 77.94% and the standard deviation decreases from 9.97% to 8.51%. Compared with SPSL, our proposed DMPCT can boost the performance in both settings (*i.e.*, 50 labeled data + 50 unlabeled data and 50 labeled data + 100 unlabeled data). Besides, the *p*-values for testing significant difference between our DMPCT (50 labeled data + 100 unlabeled data) and FCN (50 labeled data + 0 unlabeled data)

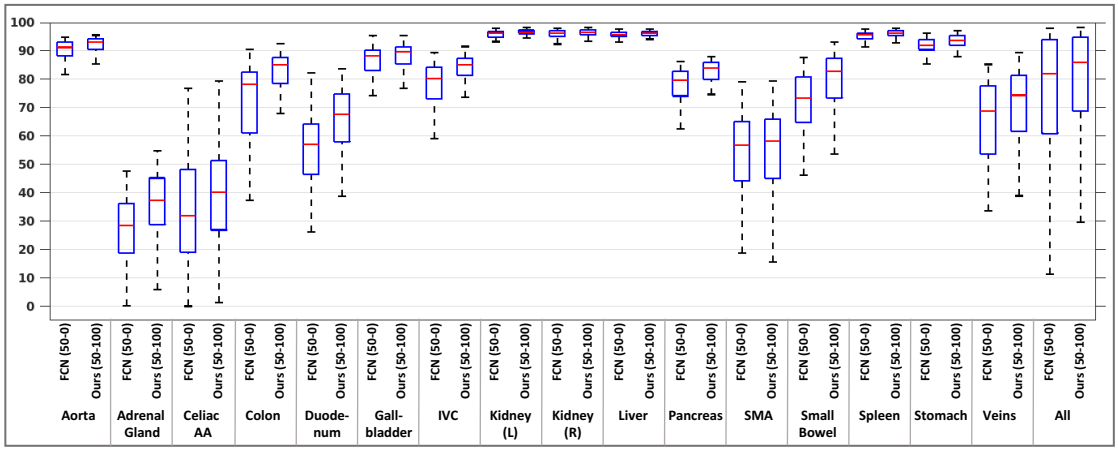


Figure 6-4. Performance comparison (DSC, %) in box plots of 16 organs by using 50 labeled data and varying the number of unlabeled data (*e.g.*, 50-0 indicates 50 labeled data and 0 unlabeled data). See Section 6.3.3 for definitions of FCN and DMPCT (Ours).

for 16 organs are shown in the last column of Table 6-I, which suggests significant statistical improvements among almost all organs. Figure 6-4 shows comparison results of our DMPCT and the fully supervised method by box plots.

It is noteworthy that greater improvements are observed especially for those difficult organs, *i.e.*, organs either small in sizes or with complex geometric characteristics. Table 6-I indicates that our DMPCT approach boosts the segmentation performance of these small hard organs by 5.54% (Pancreas), 8.72% (Colon), 9.89% (Duodenum), 8.12% (Small bowels) and 5.48% (Veins), 5.76% (IVC). This promising result indicates that our method distills a reasonable amount of knowledge from the unlabeled data. An example is shown in Figure 6-5. In this particular case, the DSCs for Celiac AA, Colon, Duodenum, IVC, Pancreas and Veins are boosted from 60.13%, 46.79%, 71.08%, 69.23%, 63.48% to 79.45%, 83.81%, 77.59%, 74.75%, 75.31% respectively.

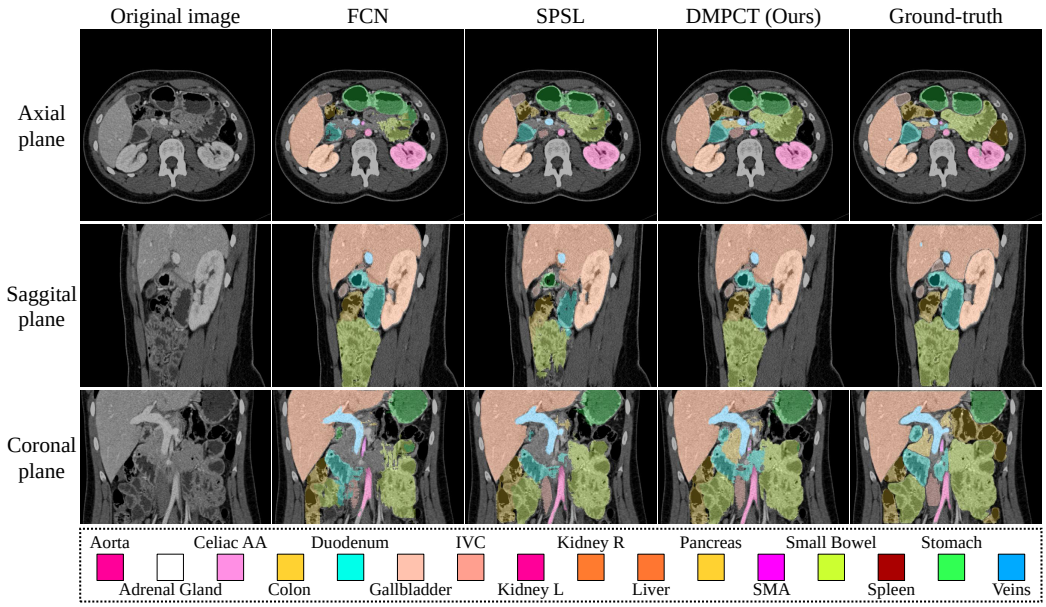


Figure 6-5. Comparisons among FCN, SPSL, and DMPCT (Ours) viewed from multiple planes. 50 labeled cases are used for all methods. 100 unlabeled cases are used for the SPSL and DMPCT. For this particular case, FCN obtains an average DSC of 72.75%, SPSL gets 78.87%, and DMPCT (Ours) gets 80.75%. See Section 6.3.3 for definitions of FCN, SPSL, and DMPCT (Ours). Best viewed in color.

6.3.4 Results and Discussion

Amount of labeled data. For ablation analysis, we enlarge the labeled training set to 100 cases and keep the rest of the settings the same. As shown in Figure 6-6, with more labeled data, the semi-supervised methods (DMPCT, SPSL) still obtain better performance than the supervised method (FCN), while the performance gain becomes less prominent. This is probably because the network is already trained well when large training set is available. We believe that if much more unlabeled data can be provided the performance should go up considerably. In addition, we find that DMPCT outperforms SPSL in every setting, which further demonstrates the usefulness of multi-planar fusion in our co-training framework.

Comparison with 3D network-based self-training. Various previous studies [15,

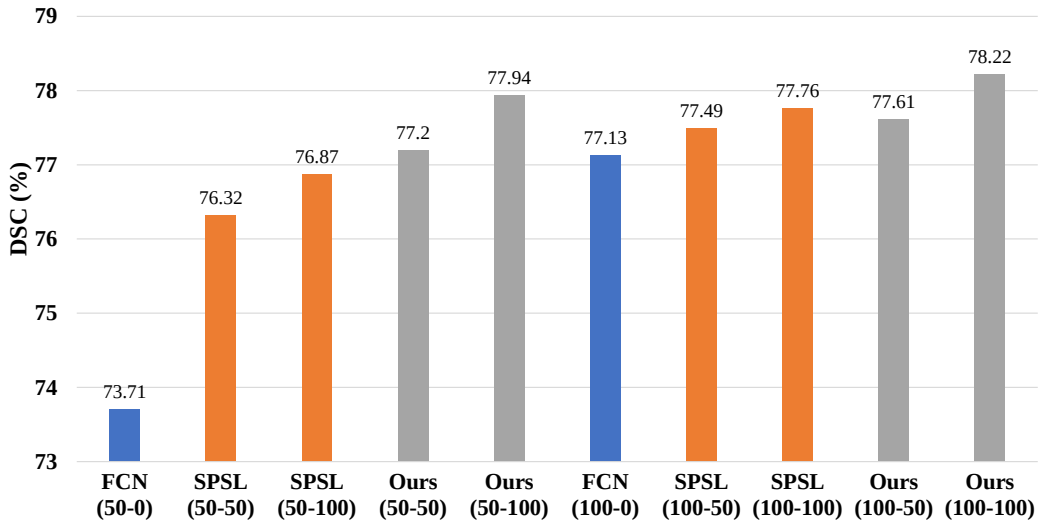


Figure 6-6. Ablation study on numbers of labeled data and unlabeled data. Mean DSC of all testing cases under all settings (*e.g.*, 50-0 indicates 50 labeled data and 0 unlabeled data). See Section 6.3.3 for definitions of FCN, SPSL, and DMPCT (Ours).

Table 6-II. Cross-dataset generalization results.

Organ	Spleen	Kidney (R)	Kidney (L)	Gall Bladder	Liver
FCN	71.85 ± 26.13	54.44 ± 20.04	54.98 ± 26.63	48.13 ± 26.07	85.46 ± 16.81
DMPCT (Ours)	**83.68** ± 16.53	**71.36** ± 20.85	**69.95** ± 20.50	**60.05** ± 26.91	**92.11** ± 6.46
Organ	Stomach	Aorta	IVC	Veins	Pancreas
FCN	38.89 ± 23.86	70.43 ± 19.70	53.67 ± 18.40	35.54 ± 18.94	39.40 ± 25.34
DMPCT (Ours)	**54.78** ± 26.57	**76.05** ± 15.99	**68.18** ± 14.58	**37.52** ± 15.86	**60.05** ± 16.61

18, 167] demonstrate that 2D multi-planar fusion outperforms directly 3D learning in the fully supervised setting. 3D CNNs come with an increased number of parameters, significant memory and computational requirements. Due to GPU memory restrictions, these 3D CNN approaches which adopt the sliding-window strategy do not act on the entire 3D CT volume, but instead on local 3D patches [48, 65, 143]. This results in the lack of holistic information and low efficiency. In order to prove that DMPCT outperforms direct 3D learning in the semi-supervised setting, we also implement a patch-based 3D UNet [143]. 3D UNet gets 69.66% in terms of mean DSC using 50 labeled data. When adding 100 unlabeled data the performance even drops to 65.21%. This clearly shows that in 3D learning the teacher model is not trained well, thus the

errors of the pseudo-labels are reinforced during student model training.

Comparison with traditional co-training. In order to show that our DMPCT outperforms traditional co-training algorithm [12], we also select only the most confident samples during each iteration. Here the confidence score is measured by the entropy of probability distribution for each voxel in one slice. Under the setting of 50 labeled cases and 50 unlabeled cases, we select top 5000 samples with the highest confidence in each iteration. The whole training process takes about 6-7 iterations for each plane. The complete training requires more than 50 hours. Compared with our approach, this method requires much more time to converge. It obtains a mean DSC of 76.52%, slightly better than SPSL but worse than our DMPCT, which shows that selecting the most confident samples during training may not be a wise choice for deep network based semi-supervised learning due to its low efficiency.

Cross dataset generalization We apply our trained DMPCT model (50 labeled data + 100 unlabeled data) and baseline FCN model (50 labeled data + 0 unlabeled data) on a public available abdominal CT datasets[1] with 13 anatomical structures labeled *without any further re-training* on new data cases. 10 out of the 13 structures are evaluated which are also manually annotated in our own dataset and we find that our proposed method improves the overall mean DSC and also reduces the standard deviation significantly, as shown in Table 6-II. The overall mean DSC as well as the standard deviation for the 10 organs is improved from $59.23 \pm 22.20\%$ to $67.38 \pm 19.64\%$. We also directly test our models on the NIH pancreas segmentation dataset of 82 cases[2] and observe that our DMPCT model achieves an average DSC of 66.16%, outperforming the fully supervised method, with an average DSC of 58.73%, by more than 7%. This may demonstrate that our approach, which leverages more unlabeled data from multiple planes, turns out to be much more generalizable than

[1] 30 training data sets at https://www.synapse.org/#!Synapse:syn3193805/wiki/217789
[2] https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

the baseline model.

Computation time. In our experiments, the teacher model training process takes about 4.94 hours on an NVIDIA TITAN Xp GPU card for $80,000$ iterations over all the training cases. The average computation time for generating pseudo-label as well as testing per volume depends on the volume of the target structure, and the average computation time for 16 organs is approximately 4.5 minutes, which is comparable to other recent methods [8, 13] even for single structure inference. The student model training process takes about 9.88 hours for $160,000$ iterations.

6.4 Summary

In this chapter, we present a systematic framework DMPCT for multi-organ segmentation in abdominal CT scans, which is motivated by the traditional co-training strategy to incorporate multi-planar information for the unlabeled data during training. The pseudo-labels are iteratively updated by inferencing comprehensively on multiple configurations of unlabeled data with a multi-planar fusion module. We evaluate our approach on our own large newly collected high-quality dataset. The results show that 1) our method outperforms the fully supervised learning approach by a large margin; 2) it outperforms the single planar method, which further demonstrates the benefit of multi-planar fusion; 3) it can learn better if more unlabeled data provided especially when the scale of labeled data is small.

Our framework can be practical in assisting radiologists for clinical applications since the annotation of multiple organs in 3D volumes requires massive labor from radiologists. Our framework is not specific to a certain structure, but shows robust results in multiple complex anatomical structures within efficient computational time. It can be anticipated that our algorithm may achieve even higher accuracy if a more powerful backbone network or an advanced fusion algorithm is employed, which we

leave as the future work.

Chapter 7

Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation

Accurate multi-organ abdominal CT segmentation is essential to many clinical applications such as computer-aided intervention. As data annotation requires massive human labor from experienced radiologists, it is common that training data are partially labeled, *e.g.*, pancreas datasets only have the pancreas labeled while leaving the rest marked as background. However, these background labels can be misleading in multi-organ segmentation since the "background" usually contains some other organs of interest. To address the background ambiguity in these partially-labeled datasets, we propose Prior-aware Neural Network (PaNN) via explicitly incorporating anatomical priors on abdominal organ sizes, guiding the training process with domain-specific knowledge. More specifically, PaNN assumes that the average organ size distributions in the abdomen should approximate their empirical distributions, prior statistics obtained from the fully-labeled dataset. As our training objective is difficult to be directly optimized using stochastic gradient descent, we propose to reformulate it in a min-max form and optimize it via the stochastic primal-dual gradient algorithm. PaNN achieves state-of-the-art performance on the MICCAI2015 challenge "Multi-Atlas Labeling Beyond the Cranial Vault", a competition on organ

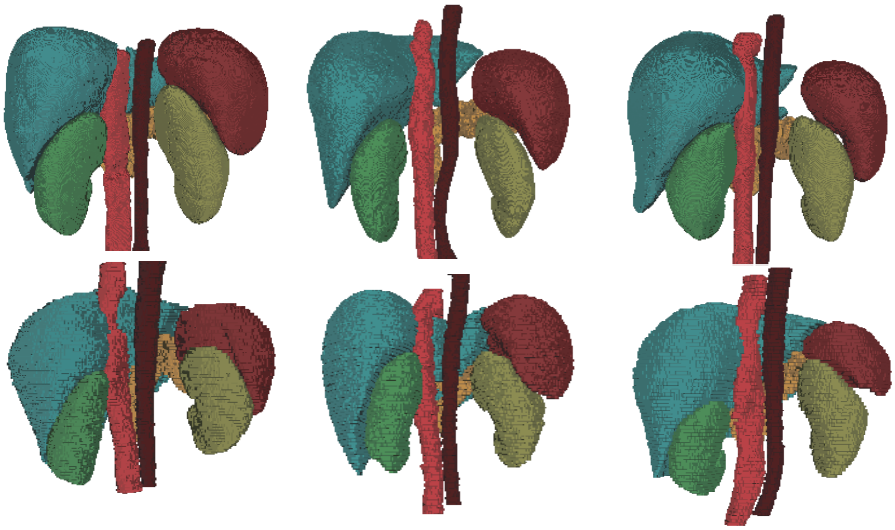


Figure 7-1. 3D Visualization of several abdominal organs (liver, spleen, left kidney, right kidney, aorta, inferior vena cava) to show the similarity of patient-wise abdominal organ size distributions.

segmentation in the abdomen. We report an average Dice score of 84.97%, surpassing the prior art by a large margin of 3.27%.

