UCLA
UCLA Electronic Theses and Dissertations

Title
Medical Note and Image Processing with Physical Models and Deep Learning Techniques

Permalink
https://escholarship.org/uc/item/45g3z07w

Author
Zhou, Hanyue

Publication Date
2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Medical Note and Image Processing with Physical Models and Deep Learning Techniques

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioengineering

by

Hanyue Zhou

2022

© Copyright by
Hanyue Zhou
2022

ABSTRACT OF THE DISSERTATION

Medical Note and Image Processing
with Physical Models and Deep Learning Techniques

by

Hanyue Zhou
Doctor of Philosophy in Bioengineering
University of California, Los Angeles, 2022
Professor Dan Ruan, Co-Chair
Professor Zhaoyang Fan, Co-Chair

We aim to perform medical note comprehension and medical image processing, with an ultimate goal of cross-domain aggregation into a cooperative disease management system. The dissertation focuses on the initial technical development of each domain utilizing both physical modeling and deep learning methods. Medical notes and images taken from patients during their clinical visits are essential for patient care management. In this thesis, natural language processing techniques are developed for patient private information removal from medical reports, and image processing techniques are developed for semantic segmentation on different imaging modalities to achieve higher accuracy and enhanced structural integrity of the segmentation. Moreover, we demonstrate the importance of the manual labels used as the ground truth for supervised learning and assessment in the biomedical applications, and further propose a refinement scheme to improve label quality. Future directions would be integrating the complementary text and image information into a single robust system.

The dissertation of Hanyue Zhou is approved.

Corey Wells Arnold

Holden H. Wu

Zhaoyang Fan, Committee Co-Chair

Dan Ruan, Committee Co-Chair

University of California, Los Angeles

2022

To my mentors . . .

who—guide me through finishing my PhD,

always inspire me, and point me to the best resources.

To Dr. Ruan . . .

who brings me into the field of deep learning,

always helps me out the most when I need help.

To my family . . .

who provide me with the best emotional support.

To my friends . . .

who are the biggest treasure I found in this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my mentors Dr. Dan Ruan and Dr. Zhaoyang Fan for guiding me through all the projects implementation and writing in PhD studies. I would also like to thank my committee Dr. Holden Wu and Dr. Corey Arnold for constructive advice and great efforts in supervision.

VITA

2016 B.Eng. (Communication Engineering), Harbin Institute of Technology.

2016–2022 Graduate Student Researcher, Bioengineering Department, UCLA.

PUBLICATIONS

Zhou, Hanyue & Li, Ying & Hsin, Yue-Loong & Liu, Wentai. (2016). Phase-amplitude Coupling Analysis for Seizure Evolvement using Hilbert Huang Transform. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference. 2016. 1022-1025. 10.1109/EMBC.2016.7590876.

Zhou, Hanyue & Wang, Yushan & Li, Ying & Ruan, Dan & Liu, Wentai. (2018). Improving EEG Source Localization with a Novel Regularization: Spatiotemporal Graph Total Variation (STGTV) Method. Conference proceedings: IEEE Engineering in Medicine and Biology Society. Conference. 2018. 4673-4676. 10.1109/EMBC.2018.8513128.

Wang, Yushan & Zhou, Hanyue & Li, Ying & Liu, Wentai. (2018). Impact of Electrode Number on the Performance of High-Definition Transcranial Direct Current Stimulation (HD-tDCS). Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference. 2018. 4182-4185. 10.1109/EMBC.2018.8513379.

Zhou, Hanyue & Ruan, Dan. (2020). An Embedding-based Medical Note De-identification Approach with Minimal Annotation. 263-268. 10.1109/BIBE50027.2020.00050.

Zhou, Hanyue & Ruan, Dan. (2020). Technical Note: An Embedding-based Medical Note

De-identification Approach with Sparse Annotation. Medical Physics. 48. 10.1002/mp.14664.

Zhou, Hanyue & Xiao, Jiayu & Fan, Zhaoyang & Ruan, Dan. (2021). Intracranial Vessel Wall Segmentation For Atherosclerotic Plaque Quantification. Proceedings. IEEE International Symposium on Biomedical Imaging. 2021. 1416-1419. 10.1109/ISBI48211.2021.9434018.

Zhou, Hanyue & Cao, Minsong & Ma, Martin & Yoon, Stephanie & Kishan, Amar & Ruan, Dan. (2022). Technical Note: Air Bubble-induced Performance Degradation in Automatic Rectum Segmentation from Cone-beam CT. Medical Physics. 10.1002/mp.15443.

Zhou, Hanyue & Cao, Minsong & Min, Yugang & Yoon, Stephanie & Kishan, Amar & Ruan, Dan. (2022). Ensemble Learning and Tensor Regularization for Cone-beam Computed Tomography-based Pelvic Organ Segmentation. Medical Physics. 10.1002/mp.15475.

Zhou, Hanyue & Xiao, Jiayu & Li, Debiao & Fan, Zhaoyang & Ruan, Dan. (2022). Intracranial Vessel Wall Segmentation with Deep Learning using a Novel Tiered Loss Function to Incorporate Class Inclusion. Proceedings. IEEE International Symposium on Biomedical Imaging. 2022.

Zhou, Hanyue & Cao, Minsong & Min, Yugang & Yoon, Stephanie & Kishan, Amar & Ruan, Dan. (2022). Ensemble Learning and Tensor Regularization for Cone-beam Computed Tomography-based Pelvic Organ Segmentation. IEEE International Symposium on Biomedical Imaging. 2022.

Zhou, Hanyue & Xiao, Jiayu & Li, Debiao & Fan, Zhaoyang & Ruan, Dan. (2022). Intracranial Vessel Wall Segmentation with Deep Learning using a Novel Tiered Loss Function to Incorporate Class Inclusion. Under review by Medical Physics.

Zhou, Hanyue & Xiao, Jiayu & Fan, Zhaoyang & Ruan, Dan. (2022). Deep Learning-guided Iterative Refinement to Improve Data Quality and Label Consistency. Under review by Biomedical and Health Informatics.

Zhou, Hanyue & Xiao, Jiayu & Ruan, Dan & Fan, Zhaoyang. (2022). Vessel Wall Imaging-Dedicated Automatic Processing Pipeline (VWI-APP): Towards Efficient and Reliable Intracranial Plaque Quantification. Preparing.

CHAPTER 1

Introduction

The high-level goal of this study is to perform text understanding and image processing in the biomedical field and ultimately integrate both data modalities into a robust disease management system.

Today, with the development of computational power in hardware and the increasing availability of public data, deep neural networks have seen much progress in various tasks involving speech, text, and image processing. In supervised learning, different from the conventional machine learning which heavily depends on feature engineering, deep learning models can take minimal preprocessed data and automatically extract and refine the useful features for improving the objective of a task as intermediate representations, which have witnessed the state-of-the-art performances. Deep learning can be even more beneficial when applied to unsupervised tasks, as unlabeled data are usually abundant but the labels are much harder to obtain.

In deep learning, speech and text are typically processed by sequence models with flexible input and output sizes, such as by recurrent neural network (RNN) either unidirectional or bidirectional [1, 2], and its common-use variants gated recurrent unit (GRU) [3] and long short-term memory (LSTM) [4], which solve the problem of vanishing gradients and can capture longer dependencies by the introduction of multiple gated functions. Further developments target on introducing more contextual information across long sequences and faster network training by enabling parallelization. Such examples include attention schemes which weigh each part of input sequence differently by its relevance to a task [5], and transformers

[6] which allow training parallelization by processing input not in sequence order but providing context for any position in the input. Common natural language processing (NLP) tasks include machine translation, speech recognition, sentiment classification, named entity recognition, and text summarization, etc. Recently, pretrained models which only require end-to-end fine-tuning are handy baselines to use for various language processing tasks. The most famous pretrained models include bidirectional encoder representations from transformers (BERT) [7], and efficiently learning an encoder that classifies token replacements accurately (ELECTRA) [8], etc., pretrained with designed language modeling tasks on large corpora.

Deep learning-based image processing are widely applied to image classification, object detection, image registration, and semantic segmentation, etc. Convolutional neural network (CNN) with its shift-invariant properties and more efficient use of trainable parameters compared to fully connected network (FCN) well solves the image tasks. Popular and useful structures of CNN are UNet and its variants [9, 10] for semantic segmentation tasks where the network can propagate context information to higher resolution layers, You-only-look-once (YOLO) for real-time object detection and classification [11], Mask R-CNN for simultaneous localization and instance segmentation [12], Residual Network (ResNet) for image recognition where its skip connections enable networks to go deeper [13], and generative adversarial network (GAN) [14] and cycle-GAN [15] as generative models for image synthesis, etc. Furthermore, most recent architectures also utilize RNN types of models [16] and transformers [17, 18] to capture long-term contextual information in images, which have been shown to outperform CNN models in some tasks.

In the biomedical field, which is the primary interest of the study in this thesis, both medical reports and scanned images play essential roles in facilitating disease diagnosis and treatment planning. Deep learning can be applied to both modalities to help with automated diagnosis and reducing human efforts. In this thesis, both lines of research have been conducted.

To protect patient privacy, the removal of patient health information (PHI) including patient name, address, phone number, and social security number, etc. is required before a medical note can be used in research. The first project proposes a novel de-identification method which combines unsupervised and supervised modules to achieve high detection accuracy of names on prostate cancer (PC) patients' clinical reports.

In the line of image processing, we focus on specific network structure and cost function design to perform semantic segmentation on two different imaging modalities, cone-beam computed tomography (CBCT) and magnetic resonance imaging (MRI), to save human efforts as well as increase contouring consistency. CBCT is taken from PC patients enrolled in stereotactic body radiotherapy (SBRT), and the contouring of organ-at-risk and target is crucial for dose calculation and ultimately for the treatment outcome of radiotherapy. The T1-weighted MRI is scanned for patients diagnosed with intracranial atherosclerosis disease (ICAD), where the vessel wall remodeling and thickening is monitored by the modality. The segmentation task focuses on the automatic segmentation of inner and outer vessel wall boundaries, based on which the remodeling ratio and plaque burden are calculated.

By analyzing the segmentation results generated by the developed neural networks, we observe that the low image quality with large presence of artifacts, as well as manual labeling variations are the primary reasons for restraining segmentation accuracy, and a refinement scheme is therefore proposed to improve the quality of manual labels.