7.1 Introduction

This work focuses on multi-organ segmentation in abdominal regions which contain multiple organs such as liver, pancreas and kidneys. The segmentation of internal structures on medical images, *e.g.*, CT scans, is an essential prerequisite for many clinical applications such as computer-aided diagnosis, computer-aided intervention and radiation therapy. Compared with other internal structures such as heart or brain, abdominal organs are much more difficult to segment due to the morphological and structural complexity, low contrast of soft tissues, *etc.*

With the development of deep convolutional neural networks (CNNs), many medical image segmentation problems have achieved satisfactory results only when full-supervision is available [8, 13, 14, 66, 101, 102]. Despite the recent progress, the annotation of medical radiology images is extremely expensive, as it must be handled

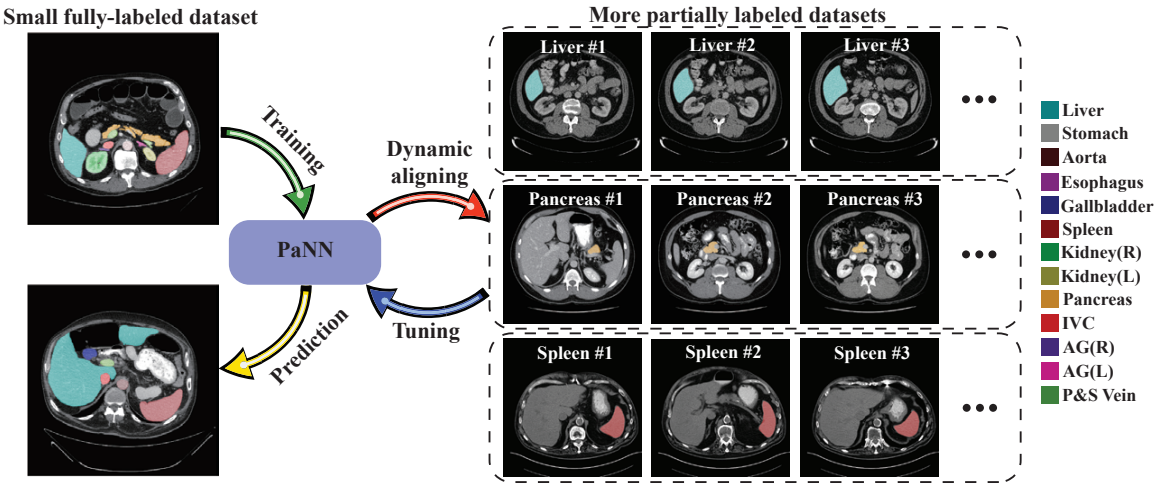


Figure 7-2. Overview of the proposed PaNN for partially-supervised multi-organ segmentation. It is trained with a small set of fully-labeled dataset and several partially-labeled datasets. The PaNN regularizes that the organ size distributions of the network output should approximate their prior statistics in the abdominal region obtained from the fully-labeled dataset.

by experienced radiologists and carefully checked by additional experts. This results in the lack of high-quality labeled training data. More critically, how to efficiently incorporate domain-specific expertise (*e.g.*, anatomical priors) with segmentation models [122, 127], such as the organ shape, size, remains an open issue.

Our key observation is that, in medical image analysis domain, instead of scribbles [168–170], points [171] and image-level tags [97, 98, 172], there exists a considerable number of datasets in the form of abdominal CT scans [7, 66, 173]. To meet different research goals or practical usages, these datasets are annotated to target different organs (a subset of abdominal organs), *e.g.*, pancreas datasets [7] only have the pancreas labeled while leaving the rest marked as background.

The aim of this work is to fully leverage these existing partially-annotated datasets to assist multi-organ segmentation, which we refer to as *partial supervision*. To address the challenge of partial supervision, an intuitive solution is to simply train a segmentation model directly on both the labeled data and the partially-labeled data in the semi-supervised manner [10, 97, 120]. However, it 1) fails to take advantages of

the fact that medical images are naturally more constrained compared with natural images [5]; 2) is intuitively misleading as it treats the unlabeled pixels/voxels as background. To overcome these issues, we propose Prior-aware Neural Network (PaNN) to handle such background ambiguity via incorporating prior knowledge on organ size distributions. We achieve this via a prior-aware loss, which acts as an auxiliary and soft constraint to regularize that the average output size distributions of different organs should approximate their prior proportions. Based on the anatomical similarities (Figure 7-1) across different patient scans [121, 122, 127], the prior proportions are estimated by statistics from the fully-labeled data. The overall pipeline is illustrated in Figure 7-2. It is important to note that the training objective is hard to be directly optimized using stochastic gradient descent. To address this issue, we propose to formulate our objective in a min-max form, which can be well optimized via the stochastic primal-dual gradient algorithm [174]. To summarize, our contributions are three-fold:

1) We propose Prior-aware Neural Network, which incorporates domain-specific knowledge from medical images, to facilitate multi-organ segmentation via using partially-annotated datasets.

2) As the training objective is difficult to be directly optimized using stochastic gradient descent, it is essential to reformulate it in a min-max form and optimize via stochastic primal-dual gradient [174].

3) PaNN significantly outperforms previous state-of-the-arts even using fewer annotations. It achieves 84.97% on the MICCAI2015 challenge "Multi-Atlas Labeling Beyond the Cranial Vault", outperforming prior arts by a large margin.

7.2 Prior-aware Neural Network

Our work aims to address the multi-organ segmentation problem with the help of multiple existing partially-labeled datasets. Given a CT scan where each element indicates the Housefield Unit (HU) of a voxel, the goal is to find the predicted labelmap of each pixel/voxel.

7.2.1 Partial Supervision

We consider a new supervision paradigm, *i.e.*, partial supervision, for multi-organ segmentation. This is motivated by the fact that there exists a considerable number of datasets with only one or a few organs labeled in the form of abdominal CT scans [7, 66, 173] in medical image analysis, which can serve as partial supervision for multi-organ segmentation (see the list in the appendix). Based on domain knowledge, our approach assumes the following characteristics of the datasets which are common in medical image analysis. First, the scanning protocols of medical images are well standardized, *e.g.*, brain, head and neck, chest, abdomen, and pelvis in CT scans, which means that the internal structures are consistent in a limited range according to the scanning protocol (see Figure 7-1). Second, internal organs have anatomical and spatial relationships such as gastrointestinal track, *i.e.*, stomach, duodenum, small intestine, and colon are connected in a fixed order.

The partially-supervised setting can be formally defined as below. Given a fully-labeled dataset $\mathbf{S}_{\mathrm{L}} = \{\mathbf{I}_{\mathrm{L}}, \mathbf{Y}_{\mathrm{L}}\}$ with the annotation $\mathbf{Y}_{\mathrm{L}}$ known and T partially-labeled datasets $\mathbf{S}_{\mathrm{P}} = \{\mathbf{S}_{\mathrm{P}_1}, \mathbf{S}_{\mathrm{P}_2}, ...\mathbf{S}_{\mathrm{P}_T}\}$ with the t-th dataset defined as $\mathbf{S}_{\mathrm{P}_t} = \{\mathbf{I}_{\mathrm{P}_t}, \mathbf{Y}_{\mathrm{P}_t}\}$. $\mathrm{L} = \{1, 2, ..., n_{\mathrm{L}}\}$ and $\mathrm{P}_t = \{1, 2, ..., n_{\mathrm{P}_t}\}$ denote the image indices for $\mathbf{S}_{\mathrm{L}}$ and $\mathbf{S}_{\mathrm{P}_t}$, respectively. For each element $y_{ij} \in \mathbf{Y}_{\mathrm{L}}$, y_{ij} denotes the annotation of the j-th pixel in the i-th image $\mathbf{I}_i \in \mathbf{I}_{\mathrm{P}_t}$ and is selected from $\mathcal{L}$, where $\mathcal{L}$ denotes the abdominal organ space, *i.e.*, $\mathcal{L} = \{\text{spleen}, \text{pancreas}, \text{liver}, ...\}$. For the t-th partially-labeled dataset $\mathbf{S}_{\mathrm{P}_t}$,

$y_{ij} \in \mathbf{Y}_{\mathrm{P}_t}$ is selected from $\mathcal{L}_{\mathrm{P}_t} \subseteq \mathcal{L}$. In 2D-based segmentation models, the i-th input $\mathbf{I}_i$ is a sliced 2D image from either Axial, Coronal or Saggital view of the whole CT scan [8, 13, 18, 19, 128]. In 3D-based segmentation models, $\mathbf{I}_i$ is a cropped 3D patch from the whole CT volume [78, 143]. Note that semi-supervision and fully-supervision are two extreme cases of partial supervision, when the set of partial labels is an empty set ($\mathcal{L}_{\mathrm{P}_t} = \oslash$) and is equal to the complete set ($\mathcal{L}_{\mathrm{P}_t} = \mathcal{L}$), respectively.

A naive solution is to simply train a segmentation network from both the fully-labeled data and the partially-labeled data and alternately update the network parameters and the segmentations (pseudo-labels) for the partially-labeled data [19][10]. While these EM-like approaches have achieved significant improvement compared with fully-supervised methods, they require high-quality pseudo-labels and fail to explicitly incorporate anatomical priors on shape or size.

To address this issue, we propose a Prior-aware Neural Network (PaNN), aiming at explicitly embedding anatomical priors without incurring any additional budget. More specifically, the anatomical priors are enforced by introducing an additional penalty which acts as a soft constraint to regularize that the average output distributions of organ sizes should mimic their empirical proportions. This prior is obtained by calculating the organ size statistics of the fully-labeled dataset. An overview of the overall framework is shown in Figure 7-2, and the detailed training procedures will be introduced in the following sections.

7.2.2 Prior-aware Loss

Consider a segmentation network parameterized by $\mathbf{\Theta}$, which outputs probabilities $\mathbf{p}$. Let $\mathbf{q} \in \mathbb{R}^{(|\mathcal{L}|+1)\times 1}$ be the label distribution in the fully-labeled dataset, with q^l describing the proportion of the l-th label (organ). Then, we estimate the average

predicted distribution of the pixels in the partially-labeled datasets as

$$\bar{\mathbf{p}} = \frac{1}{N} \sum_{t=1}^{T} \sum_{i \in \mathrm{P}_t} \sum_{j} \mathbf{p}_{ij}, \tag{7.1}$$

where $\mathbf{p}_{ij} = [p_{ij}^0, p_{ij}^1, ..., p_{ij}^{|\mathcal{L}|}]$ denotes the probability vector of the j-th pixel in the i-th input slice $\mathbf{I}_i$, and N is the total number of pixels/voxels. Recall that T is the total number of partially-labeled datasets.

To embed the prior knowledge, the prior-aware loss is defined as

$$\begin{aligned} &\mathrm{KL}_{\mathrm{marginal}}(\mathbf{q}|\bar{\mathbf{p}}) \triangleq \textstyle\sum_l \mathrm{KL}(q^l|\bar{p}^l) \\ &= -\textstyle\sum_l \Big(q^l \log \bar{p}^l + (1-q^l)\log(1-\bar{p}^l)\Big) + const \\ &= -\{\mathbf{q} \log \bar{\mathbf{p}} + (1-\mathbf{q})\log(1-\bar{\mathbf{p}})\} + const, \end{aligned} \tag{7.2}$$

which measures the matching probability of the two distributions $\mathbf{q}$ and $\bar{\mathbf{p}}$ via Kullback-Leibler divergence. Note that each class is treated as one vs. rest when calculating the matching probabilities. Therein, the rationale of Eqn. (7.2) is that the output distributions $\bar{\mathbf{p}}$ of different organ sizes should approximate their empirical marginal proportions $\mathbf{q}$, which generally reflects the domain-specific knowledge.

Note that $\mathbf{q}$ is a global estimation of label distribution of the fully-labeled training data, which remains unchanged. Consequently, $\mathrm{H}(\mathbf{q})$ is constant which can be omitted during the network training. Nevertheless, we observe that it is still problematic to directly apply stochastic gradient descent, as we will detail in Section 7.2.3.

Specifically in our case, our final training objective is

$$\min_{\boldsymbol{\Theta}, \mathbf{Y}_\mathrm{P}} \mathcal{J}_\mathrm{L}(\boldsymbol{\Theta}) + \lambda_1 \mathcal{J}_\mathrm{P}(\boldsymbol{\Theta}, \mathbf{Y}_\mathrm{P}) + \lambda_2 \mathcal{J}_\mathrm{C}(\boldsymbol{\Theta}), \tag{7.3}$$

where $\mathcal{J}_\mathrm{L}(\boldsymbol{\Theta})$ and $\mathcal{J}_\mathrm{P}(\boldsymbol{\Theta}, \mathbf{Y}_\mathrm{P})$ are the cross entropy loss on the fully-labeled data and the partially-labeled data, respectively. And $\mathbf{Y}_\mathrm{P}$ denotes the computed pseudo-labels as well as existing partial labels from the partially-labeled dataset(s). Note that the prior-aware loss $\mathcal{J}_\mathrm{C}$ is used as a soft global constraint to stablize the training process. Concretely, $\mathcal{J}_\mathrm{L}(\boldsymbol{\Theta})$ is defined as

$$\mathcal{J}_\mathrm{L} = -\frac{1}{N} \sum_{i \in \mathrm{L}} \sum_{j} \sum_{l=0}^{|\mathcal{L}|} \mathbb{1}(y_{ij} = l) \log p_{ij}^l, \tag{7.4}$$

where p_{ij}^{l} denotes the softmax probability of the j-th pixel in the i-th image to the l-th category. $\mathcal{J}_{\mathrm{P}}(\boldsymbol{\Theta}, \mathbf{Y}_{\mathrm{P}})$ is given by

$$\begin{aligned}\mathcal{J}_{\mathrm{P}} = -\frac{1}{N}\sum_{t=1}^{T}\sum_{i\in \mathrm{P}_t}\sum_{j}\sum_{l=0}^{|\mathcal{L}|}\{\mathbb{1}(y_{ij}=l)\log p_{ij}^{l} \\ +\mathbb{1}(y'_{ij}=l)\log p_{ij}^{l}\},\end{aligned} \tag{7.5}$$

where the first term corresponds to the pixels with their labels $\mathbf{Y}_{\mathrm{P}}$ given, *i.e.*, $y_{ij} \in \mathcal{L}_{\mathrm{P}_t}$. The second term corresponds to unlabeled background pixels, and $\mathbf{Y}_{\mathrm{P}}$ needs to be estimated during the model training as a kind of pseudo-supervision, *i.e.*, $y'_{ij} \in \mathcal{L} - \mathcal{L}_{\mathrm{P}_t}$.

7.2.3 Derivation

By substituting Eqn. (7.1) into Eqn. (7.2) and expanding $\mathbf{q}, \bar{\mathbf{p}}$ into scalars, we rewrite Eqn. (7.2) as

$$\begin{aligned}\mathcal{J}_{\mathrm{C}} = -\sum_{l=0}^{|\mathcal{L}|}\{q^{l}\log\frac{1}{N}\sum_{t=1}^{T}\sum_{i\in \mathrm{P}_t}\sum_{j}p_{ij}^{l} + \\ (1-q^{l})\log(1-\frac{1}{N}\sum_{t=1}^{T}\sum_{i\in \mathrm{P}_t}\sum_{j}p_{ij}^{l})\} + const.\end{aligned} \tag{7.6}$$

From Eqn. (7.2) and Eqn. (7.6) we can see that the average distribution $\bar{\mathbf{p}}$ of organ sizes is inside the logarithmic loss, which is very different from standard machine learning loss such as Eqn. (7.4) and Eqn. (7.5) where the average is outside logarithmic loss. And directly minimizing by stochastic gradient descent is very difficult as the true gradient induced by Eqn. (7.2) is not a summation of independent terms, the stochastic gradients would be intrinsically biased [174].