This thesis contributes to medical text and image processing by developing physical-driven rationales which are further integrated with deep learning techniques. Our study also shows the importance of assessing and improving the quality of manual labels for the learning and assessment of neural networks in supervised setting, beyond merely accumulating data quantity. The proposal of various techniques in pre- and post-processing of medical text and images, the assessment of performance degradation, as well as a scheme for manual label refinement lead to a rich and comprehensive study.

The contents of the thesis are outlined as follows. Chapter 2 focuses on the note de-

identification approach for the text processing track. Chapter 3 proposes an organ segmentation method based on CBCT and analyzes the relationship between image quality and the segmentation performance. Chapter 4 investigates the intracranial vessel wall segmentation task based on MRI, and proposes two developed segmentation neural networks. Chapter 5 explains the proposed manual label refinement process with intracranial vessel wall segmentation as a particular use case. Finally, chapter 6 summarizes the whole thesis and points out the future directions to pursue to extend the current studies.

CHAPTER 2

An Embedding-based Medical Note De-identification Approach with Sparse Annotation

2.1 INTRODUCTION

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) specifies 18 categories of protected health information (PHI) that must be removed before a medical note can be used in research [19]. The categories include names, locations, dates, telephone numbers, social security numbers, medical record numbers, and other sensitive information of a patient. For both legal and ethical reasons, removing PHI from medical notes is essential for patient privacy protection and community data sharing.

Among all the categories, names can be the most difficult to remove compared to others. The numeric format can be exploited to detect phone numbers, SSN, and MRN, etc., and specific semantic format can be utilized for detecting addresses that usually occur in particular locations in a note. On the other hand, names can appear multiple times in arbitrary locations, and with different combinations of first, middle, and last names referring to the same person. Furthermore, certain name strings can coincide with other non-sensitive text contents, such as "Mr. Parkinson" vs. "Parkinson's disease". Therefore, processing on the individual word level, which most of the pattern matching approaches adopt, will result in an inevitable trade-off between type I and type II errors [20]. Also, there can be multiple "target" names within a single note, such as relative names in family history and practitioner names, and this makes approaches trying to create PHI dictionaries for each patient

inefficient [21].

Apart from the pattern matching approaches, there also exist de-identification methods based on natural language processing (NLP) techniques [22, 23]. The NLP-based methods usually utilize entity tagging mechanisms, as they run through a whole report or an entire sentence and classify each word in a sequence into different categories. However, these methods require a large number of word-wise annotated texts to train a good model. To address the major hurdle of the NER models that require a large number of training labels, this study investigates an alternative to automate name detection with as few training tagging as possible. Our method applies to other strings as well, and in this paper, we use name removal as a specific de-identification task.

While admitting the variations of name format as well as the usage in multiple places in a note, e.g., from conversation recording to symptom and history review, the contexts around names still fall into a few typical cases. Therefore, we hypothesize that using a vector embedding model which implicitly captures the word context would benefit in 1) addressing name format variations, e.g., "John Grant Doe" and "John G. Doe"; and 2) distinguishing names from non-name words that share spellings, e.g., "Parkinson".

In this work [24], we pre-trained word embeddings by the word2vec model using our clinical reports [25]. Then the contextual embedding of a word was obtained as the cooccurrence-weighted sum of the word2vec embeddings within a context window [26]. This simple modification assigned a polysemy with different embeddings depending on its contexts. This alleviated the pattern matching issues from neglecting contexts when classifying words. A small number of name instances identified from a single medical record, which we call landmarks, were the references for name sub-clusters in the contextualized embedding space. Each word was classified based on its adjacency, defined as the cosine similarity in the contextualized embedding space, to the elements in the landmark set. The specific binary classification of name vs. non-name was obtained with a multilayer perceptron model (MLP), optimized by supervised learning from the pre-constructed name and non-name contextual instances. The

overall pipeline is illustrated in Fig. 2.1.

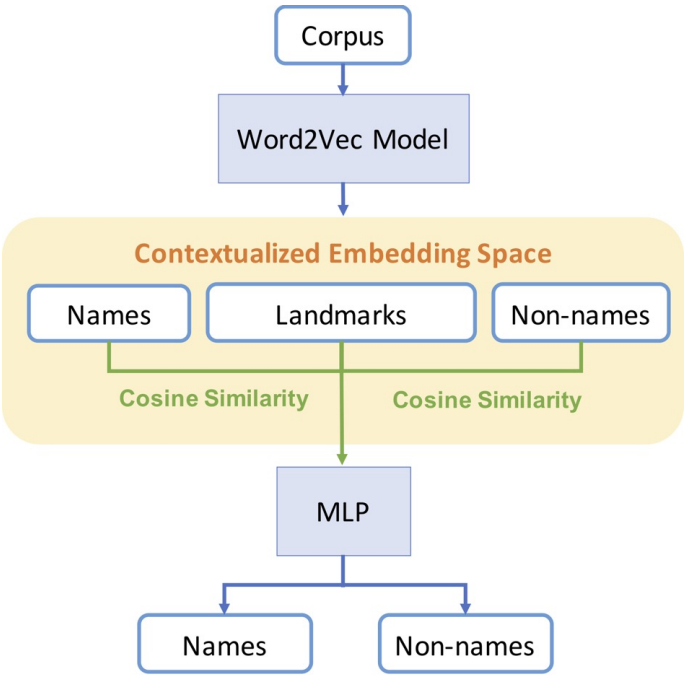


Figure 2.1: Overall pipeline of the proposed note de-identification method

The labeled instances of name and non-name words were partitioned into four distinct subsets. Among the subsets, the first one was used for training the MLP, the second one was used as validation for optimizing the number of landmarks, and the third one was used as validation for optimizing the size of the hidden layer and activation functions in the MLP. Finally, the fourth subset was held out for testing. Our model achieved > 0.99 in accuracy for each set. We illustrate two example reports of our collection, and report the sensitivity and specificity of the two reports for our proposed model and the Stanford 3-class NER model. We also report the F1 score of six randomly picked reports for both models. We show that our name detection and removal performance is superior to the Stanford NER, with a significantly higher F1 score achieved.

2.2 RELATED WORK

The existing de-identification approaches mainly fall into two categories: pattern matching models [20, 21], and named entity recognition models [22, 23, 27]. There are also hybrid models combining the two approaches [28].

2.2.1 Pattern-matching Models

The pattern-matching de-identification approaches are intuitive as they perform lexical matching with look-up tables and regular expressions. A very comprehensive note de-identification work was performed by Neamatullah et al. to remove every PHI aspect from medical records [20]. They identified numeric PHI instances by finding numeric patterns and contextual keywords. Non-numeric PHI, such as names and locations, were detected by contextual keywords and dictionary look-ups. Specifically, four look-up tables were constructed: full patient names, generic male and female names, key name indicators like prefixes and suffixes, and non-PHI tables. Another work created dictionaries using patient identifiers, including first, middle and last name, current and old address, postcode, phone number, email address, date of birth, and ID numbers [21]. The patient identifiers were identified and masked out when they appeared in medical notes. The drawbacks of the pattern-matching models are that they need to be fine-tuned for each specific dataset, and they enforce a single tag for a word with polysemy without differentiation of contexts.

2.2.2 Named Entity Recognition Models

Named entity recognition (NER) models tag each word in a sequence as an entity, e.g., PHI or not PHI; name, location, dates, or others. He et al. proposed a method based on conditional random fields (CRF) [22]. They extracted lexical, orthographic, and dictionary features, which contained the information of word casing, lengths, and common-use country and city vocabularies, etc. A CRF classifier uses the extracted features and predicts a sequence of

possible labels y given a sequence of tokens x, by maximizing the conditional probability $P(y|x)$.

The recurrent neural network (RNN) was utilized to automate the feature extraction with a token classification, which required no handcrafted features or rules [23]. The de-identification system was composed of 3 layers: an embedding layer, a label prediction layer, and a label-sequence optimization layer. The embedding layer maps each token to a vector, the label prediction layer outputs the probabilities of each label for each token in a sequence, and the final sequence-optimization layer outputs the most likely sequence of predicted labels.

2.2.3 Named Entity Recognition Models using Distributional Features

Within the class of NER models, there exist models that perform clustering in semantic space, and they usually incorporate clustering results as an additional feature to the final classification model [29].

Henriksson et al. learned vector representations of each named entity class in the semantic space by the utilization of a random indexing model over a large corpus [29]. For each class, the semantic feature was a binary value based on whether the cosine similarity between a target word and the prototype vector exceeded a certain threshold. It was shown that the use of additional semantic features improved de-identification accuracy, compared to its counterpart with orthographic and syntactic features only. The major drawback of the NER type of models is that training a tagger model requires a large number of labeled data, which is demanding of manual effort.

2.3 CLINICAL REPORT DESCRIPTION

Under IRB approval, 5200 clinical reports were collected across various stages of prostate cancer management, including consultations, on-treatment visits, phone encounters, treatment records, and follow-ups. The reports take on various formats and contain different

types of information written in no specific standard. In addition, they may contain one or more name instances of the patient, family members, and care-team members.

2.4 METHODS

2.4.1 Preprocessing and Data Preparation

We first constructed a large corpus using all the reports of our collection. We tokenized the corpus by the NLTK tokenizer [30]. We then removed punctuation, numbers that were more than four digits (which included the note ID and MRN in our case), tokens containing ":" and tokens with a format of "number/number", which mainly represent times and dates respectively. We did not perform case conversion because leading upper cases could be a good indicator of names. This preprocessing step managed to eliminate most of the ID numbers, dates, and times by format matching.

We arbitrarily picked a medical note and took all the word instances of the first, middle, and last names of the patient and care-givers as potential landmarks. We constructed 1400 word instances of names and 1400 word instances of non-names, and split them into 1100 name word instances / 1100 non-name word instances as the training set, 100/100 as a validation set A, 100/100 as a validation set B, and 100/100 for testing. Each word instance can be seen as a trigram with a name word in the middle with the left and right neighbors being its context words. The validation set A would be used for investigating the optimal number of landmarks, and set B would be used to optimize the hyper-parameters of the classifier. The details of data construction and hyper-parameter tuning will be elaborated further in the experiment section.

2.4.2 Initial Embeddings with the Word2vec Model

The word2vec model has two variations: CBOW predicts a target word from a window of its context words, while the skip-gram predicts the surrounding words by a target word, shown in Fig. 2.2 [25]. Between the two models, skip-gram assigns the nearby context words more weights than distant context words and was adopted in our pipeline.

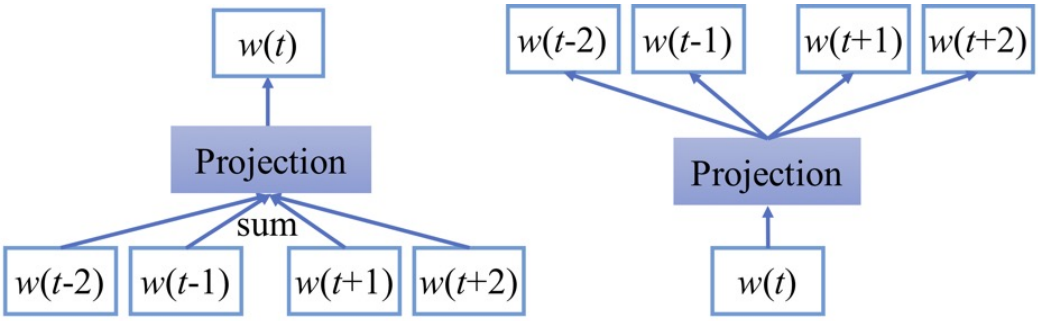


Figure 2.2: CBOW (left) and skip-grams (right) model architecture

In particular, we pre-trained word embeddings of the preprocessed corpus with the ngrams incorporated skip-gram model - the Fasttext method [31]. Fasttext represents each word as a bag of character ngrams to incorporate subword information, such that a word embedding is represented as the summation of the included ngram embeddings. Specifically, we constructed 50-dimensional word embeddings using a window size of three, and the subwords ranged between three and six characters in length. Besides, only words that appeared more than five times were retained in the vocabulary. The thresholding improves model robustness to typos and helps remove uncommon tokens.

The pre-training step assigns each word a vector representation based on its context. Therefore, words that have similar contexts should reside in close vicinity in the embedding space. Fasttext additionally encourages words that share common subwords to reside closer in the embedding space, such that names like “Nicholas” and “Nicolas” should be close to each other.

2.4.3 Extension to Contextual Embeddings

The word2vec model assigns a single vector representation to each word and is incapable of differentiating polysemy. To address this limitation, we adopted the contextual extension approach to associate a contextual embedding to each word based on its context [26]. Firstly, a symmetric cooccurrence matrix C is constructed between each pair of words in the vocabulary

$$C_{ij} = \frac{\#cooccurrence(word_i, word_j)}{freq(word_i) \times freq(word_j)}, \tag{2.1}$$

where the diagonal element

$$C_{ii} = \frac{1}{freq(word_i)}. \tag{2.2}$$

Then the contextualized word2vec embedding of $word_i$ is:

$$e_i = \frac{1}{|window|} \sum_{word_j \in window} C_{ji} v_j, \tag{2.3}$$

where v_j is the original word2vec embedding of $word_j$ in the vocabulary. The contextual embedding is the weighted sum of the word2vec embeddings of all the words in a window normalized by the window size. In our method, we took a window size of three.

2.4.4 Binary Classification

The extended contextual word2vec generates vector representations for words that form natural clusters in the embedding space, accounting for their contexts. It prepares for further classification between name and non-name classes. This motivated us to design a supervised classification scheme on top of the unsupervised contextual representation.

We performed a binary classification between the name and non-name classes by assessing word adjacency to each element in the pre-constructed landmark set in the contextualized embedding space. Here, cosine similarity, quantifying the angle of two vectors, was used to measure the resemblance level. It is a variation of Euclidean distance that is invariant to

scale that facilitates the training process and is more robust against the outliers. The idea is illustrated in Fig. 2.3.

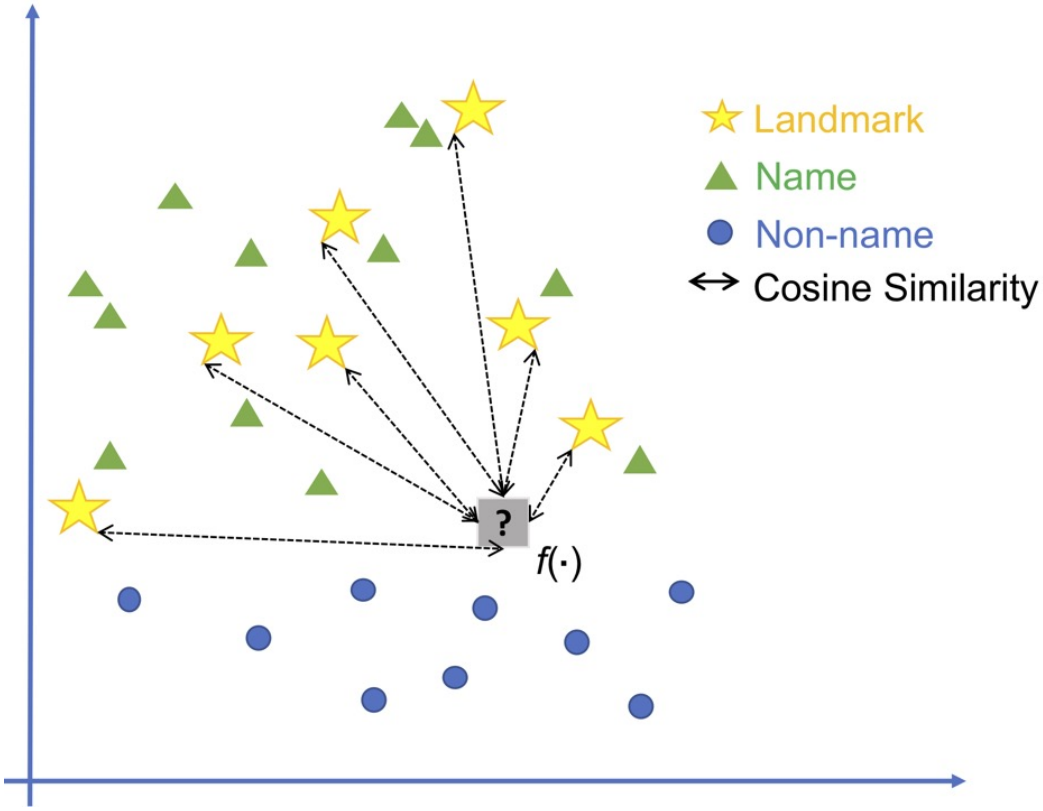


Figure 2.3: Binary classification scheme for note de-identification

We used the embedding adjacency of each candidate word to the landmark set as a feature vector for the subsequent binary classification:

$$\vec{s}(e_c) = [s_{l_1}, s_{l_2}, ..., s_{l_N}], \tag{2.4}$$

where

$$s_{l_i} = CosineSimilarity(e_c, e_{l_i}) = \frac{e_c \cdot e_{l_i}}{\|e_c\|\|e_{l_i}\|}, i = 1, 2, ..., N \tag{2.5}$$

The vectors $e_c, e_{l_i} \in \mathcal{R}^{50}$ in our case are the contextual embeddings of a candidate word and each of the N landmarks, respectively.

2.4.5 Multilayer Perceptron Classifier

An MLP with a single hidden-layer was applied to learn a function which took the vectorized adjacency measure as input, and generated the class estimate of a candidate word:

$$f(\vec{s}(e_c)) = \delta^{(2)}(b^{(2)} + A^{(2)}(\delta^{(1)}(b^{(1)} + A^{(1)}\vec{s}(e_c)))), \tag{2.6}$$

where A and b are the learning parameters, and δ is the activation function in MLP. In the step of supervised learning, we minimized the binary cross-entropy loss between the predicted label and the true label.

2.5 EXPERIMENT

In this section, we provide the details for the data construction process and the hyper-parameters used in our design.

2.5.1 Landmark Construction

There were 14 unique words of names (excluding suffix and prefix) in the report that we arbitrarily selected. Some of them appeared multiple times under different contexts, leading to a total of 28 word instances. For example, "Patient: John Grant Doe, Mr. Doe is fine today." has four name word instances. We treated the 28 word instances as the full landmark set, each had a different contextual word embedding.

2.5.2 Data Construction

We constructed 1400 name word instances and 1400 non-name word instances for model training, validation, and testing. While the names and non-names can be handcrafted from the reports or external sources, we adopted a "lazy" approach to avoid handcrafting. As a byproduct of the landmark construction process, we had a report where all the name

instances were labeled in a sparse effort (in our case, only 28 instances). With each word in the report already labeled, the report was used to generate non-name trigrams, and each word's associated contextual embeddings were calculated. A partial illustration of non-name trigrams is in Fig. 2.4. Note that it is still preferable to construct an independent non-name list not based on the landmarks, and we used the trick simply to reduce the manual load. A few names can be manually picked by going over a few reports, and then we constructed 1400 name word instances (or tri-grams with a name word in the middle) by finding their various contexts.

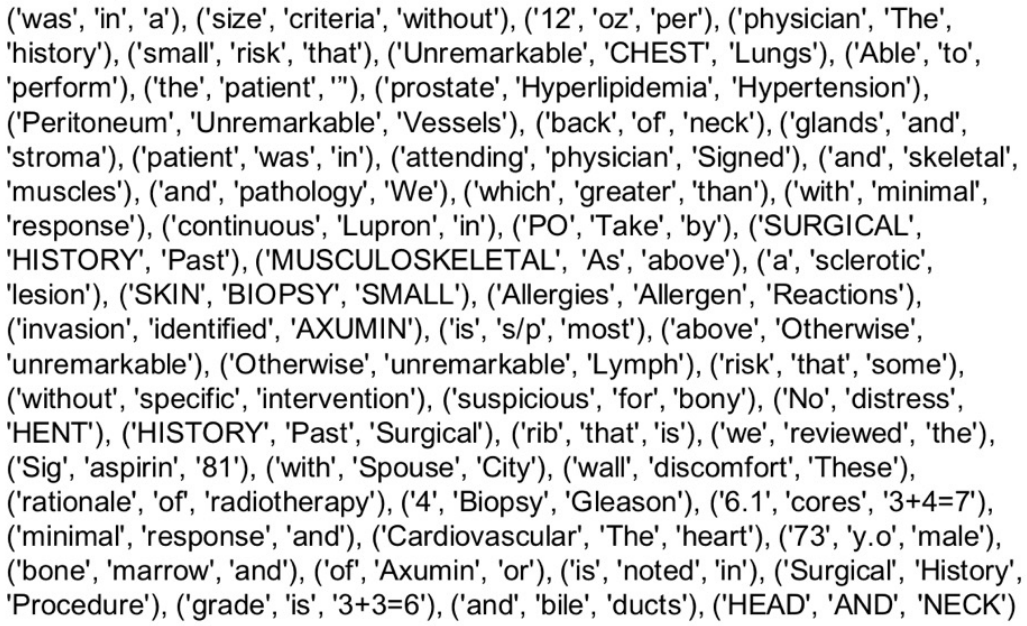


Figure 2.4: A partial illustration of the constructed list of non-names with contexts

2.5.3 Hyper-parameter Tuning

The validation set A was used for finding the optimal number of landmarks N. We performed an exhaustive search by varying the number of name word instances from 1 to 28, and assessed its utility as the landmark set. We ran ten repetitions of model training and validation with the validation set A, and the average model accuracy for the validation set is reported in

Fig. 2.5. It shows that the validation accuracy experiences a drastic improvement followed by a gradual enhancement-to-saturation pattern as the number of landmarks increases; thus we used all the 28 instances as landmarks in our experiment.