To remedy this, we propose to optimize the KL divergence term using stochastic primal-dual gradient [174]. Our goal here is to transform the prior-aware loss into an equivalent min-max problem by taking the sample average out of the logarithmic loss. We introduce two auxiliary variables to assist the optimization, *i.e.*, the primal variable α and the dual variable β. First, the following identity holds

$$-\log\alpha = \max_{\beta}\left(\alpha\beta + 1 + \log(-\beta)\right) \tag{7.7}$$

Algorithm 4 The training procedure of PaNN

Require: Fully-labeled training data $\mathbf{S}_{\mathrm{L}}$;
Require: Partially-labeled training data $\mathbf{S}_{\mathrm{P}}$;
Require: Hyperparameters: λ_1, λ_2;
Require: Maximum training iteration K;
Ensure: Segmentation model $\boldsymbol{\Theta}$;
  Train the segmentation model $\boldsymbol{\Theta}$ on $\mathbf{S}_{\mathrm{L}}$;
  Compute the prior distribution $\mathbf{q}$ on $\mathbf{S}_{\mathrm{L}}$;
  Initialize $\boldsymbol{\nu} = -1/\mathbf{q}$ and $\boldsymbol{\mu} = 1/(1-\mathbf{q})$;
  repeat
    Estimate pesudo-labels $\mathbf{Y}_{\mathrm{P}}$ with $\boldsymbol{\Theta}$;
    Update $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ via stochastic gradient ascent;
    Update $\boldsymbol{\Theta}$ via stochastic gradient descent;
  until Training iteration reaches K;
  return $\boldsymbol{\Theta}$

due to the property of the log function. Based on Eqn. (7.7), we define $\boldsymbol{\nu} \in \mathbb{R}^{|\mathcal{L}|\times 1}$ as the dual variable associated to the primal variable $\bar{\mathbf{p}}$, and define $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{L}|\times 1}$ as the dual variable associated to the primal variable $(1-\bar{\mathbf{p}})$. Then, we have

$$\begin{aligned} -\log \bar{p}^l &= \max_{\nu^l} \left(\bar{p}^l \nu^l + 1 + \log(-\nu^l) \right) \\ -\log(1-\bar{p}^l) &= \max_{\mu^l} \left((1-\bar{p}^l)\mu^l + 1 + \log(-\mu^l) \right), \end{aligned} \tag{7.8}$$

where ν^l (or μ^l) denotes the l-th element of $\boldsymbol{\nu}$ (or $\boldsymbol{\mu}$). Substituting them into Eqn. (7.2)/Eqn. (7.6), maximizing the KL divergence is equivalent to the following min-max optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\Theta}} \max_{\boldsymbol{\nu},\boldsymbol{\mu}} & \sum_l q^l \left(\bar{p}^l \nu^l + 1 + \log(-\nu^l) \right) \\ & + \sum_l (1-q^l) \left((1-\bar{p}^l)\mu^l + 1 + \log(-\mu^l) \right) \\ \Leftrightarrow \min_{\boldsymbol{\Theta}} \max_{\boldsymbol{\nu},\boldsymbol{\mu}} & \sum_l \left(q^l \nu^l - (1-q^l)\mu^l \right) \bar{p}^l + q^l \log(-\nu^l) \\ & + \sum_l (1-q^l) \left(\mu^l + \log(-\mu^l) \right), \end{aligned} \tag{7.9}$$

which brings the sample average out of the logarithmic loss. Note that we ignore the constant in the above formulas.

7.2.4 Model Training

We consider training a fully convolutional network [41, 42, 101] for multi-organ segmentation, where the input images are either 2D slices [8, 13, 19, 128] or 3D cropped patches [78, 143]. The training procedure can be divided into two stages.

In the first stage, we only train on the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$ by optimizing Eqn. (7.4) via stochastic gradient descent (also means $\lambda_1 = 0$ and $\lambda_2 = 0$ in Eqn. (7.3)). The goal of this stage is to find a proper initialization $\boldsymbol{\Theta_0}$ for the network weights, which stabilizes the later training procedure.

In the second stage, we train the model on the union of the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$ and partially-labeled dataset(s) $\mathbf{S}_{\mathrm{P}}$ via Eqn. (7.3). As can be drawn, we have two groups of variables, *i.e.*, the network weights $\boldsymbol{\Theta}$ and the three auxiliary variables $\{\boldsymbol{\nu}, \boldsymbol{\mu}, \mathbf{Y}_{\mathrm{P}}\}$. We adopt an alternating optimization, which can be decomposed into two subproblems:

• **Fixing $\boldsymbol{\Theta}$, Updating $\{\boldsymbol{\nu}, \boldsymbol{\mu}, \mathbf{Y_P}\}$.** With the network weights $\boldsymbol{\Theta}$ given, we can first estimate the pesudo-labels $\mathbf{Y}_{\mathrm{P}}$ of background pixels in the partially-labeled dataset(s) $\mathbf{S}_{\mathrm{P}}$. Meanwhile, the optimization of $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ is a maximization problem. Hence, we do stochastic gradient *ascent* to learn $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$. As for the initialization, we set $\boldsymbol{\nu}$ to $-1/\mathbf{q}$ and set $\boldsymbol{\mu}$ to $-1/(1-\mathbf{q})$, respectively.

• **Fixing $\{\boldsymbol{\nu}, \boldsymbol{\mu}, \mathbf{Y_P}\}$, Updating $\boldsymbol{\Theta}$.** By fixing the three auxiliary variables, we can then update the network weights $\boldsymbol{\Theta}$ via the standard stochastic gradient *descent*.

As can be seen, our algorithm is formulated as a min-max optimization. We summarize the detailed procedure of optimization in Algorithm 4.

7.3 Experiments

7.3.1 Experiment Setup

Datasets and Evaluation Metric. We use the training set released in the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge as the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$, which contains 30 abdominal CT scans with 3779 axial contrast-enhanced abdominal clinical CT images in total. For each case, 13 anatomical structures are annotated, including spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava (IVC), portal vein & splenic vein, pancreas, left adrenal gland, right adrenal gland. Each CT volume consists of $85 \sim 198$ slices of 512×512 pixels, with a voxel spatial resolution of $([0.54 \sim 0.54] \times [0.98 \sim 0.98] \times [2.5 \sim 5.0])\mathrm{mm}^3$.

As for the partially-labeled dataset(s) $\mathbf{S}_{\mathrm{P}}$, we use a spleen segmentation dataset[1] (referred as **A**), a pancreas segmentation dataset[2] (referred as **B**) and a liver segmentation dataset[1] (referred as **C**). To make these partially-labeled datasets balanced, 40 cases are evenly selected from each dataset to constitute the partial supervision.

Following the standard cross-validation evaluation [8, 13, 66, 128, 175], we randomly partition the fully-labeled dataset $\mathbf{S}_L$ into 5 complementary folds, each of which contains 6 cases, then apply the standard 5-fold cross-validation. For each fold, we use 4 folds (*i.e.*, 24 cases) as full supervision and test on the remaining fold.

The evaluation metric we use is the Dice-Sørensen Coefficient (DSC), which measures the similarity between the prediction voxel set $\mathcal{Z}$ and the ground-truth set $\mathcal{Y}$. Its mathematical definition is $\mathrm{DSC}(\mathcal{Z}, \mathcal{Y}) = \frac{2\times|\mathcal{Z}\cap\mathcal{Y}|}{|\mathcal{Z}|+|\mathcal{Y}|}$. We report an average DSC of all the testing cases over the 13 labeled anatomical structures for performance evaluation.

Implementation Details. Similar to [8, 13, 15, 19, 66], we use the soft tissue CT window range of $[-125, 275]$ HU. The intensities of each slice are then rescaled

[1] Available at http://medicaldecathlon.com
[2] Available at https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

to $[0.0, 255.0]$. Random rotation of $[0, 15]$ is used as an online data augmentation. Our implementations are based on the current state-of-the-art 2D[3] [94, 95] and 3D models[4] [101, 176]. We provide an extensive study about how partially-labeled datasets facilitate multi-organ segmentation task and list thorough comparisons under different settings.

As described in Section 7.2.4, the whole training procedure is divided into two stages. The first stage is the same as fully-supervised training, *i.e.*, we train exclusively on the fully-labeled dataset $\mathcal{S}_{\mathrm{L}}$ for a certain number of iterations M1.

In the second stage, we switch to the min-max optimization on the union of the fully-labeled dataset and partially-labeled datasets for M2 iterations. In each mini-batch, the sampling rate of labeled data and partially-labeled data is $3:1$. It has been suggested [10] that it is less necessary to update the pseudo-label $\mathbf{Y}_{\mathrm{P}}$ per iteration. Hence, $\mathbf{Y}_{\mathrm{P}}$ is updated every 10K iterations in practice. In addition, the hyperparameters λ_1 and λ_2 are set to be 1.0 and 0.1, respectively. The same decay policy of learning rate is utilized as that used in the first stage. In the second stage, the initial learning rate for the minimization step and the maximization step are set as 10^{-5} and 2×10^{-5}, respectively.

For 2D implementations, the initial learning rate of the first stage is 2×10^{-5} and a *poly* learning rate policy is employed. M1 and M2 are set as 40K and 30K, respectively. Following [6, 66, 95], we apply multi-scale inputs (scale factors are $\{0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}$) in both training and testing phase. For 3D implementations, the initial learning rate of the first stage is $5e^{-4}$ and a fixed learning rate policy is employed. M1 and M2 are set as 80K and 100K, respectively.

Model	Supervision	Partially-labeled dataset			Average Dice
		A	B	C	
ResNet50 [37]	Full				0.7535
	Semi [10]	✓			0.7593
			✓		0.7632
				✓	0.7596
		✓	✓	✓	**0.7669**
	Partial **(ours)**	✓			0.7650
			✓		0.7662
				✓	0.7631
		✓	✓	✓	**0.7705**
	PaNN **(ours)**	✓			0.7716
			✓		0.7712
				✓	0.7705
		✓	✓	✓	**<u>0.7833</u>**
ResNet101 [37]	Full				0.7614
	Semi [10]	✓			0.7637
			✓		0.7649
				✓	0.7647
		✓	✓	✓	**0.7719**
	Partial **(ours)**	✓			0.7714
			✓		0.7695
				✓	0.7684
		✓	✓	✓	**0.7735**
	PaNN **(ours)**	✓			0.7770
			✓		0.7819
				✓	0.7748
		✓	✓	✓	**<u>0.7904</u>**
3D-UNet [143]	3D-UNet-fully-sup				0.7066
	Semi [10]	✓	✓	✓	0.7193
	Partial **(ours)**	✓	✓	✓	0.7163
	PaNN **(ours)**	✓	✓	✓	**<u>0.7208</u>**

Table 7-I. Performance comparison (DSC) with fully-supervised and semi-supervised methods. **<u>Bold underline</u>** denotes the best results, **bold** denotes the second best results.

7.3.2 Experimental Comparison

We compare the proposed PaNN with a series of state-of-the-art algorithms, including 1) the fully-supervised approach (denoted as "-fully-sup"), where we train exclusively only on the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$, 2) the semi-supervised approach (denoted as "-semi-sup"), where we train the network on both the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$ and the partially-labeled dataset(s) $\mathbf{S}_{\mathrm{P}}$ while treating $\mathbf{S}_{\mathrm{P}}$ as unlabeled following the representative method [10], and 3) the naive partially-supervised approach (denoted as "-partial-

[3] https://github.com/tensorflow/models/tree/master/research/deeplab
[4] https://github.com/DLTK/DLTK

sup"), where we also train the network on both $\mathbf{S}_{\mathrm{L}}$ and $\mathbf{S}_{\mathrm{P}}$ while treating the partial labels as they are. Different from PaNN, we set $\lambda_2 = 0$ in Eqn. (7.3) to verify the efficacy of the prior-aware loss.

Benefit of Partial Supervision. As shown from Table 7-I, among three kinds of supervisions, partial supervision obtains the best performance followed by the semi-supervision and full supervision. It is no surprise to observe such a phenomenon for two reasons. First, compared with full supervision, semi-supervision has more training data, though part of them is not annotated. Second, compared with semi-supervision, partial supervision involves more annotated pixels in the organ of interest.

Effect of PaNN. From Table 7-I, PaNN generally achieves better performance than the naive partially-supervised methods, which demonstrates the effectiveness of our proposed PaNN. For example, when setting the partial dataset as the union of **A**, **B** and **C**, PaNN achieves the best result either using 2D models or 3D models. 2D models generally observe a better performance in each setting compared with 3D models. This is probably due to the fact that current 3D models only act on local patches (*e.g.*, $64 \times 64 \times 64$), which results in lacking holistic information [18]. A detailed discussion of 2D and 3D models is listed in [80]. More specifically, PaNN outperforms the naive partially-supervised method by 1.28% with ResNet-50 and by 1.69% with ResNet-101 as the backbone model, respectively. Additionally, we also observe a convincing performance gain of 0.45% using 3D UNet [101, 143] as the backbone model.

Meanwhile, by increasing the number of partially-labeled datasets (from using only **A**, **B** or **C** to the union of three), the performance improvements of different methods are also different. For example, with the ResNet-101 as the backbone, the largest improvement obtained under semi-supervision is 0.82% (from 76.37% to 77.19%), and that of partial supervision is 0.51% (from 76.84% to 77.35%). By contrast, PaNN

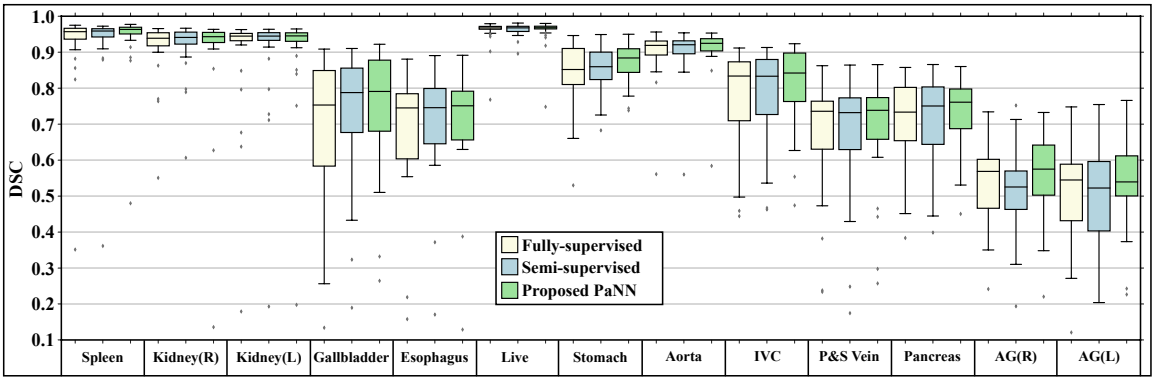


Figure 7-3. Performance comparison (DSC) in box plots of 13 abdominal structures, where the partially-labeled dataset **C** is used with ResNet-50 as the backbone model. Our proposed PaNN improves the overall mean DSC and also reduces the standard deviation. Kidney/AG (R), Kidney/AG (L) stand for the right and left kidney/adrenal gland, respectively.

Name	Spleen	Kidney(R)	Kidney(L)	Gallbladder	Esophagus	Aorta	IVC	Average Dice	Mean Surface Distance	Hausdorff Distance
AutoContext3DFCN [66]	0.926	0.866	0.897	0.629	0.727	0.852	0.791	0.782	1.936	26.095
deedsJointCL [177]	0.920	0.894	0.915	0.604	0.692	0.857	0.828	0.790	2.262	25.504
dltk0.1_unet_sub2 [176]	0.939	0.895	0.915	**0.711**	0.743	0.891	0.826	0.815	1.861	62.872
results_13organs_p0.7	0.890	0.898	0.883	0.685	0.754	0.870	0.819	0.817	4.559	38.661
PaNN* (**ours**)	**0.961**	**0.901**	**0.943**	0.704	**0.783**	**0.913**	**0.835**	**0.832**	**1.641**	**25.176**
PaNN (**ours**)	**0.968**	**0.920**	**0.953**	**0.729**	**0.790**	**0.925**	**0.847**	**0.850**	**1.450**	**18.468**

Table 7-II. Performance comparison on the 2015 MICCAI Multi-Atlas Abdomen Labeling challenge leaderboard. Our method achieves the largest Dice score and the smallest average surface distances and Hausdorff distances. PaNN* only uses 80% of the training data as the fully-supervised dataset and use the rest 20% data as partially-labeled data (by randomly removing labels of 7/13 organs), without using extra data. In this table, we only show 7/13 organs' average Dice scores due to the space limit.

obtains a much more remarkable improvement of 1.56% (from 77.48% to 79.04%). Such an observation suggests that PaNN is capable of handling more partially-labeled training data and is less susceptible to the background ambiguity.

Organ-by-organ Analysis. To reveal the detailed effect of PaNN, we present an organ-by-organ analysis in Figure 7-3. We use ResNet-50 as the backbone model (ResNet-101 has a similar trend) and the partially-labeled dataset **C** (indicates that the liver is the target organ).