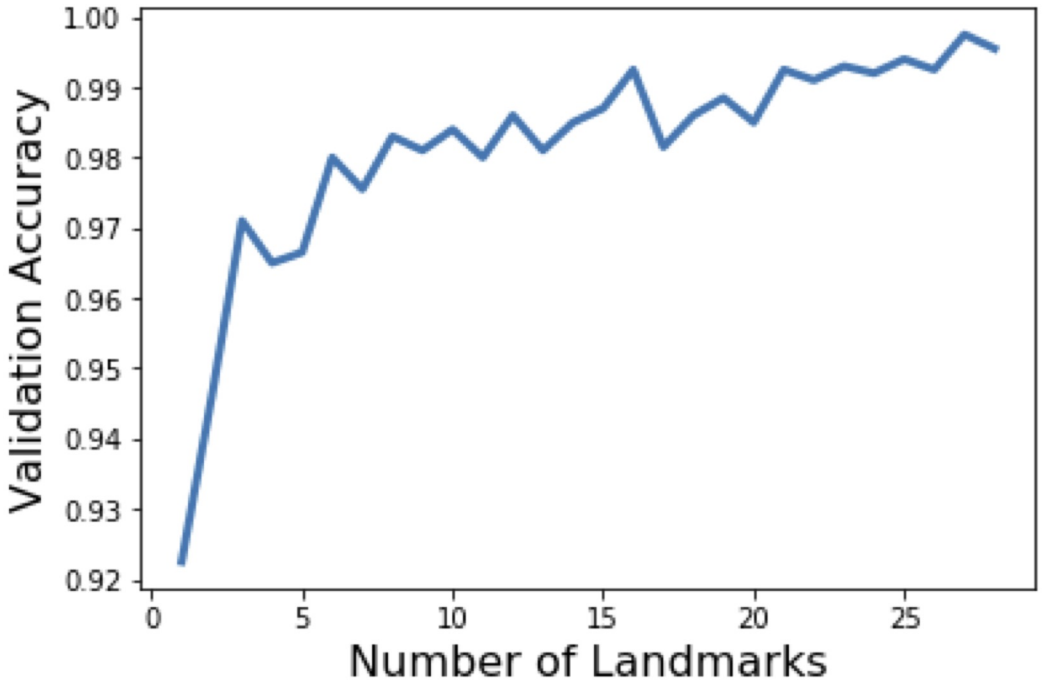


Figure 2.5: Optimal number of landmarks searching for note de-identification

We tuned the hyper-parameters for the single-layer MLP using the validation set B. Validation performance on the validation set B led to a decision of using 50 units on the hidden layer of the MLP, and the sigmoid activation was used for both the hidden layer and the output layer. We trained the model for 200 epochs with a batch size of 64. The learning rate was also tuned by the validation set B, where 5e-3 was chosen with the Adam optimizer.

2.5.4 Inference and Name Removal

The final inference process was to compute the contextual embedding for each word in a report, calculate the cosine similarity to each landmark in the contextual embedding space,

and feed the similarity vector into the trained classifier. Lastly, words classified as names were removed from each report.

2.5.5 Benchmark Comparison

The publicly available Stanford NER toolbox, which is the representative benchmark in this work, is feature-based and uses a CRF classifier [27]. The features it extracts include word and orthographic features, prefixes and suffixes, as well as distributional similarity features, etc [32]. The distributional similarity features cluster words based on their context distributions in unlabeled texts. The similarity of words is measured by the Kullback-Leibler (KL) divergence between their context distributions, where the context distributions are estimated from the observed contexts in a corpus.

Specifically, we compared with the 3-class (location, person, and organization) classifier with distributed similarity function, which was pre-trained on large English corpus of both CoNLL 2003 and MUC datasets. We only used the "person" category, which we called "name" instead, and the other categories were all considered as "non-name". To maintain full sequence characteristics for feature capturing in the Stanford NER, we applied Stanford NER to the raw reports.

2.6 RESULTS

The training, validation and testing sets achieved 1.0/1.0/1.0, 0.99/1.0/0.99, 0.99/0.99/0.99, and all 1.0 respectively in accuracy/precision/recall. To visualize the different performance between our model and the benchmark Stanford NER, we randomly picked two reports and compared the name removal performances. Fig. 2.6 illustrates partial results on report 1 and report 2 from both the Stanford NER and our proposed method. For both models, we report the incidence-based detection performance, where "Mr. Doe" and "John G. Doe" are considered as two samples. This is appropriate as it aligns with the goal of detecting all

the occurrences and variations of names across medical notes. The detection requirement is relaxed only when a middle initial appears alone, as it contains minimal private information and cannot be interpreted in isolation. We also performed a one-sided paired t-test on the F1 score of six randomly picked reports, and reported the mean and the standard deviation (STD) for both models. We marked whether the result is significant under the t-test with $\alpha = 0.1$.

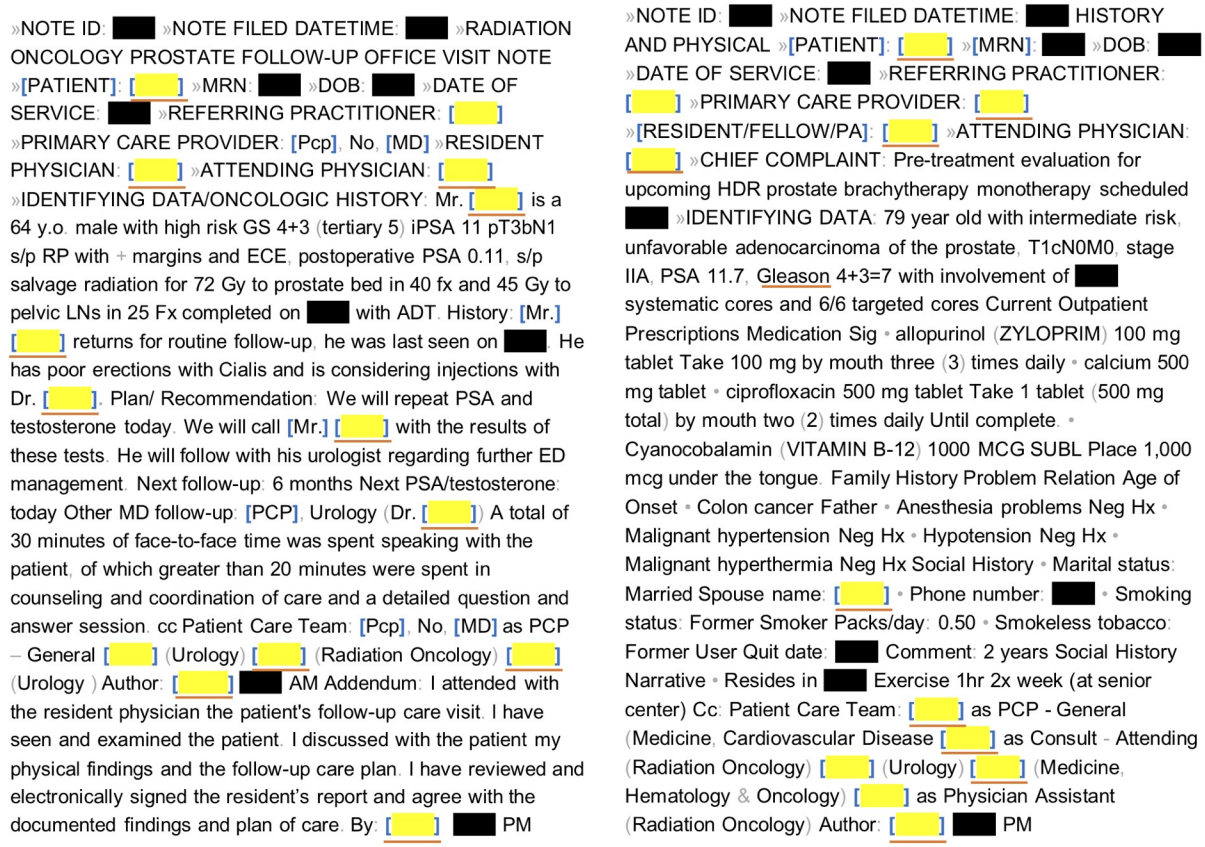


Figure 2.6: Partial illustration of report 1 (left) and report 2 (right) after note de-identification: light grey text represents the punctuation removed in preprocessing; blocks are PHI (light yellow: names, dark black: other non-name PHI); brackets are the name detections by the proposed method; and underlines denote the name detections by the Stanford NER

Table 2.1 reports the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) on the two reports, as well as the sensitivity and specificity.

Table 2.2 reports the mean and the STD of the F1 score on the six reports with significant results marked.