In Figure 7-3, we observe clear statistical improvements over the fully-supervised method for almost every organ (p-values $p < 0.001$ hold for 11/13 of all abdominal

organs). Great improvements are also observed for those difficult organs, *i.e.*, organs either in small sizes or with complex geometric characteristics such as gallbladder (from 67.26% to 72.26%), esophagus (from 69.35% to 71.21%), stomach (from 84.09% to 87.21%), IVC (from 77.34% to 80.70%), portal vein & splenic vein (from 66.74% to 68.75%), pancreas (from 71.45% to 73.62%), right adrenal gland (from 53.65% to 55.56%) and left adrenal gland (from 49.51% to 53.63%). This promising result indicates that our method distills a reasonable amount of knowledge from additional partially-labeled data and the regularization loss can help facilitate the network to enhance the discriminative information to a certain degree.

Meanwhile, we also observe a distinct performance improvement for organs other than the partially-labeled structures (*i.e.*, the liver). For instance, the performance of gallbladder, stomach, IVC, pancreas are boosted from 68.97%, 85.57%, 78.59%, 71.94% to 72.26%, 87.21%, 80.70%, 73.62%, respectively. This suggests that the superiority of PaNN not only originates from more training data, but also from the fact that PaNN can effectively incorporate anatomical priors on organ sizes in abdominal regions, which is helpful for multi-organ segmentation.

Qualitative Evaluation. We also show a set of qualitative examples, *i.e.*, 5 slices from 3 cases, in Figure 7-4, where we zoom in to visualize the finer details of the improved region.

In these samples, we observe that PaNN is the only method that successfully detects the pancreatic tail in Figure 7-4(a). In Figure 7-4(b), all other methods fail to detect the portal vein and splenic vein while PaNN demonstrates an almost perfect detection of these veins. For Figure 7-4(c) to Figure 7-4(e), apart from the evident improvements of the pancreas, left adrenal gland, one of the smallest abdominal organs, is also clearly segmented by PaNN.

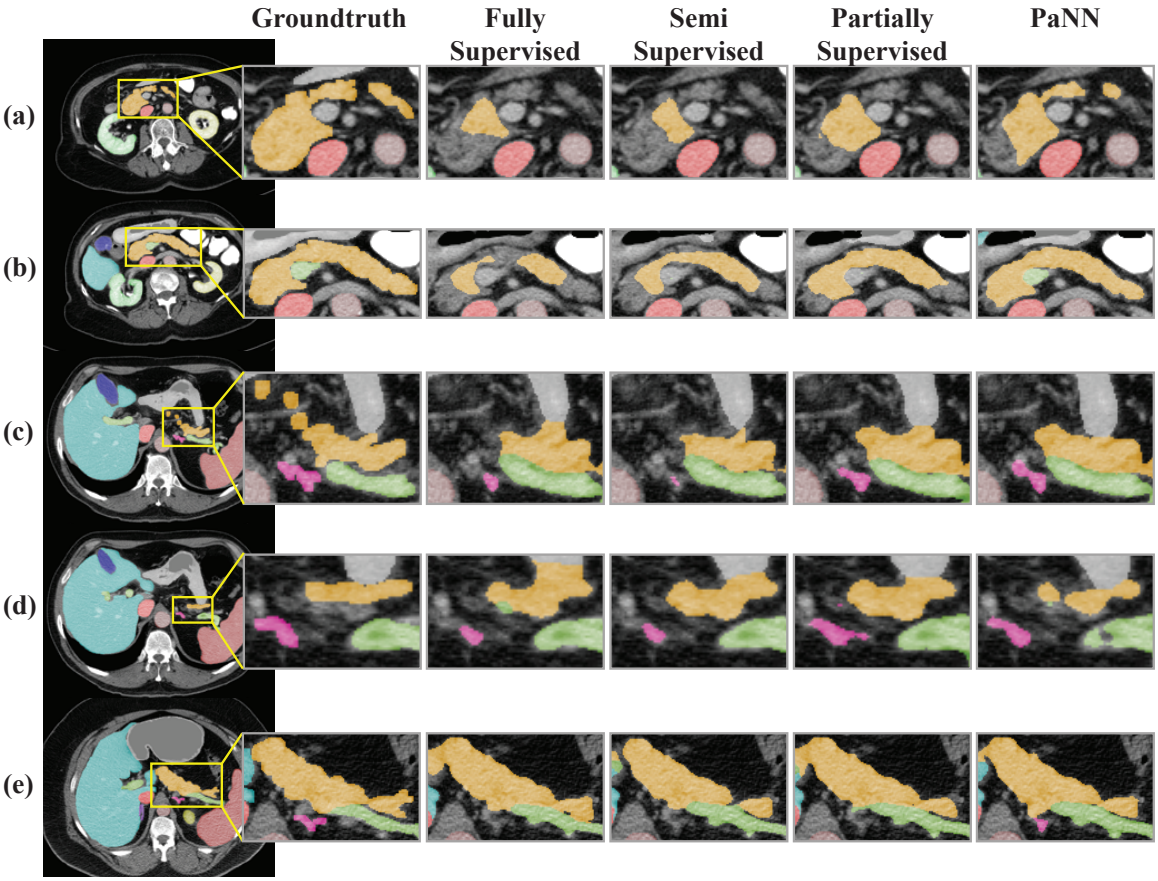


Figure 7-4. Qualitative comparison of different methods, where the partially-labeled dataset **C** is used as partial supervision with ResNet-101 as the backbone model. We exhibit 3 cases (5 slices) as examples. Improved segmentation regions are zoomed in from the axial view to demonstrate finer details.

7.3.3 MICCAI 2015 Multi-Atlas Labeling Challenge

We test our model in the 2015 MICCAI Multi-Atlas Abdomen Labeling challenge. The top model (denoted as "PaNN" in Table 7-II) we submit is based on ResNet-101, and trained on all 30 cases of the fully-labeled dataset $\mathbf{S}_L$ and the union of three partially-labeled datasets **A**, **B** and **C**. The evaluation metric employed in this challenge includes the Dice scores, average surface distances [8] and Hausdorff distances [78]. We compare PaNN with the other top submissions of the challenge leaderboard in Table 7-II. As it shows, the proposed PaNN achieves the best performance under all the three evaluation metrics, easily surpassing prior best result by a large margin.

Without using any additional data and even randomly removing partial labels from the challenge data, our method (denoted as "PaNN*" in Table 7-II) stills obtains the state-of-the-art result of 83.17%, outperforming the previous best result of DLTK UNet [176] by 2% in average Dice. It is noteworthy that our method is far from its potential maximum performance as we only use 2D single view algorithms. It is suggested [13, 16, 18, 19] that using multi-view algorithms or model ensemble can boost the performance further.

7.3.4 Generalization to Other Datasets

Organ	Fully Supervised	Semi Supervised	Partially Supervised (**ours**)	PaNN (**ours**)
Spleen	0.9640	0.9651	0.9673	0.9666
Right kidney	0.9626	0.9627	0.9625	0.9615
Left kidney	0.9530	0.9547	0.9526	0.9541
Gallbladder	0.8225	0.8399	0.8465	0.8467
Liver	0.9684	0.9691	0.9691	0.9689
Stomach	0.9344	0.9363	0.9396	0.9361
Aorta	0.9110	0.9096	0.9121	0.9133
IVC	0.8083	0.8175	0.7995	0.8266
Pancreas	0.7831	0.7994	0.8079	0.8193
avg. Dice	0.9008	0.9060	0.9063	**0.9103**

Table 7-III. Performance comparison on a newly collected high-quality abdominal dataset, where our method achieves the best result.

We also apply our algorithm to a different set of abdominal clinical CT images, where 20 cases are used for training and 15 cases are used for testing. A total of 9 structures (spleen, right kidney, left kidney, gallbladder, liver, stomach, aorta, IVC, pancreas) are manually labeled. Each case was segmented by four experienced radiologists, and confirmed by an independent senior expert. Each CT volume consists of $319 \sim 1051$ slices of 512×512 pixels, and has voxel spatial resolution of $([0.523 \sim 0.977] \times [0.523 \sim 0.977] \times 0.5)mm^3$. We use the union of all 3 datasets **A**, **B**, and **C** as the partial supervision. The results are summarized in Table 7-III, where the proposed PaNN also achieves better results compared with existing methods.

7.4 Summary

In this work, we have presented PaNN, for multi-organ segmentation, as a way to better utilize existing partially-labeled datasets. In this method, we intend to make the learned model both data-efficient and knowledge-aware by incorporating prior knowledge in the self-training framework on multiple heterogeneous datasets. To handle the background ambiguity brought by the partially-labeled data, the proposed PaNN exploits the anatomical priors by regularizing that the organ size distributions of the network output should approximate their prior statistics in the abdominal region.

Also, we want to address that knowledge can arise from various sources such as imaging physics, statistical constraints, and task specifics [1]. And this makes the knowledge priors take in various forms, such as shape models; spatial priors; topology specification; geometrical interaction and distance prior between different regions/labels; and atlas or pre-known models [5]. We use the size prior here, which is one of the simplest form to embed in the learning. But we definitely encourage more types of prior information to be studied for various learning tasks in the future.

Our proposed PaNN shows promising results using state-of-the-art models. And we believe that the use of priors can also stabilize training and make the learned presentations more generalized. We hope that our effort can offer some insights along this direction.

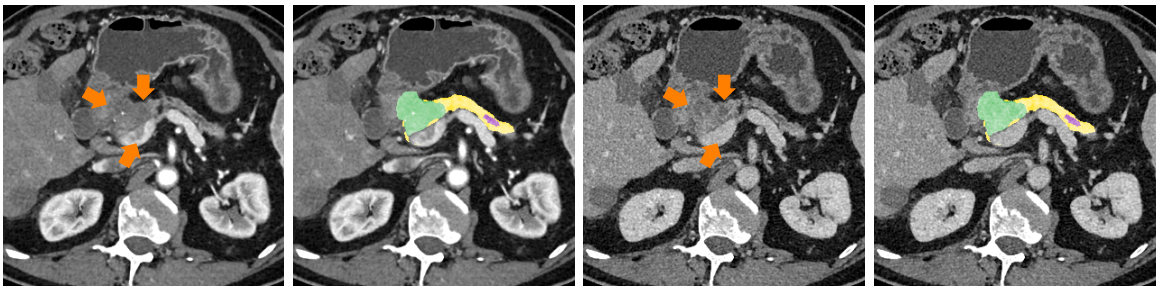
Chapter 8

Hyper-Pairing Network for Multi-Phase Pancreatic Ductal Adenocarcinoma Segmentation

Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal cancers with an overall five-year survival rate of 8%. Due to subtle texture changes of PDAC, pancreatic dual-phase imaging is recommended for better diagnosis of pancreatic disease. In this study, we aim at enhancing PDAC automatic segmentation by integrating multi-phase information (*i.e.*, arterial phase and venous phase). To this end, we present Hyper-Pairing Network (HPN), a 3D fully convolution neural network which effectively integrates information from different phases. The proposed approach consists of a dual path network where the two parallel streams are interconnected with hyper-connections for intensive information exchange. Additionally, a pairing loss is added to encourage the commonality between high-level feature representations of different phases. Compared to prior arts which use single phase data, HPN reports a significant improvement up to 7.73% (from 56.21% to 63.94%) in terms of DSC.

8.1 Introduction

Pancreatic ductal adenocarcinoma (PDAC) is the 4th most common cancer of death with an overall five-year survival rate of 8%. Currently, detection or segmentation



(a) Arterial Image (b) Arterial Label (c) Venous Image (d) Venous Label

Figure 8-1. Visual comparison of arterial and venous images (after alignment) as well as the manual segmentation of normal pancreas tissues (yellow), pancreatic duct (purple) and PDAC mass (green). Orange arrows indicate the ambiguous boundaries and differences of the abnormal appearances between the two phases (Best viewed in color).

at localized disease stage followed by complete resection can offer the best chance of survival, *i.e.*, with a 5-year survival rate of 32%. The accurate segmentation of PDAC mass is also important for further quantitative analysis, *e.g.*, survival prediction [178]. Computed tomography (CT) is the most commonly used imaging modality for the initial evaluation of PDAC. However, textures of PDAC on CT are very subtle (Figure 8-1) and therefore can be easily neglected by even experienced radiologists. To our best knowledge, the state-of-the-art on this matter is [179], which reports an average Dice of 56.46%. For better detection of PDAC mass, dual-phase pancreas protocol using contrast-enhanced CT imaging, which is comprised of arterial and venous phases with intravenous contrast delay, are recommended.

In recent years, deep learning has largely advanced the field of computer-aided diagnosis (CAD), especially in the field of biomedical image segmentation [8, 13, 18, 48, 101, 180–182]. However, there are several challenges for applying existing segmentation algorithms to dual-phase images. First, segmentation of pancreatic lesion, *e.g.*, cysts [16], is more difficult than organ segmentation due to its smaller sizes, lower contrast and texture similarity, *etc.* Secondly, these algorithms are optimized for segmenting only one type of input, and therefore cannot be directly applied to handle multi-phase data. More importantly, how to properly handle the variations between

different views requires a smart information exchange strategy between different phases. While how to efficiently integrate information from multi-modalities has been widely studied [183–185], the direction on learning multi-phase information has been rarely explored, especially for tumor detection and segmentation purposes.

To address these challenges, we propose a multi-phase segmentation algorithm, Hyper-Pairing Network (HPN), to enhance the segmentation performance especially for pancreatic abnormality. Following HyperDenseNet [183] which is effective on multi-modal image segmentation, we construct a dual-path network for handling multi-phase data, where each path is intended for one phase. To enable information exchange between different phases, we apply skip connections across different paths of the network [183], referred as *hyper-connections*. Moreover, noticing that a standard segmentation loss (cross-entropy loss, Dice loss [78]) only aims at minimizing the differences between the final prediction and the groundtruth thus cannot well handle the variance between different views, we introduce an additional *pairing loss term* to encourage the commonality between high-level features across both phases for better incorporation of multi-phase information. We exploit three structures together in HPN including PDAC mass, normal panreatic tissues, and pancreatic duct, which serves as an important clue for localizing PDAC. Extensive experiments demonstrate that the proposed HPN significantly outperforms prior arts by a large margin on all 3 targets.

8.2 Methodology

We hereby focus on dual-phase inputs while our approach can be generalized to multi-phase scans. With phase A and aligned phase B by the deformable registration, we have the set $\mathcal{S} = \{\left(\mathbf{X}_i^{\mathrm{A}}, \mathbf{X}_i^{\mathrm{B}}, \mathbf{Y}_i\right) | i = 1, ..., M\}$, where $\mathrm{X}_i^{\mathrm{A}} \in \mathbb{R}^{W_i \times H_i \times L_i}$ is the i-th 3D volumetric CT images of phase A with the dimension $(W_i \times H_i \times L_i) = \mathcal{D}_i$ and $\mathrm{X}_i^{\mathrm{B}} \in \mathbb{R}^{\mathcal{D}_i}$ is the corresponding aligned volume of phase B. $\mathbf{Y}_i = \{y_{ij} | j = 1, ..., \mathcal{D}_i\}$

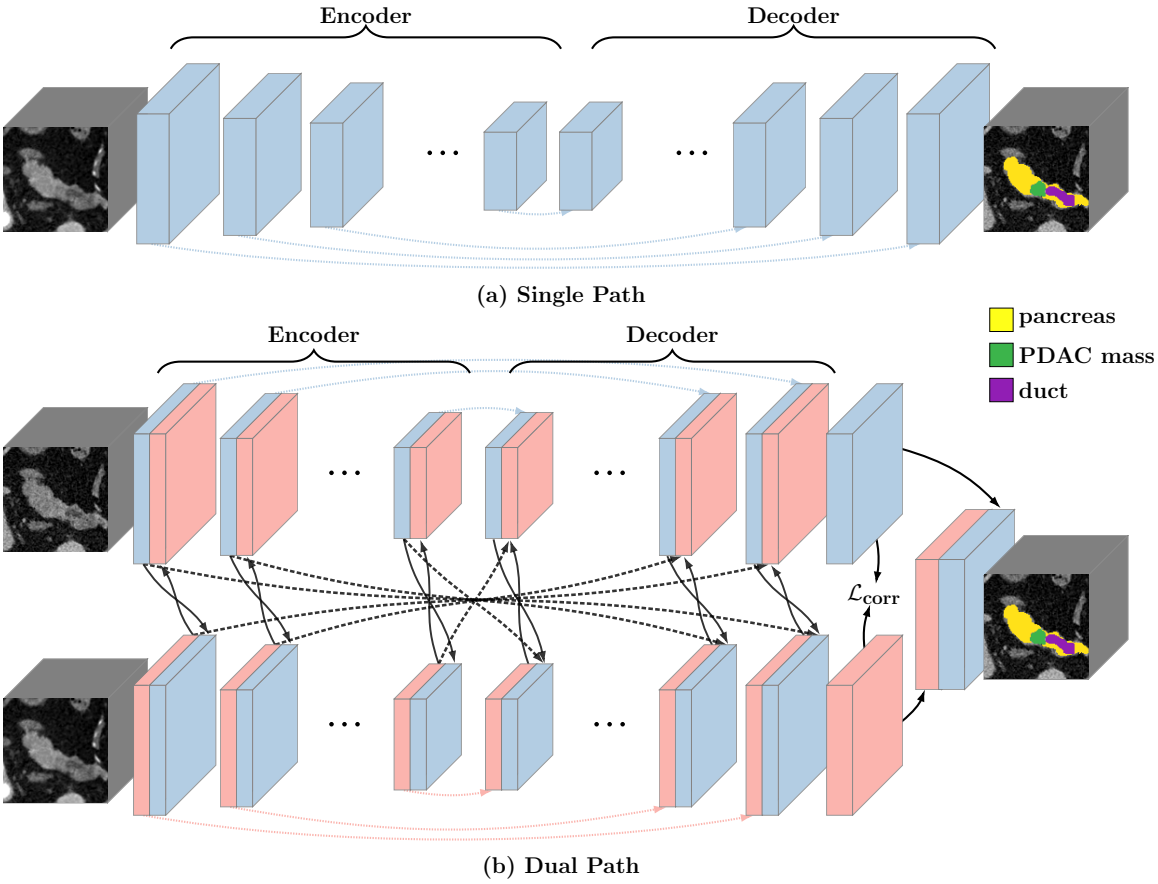


Figure 8-2. (a) The single path network where only one phase is used. The dash arrows denote skip connections between low-level features and high-level features. (b) HPN structure where multiple phases are used. The black arrows between the two single path networks indicate hyper-connections between the two streams. An additional pairing loss is employed to regularize view variations, therefore can benefit the integration between different phases. Blue and pink stand for arterial and venous phase, respectively.

denotes the corresponding voxel-wise label map of the i-th volume, where $y_{ij} \in \mathcal{L}$ is the label of the j-th voxel in the i-th image, and $\mathcal{L}$ denotes the label of the target structures. In this study, $\mathcal{L} = \{$normal panreatic tissues, PDAC mass, pancreatic duct$\}$. The goal is to learn a model to predict label of each voxel $\hat{\mathbf{Y}} = f(\mathbf{X}^{\text{A}}, \mathbf{X}^{\text{B}})$ by utilizing multi-phase information.

8.2.1 Hyper-connections

Segmentation networks (*e.g.*, UNet [101, 143], FCN [41]) usually contain a contracting encoder part and a successive expanding decoder part to produce a full-resolution

segmentation result as illustrated in Figure 8-2(a). As the layer goes deeper, the output features evolve from low-level detailed representations to high-level abstract semantic representations. The encoder part and the decoder part share an equal number of resolution steps [101, 143].

However, this type of network can only handle single-phase data. We construct a dual path network where each phase has a branch with a U-shape encoder-decoder architecture as mentioned above. These two branches are connected via hyper-connections which enrich feature representations by learning more complex combinations between the two phases. Specifically, hyper-connections are applied between layers which output feature maps of the same resolution across different paths as illustrated in Figure 8-2(b). Let $\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_\mathrm{T}$ denote the intermediate feature maps of a general segmentation network, where $\mathbf{R}_t$ and $\mathbf{R}_{\mathrm{T}-t}$ share the same resolution ($\mathbf{R}_t$ is on the encoder path and $\mathbf{R}_{\mathrm{T}-t}$ is on the decoder path). Hyper-connections are applied as follows: $\mathbf{R}_t^\mathrm{A} \longrightarrow \mathbf{R}_\mathrm{t}^\mathrm{B}$, $\mathbf{R}_t^\mathrm{B} \longrightarrow \mathbf{R}_\mathrm{t}^\mathrm{A}$, $\mathbf{R}_t^\mathrm{A} \longrightarrow \mathbf{R}_{\mathrm{T}-t}^\mathrm{B}$, $\mathbf{R}_t^\mathrm{B} \longrightarrow \mathbf{R}_{\mathrm{T}-t}^\mathrm{A}$,$\mathbf{R}_{\mathrm{T}-t}^\mathrm{A} \longrightarrow \mathbf{R}_{\mathrm{T}-t}^\mathrm{B}$, $\mathbf{R}_{\mathrm{T}-t}^\mathrm{B} \longrightarrow \mathbf{R}_{\mathrm{T}-t}^\mathrm{A}$, while maintaining the original skip connections that already occur within the same path, *i.e.*, $\mathbf{R}_t^\mathrm{A} \longrightarrow \mathbf{R}_{\mathrm{T}-t}^\mathrm{A}$, $\mathbf{R}_t^\mathrm{B} \longrightarrow \mathbf{R}_{\mathrm{T}-t}^\mathrm{B}$.

8.2.2 Pairing loss

The standard loss for segmentation networks only aims at minimizing the difference between the ground-truth and the final estimation, which cannot well handle the variance between different views. Applying this loss alone is inferior in our situation since the training process involves heavy integration of both arterial information and venous information. To this end, we propose to apply an additional pairing loss, which encourages the commonality between the two sets of high-level semantic representations, to reduce view divergence.

We instantiate this additional objective as a correlation loss [186]. Mathematically, for any pair of aligned images (X_i^A, X_i^B) passing through the corresponding view sub-

network, the two sets of high-level semantic representations (feature responses in later layers) corresponding to the two phases are denoted as $f_1(\mathrm{X}_i^{\mathrm{A}};\boldsymbol{\Theta}_1)$ and $f_2(\mathrm{X}_i^{\mathrm{B}};\boldsymbol{\Theta}_2)$, where the two sub-networks are parameterized by $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ respectively. The outputs of two branches will be simultaneously fed to the final classification layer. In order to better integrate the outcomes from the two branches, we propose to use a pairing loss which exploits the consensus of $f_1(\mathrm{X}_i^{\mathrm{A}};\boldsymbol{\Theta}_1)$ and $f_2(\mathrm{X}_i^{\mathrm{B}};\boldsymbol{\Theta}_2)$ during training. The loss is formulated as following:

$$\mathcal{L}_{corr}(\mathrm{X}_i^{\mathrm{A}},\mathrm{X}_i^{\mathrm{B}};\boldsymbol{\Theta}) = -\frac{\sum_{j=1}^{N}\left(f_1(\mathrm{X}_{ij}^{\mathrm{A}})-\overline{f_1(\mathrm{X}_i^{\mathrm{A}})}\right)\left(f_2(\mathrm{X}_{ij}^{\mathrm{B}})-\overline{f_2(\mathrm{X}_i^{\mathrm{B}})}\right)}{\sqrt{\sum_{j=1}^{N}\left(f_1(\mathrm{X}_{ij}^{\mathrm{A}})-\overline{f_1(\mathrm{X}_i^{\mathrm{A}})}\right)^2\sum_{j=1}^{N}\left(f_2(\mathrm{X}_{ij}^{\mathrm{B}})-\overline{f_2(\mathrm{X}_i^{\mathrm{B}})}\right)^2}}, \tag{8.1}$$

where N denotes the total number of voxels in the i-th sample and $\boldsymbol{\Theta}$ denotes the parameters of the entire network. During the training stage, we impose this additional loss to further encourage the commonality between the two intermediate outputs. The overall loss is the weighted sum of this additional penalty term and the standard voxel-wise cross-entropy loss:

$$\mathcal{L}_{total} = -\frac{1}{N}\left[\sum_{j=1}^{N}\sum_{k=0}^{K}\mathbb{1}(y_{ij}=k)\log p_{ij}^{k}\right] + \lambda\mathcal{L}_{corr}(\mathrm{X}_i^{\mathrm{A}},\mathrm{X}_i^{\mathrm{B}};\boldsymbol{\Theta}), \tag{8.2}$$

where p_{ij}^{k} denotes the probability of the j-th voxel be classified as label k on the i-th sample and $\mathbb{1}(\cdot)$ is the indicator function. K is the total number of classes. The overall objective function is optimized via stochastic gradient descent.

8.3 Experiments

8.3.1 Experiment setup

Data acquisition. This is an institutional review board approved HIPAA compliant retrospective case control study. 239 patients with pathologically proven PDAC were retrospectively identified from the radiology and pathology databases from 2012 to 2017 and the cases with $\leq$ 4cm tumor (PDAC mass) diameter were selected for the

experiment. PDAC patients were scanned on a 64-slice multidetector CT scanner (Sensation 64, Siemens Healthineers) or a dual-source multidetector CT scanner (FLASH, Siemens Healthineers). PDAC patients were injected with 100-120 mL of iohexol (Omnipaque, GE Healthcare) at an injection rate of 4-5 mL/Section Scan protocols were customized for each patient to minimize dose. Arterial phase imaging was performed with bolus triggering, usually 30 seconds post-injection, and venous phase imaging was performed 60 seconds.

Evaluation. Denote $\mathcal{Y}$ and $\mathcal{Z}$ as the set of foreground voxels in the ground-truth and prediction, *i.e.*, $\mathcal{Y} = \{i \mid y_i = 1\}$ and $\mathcal{Z} = \{i \mid z_i = 1\}$. The accuracy of segmentation is evaluated by the Dice-Sørensen coefficient (DSC): $\text{DSC}(\mathcal{Y}, \mathcal{Z}) = \frac{2\times|\mathcal{Y}\cap\mathcal{Z}|}{|\mathcal{Y}|+|\mathcal{Z}|}$. We evaluate DSCs of all three targets, *i.e.*, abnormal pancreas, PDAC mass and pancreatic duct. All experiments are conducted by three-fold cross-validation, *i.e.*, training the models on two folds and testing them on the remaining one. Through our experiment, abnormal pancreas stands for the union of normal pancreatic tissues, PDAC mass and pancreatic duct. The average DSC of all cases as well as the standard deviations are reported.

Implementation details Our experiments were performed on the whole CT scan and the implementations are based on PyTorch. We adopt a variation of diffeomorphic demons with direction-dependent regularizations [187, 188] for accurate and efficient deformable registration between the two phases. For data pre-processing, we truncated the raw intensity values within the range [-100, 240] HU and normalized each raw CT case to have zero mean and unit variance. The input sizes of all networks are set as $64\times64\times64$. The coefficient of the correlation loss λ is set as 0.5. No further post-processing strategies were applied.

We also used data augmentation during training. Different from single-phase segmentation which commonly uses rotation and scaling [6, 179], virtual sets [189]

Method	Abnormal pancreas	PDAC mass	pancreatic duct
3D-UNet-single-phase (Arterial)	78.35 ± 11.89	52.40 ± 27.53	38.35 ± 28.98
3D-UNet-single-phase (Venous)	79.61 ± 10.47	53.08 ± 27.06	40.25 ± 27.89
3D-UNet-multi-phase (fusion)	80.05 ± 10.56	52.88 ± 26.97	39.06 ± 27.33
3D-UNet-multi-phase-HyperNet	82.45 ± 9.98	54.36 ± 26.34	43.27 ± 26.33
3D-UNet-multi-phase-HyperNet-aug	83.67 ± 8.92	55.72 ± 26.01	43.53 ± 25.94
3D-UNet-multi-phase-HPN (Ours)	**84.32 ± 8.59**	**57.10 ± 24.76**	**44.93 ± 24.88**
3D-ResDSN-single-phase (Arterial)	83.85 ± 9.43	56.21 ± 26.33	47.04 ± 26.42
3D-ResDSN-single-phase (Venous)	84.92 ± 7.70	56.86 ± 26.67	49.81 ± 26.23
3D-ResDSN-multi-phase (fusion)	85.52 ± 7.84	57.59 ± 26.63	48.49 ± 26.37
3D-ResDSN-multi-phase-HyperNet	85.79 ± 8.86	60.87 ± 24.95	54.18 ± 24.74
3D-ResDSN-multi-phase-HyperNet-aug	85.87 ± 7.91	61.69 ± 23.24	54.07 ± 24.06
3D-ResDSN-multi-HPN (Ours)	**86.65 ± 7.46**	**63.94 ± 22.74**	**56.77 ± 23.33**

Table 8-I. DSC (%) comparison of abnormal pancreas, PDAC mass and pancreatic duct. We report results in the format of mean ± standard deviation.

are also utilized in this work. Even though arterial and venous phase scanning are customized for each patient, the level of enhancement can be different from patients by variation of blood circulation, which causes inter-subject enhancement variations on each phase. Therefore we construct virtual examples by interpolating between venous and arterial data, similar to [189]. The i-th augmented training sample pair can be written as: $\tilde{\mathrm{X}}_i^{\mathrm{A}} = \lambda \mathrm{X}_i^{\mathrm{A}} + (1-\lambda)\mathrm{X}_i^{\mathrm{B}}, \quad \tilde{\mathrm{X}}_i^{\mathrm{B}} = \lambda \mathrm{X}_i^{\mathrm{B}} + (1-\lambda)\mathrm{X}_i^{\mathrm{A}}$, where $\lambda \sim \mathrm{Beta}(\alpha, \alpha) \in [0, 1]$. The final outcome of HPN is obtained by taking the union of predicted regions from models trained with the original paired sets and the virtual paired sets. We set the hyper-parameter $\alpha = 0.4$ following [189].

8.3.2 Results and Discussions

All results are summarized in Table 8-I. We compare the proposed HPN with the following algorithms: 1) single-phase algorithms which are trained exclusively on one phase (denoted as "single-phase"); 2) multi-phase algorithm where both arterial and venous data are trained using a dual path network bridged with hyper connections (denoted as "HyperNet"). In general, compared with single-phase algorithms, multi-phase algorithms (*i.e.*, HyperNet, HPN) observe significant improvements for all target structures. It is no surprise to observe such a phenomenon as more useful information

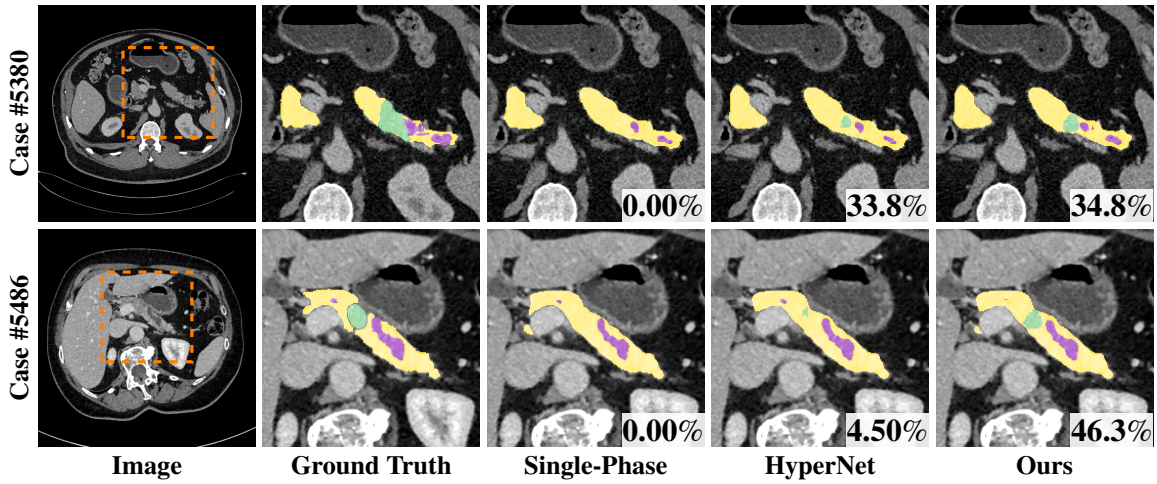


Figure 8-3. Qualitative comparison of different methods, where HPN enhances PDAC mass segmentation (green) significantly compared with other methods (Best viewed in color).

is distilled for multi-phase algorithms.