Table 2.1: Quantitative Name Detection Comparison with Stanford NER on Two Reports

Metrics		Stanford NER	Proposed
Report 1: 15 samples	TP	13	15
	TN	680	677
	FP	1	4
	FN	2	0
Report 2: 19 samples	TP	16	19
	TN	1862	1866
	FP	7	3
	FN	3	0
Sensitivity:$\frac{TP}{TP+FN}$		0.8529	1.0
Specificity:$\frac{TN}{TN+FP}$		0.9969	0.9973

Table 2.2: Statistical Name Detection Comparison with Stanford NER on Six Reports

Metric	Stanford NER	Proposed
F1 score:$\frac{TP}{TP+0.5(FP+FN)}$	0.822 ± 0.103*	0.889 ± 0.046

*significant under a one-sided t-test with $\alpha = 0.1$, compared to the proposed method

2.7 DISCUSSION AND CONCLUSIONS

2.7.1 Comparison with Stanford NER

As shown by the results, our method achieved a significantly higher F1 score compared to the Stanford NER. From the illustration, our method successfully removed all the person names from the two notes, while the Stanford NER missed some of the names and mistakenly tagged on some medical terms such as "Gleason". Our model has a moderate tendency to

mislabel context words around a name, e.g., "PATIENT" in "NOTE PATIENT John Doe". In fact, this was the major source of false positives. It is likely due to the fact that the word "PATIENT" appears quite frequently in conjunction with the patient name, and the contextual embedding scheme attributes a large portion of name embedding to it. While this might be tolerable for our task for note de-identification since the removal of such words does not affect the integrity of information in the de-identified notes, this problem may be further alleviated by applying a more sophisticated contextual embedding scheme.

The pre-trained Stanford NER was trained on more literal texts (i.e., newswire), and may not be suitable for tagging clinical reports that are typically written in a different style and using a specific vocabulary set. In our experiment, we found that the Stanford NER performance was good when detecting names on narrative texts, but its performance degraded when detecting names on a table and bullet template like Header: Name, which violates the flow of the natural narrative language. Unfortunately, this concise pattern is very common in medical notes. It also had a hard time to differentiate medical terms such as "MR" and "Gleason" from real person names in narrative texts.

We tried to use the same amount of labeled data, i.e., one annotated report to train the Stanford NER model, and the performance was worse than the pre-trained 3-class model, as shown from Table 2.3 and 2.4 This indicates the need for extensive labeling for the NER model and shows the benefit of our method in coping with limited tagging.

2.7.2 Comparison with Naïve Bayes Classifier

We compared our method with the Naïve Bayes classifier with frequency-inverse document frequency (TF-IDF) features, which also has the advantage of being compatible with a small number of training samples. Table 2.3 and 2.4 shows that the proposed method achieved a significantly better F1 score than the Naïve Bayes model, demonstrating the advantage of proposed contextual embedding over the native words and ngrams.

Table 2.3: Customized Stanford NER and Naïve Bayes Name Detection Performance on Two Reports

Metrics		Customized NER	Naïve Bayes
Report 1: 15 samples	TP	1	14
	TN	681	600
	FP	0	81
	FN	14	1
Report 2: 19 samples	TP	0	16
	TN	1869	1786
	FP	0	83
	FN	19	3
Sensitivity:$\frac{TP}{TP+FN}$		0.0294	0.8823
Specificity:$\frac{TN}{TN+FP}$		1.0	0.9357

Table 2.4: Customized Stanford NER and Naïve Bayes Name Detection Performance on Six Reports

Metric	Customized NER	Naïve Bayes
F1 score:$\frac{TP}{TP+0.5(FP+FN)}$	0.150 ± 0.174*	0.298 ± 0.043*

*significant under a one-sided t-test with alpha = 0.1, compared to the proposed method

2.7.3 Landmark Robustness

To test the robustness of the proposed method to landmark selection, we drew landmarks from another report and repeated the same pipeline on the same datasets. The report contained 36 name word instances from the patient and seven care-providers. We also achieved an accuracy of almost 1.0 for each set. It shows that our method is robust to the selection of landmarks.

2.7.4 Conclusions and Future Development

In this study, we have proposed a novel and simple de-identification approach that utilizes a combination of unsupervised and supervised learning modules and achieved high classification performance. We used name removal in this paper to demonstrate the logic and pipeline, yet the method generally applies to tasks of extracting strings that share similar contexts. The most important contribution of this work is its requirement of very sparse manual tagging, which addresses the major hurdle in the current NER approaches. Our quantitative results show that our method achieves a significantly higher F1 score than the pretrained 3-class Stanford NER model.

Pattern matching approaches, similar to the one we adopted in preprocessing, can be proficient in removing the numeric PHI, such as patient ID numbers, dates and times, and phone numbers. The numeric PHI can be removed by exploiting the format such as digit lengths and patterns like "xxx-xxx-xxxx". To detect other non-numeric PHI, such as addresses, the pattern and key contextual word matching can be combined. Location fragments that distribute across a note such as district and city names can be detected by a location vs. non-location classification scheme, similar to the one proposed in this work.

CHAPTER 3

Cone-beam Computed Tomography-based Pelvic Organ Segmentation

3.1 BACKGROUND

Structure segmentation is a common medical imaging task. There are many applications where daily variant anatomy needs to be tracked and quantified. An important application discipline is radiation therapy where tumor targets and critical structures are contoured based on a high-quality computed tomography (CT) image, and treatment parameters are optimized with respect to (w.r.t.) this image, a procedure known as planning. At each subsequent treatment fraction, patients are set up to best conform to the planned position and the therapeutic radiation is delivered as planned.

However, organs deform and move, and change in filling status, leading to discrepancy between the expected exposure and the actual delivered dose pattern, impairing conformality and ultimately treatment efficacy and safety. Therefore, it is critical to monitor and quantify morphological variations, and to trigger corresponding adjustment when clinically significant deviation is detected [33].

Motion occurs in different scales - while intra-fraction motion caused by breathing, cardiac, and musculoskeletal motion can usually be characterized with stochastic processes, inter-fraction morphological variation can induce larger systematic discrepancies in certain body sites [34]. For example, in the pelvic region, inter-fractional rectum volume changes as large as from -14% to 39.8%, and bladder volume varies from -46% to 127.2% relative

to the original planning volume [35], and it could be even larger for post-prostatectomy patients, reaching 50%-270% for the rectum, and 30%-180% for the bladder [36].

Inter-fraction anatomy variation can be monitored with cone-beam CT (CBCT), which is a low-dose, widely available imaging modality offered on-board radiotherapy platforms. Unfortunately, despite CBCT offers 3D attenuation distribution, its current clinical use is primarily for rigid setup adjustment or visual verification. This is because CBCT exhibits low image contrast and low signal-to-noise behavior, and is subject to stronger motion and Feldkamp artifacts, resulting in a high level of uncertainty and inconsistency in manual contours, compared to those based on CT or magnetic resonance imaging (MRI). In fact, it has been reported that manual organ delineation on CBCT images varied significantly among physicians with an overall mean Dice index of only 0.7 among ten sets of manual delineation [37].

In this study, we focus on developing automatic pelvic organ segmentation for post-prostatectomy patients, where the prostate surgery induces additional challenges in that (1) generally compromised genitourinary control leads to larger variations in the filling status of the bladder [38], (2) compromised bowel movements and a much higher chance of constipation result in a large presence of air pockets in the bowel and rectum.

3.2 CBCT AND AUXILIARY CT DESCRIPTION

Under phase II trial, post-prostatectomy patients were enrolled to receive five-fraction stereotactic body radiotherapy (SBRT). Under IRB approval, planning CT and per-fraction CBCT were obtained from 17 patients. Another 65 patients with planning CT only were added to the dataset as augmentation to training data to prevent overfitting, and the number of CT and CBCT training data were approximately equal. The original image resolution for both CBCT and CT was mostly 1.17 mm $\times$ 1.17 mm in-plane and 1.5 mm thickness between slices. Both CBCT and CT were manually contoured by physicians. For preprocessing, all

images were first resampled to isotropic 1.17 mm × 1.17 mm × 1.17 mm in resolution. The Hounsfield unit values that were greater than 2000 were clipped to 2000. The images were then globally min-max normalized.

With the prostate surgically removed, the pelvic organs present high variations in geometry and spatial relations measured in standard deviation (std). The inter-subject and intra-subject (across the five CBCT fractions) volume statistics for the 17 subjects are summarized in Table 3.1.

Table 3.1: Rectum and Bladder Volume Characteristics

	Inter-subject			Intra-subject	
	Min	Max	Mean ± std	Min std	Max std
Rectum	21.9	181.2	82.8 ± 38.9	3.4	53.3
Bladder	50.3	759.2	277.9 ± 161.7	20.0	213.5

All units are cm^3; max std and min std are the maximum and minimum of inter-fractional std, respectively.

3.3 ENSEMBLE LEARNING AND TENSOR REGULARIZATION

3.3.1 Related Deep Learning Methods for CBCT Segmentation

Automatic approaches have been extensively studied for segmentation, ranging from gradient-driven boundary detection-based approaches to active contours, and to more recent deep learning based approaches utilizing contextual structures which have witnessed much success [9, 13, 39, 40]. Realizing the intrinsic limitation associated with CBCT modality, transfer learning and common domain embedding techniques have been investigated to incorporate priors from other modalities or training instances [41, 42, 43, 44, 45, 46].

These approaches can be roughly categorized as (1) direct augmentation of training data

with another modality: an example being a 3D UNet trained with additional CT scans as to augment CBCT scans which improved the resultant Dice similarity coefficient (DSC) to 0.874 ± 0.096 and 0.814 ± 0.055 for the bladder and rectum, respectively, from 0.796 ± 0.128 and 0.680 ± 0.117 [42]; and (2) modality translation based on generative adversarial network (GAN) techniques. CT was synthesized from CBCT using Cycle-consistent GAN (CycleGAN) to improve the virtual input image quality, and the resulting bladder and rectum DSC was 0.916 ± 0.005 and 0.872 ± 0.201, respectively [44]. A related work used a similar logic to synthesize CBCT from CT and trained a segmentation network for CBCT using CT and CBCT images and labels on CT, to alleviate the burden of direct contouring on CBCT [43]. MRI can be utilized similarly: under more demanding CBCT-MRI data pairing preprocessing, CycleGAN was trained to synthesize MRI (sMRI), and CBCT and sMRI inputs were processed separately to extract modality specific features which were combined in a late-fusion attention pyramid network. The achieved DSC was 0.96 ± 0.03 and 0.93 ± 0.04 for the bladder and rectum, respectively, thanks to the superior soft-tissue contrast of MRI [46].