Efficacy of hyper-connections. To show the effectiveness of hyper-connections, output from different phases (using single-phase algorithms) are fused by taking at each position the average probability (denoted as "fusion"). However, we observe that simply fusing the outcomes from different phases usually yield either similar or slightly better performances compared with single-phase algorithms. This indicates that simply fusing the estimations during the inference stage cannot effectively integrate multi-phase information. By contrast, hyper-connections enable the training process to be communicative between the two phase branches and thus can efficiently elevate the performance. Note that directly applying [183] yield unsatisfactory results. Our hyper-connections are not densely connected but are carefully designed based on previous state-of-the-art on PDAC segmentation [179] for better segmentation of PDAC. Meanwhile, we show much better performance of 63.94% compared to 56.46% reported in [179].

Efficacy of data augmentation. From Table 8-I, compared with HyperNet,

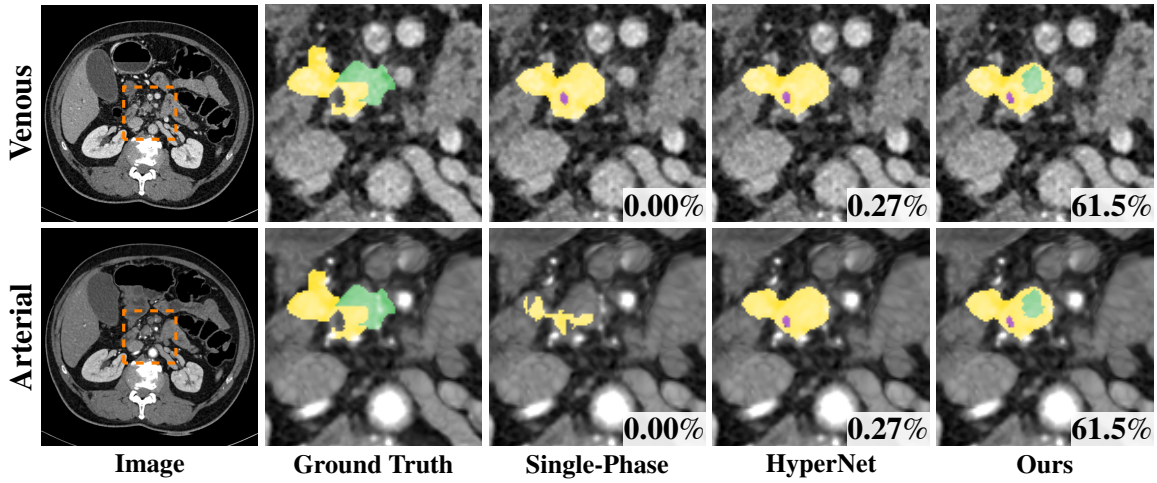


Figure 8-4. Qualitative example where HPN detects the PDAC mass (green) while single-phase methods for both phases fail. From left to right: venous and arterial images (aligned), groundtruth, predictions of single-phase algorithms, HyperNet prediction, HPN prediction (overlayed with venous and arterial images). Best viewed in color.

HyperNet-aug witnesses performance gain especially for PDAC mass (*i.e.*, from 60.87% to 61.69% for 3D-ResDSN; from 54.36% to 55.72% for 3D-UNet), which validates the usefulness of using virtual paired sets as data augmentation.

Efficacy of HPN. We can observe additional benefit of our HPN over hyperNet-aug (*e.g.*, abnormal pancreas: 85.87% to 86.65%, PDAC mass: 61.69% to 63.94%, pancreatic duct: 54.07% to 56.77%, 3D-ResDSN). Overall, HPN observes an evident improvement compared with HyperNet, *i.e.*, abnormal pancreas: 85.79% to 86.65%, PDAC mass: 61.69% to 63.94%, pancreatic duct: 54.07% to 56.77% (3D-ResDSN). The *p-value*s for testing significant difference between hyperNet and our HPN of all 3 targets are $p < 0.0001$, which suggests a general statistical improvement. We also show two qualitative examples in Figure 8-3, where HPN shows much better segmentation accuracy especially for PDAC mass.

Another noteworthy fact is that 11/239 cases are false negatives which failed to detect any PDAC mass using either phase (Dice = 0%). Out of these 11 cases, 7 cases are successfully detected by HPN. An example is shown in Figure 8-4 — the

PDAC mass is missing from both single phases and almost missing in the original HyperNet (DSC=0.27%), but our HPN can detect a reasonable portion of the PDAC mass (DSC=61.5%).

The deformable registration error by computing pancreas surface distances between two phases is $1.01 \pm 0.52mm$ (mean $\pm$ standard deviations) which can be considered as acceptable for this study. However, the effects between different alignments can be described as a further study.

8.4 Summary

Motivated by the fact that radiologists usually rely on analyzing multi-phase data for better image interpretations, we develop an end-to-end framework, HPN, for multi-phase image segmentation. Specifically, HPN consists of a dual path network where different paths are connected for multi-phase information exchange, and an additional loss is added for removing view divergence. Extensive experiment results demonstrate that the proposed HPN can substantially and significantly improve the segmentation performance, *i.e.*, HPN reports an improvement up to 7.73% in terms of DSC compared to prior arts which use single phase data. In the future, we plan to examine the behaviour of HPN when using different alignment strategies and try to extend the current approach to other multi-phase learning problems.

Chapter 9

Conclusion and Discussion

9.1 Summary

In this thesis, we focus on designing *data-efficient and knowledge-aware* deep learning techniques towards medical machine intelligence. Our studies [190–192] summarize further clinical evaluations of various proposed approaches. In particular, we consider several existing challenges in the field of medical image analysis, *i.e.*, unsatisfactory performance regarding challenging small targets, insufficient training data, high annotation cost, and the lack of domain-specific knowledge. To effectively handle small target (*e.g.*, the pancreas) segmentation, in Chapter 3, we introduce multi-stage coarse-to-fine frameworks where the coarse stage generates attention-related regions which can later facilitate the fine stage to output more accurate results. Then in Chapter 4 and Chapter 5, we show that this type of attention mechanism can be also well adapted to cystic lesion segmentation and multi-organ segmentation, respectively. This further shows the generalization of the proposed coarse-to-fine frameworks beyond small target segmentation. In Chapter 6, we present a *data-efficient* deep learning method by leveraging the power of unlabeled data. The method belongs to the semi-supervised regime, which co-trains the deep model to mine consensus information from multiple viewing directions.

In addition to *data-efficiency*, we also rethink existing deep-learning-based strategies

in terms of *knowledge-awareness*. When clinical experts interpret medical images, they often rely on prior knowledge about anatomy, and may use a template of the structures to constrain the task. By contrast, convolutional methods are often limited in incorporating such domain-specific knowledge. Therefore, we further discuss how to make deep neural networks aware of knowledge priors in Chapter 7, so as to approach the real clinical expertise. Similarly, we design a multi-phase learning algorithm for detecting pancreatic ductal adenocarcinoma in Chapter 8, since dual-phase imaging is clinically recommended for better diagnosis of pancreatic disease. In addition, we note that deep supervision for pancreatic cyst segmentation (Chapter 4) can be also deemed as making use of the spatial prior of medical images, since it is based on the high relevance between the location of a pancreas and its cystic region.

9.2 Future Works

The recent success in machine learning (especially deep learning) has led to remarkable achievements in the field of medical image analysis. However, the gap between research settings and real-world clinical settings remains large. We hope our proposed data-efficient and knowledge-aware techniques offer some insights for enhancing the applicability of existing medical image analysis systems. In the future, we aim to investigate the following research directions to further bridge this gap.

Learning from heterogeneous and isolated data. Due to different scanners, image acquisition protocols or different patient populations and ethnicities, medical datasets are usually scattered and disjoint. To maximize data utilization and make the learned representations more robust and universal, the following strategies can be sought to handle heterogeneous data: 1) domain adaptation techniques, such as our recent study [193], which attempt to bridge the gap between multiple domains (*e.g.*, heterogeneous datasets) by either learning a latent representation that is shared by

these different domains or by translating images from one domain to the other. 2) dataset fusion techniques, which leverage multiple datasets to train a universal model. These strategies can be furthered by:

- **Multi-task learning.** Different levels of annotations can be formulated tailored to different learning tasks. Then Multi-task learning, which refers to the paradigm where multiple tasks are derived from a single learned representation, can encourage the encoder network to learn a latent representation that is generic across different tasks and datasets.

- **Self-supervised learning.** In terms of unlabeled data, we can also formulate other proxy tasks (*e.g.*, jigsaw puzzles, image reconstruction) where the data provides the supervision. This enables the unlabeled data to be either used for pre-training or jointly learned to enhance the model scalability and the generalization.

Multi-task learning and self-supervised learning can be also combined with our previous studies [19, 20] to deliver more effective models. Many efforts have been devoted to related research areas [121, 194–218].

Learning with noisy labels. Despite the recent advance of medical image analysis, existing works mostly assume that reliable ground-truth annotations are abundant, which is not always the case in practice: 1) collecting annotations can be time-consuming; 2) human-annotations are inherently noisy. Further, annotations generally suffer from inter/intra-observer variation even among experts. Therefore, it becomes important to acquire a more generalized model which can be robust against noisy annotations. To achieve this, our domain adaptation, self-supervised learning, semi-supervised learning, weakly/partially supervised learning strategies [19, 20, 193] can be applied. In addition, we also aim to alleviate the negative effects induced by the

heavily noisy samples by employing our proposed weighted sampling strategy [128] based on the upper confident bound, an exploitation-exploration strategy.

Leveraging priors of medical images. As aforementioned, our proposed strategies in Chapter 4, Chapter 7 and Chapter 8 are different ways to make use of domain-specific knowledge. Besides using the prior information on spatial location (Chapter 4) and size (Chapter 7), we have also demonstrated that the cylinder-like shape of tubular structures (*e.g.*, vessels) can be well represented by its skeleton and cross-sectional radii, and such shape (geometry) priors can be used to effectively refine the segmentation outcome [219]. Similar conclusions are also drawn in [220, 221]. Given different types of prior information, one promising direction is combining them to jointly guide the model training via curriculum learning. The learned priors are intended to make trained models more generalizable and aligned with clinician's perceptions.

Efficient Neural Networks for Medical Image Analysis. Medical images (*e.g.*, CT) are often high-dimensional 3D volumes. However, conventional 3D convolution layers typically result in expensive computation and suffer from convergence problems due to over-fitting issues and lack of pre-trained weights. [222–226] discuss the advantages and disadvantages of 2D/3D architectures. In [57], we propose to use a hybrid 2D-3D architecture which can largely reduce the compute budget while enjoying the benefit of 3D context. Our later efforts on neural architecture search [227, 228] further show that light-weight networks can be obtained for various imaging tasks. How to further reduce the computation cost and make the searched architecture more practical for clinical usage should be explored.

Multi-modal medical machine learning. Many medical image analysis problems heavily rely on information integration of different imaging modalities. Multi-modalities, generally speaking, can refer to any different types of input. In the area

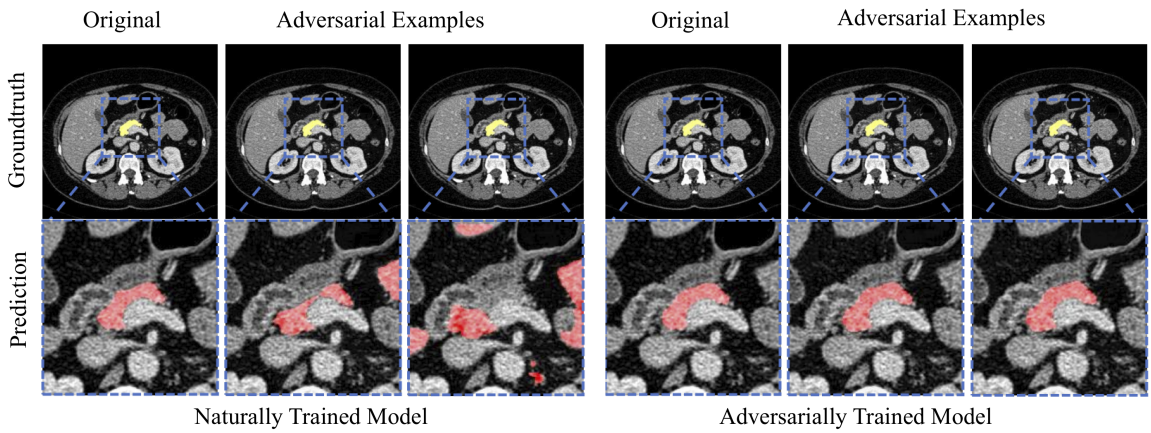


Figure 9-1. Medical image analysis systems are vulnerable to adversarial examples, and adversarial training can improve model robustness.

of computer vision and natural language processing, multi-modality fusion can be applied to different types of data (*e.g.*, image, motion, aural, speech) since they are usually relevant and complementary to each other. In the medical domain, we have demonstrated the importance of information integration from different phases (arterial and venous phase), from the shape and texture features [21, 229, 230] for abnormality detection. How to discover and utilize the intrinsic relationship between different modalities for designing a generalized and robust learning system is an important task for pushing forward medical machine intelligence.

Towards trustworthy medical image analysis systems. Adversarial examples have raised concerns about the practical deployment of deep learning systems for image classification, person re-identification and object detection in the wild as illustrated in some of our previous studies [231–235]. Beyond the computer vision area, the existence of adversarial examples is ubiquitous for essentially every type of machine learning model ever studied and across a wide range of data types, including images, audio and text. Here we select a few CT images as well as their adversarial examples in Figure 9-1, to demonstrate the vulnerabilities of existing medical machine learning systems. However, compared with other applications, the clinical practice of deep learning requires a higher level of safety and security standard. Our previous attempts

of adversarial training, in which the defense mechanism augments each training minibatch with adversarial examples, suggest a solid improvement in the robustness of medical machine learning systems [236]. As shown in Figure 9-1, our results show quite promising defenses against adversarial attacks. How to build medical defenses which can secure against conceivable present and future attacks should be further explored.

References

1. Zhou, S. K. *et al.* A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises. *arXiv preprint arXiv:2008.09104* (2020).
2. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annual review of biomedical engineering* **19,** 221–248 (2017).
3. Shen, C. *et al.* An introduction to deep learning in medical physics: advantages, potential, and challenges. *Physics in Medicine & Biology* **65,** 05TR01 (2020).
4. Tajbakhsh, N. *et al.* Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis,* 101693 (2020).
5. Nosrati, M. S. & Hamarneh, G. Incorporating prior knowledge in medical image segmentation: a survey. *arXiv preprint arXiv:1607.01092* (2016).
6. Kamnitsas, K. *et al.* Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *Medical Image Analysis* **36,** 61–78 (2017).
7. Roth, H. *et al.* DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015).
8. Roth, H., Lu, L., Farag, A., Sohn, A. & Summers, R. Spatial Aggregation of Holistically-Nested Networks for Automated Pancreas Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2016).
9. Greenspan, H., Van Ginneken, B. & Summers, R. M. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging* **35,** 1153–1159 (2016).
10. Bai, W. *et al. Semi-supervised learning for network-based cardiac MR image segmentation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017), 253–260.
11. Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. *Self-training with noisy student improves imagenet classification* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 10687–10698.
12. Blum, A. & Mitchell, T. *Combining labeled and unlabeled data with co-training* in *Proceedings of the eleventh annual conference on Computational learning theory* (1998), 92–100.
13. Zhou, Y. *et al.* A Fixed-Point Model for Pancreas Segmentation in Abdominal CT Scans. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017).

14. Yu, Q. *et al. Recurrent Saliency Transformation Network: Incorporating Multi-Stage Visual Cues for Small Organ Segmentation* in *Computer Vision and Patter Recognition* (2018).

15. Zhou, Y. *et al.* in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics* 43–67 (Springer, 2019).

16. Zhou, Y., Xie, L., Fishman, E. & Yuille, A. Deep Supervision for Pancreatic Cyst Segmentation in Abdominal CT Scans. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017).

17. Xie, L. *et al.* Recurrent Saliency Transformation Network for Tiny Target Segmentation in Abdominal CT Scans. *IEEE transactions on medical imaging* **39,** 514–525 (2019).

18. Wang, Y. *et al.* Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical image analysis* **55,** 88–102 (2019).

19. Zhou, Y. *et al.* Semi-Supervised Multi-Organ Segmentation via Multi-Planar Co-Training. *IEEE Winter Conference on Applications of Computer Vision* (2019).

20. Zhou, Y. *et al. Prior-aware neural network for partially-supervised multi-organ segmentation* in *Proceedings of the IEEE International Conference on Computer Vision* (2019), 10672–10681.

21. Zhou, Y. *et al. Hyper-pairing network for multi-phase pancreatic ductal adenocarcinoma segmentation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), 155–163.

22. Brosch, T. *et al.* Deep 3D Convolutional Encoder Networks with Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation. *IEEE Transactions on Medical Imaging* **35,** 1229–1239 (2016).

23. Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. Deep Learning for Identifying Metastatic Breast Cancer. *arXiv preprint arXiv:1606.05718* (2016).

24. Havaei, M. *et al.* Brain Tumor Segmentation with Deep Neural Networks. *Medical Image Analysis* (2017).

25. Ling, H. *et al.* Hierarchical, Learning-based Automatic Liver Segmentation. *Computer Vision and Pattern Recognition* (2008).

26. Heimann, T. *et al.* Comparison and Evaluation of Methods for Liver Segmentation from CT Datasets. *IEEE Transactions on Medical Imaging* **28,** 1251–1265 (2009).

27. Linguraru, M., Sandberg, J., Li, Z., Shah, F. & Summers, R. Automated Segmentation and Quantification of Liver and Spleen from CT Images Using Normalized Probabilistic Atlases and Enhancement Estimation. *Medical Physics* **37,** 771–783 (2010).

28. Lin, D., Lei, C. & Hung, S. Computer-Aided Kidney Segmentation on Abdominal CT Images. *IEEE Transactions on Information Technology in Biomedicine* **10,** 59–65 (2006).

29. Ali, A., Farag, A. & El-Baz, A. Graph Cuts Framework for Kidney Segmentation with Prior Shape Constraints. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2007).

30. Hu, S., Hoffman, E. & Reinhardt, J. Automatic Lung Segmentation for Accurate Quantitation of Volumetric X-ray CT Images. *IEEE Transactions on Medical Imaging* **20,** 490–498 (2001).

31. Chu, C. *et al.* Multi-organ Segmentation based on Spatially-Divided Probabilistic Atlas from 3D Abdominal CT Images. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2013).

32. Wang, Z. *et al.* Geodesic Patch-based Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2014).

33. Dmitriev, K., Gutenko, I., Nadeem, S. & Kaufman, A. *Pancreas and Cyst Segmentation* in *Medical Imaging 2016: Image Processing* **9784** (2016), 97842C.

34. Zhang, L., Lu, L., Summers, R. M., Kebebew, E. & Yao, J. *Personalized Pancreatic Tumor Growth Prediction via Group Learning* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017).

35. Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems* (2012).

36. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations* (2015).

37. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition* (2016).

38. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Computer Vision and Pattern Recognition* (2014).

39. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems* (2015).

40. Tang, P. *et al.* PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence* (2018).

41. Long, J., Shelhamer, E. & Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *Computer Vision and Pattern Recognition* (2015).

42. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *International Conference on Learning Representations* (2015).

43. Graves, A., Mohamed, A. & Hinton, G. Speech Recognition with Deep Recurrent Neural Networks. *International Conference on Acoustics, Speech and Signal Processing* (2013).

44. Socher, R., Lin, C., Manning, C. & Ng, A. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. *International Conference on Machine Learning* (2011).

45. Shen, W., Wang, B., Jiang, Y., Wang, Y. & Yuille, A. Multi-Stage Multi-Recursive-Input Fully Convolutional Networks for Neuronal Boundary Detection. *International Conference on Computer Vision* (2017).

46. Liang, M. & Hu, X. Recurrent Convolutional Neural Network for Object Recognition. *Computer Vision and Pattern Recognition* (2015).

47. Pinheiro, P. & Collobert, R. Recurrent Convolutional Neural Networks for Scene Labeling. *International Conference on Machine Learning* (2014).

48. Dou, Q. *et al.* 3D Deeply Supervised Network for Automatic Liver Segmentation from CT Volumes. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2016).

49. Harrison, A. *et al.* Progressive and Multi-Path Holistically Nested Neural Networks for Pathological Lung Segmentation from CT Images. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017).

50. Cai, J., Lu, L., Xie, Y., Xing, F. & Yang, L. Improving Deep Pancreas Segmentation in CT and MRI Images via Recurrent Neural Contextual Learning and Direct Loss Function. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017).

51. Christ, P. F. *et al.* Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. *arXiv preprint arXiv:1702.05970* (2017).

52. Li, X. *et al.* H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes. *IEEE Transactions on Medical Imaging* (2018).

53. Sun, C. *et al.* Automatic segmentation of liver tumors from multiphase contrast-enhanced CT images based on FCNs. *Artificial intelligence in medicine* **83,** 58–66 (2017).

54. Yan, K. *et al. Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), 9261–9270.

55. Yan, K., Wang, X., Lu, L. & Summers, R. M. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging* **5,** 036501 (2018).

56. Yan, K., Bagheri, M. & Summers, R. M. *3D context enhanced region-based convolutional neural network for end-to-end lesion detection* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), 511–519.

57. Zhang, Z., Zhou, Y., Shen, W., Fishman, E. & Yuille, A. *Lesion Detection by Efficiently Bridging 3D Context* in *International Workshop on Machine Learning in Medical Imaging* (2019), 470–478.

58. Iglesias, J. E. & Sabuncu, M. R. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis* **24,** 205–219. eprint: 1412.3421 (2015).

59. Asman, A. J. & Landman, B. A. Non-local statistical label fusion for multi-atlas segmentation. *Medical Image Analysis* **17,** 194–208. arXiv: NIHMS150003 (2013).

60. Chu, C. *et al.* Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images. *Lecture Notes in Computer Science* **8150 LNCS,** 165–172 (2013).

61. Wolz, R. *et al.* Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE Transactions on Medical Imaging* **32,** 1723–1730 (2013).

62. Kada, T. *et al.* Abdominal multi-organ segmentation from CT images using conditional shape-location and unsupervised intensity priors. *Medical Image Analysis* **26,** 1–18 (2015).

63. Zhuang, X. & Shen, J. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical Image Analysis* **31,** 77–87 (2016).

64. Karasawa, K. *et al.* Multi-atlas pancreas segmentation: Atlas selection based on vessel structure. *Medical Image Analysis* **39,** 18–28 (2017).

65. Gibson, E. *et al. Towards image-guided pancreas and biliary endoscopy: Automatic multi-organ segmentation on abdominal CT with dense dilated networks* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017), 728–736.

66. Roth, H. R. *et al. A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation* in *MICCAI* (2018).

67. Brosch, T. & Saalbach, A. *Foveal fully convolutional nets for multi-organ segmentation* in *Medical Imaging 2018: Image Processing* **10574** (2018), 105740U.

68. Weston, A. D. *et al.* Complete Abdomen and Pelvis Segmentation using U-Net Variant Architecture. *Medical Physics* (2020).

69. Akhter, M. E. *et al. An Analysis of Multi-organ Segmentation Performance of CNNs on Abdominal Organs with an Emphasis on Kidney* in *International Conference on Medical Imaging and Computer-Aided Diagnosis* (2020), 229–241.

70. Lei, Y. *et al.* Deep Learning in Multi-organ Segmentation. *arXiv preprint arXiv:2001.10619* (2020).

71. Rehman, A. & Khan, F. G. A deep learning based review on abdominal images. *Multimedia Tools and Applications,* 1–32 (2020).

72. Valindria, V. V. *et al. Small organ segmentation in whole-body MRI using a two-stage FCN and weighting schemes* in *International Workshop on Machine Learning in Medical Imaging* (2018), 346–354.

73. Wang, Y., Zhao, L., Wang, M. & Song, Z. Organ at risk segmentation in head and neck ct images using a two-stage segmentation framework based on 3D U-Net. *IEEE Access* **7,** 144591–144602 (2019).

74. Hu, P. *et al.* Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. *International journal of computer assisted radiology and surgery* **12,** 399–411 (2017).

75. Roth, H. R. *et al.* Hierarchical 3D fully convolutional networks for multi-organ segmentation. *arXiv preprint arXiv:1704.06382* (2017).

76. Zhou, X. *et al.* in *Deep Learning and Data Labeling for Medical Applications* 111–120 (Springer, 2016).

77. Merkow, J., Kriegman, D., Marsden, A. & Tu, Z. Dense Volume-to-Volume Vascular Boundary Detection. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2016).

78. Milletari, F., Navab, N. & Ahmadi, S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *International Conference on 3D Vision* (2016).

79. Yu, L., Yang, X., Chen, H., Qin, J. & Heng, P. Volumetric ConvNets with Mixed Residual Connections for Automated Prostate Segmentation from 3D MR Images. *AAAI Conference on Artificial Intelligence* (2017).

80. Lai, M. Deep Learning for Medical Image Segmentation. *arXiv preprint arXiv:1505.02000* (2015).

81. Xia, F., Wang, P., Chen, L. & Yuille, A. Zoom Better to See Clearer: Human and Object Parsing with Hierarchical Auto-Zoom Net. *European Conference on Computer Vision* (2016).

82. Chen, H., Dou, Q., Wang, X., Qin, J. & Heng, P. Mitosis Detection in Breast Cancer Histology Images via Deep Cascaded Networks. *AAAI Conference on Artificial Intelligence* (2016).

83. Kuen, J., Wang, Z. & Wang, G. Recurrent Attentional Networks for Saliency Detection. *Computer Vision and Pattern Recognition* (2016).

84. Li, G., Xie, Y., Lin, L. & Yu, Y. Instance-Level Salient Object Segmentation. *Computer Vision and Pattern Recognition* (2017).

85. Lin, G., Milan, A., Shen, C. & Reid, I. RefineNet: Multi-Path Refinement Networks with Identity Mappings for High-Resolution Semantic Segmentation. *Computer Vision and Pattern Recognition* (2017).

86. Wang, X., Girshick, R., Gupta, A. & He, K. *Non-local neural networks* in *VPR* (2018), 7794–7803.

87. Wang, F. *et al. Residual attention network for image classification* in *CVPR* (2017), 3156–3164.

88. Hu, J., Shen, L. & Sun, G. *Squeeze-and-excitation networks* in *CVPR* (2018), 7132–7141.

89. Wang, Y. *et al.* Deep attentive features for prostate segmentation in 3d transrectal ultrasound. *TMI* **38,** 2768–2778 (2019).

90. Schlemper, J. *et al.* Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis* **53,** 197–207 (2019).

91. Zhou, Y. *et al. Multi-Scale Attentional Network for multi-focal segmentation of active bleed after pelvic fractures* in *International Workshop on Machine Learning in Medical Imaging* (2019), 461–469.

92. Roy, A. G., Navab, N. & Wachinger, C. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE transactions on medical imaging* **38,** 540–549 (2018).

93. Chen, H., Qi, X., Cheng, J.-Z. & Heng, P.-A. *Deep contextual networks for neuronal structure segmentation* in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (2016), 1167–1173.

94. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* **40,** 834–848 (2018).

95. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation (2018).

96. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. *Pyramid scene parsing network* in *CVPR* (2017).

97. Papandreou, G., Chen, L.-C., Murphy, K. & Yuille, A. L. *Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation* in *ICCV* (2015).

98. Pathak, D., Shelhamer, E., Long, J. & Darrell, T. Fully convolutional multi-class multiple instance learning (2015).

99. Dai, J., He, K. & Sun, J. *BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation* in *ICCV* (2015).

100. Souly, N., Spampinato, C. & Shah, M. *Semi supervised semantic segmentation using generative adversarial network* in *ICCV* (2017).

101. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015).

102. Chen, H., Dou, Q., Yu, L., Qin, J. & Heng, P.-A. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage* (2017).

103. Cheplygina, V., de Bruijne, M. & Pluim, J. P. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis* **54,** 280–296 (2019).

104. Blum, A. & Chawla, S. Learning from labeled and unlabeled data using graph mincuts (2001).

105. Wang, J., Jebara, T. & Chang, S.-F. Semi-supervised learning using greedy max-cut. *Journal of Machine Learning Research* **14,** 771–800 (2013).

106. Rosenberg, C., Hebert, M. & Schneiderman, H. Semi-supervised self-training of object detection models.

107. Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G. & He, K. *Data distillation: Towards omni-supervised learning* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 4119–4128.

108. Xu, C., Tao, D. & Xu, C. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634* (2013).

109. Ciurte, A. *et al.* Semi-supervised segmentation of ultrasound images based on patch representation and continuous min cut. *PloS one* **9,** e100972 (2014).

110. Gu, L. *et al. Semi-supervised learning for biomedical image segmentation via forest oriented super pixels (voxels)* in *MICCAI* (2017), 702–710.

111. Qin, C. *et al. A Semi-supervised Large Margin Algorithm for White Matter Hyperintensity Segmentation* in *MICCAI* (2016), 104–112.

112. Sousa, R. T. & Gama, J. Comparison between Co-training and Self-training for single-target regression in data streams using AMRules (2017).

113. Qiao, S., Shen, W., Zhang, Z., Wang, B. & Yuille, A. *Deep co-training for semi-supervised image recognition* in *Proceedings of the european conference on computer vision (eccv)* (2018), 135–152.

114. Zhang, Y. *et al. Deep adversarial networks for biomedical image segmentation utilizing unannotated images* in *MICCAI* (2017).

115. Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I. & de Bruijne, M. *Semi-supervised medical image segmentation via learning consistency under transformations* in *MICCAI* (2019), 810–818.

116. Li, X. *et al.* Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

117. Liu, Q., Yu, L., Luo, L., Dou, Q. & Heng, P. A. Semi-supervised medical image classification with relation-driven self-ensembling model. *TMI* (2020).

118. Yu, L., Wang, S., Li, X., Fu, C.-W. & Heng, P.-A. *Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation* in *MICCAI* (2019), 605–613.

119. Wang, X. *et al.* UD-MIL: Uncertainty-driven Deep Multiple Instance Learning for OCT Image Classification. *IEEE Journal of Biomedical and Health Informatics* (2020).

120. Rajchl, M. *et al.* Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging* **36,** 674–683 (2017).

121. Kervadec, H. *et al.* Constrained-CNN losses for weakly supervised segmentation. *Medical Image Analysis* (2019).

122. Dalca, A. V., Guttag, J. & Sabuncu, M. R. *Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), 9290–9299.

123. BenTaieb, A. & Hamarneh, G. *Topology aware fully convolutional networks for histology gland segmentation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2016), 460–468.

124. Chen, H. *et al.* DCAN: Deep contour-aware networks for object instance segmentation from histology images. *Medical image analysis* **36,** 135–146 (2017).

125. Wachinger, C., Reuter, M. & Klein, T. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage* **170,** 434–445 (2018).

126. Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P. & Vaidya, V. *Learning and incorporating shape models for semantic segmentation* in *International conference on medical image computing and computer-assisted intervention* (2017), 203–211.

127. Oktay, O. *et al.* Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging* **37,** 384–395 (2018).

128. Wang, Y. *et al.* Training Multi-organ Segmentation Networks with Sample Selection by Relaxed Upper Confident Bound. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018).

129. Li, Q., Wang, J., Wipf, D. & Tu, Z. Fixed-Point Model For Structured Labeling. *International Conference on Machine Learning* (2013).

130. Zhang, Y., Ying, M., Yang, L., Ahuja, A. & Chen, D. Coarse-to-Fine Stacked Fully Convolutional Nets for Lymph Node Segmentation in Ultrasound Images. *IEEE International Conference on Bioinformatics and Biomedicine* (2016).

131. Everingham, M., Van Gool, L., Williams, C., Winn, J. & Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* **88,** 303–338 (2010).

132. Zhu, Z., Xie, L. & Yuille, A. Object Recognition with and without Objects. *International Joint Conference on Artificial Intelligence* (2017).

133. Dmitriev, K., Gutenko, I., Nadeem, S. & Kaufman, A. *Pancreas and cyst segmentation* in *Medical Imaging 2016: Image Processing* **9784** (2016), 97842C.

134. Al-Ayyoub, M., Alawad, D., Al-Darabsah, K. & Aljarrah, I. Automatic detection and classification of brain hemorrhages. *WSEAS transactions on computers* **12,** 395–405 (2013).

135. Gintowt, A., Hac, S., Dobrowolski, S. & Śledziński, Z. An unusual presentation of pancreatic pseudocyst mimicking cystic neoplasm of the pancreas: a case report. *Cases journal* **2,** 9138 (2009).