3.3.2 Development of Segmentation with Ensemble Learning

To address the specific challenges in CBCT-based organ segmentation for post-prostatectomy patients, we propose a development with a consistent ensemble logic at various levels. On the estimator level, to define the region of interest (ROI) and later segmentation structures, multiple networks from different views are developed in parallel and ensembled to generate a final 3D estimation with enhanced performance; on the feature level, coupled filters are imposed on feature learning from CBCT and the auxiliary CT to achieve intrinsic information ensemble [47]. Specifically, major contribution and novelty of this work include:

(1) a simple ensemble 2.5D You-only-look-once (YOLO) from multiple views to consistently define a 3D ROI for segmentation;

(2) view-specific two-stream 2.5D segmentation networks with coupled encoder and early

inference layers, using auxiliary high-quality planning CT;

(3) a novel tensor-regularized ensemble scheme to aggregate the estimates from multiple segmentation networks while regularizing the spatial integrity and continuity of the final segmentation contours.

All the central modules are light-weight 2.5D networks, and the general structure of developing independent weaker learners followed by a fusion makes the training and testing friendly for parallelization. The overall sequential structure also offers a potential pathway for a modular interpretation of the overall pipeline. Fig. 3.1 introduces the proposed pipeline.

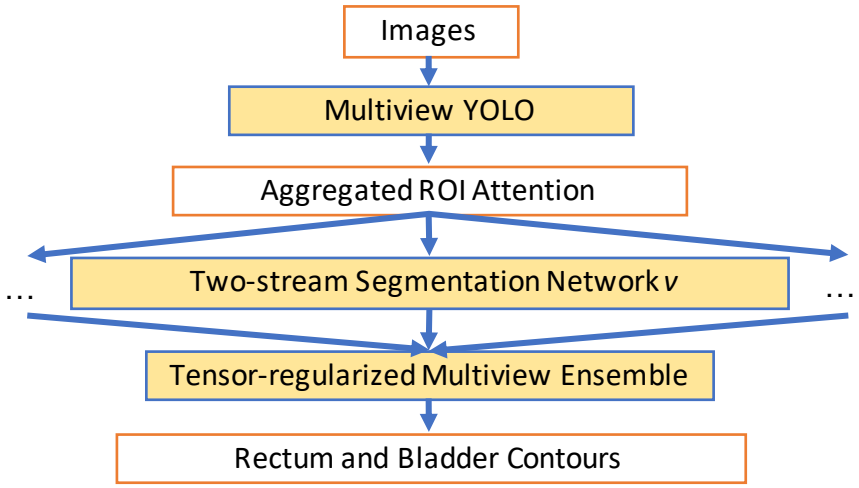


Figure 3.1: Overall pipeline for the proposed segmentation method: a consistent 3D ROI is first localized by a customized multiview ensemble 2.5D YOLO detector. Then view-specific segmentation estimates of the rectum and bladder are obtained by a set of two-stream UNet models using unpaired CT and CBCT data. The predicted contours from the multiple views are further aggregated with a tensor-regularized optimization scheme to yield a set of geometrically enhanced final contours.

3.3.2.1 3D Localization with Multiview 2.5D YOLO

ROI localization is important to standardize input to a manageable size with coarse attention to the segmentation objective. While there are many approaches involving various levels of coarse segmentation as the localization step [48, 49], we have chosen a simple and stable approach by modifying the original YOLO [11] to a multiview 2.5D version [50] to generate a single 3D ROI bounding box, as illustrated in Fig. 3.2. The proposed 3D ensemble YOLO localizer provides consistent, reproducible, and robust organ coverage for subsequent tasks.

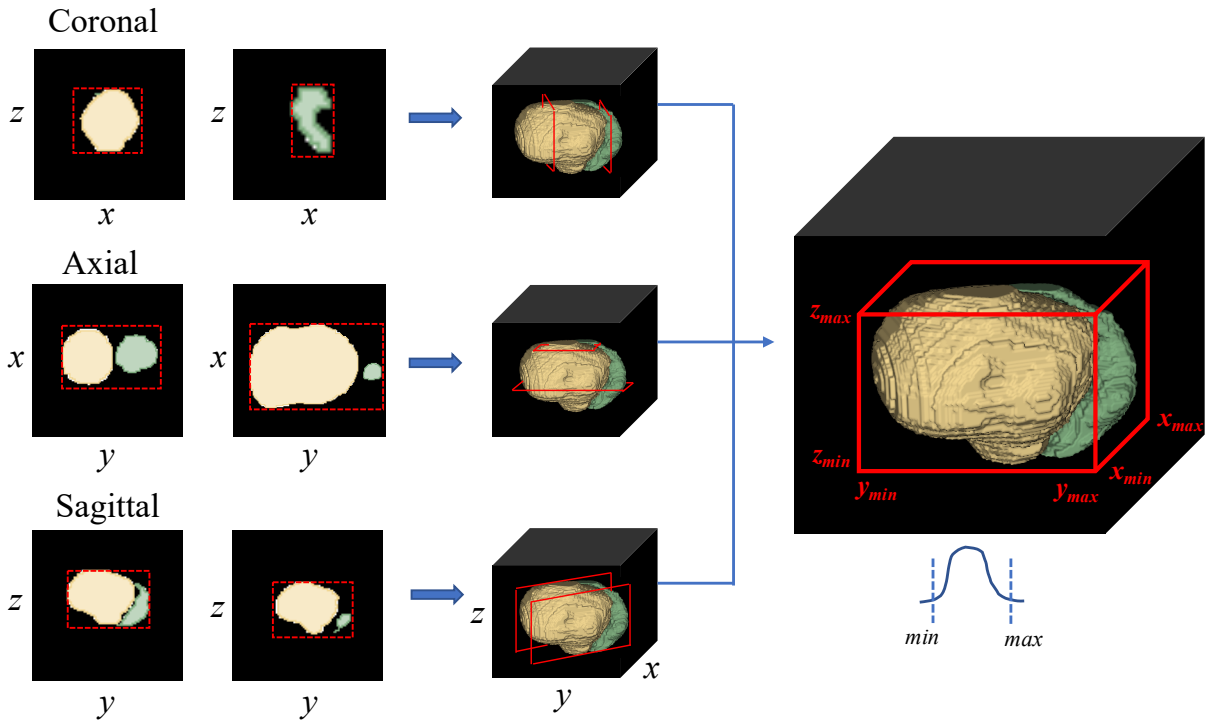


Figure 3.2: Illustration of pelvic ROI localization for CBCT segmentation: a 3D bounding box including the rectum and bladder is generated by taking the tails of 2D YOLO bounding boxes from x, y, z directions, and obtaining the smaller minimums and larger maximums between any two of the axial, sagittal, and coronal views. Yellow is the bladder and green the rectum.

Specifically, 2D YOLO localizers are first trained for each view. During inference, a 3D

bounding box is generated by taking the minimum of the 0.05 quantiles and the maximum of the 0.95 quantiles for each coordinate axis x, y, and z. Using the robust tails of directional estimates across the slices reduces the risk of missing structure boundary or being lead astray by noisy estimates from a small subset of 2D YOLO. For example, a 2D YOLO bounding box for the axial view is defined:

$$\begin{cases} x_{\min}^{\text{axial}} = 5\%(\{x_i\}), \\ x_{\max}^{\text{axial}} = 95\%(\{x_i\}), \\ y_{\min}^{\text{axial}} = 5\%(\{y_i\}), \\ y_{\max}^{\text{axial}} = 95\%(\{y_i\}), \end{cases} \quad i = 1, 2, \ldots, N_Z. \tag{3.1}$$

where N_Z is the total number of slices in the superior-inferior direction, and the final 3D bounding box is given by:

$$\begin{aligned} x_{\min} &= \min(x_{\min}^{\text{axial}}, x_{\min}^{\text{coronal}}), \\ x_{\max} &= \max(x_{\max}^{\text{axial}}, x_{\max}^{\text{coronal}}), \\ y_{\min} &= \min(y_{\min}^{\text{axial}}, y_{\min}^{\text{sagittal}}), \\ y_{\max} &= \max(y_{\max}^{\text{axial}}, y_{\max}^{\text{sagittal}}) \\ z_{\min} &= \min(z_{\min}^{\text{coronal}}, z_{\min}^{\text{sagittal}}), \\ z_{\max} &= \max(z_{\max}^{\text{coronal}}, z_{\max}^{\text{sagittal}}). \end{aligned} \tag{3.2}$$

3.3.2.2 View-specific 2.5D Segmentation Network

For each chosen angle of view, the corresponding data stream is generated by applying the YOLO localization and reorient the image data. A two-stream segmentation model [51] based on 2.5D UNet with ResNet backbone is adopted [9, 13], which takes CT and CBCT on two input data streams and shares a subset of model structures to induce coupling between the two UNets corresponding to CT and CBCT, respectively, as illustrated in Fig. 3.3.

During training, we alternate between the back propagation using CT batches and CBCT batches. Upon testing various sharing structures of the encoder and decoder, we propose a final model which shares all layers of the encoder and the first two layers of the decoder,

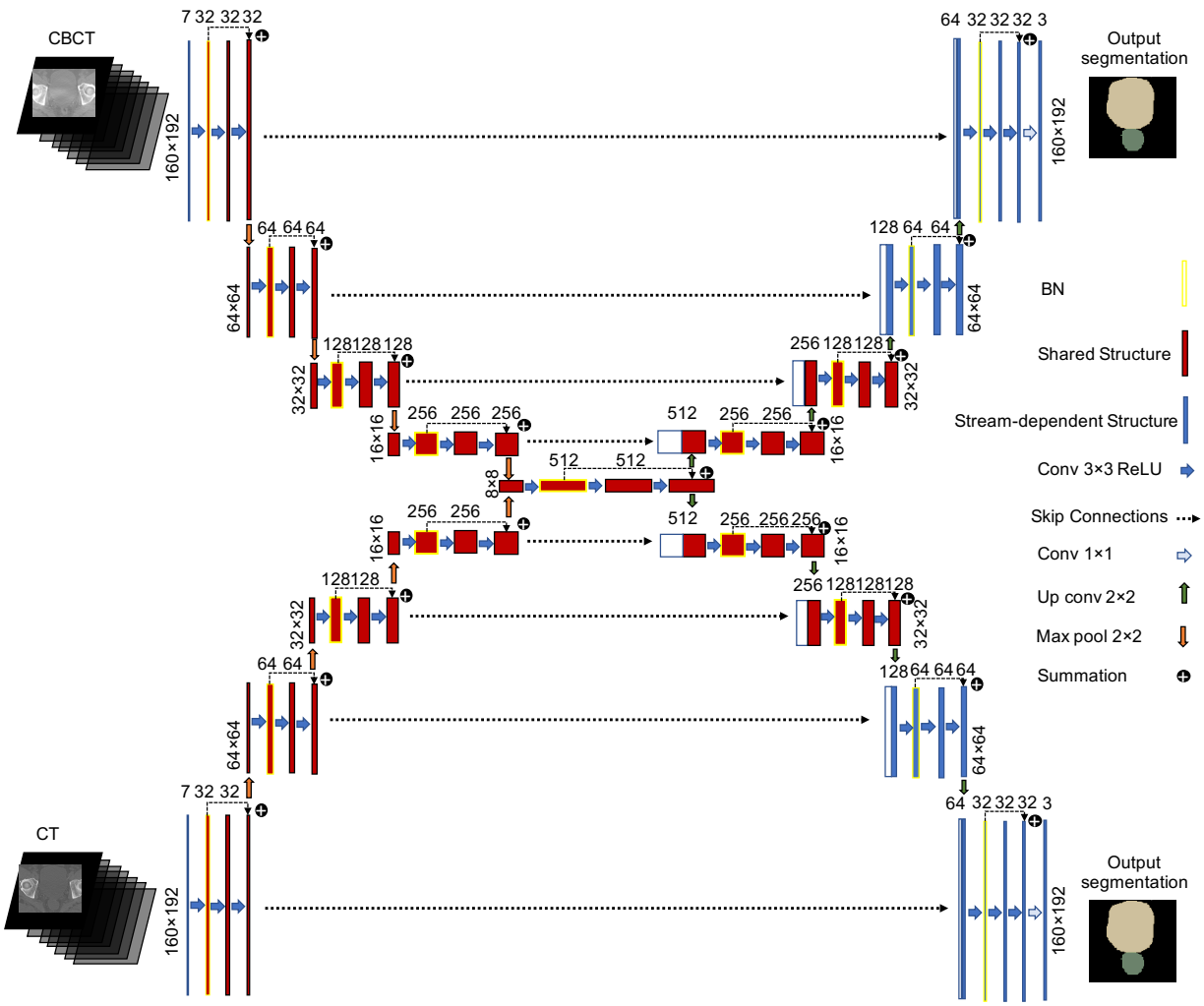


Figure 3.3: Structure of 2.5D two-stream UNet with ResNet backbone for pelvic organ segmentation: the input images are the cropped images (axial view for illustration) by the proposed 2.5D YOLO localizer; the model takes an image stack and outputs the class predictions of the background (black), rectum (green), and bladder (yellow) for the middle slice; skip connections are inserted after the first convolution layer where element-wise summation is used to incorporate information from the previous layer within each convolution block to enhance model convergence during training.

whereas the last two layers of the decoder are stream-dependent. Sharing the encoder stabilizes feature extraction from CBCT with the most consistent appearance of high-quality CT, and the stream-dependent late inference layers allow for domain-specific contour inference.

Soft Dice coefficient loss is used as the training objective:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{CT}} + \mathcal{L}_{\text{CBCT}},$$

where

$$\mathcal{L}_d = \sum_{c=1}^{C}(1 - \frac{1}{N_d}\sum_n \frac{2p_d^{n,c}y_d^{n,c}}{(p_d^{n,c})^2 + (y_d^{n,c})^2}), d = \text{CT}, \text{CBCT}.$$

$y_d^{n,c}$ and $p_d^{n,c}$ are the prediction and the ground truth for the nth pixel of the cth class, respectively, and N_d is the total number of pixels in a batch, for domain $d = \text{CT}$ and CBCT, respectively. The predicted classes include the background, rectum, and bladder.

3.3.2.3 Ensemble with Tensor Regularization

While 2.5D networks can access image contexts via stack inputs, there may still be spatial non-smoothness across stack directions during inference. Furthermore, it is quite common for post-prostatectomy patients to experience constipation due to abnormal bowel movements, increasing the frequency and severity of air pocket presence in CBCT. The distribution and shape of air pockets can be roughly considered as random across space, and may affect data quality in a view-dependent way. Similarly, the relative geometric configuration and motion artifacts could also impact views differently. To this end, we propose to perform independent multi-class segmentation for each view, and consider each as a contributing classifier in an overall ensemble framework. We further develop an aggregation scheme by solving a tensor optimization problem with shape regularization and probability constraints:

$$\underset{y\in\mathcal{S}}{\text{argmin}} \underbrace{\sum_v \frac{1}{2}||y - Y_v||_2^2 + \sum_c \lambda_c TV_{3D}(y_{\cdot,c})}_{f_1} + \underbrace{\eta \sum_{\mathbf{x},c} |y_{\mathbf{x},c}|^\alpha}_{f_2} \tag{3.3}$$

where $\mathcal{S}$ is the probability simplex such that

$$\begin{aligned} y &\geq \mathbf{0}, \\ \textstyle\sum_c y_{\mathbf{x},c} &= 1, \forall \mathbf{x}. \end{aligned} \tag{3.4}$$

The 4D tensor $y \in |\Omega| \times \Re^C$ presents the aggregated soft class membership estimate as a spatial-class map, where Ω is the 3D ROI and C is the number of classes. The input data Y_v is the multi-class prediction from view v. Hyperparameter λ_c controls the regularization strength of spatial smoothness of the membership by 3D total variation (TV) for class c. With $\alpha < 1$, $f_2 = \sum_{\mathbf{x},c} |y_{\mathbf{x},c}|^\alpha$ encourages label sparsity at each spatial location to drive a unique class association with strength η.

Constraining the resultant y to the probability simplex preserves the class association per pixel as that from multi-class inference, and the class prediction for a pixel is naturally given by the one with the highest odds. This approach rids threshold selection, and avoids the risk of overlapping membership assignment for the same pixel and the needs to address such ambiguity.

We propose in Algorithm 1 a nested Douglas-Rachford (DR) splitting scheme to solve the optimization problem (3.3) by the splitting strategy. Although the DR splitting theoretically only works with convex functions, studies have shown that it works empirically for reasonable nonconvex functions [52, 53]. The nested splitting is formalized as an outer splitting to address the simplex constraint g [54], and an inner splitting for the TV and the nonconvex α-norm terms.

$$\begin{aligned} &\text{Outer DR } J = f(y) + g(y) : \\ &g(y) = \delta_S(y) := \begin{cases} 0, \text{iff } y \in S \\ +\infty, \text{ else.} \end{cases} \\ &\text{Inner DR } f(y) = f_1(y) + f_2(y). \end{aligned} \tag{3.5}$$

Each sub-minimization problem in the nested DR corresponds to evoking a proximal operation w.r.t. the TV, sparse α-norm, and the impulse barrier on the simplex set term.

Algorithm 1 nested Douglas-Rachford Splitting Algorithm

Initialize $q = 1$, $t^q = \mathbf{0}$, choose values for P, λ_c, η, δ_{thr}, α

while 1 **do**

$z^1 \leftarrow t^q$

for $p = 1, 2, ..., P$ **do**

$$f \begin{cases} y^p \leftarrow \underset{y}{\text{argmin}} \sum_v \frac{1}{2}||y - (Y_v + z^p)||_2^2 + \sum_c \lambda_c TV(y) \\ w^p \leftarrow \underset{w}{\text{argmin}}\ \eta|w|^\alpha + \frac{1}{2}(w - (2y^p - z^p))^2 \\ z^{p+1} \leftarrow z^p + w^p - y^p \end{cases}$$

end for

$$g \begin{cases} u^q \leftarrow \underset{u}{\text{argmin}}\ \frac{1}{2}||u - (2y^{q,P} - t^q)||_2^2 + \delta_S(u) \\ t^{q+1} \leftarrow t^q + u^q - y^{q,P} \end{cases}$$

if $||y^{q+1,P} - y^{q,P}||_\infty < \delta_{thr}$ **then**

break

end if

end while

The proximal operator of a function h with step size τ is defined as:

$$\text{Prox}_{\tau h}(z) = \underset{x}{\text{argmin}}(h(x) + \frac{1}{2\tau}||x - z||_2^2). \tag{3.6}$$

For simplicity, we present all the proximal operators with $\tau = 1$ and write them as $\text{Prox}_h(z)$.

The outer split induces proximal operator on g, and can be efficiently solved by sorting and thresholding the input vector, with the procedure presented in Algorithm 2 [55].

Algorithm 2 Proximal Operator of Simplex constraint g

Input $x \in \Re^M \times \Re^C, M = |\Omega|$ number of pixels.

for $m = 1, 2, ..., M$ **do**

Sort $x_{m,\cdot}$ to $z : z_1 \geq z_2 \geq ... \geq z_C$

Find $Q_m = \max\{1 \leq c \leq C : z_c + \frac{1}{c}(1 - \sum_{q=1}^{c} z_q)\}$

for $c = 1, 2, ..., C$ **do**

$\text{Prox}_g(x_{m,c}) = \max\{x_{m,c} + \frac{1}{Q_m}(1 - \sum_{q=1}^{Q_m} z_q), 0\}$

end for

end for

Output $\text{Prox}_g(x)$

The 3D TV-$l2$ minimization is solved with the proxTV toolbox [56], where parallel-proximal Dykstra is utilized for solving multi-dimensional TV. The m-dimensional TV is regarded as the sum of m proximal operator terms, each of which is further decomposed into a number of inner 1D TV terms. The dual to 1D TV-$l2$ proximal operator is the well-known trust region subproblem, and is solved by a method based on the More-Sorensen Newton method [56, 57].

We set $\alpha = \frac{1}{2}$ in the label sparsity f_2, which has been shown to empirically work well [58, 59]. The closed-form analytic proximal operator is pixel-wise [60]:

$$\text{Prox}_{f_2}(x) = \begin{cases} 0, \text{ if } y > \frac{2\sqrt{6}}{9} \\ \frac{4}{3}\sin^2(\frac{1}{3}\arccos(\frac{3\sqrt{3}}{4}x) + \frac{\pi}{2}) \cdot y, \text{ else,} \end{cases} \tag{3.7}$$

where $y := \eta|x|^{-\frac{3}{2}}$.