136. Lasboo, A. A., Rezai, P. & Yaghmai, V. Morphological analysis of pancreatic cystic masses. *Academic radiology* **17,** 348–351 (2010).

137. Klauß, M. *et al.* Value of three-dimensional reconstructions in pancreatic carcinoma using multidetector CT: initial results. *World journal of gastroenterology: WJG* **15,** 5827 (2009).

138. Lee, C., Xie, S., Gallagher, P., Zhang, Z. & Tu, Z. Deeply-Supervised Nets. *International Conference on Artificial Intelligence and Statistics* (2015).

139. Mharib, A. M., Ramli, A. R., Mashohor, S. & Mahmood, R. B. Survey on liver CT image segmentation methods. *Artificial Intelligence Review* **37** (2012).

140. Li, G. *et al.* Automatic liver segmentation based on shape constraints and deformable graph cut in CT images. *IEEE Transactions on Image Processing* **24,** 5315–5329 (2015).

141. Kirbas, C. & Quek, F. A review of vessel extraction techniques and algorithms. *ACM Computing Surveys* **36,** 81–121 (2004).

142. Lesage, D., Angelini, E. D., Bloch, I. & Funka-Lea, G. A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes. *Medical Image Analysis* **13,** 819–845 (2009).

143. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. *3D U-Net: learning dense volumetric segmentation from sparse annotation* in *International conference on medical image computing and computer-assisted intervention* (2016), 424–432.

144. Nascimento, J. & Carneiro, G. Multi-atlas segmentation using manifold learning with deep belief networks. *Proceedings - International Symposium on Biomedical Imaging* **2016-June** (2016).

145. Setio, A. A. A. *et al.* Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Transactions on Medical Imaging* **35,** 1160–1169 (2016).

146. Zu, C. *et al.* Robust multi-atlas label propagation by deep sparse representation. *Pattern Recognition* **63,** 511–517 (2017).

147. Chen, L., Yang, Y., Wang, J., Xu, W. & Yuille, A. Attention to Scale: Scale-aware Semantic Image Segmentation. *Computer Vision and Pattern Recognition* (2016).

148. Kong, T. *et al. RON: Reverse Connection with Objectness Prior Networks for Object Detection* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

149. Warfield, S. K., Zou, K. H. & Wells, W. M. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging* **23,** 903–921. arXiv: arXiv:1011.1669v3 (2004).

150. Lugo-Fagundo, C., Vogelstein, B., Yuille, A. & Fishman, E. K. Deep learning in radiology: Now the real work begins. *Journal of the American College of Radiology* **15,** 364–367 (2 2018).

151. Xie, S. & Tu, Z. *Holistically-Nested Edge Detection* in *IEEE International Conference on Computer Vision (ICCV)* (2015).

152. Asman, A. J. & Landman, B. A. Formulating spatially varying performance in the statistical fusion framework. *IEEE Transactions on Medical Imaging* **31,** 1326–1336 (2012).

153. Sabuncu, M., Yeo, B., van Leemput, K., Fischi, B. & Golland, P. A generative model for image segmentation based on label fusion. *IEEE Transactions on Medical Imaging* **29,** 1714–1729 (10 2010).

154. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transcations on Image Processing* **13,** 600–612 (2004).

155. Boykov, Y., Veksler, O. & Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23,** 1222–1239 (2001).

156. Shelhamer, E., Long, J. & Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39,** 640–651 (Apr. 2017).

157. Saito, A., Nawano, S. & Shimizu, A. Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs. *Medical Image Analysis* **28,** 46–65 (2106).

158. Tang, P., Wang, X., Bai, X. & Liu, W. *Multiple instance detection network with online instance classifier refinement* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), 2843–2851.

159. Tang, P. *et al.* Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence* **42,** 176–191 (2018).

160. Tang, P. *et al. Weakly supervised region proposal network and object detection* in *Proceedings of the European conference on computer vision (ECCV)* (2018), 352–368.

161. Ciresan, D., Giusti, A., Gambardella, L. M. & Schmidhuber, J. *Deep neural networks segment neuronal membranes in electron microscopy images* in *Advances in neural information processing systems* (2012), 2843–2851.

162. Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D. & Hammers, A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage* **33,** 115–126 (2006).

163. Rohlfing, T., Brandt, R., Menzel, R. & Maurer Jr, C. R. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage* **21,** 1428–1442 (2004).

164. Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V. & Rueckert, D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage* **46,** 726–738 (2009).

165. Asman, A. J. & Landman, B. A. Non-local statistical label fusion for multi-atlas segmentation. *Medical image analysis* **17,** 194–208 (2013).

166. Harrison, A. P. *et al. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images* in *International conference on medical image computing and computer-assisted intervention* (2017), 621–629.

167. Prasoon, A. *et al. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network* in *International conference on medical image computing and computer-assisted intervention* (2013), 246–253.

168. Lin, D., Dai, J., Jia, J., He, K. & Sun, J. *Scribblesup: Scribble-supervised convolutional networks for semantic segmentation* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 3159–3167.

169. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y. & Schroers, C. *Normalized cut loss for weakly-supervised CNN segmentation* in *CVPR* (2018), 1818–1827.

170. Tang, M. *et al. On regularized losses for weakly-supervised cnn segmentation* in *ECCV* (2018), 507–522.

171. Bearman, A., Russakovsky, O., Ferrari, V. & Fei-Fei, L. *What's the point: Semantic segmentation with point supervision* in *European conference on computer vision* (2016), 549–565.

172. Xu, J., Schwing, A. G. & Urtasun, R. *Tell me what you see and i will show you where it is* in *CVPR* (2014), 3190–3197.

173. Simpson, A. L. *et al.* A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019).

174. Liu, Y., Chen, J. & Deng, L. *An Unsupervised Learning Method Exploiting Sequential Output Statistics* in *NIPS* (2017).

175. Nogues, I. *et al. Automatic lymph node cluster segmentation using holistically-nested neural networks and structured optimization in CT images* in *MICCAI* (2016).

176. Pawlowski, N. *et al.* Dltk: State of the art reference implementations for deep learning on medical images. *arXiv preprint arXiv:1711.06853* (2017).

177. Heinrich, M. P. Multi-Organ Segmentation using deeds, Self-Similarity Context and Joint Fusion (2015).

178. Attiyeh, M. A., Chakraborty, J., Doussot, A., Langdon-Embry, L., Mainarich, S., *et al.* Survival prediction in pancreatic ductal adenocarcinoma by quantitative computed tomography image analysis. *Annals of surgical oncology* **25** (2018).

179. Zhu, Z., Xia, Y., Xie, L., Fishman, E. K. & Yuille, A. L. Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019).

180. Zhu, W. *et al.* AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics* **46,** 576–589 (2019).

181. Liu, F., Xie, L., Xia, Y., Fishman, E. & Yuille, A. *Joint shape representation and classification for detecting PDAC* in *International Workshop on Machine Learning in Medical Imaging* (2019), 212–220.

182. Zhu, Z., Xia, Y., Shen, W., Fishman, E. & Yuille, A. *A 3D coarse-to-fine framework for volumetric medical image segmentation* in *2018 International Conference on 3D Vision (3DV)* (2018), 682–690.

183. Dolz, J. *et al.* HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE Transactions on Medical Imaging* (2018).

184. Li, Y. *et al.* Multimodal hyper-connectivity of functional networks using functionally-weighted LASSO for MCI classification. *Medical image analysis* **52,** 80–96 (2019).

185. Zhang, W. *et al.* Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage* **108,** 214–224 (2015).

186. Yao, J., Zhu, X., Zhu, F. & Huang, J. *Deep correlational learning for survival prediction from multi-modality data* in *MICCAI* (2017), 406–414.

187. Vercauteren, T., Pennec, X., Perchange, A. & Ayache, N. Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage* **45,** S61–S82 (2009).

188. Reaungamornrat, S. *et al.* MIND demons: symmetric diffeomorphic deformable registration of MR and CT for image-guided spine surgery. *IEEE Transactions on Medical Imaging* **35,** 2413–2424 (2016).

189. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. *mixup: Beyond empirical risk minimization* in *ICLR* (2018).

190. Chu, L. C. *et al.* Application of deep learning to pancreatic cancer detection: Lessons learned from our initial experience. *Journal of the American College of Radiology* **16,** 1338–1342 (2019).

191. Dreizin, D. *et al.* Deep learning-based quantitative visualization and measurement of extraperitoneal hematoma volumes in patients with pelvic fractures: Potential role in personalized forecasting and decision support. *Journal of Trauma and Acute Care Surgery* **88,** 425–433 (2020).

192. Dreizin, D., Zhou, Y., Zhang, Y., Tirada, N. & Yuille, A. L. Performance of a deep learning algorithm for automated segmentation and quantification of traumatic pelvic hematomas on CT. *Journal of Digital Imaging* **33,** 243–251 (2020).

193. Fu, S. *et al. Domain Adaptive Relational Reasoning for 3D Multi-Organ Segmentation* in *MICCAI* (2020).

194. Liu, D. *et al.* Dynamic Graph Correlation Learning for Disease Diagnosis with Incomplete Labels. *arXiv preprint arXiv:2002.11629* (2020).

195. Zheng, H. *et al. Cartilage Segmentation in High-Resolution 3D Micro-CT Images via Uncertainty-Guided Self-training with Very Sparse Annotation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 802–812.

196. Xia, Y. *et al.* Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical Image Analysis* **65,** 101766 (2020).

197. Raju, A. *et al.* Co-Heterogeneous and Adaptive Segmentation from Multi-Source and Multi-Phase CT Imaging Data: A Study on Pathological Liver and Lesion Segmentation. *arXiv preprint arXiv:2005.13201* (2020).

198. Huo, X., Xie, L., He, J., Yang, Z. & Tian, Q. ATSO: Asynchronous Teacher-Student Optimizationfor Semi-Supervised Medical Image Segmentation. *arXiv preprint arXiv:2006.13461* (2020).

199. Mondal, A. K., Dolz, J. & Desrosiers, C. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv preprint arXiv:1810.12241* (2018).

200. Peng, J., Estrada, G., Pedersoli, M. & Desrosiers, C. Deep co-training for semi-supervised image segmentation. *Pattern Recognition,* 107269 (2020).

201. Kervadec, H., Dolz, J., Granger, É. & Ayed, I. B. *Curriculum semi-supervised segmentation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), 568–576.

202. Tang, P., Ramaiah, C., Xu, R. & Xiong, C. Proposal Learning for Semi-Supervised Object Detection. *arXiv preprint arXiv:2001.05086* (2020).

203. Kervadec, H. *et al.* Constrained-cnn losses forweakly supervised segmentation. *arXiv preprint arXiv:1805.04628* (1805).

204. Peng, J., Pedersoli, M. & Desrosiers, C. *Mutual information deep regularization for semi-supervised segmentation* in *Medical Imaging with Deep Learning* (2020), 601–613.

205. Tang, Y. *et al.* Learning from dispersed manual annotations with an optimized data weighting policy. *Journal of Medical Imaging* **7,** 044002 (2020).

206. Shi, G., Xiao, L., Chen, Y. & Zhou, S. K. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *arXiv preprint arXiv:2007.03868* (2020).

207. Peng, J. *et al.* Discretely-constrained deep network for weakly supervised segmentation. *Neural Networks* **130,** 297–308 (2020).

208. Luo, X. *et al.* Semi-supervised Medical Image Segmentation through Dual-task Consistency. *arXiv preprint arXiv:2009.04448* (2020).

209. Valvano, G., Leo, A. & Tsaftaris, S. A. Weakly Supervised Segmentation with Multi-scale Adversarial Attention Gates. *arXiv preprint arXiv:2007.01152* (2020).

210. Huang, R., Zheng, Y., Hu, Z., Zhang, S. & Li, H. *Multi-organ Segmentation via Co-training Weight-Averaged Models from Few-Organ Datasets* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 146–155.

211. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H. & Ayed, I. B. Source-Relaxed Domain Adaptation for Image Segmentation. *arXiv preprint arXiv:2005.03697* (2020).

212. Jacobzon, G. *Multi-site Organ Detection in CT Images using Deep Learning* 2020.

213. Ouyang, C. *et al. Self-supervision with Superpixels: Training Few-Shot Medical Image Segmentation Without Annotation* in *European Conference on Computer Vision* (2020), 762–780.

214. Zou, D., Zhu, Q. & Yan, P. Unsupervised Domain Adaptation with Dual-Scheme Fusion Network for Medical Image Segmentation.

215. Fang, X. & Yan, P. Multi-organ Segmentation over Partially Labeled Datasets with Multi-scale Feature Abstraction. *arXiv preprint arXiv:2001.00208* (2020).

216. Dou, Q., Ouyang, C., Chen, C., Chen, H. & Heng, P.-A. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv preprint arXiv:1804.10916* (2018).

217. Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J. & Rueckert, D. *Data efficient unsupervised domain adaptation for cross-modality image segmentation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), 669–677.

218. Chen, S., Bortsova, G., Juárez, A. G.-U., van Tulder, G. & de Bruijne, M. *Multi-task attention-based semi-supervised learning for medical image segmentation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), 457–465.

219. Wang, Y. *et al. Deep distance transform for tubular structure segmentation in ct scans* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 3833–3842.

220. Chen, C. *et al. Learning shape priors for robust cardiac MR segmentation from multi-view images* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), 523–531.

221. Xie, X., Niu, J., Liu, X., Chen, Z. & Tang, S. A Survey on Domain Knowledge Powered Deep Learning for Medical Image Analysis. *arXiv preprint arXiv:2004.12150* (2020).

222. Song, Y. *et al. Learning 3D Features with 2D CNNs via Surface Projection for CT Volume Segmentation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 176–186.

223. Román, K. L.-L., Ocaña, M. I. G., Urzelai, N. L., Ballester, M. Á. G. & Oliver, I. M. in *Deep Learning in Healthcare* 17–31 (Springer, 2020).

224. Yu, Q., Xia, Y., Xie, L., Fishman, E. K. & Yuille, A. L. Thickened 2D Networks for Efficient 3D Medical Image Segmentation. *arXiv preprint arXiv:1904.01150* (2019).

225. Xia, Y. *et al. Bridging the gap between 2d and 3d organ segmentation with volumetric fusion net* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), 445–453.

226. Ni, T., Xie, L., Zheng, H., Fishman, E. K. & Yuille, A. L. *Elastic boundary projection for 3D medical image segmentation* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), 2109–2118.

227. Li, Y. *et al. Neural Architecture Search for Lightweight Non-Local Networks* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 10297–10306.

228. Yu, Q., Li, Y., Mei, J., Zhou, Y. & Yuille, A. L. CAKES: Channel-wise Automatic KErnel Shrinking for Efficient 3D Network. *arXiv preprint arXiv:2003.12798* (2020).

229. Xia, Y. *et al. Detecting Pancreatic Ductal Adenocarcinoma in Multi-phase CT Scans via Alignment Ensemble* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 285–295.

230. Liu, F., Zhou, Y., Fishman, E. & Yuille, A. *FusionNet: Incorporating Shape and Texture for Abnormality Detection in 3D Abdominal CT Scans* in *International Workshop on Machine Learning in Medical Imaging* (2019), 221–229.

231. Xie, C. *et al. Adversarial Examples for Semantic Segmentation and Object Detection* in *IEEE International Conference on Computer Vision (ICCV)* (2017).

232. Xie, C. *et al. Improving transferability of adversarial examples with input diversity* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 2730–2739.

233. Li, Y. *et al. Learning Transferable Adversarial Examples via Ghost Networks* in *AAAI Conference on Artificial Intelligence (AAAI)* (2020).

234. Huang, L. *et al. UPC: Learning Universal Physical Camouflage Attacks on Object Detectors* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).

235. Bai, S., Li, Y., Zhou, Y., Li, Q. & Torr, P. H. Metric Attack and Defense for Person Re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

236. Li, Y. *et al.* in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics* 69–91 (Springer, 2019).

Vita

Yuyin Zhou is completing her Ph.D. degree of Computer Science at the Johns Hopkins University, under the supervision of Bloomberg Distinguished Professor Alan L. Yuille. Yuyin received her M.S. degree from University of California, Los Angeles (UCLA) in 2016, and the B.S. degree from Huazhong University of Science and Technology in 2014. Yuyin's research interests span in the fields of medical image computing, computer vision, and machine learning, especially the intersection of them. She also worked at Google Cloud AI and Google Brain.