3.3.3 Assessment and Method Specifications

3.3.3.1 Evaluation Criteria

Segmentation performance was measured quantitatively with DSC and mean surface distance (MSD) between the prediction and the ground truth contours in 3D. A four-fold cross validation was applied such that each CBCT volume of the 17 patients had been treated as a test sample exactly once. First, an initial splitting of $13:2:2$ patients (i.e., $65:10:10$ CBCT volumes) was performed for training, validation, and testing, where the network hyperparameters (training epochs, network depth, learning rate, etc.) were optimized w.r.t. the validation set. Intra-subject inter-fraction mean and std across the five fractions were assessed for each of the two test subjects, and one-sided paired t-tests with significance level $\alpha = 0.05$ were performed between each measure and the best measure across methods for each class. The structural parameters were held constant and only network parameter values were updated with the next three folds, each had 12 patients for training and five patients for testing. Eventually, the mean and std of each segmented structure across all the 85 volumes were evaluated, where one-sided t-tests with $\alpha = 0.01$ were performed.

3.3.3.2 Network Specifications

Both CT and CBCT were used for training YOLO [61]. The output bounding boxes from YOLO were used to crop the raw data from around $512 \times 512 \times 400$ to a consistent dimension of around $160 \times 192 \times 100$ as the input to the segmentation networks for training and inference. In this specific implementation, axial, coronal, and sagittal views were used for ensemble in

both YOLO and the segmentation neural networks.

In our design, the 2.5D UNets took seven consecutive slices as the input stack. All the UNets had a depth of four and a base number of channels of 32, with convolution blocks each contained three 3×3 convolution layers with ReLU activation, where a batch normalization (BN) layer was inserted after the first convolution layer, and a final layer- 1×1 convolution with softmax activation. A skip-connection was added between the first convolution layer and the last convolution layer in each block, where element-wise summation was used to incorporate information from the previous layer. The learning rate for all segmentation models started with 10^{-3} and decreased to 5×10^{-4} over 60 epochs. We used Adam optimizer and a batch size of 32, and trained on GPU GTX 1080 Ti, and the code implementation was with Tensorflow 2.0.

3.3.3.3 Segmentation Network Design and Comparison

To assess the role of the two-stream structure, we compared it against a 2.5D UNet with CBCT as the only input, a 2.5D UNet [42], and a 3D UNet with a brute-force inclusion of both CT and CBCT as input that entered the network in a random order. A comparative study on two-stream models with different sharing structures, i.e., shared encoder, shared decoder, and the proposed shared encoder and early decoder was performed.

To further examine the differential behavior of the paths in the two-stream model to appreciate its distinction from typical transfer learning or a single UNet with mixed modality samples, we conducted experiments by inferring test segmentation by channeling the test data into the path corresponding to its own domain or the different one.

3.3.3.4 Contour Regularization

To analyze the role of regularization and ensemble, performance was compared qualitatively and quantitatively by using different views, as well as ensemble either with an "average-

winner-takes-it" scheme or the proposed tensor regularization. The "average-winner-takes-it" scheme took the average of the soft memberships from all views, and assigned each pixel the class with the highest membership value.

The tensor-regularized ensemble scheme was also compared with slice-wise morphological operations, which included a combination of dilation operation with two kernel sizes, i.e., 2 and 6, with connected component analysis (CCA) as preprocessing. Additional to the measurement of the agreement to the ground truth in DSC and MSD, the number of hole pixels inside organs and 2D TV were utilized as two morphological measures reflecting the geometric integrity and smoothness.

3.3.3.5 Hyperparameter Selection

We found the segmentation accuracy to be robust w.r.t. the sparsity weight η, and a typical good range of η was $[0.1, 0.6]$. To test the robustness of contour accuracy against the selection of λ_1 and λ_2, corresponding to the TV regularization strengths for the rectum and bladder in Algorithm 1, respectively, we performed a grid-search w.r.t. the performance of the 10 validation points of the initial $13:2:2$ splitting fold. We had the prior knowledge that λ_1 approximately ranged from 0.1 to 1 and we searched the optimized ratio between λ_2 and λ_1.

In our experiment, the class-specific TV weight λ_c was set 0.5 and 2 for the rectum and bladder to impose a different level of smoothness. A stopping threshold δ_{thr} of 0.8 was used. The label sparsity regularization hyperparameter η was set 0.5, and the number of iterations for the inner loop was $P = 1$.

3.3.4 Segmentation Results

3.3.4.1 Segmentation Network Design and Comparison

Fig. 3.4 visualizes the impact of CT augmentation approach and different coupling schemes in the two-stream model, with one test example. Under comparison are (1) 2.5D UNet with

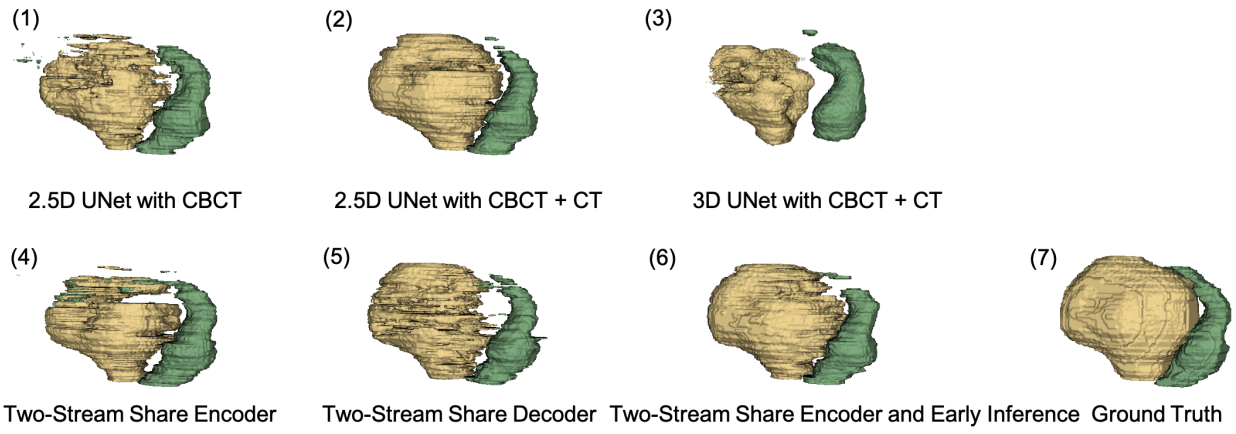


Figure 3.4: Rectum and bladder contour predictions of one CBCT fraction of one testing patient in 3D illustration. The results are from the models in comparison, all trained with axial slices except the 3D model: (1) a baseline 2.5D UNet with CBCT input, (2) a baseline 2.5D UNet with randomly mixed CBCT and CT as input, (3) a 3D UNet with randomly mixed CBCT and CT as input, (4) a two-stream UNet with the shared encoder, (5) a two-stream UNet with the shared decoder, (6) the proposed two-stream UNet with the shared encoder and early inference, and (7) the ground truth. Green: rectum, yellow: bladder.

only CBCT input, (2) 2.5D UNet with randomly mixed CBCT and CT input, (3) 3D UNet with randomly mixed CBCT and CT input, (4) a two-stream UNet sharing encoder between streams, (5) a two-stream UNet sharing decoder, and (6) the proposed two-stream UNet that shares encoder and two early inference layers, all 2.5D models are with a single axial view configuration, without the impact of tensor regularization. This example illustrates improved segmentation with the additional coupled CT path from the two-stream model, but also reveals the need for further enhancing geometric integrity in the prediction. Table 3.2 reports the quantitative performances. We observed moderate improvement of the proposed two-stream UNet over a 2.5D UNet with CT + CBCT and generally significant improvement over other models. We fed the output of the proposed two-stream UNet model to tensor regularization for further ensemble processing.

Table 3.2: Assessment of segmentation performance dependency on CT augmentation approach and coupling structure

Testing Subject/Model/Metric		DSC		MSD (mm)	
		Rectum	Bladder	Rectum	Bladder
Subject1	2.5D UNet with CBCT	0.697 ± 0.076	0.914 ± 0.023*	3.279 ± 0.593	2.076 ± 0.610
	2.5D UNet with CBCT + CT	0.717 ± 0.102	0.919 ± 0.022	3.146 ± 1.039	1.961 ± 0.658
	3D UNet with CBCT + CT	0.723 ± 0.126	**0.934 ± 0.017**	3.012 ± 1.295	**1.611 ± 0.456**
	Two-stream UNet shared encoder	0.717 ± 0.085	0.916 ± 0.028	3.833 ± 0.589*	2.030 ± 0.761
	Two-stream UNet shared decoder	0.715 ± 0.086	0.923 ± 0.015	**2.888 ± 0.848**	1.879 ± 0.427
	Proposed Two-Stream UNet	**0.726 ± 0.104**	0.922 ± 0.015	3.033 ± 1.144	1.977 ± 0.398
Subject2	2.5D UNet with CBCT	0.787 ± 0.043*	0.784 ± 0.024*	3.582 ± 0.987*	6.153 ± 1.023*
	2.5D UNet with CBCT + CT	0.827 ± 0.038	0.900 ± 0.009	2.787 ± 1.300	3.969 ± 0.775*
	3D UNet with CBCT + CT	0.781 ± 0.037*	0.686 ± 0.119*	4.285 ± 1.666*	7.490 ± 1.422*
	Two-stream UNet shared encoder	0.775 ± 0.036*	0.797 ± 0.019*	7.784 ± 1.794*	5.789 ± 0.764*
	Two-stream UNet shared decoder	0.798 ± 0.042*	0.819 ± 0.039*	3.577 ± 2.483	5.042 ± 0.988*
	Proposed Two-Stream UNet	**0.841 ± 0.031**	**0.916 ± 0.016**	**2.116 ± 0.564**	**2.754 ± 0.560**

Bold numbers denote the best measure in each column for each class across methods, and * indicates statistical significance under one-sided paired t-tests with $p < 0.05$ w.r.t. the best performance.

Table 3.3: Segmentation results of CBCT entering each path of the two-stream model

Testing Subject/Input/Metric		DSC		MSD (mm)	
		Rectum	Bladder	Rectum	Bladder
Subject1	CBCT -> CBCT	**0.726 ± 0.104**	**0.922 ± 0.015**	**3.033 ± 1.144**	**1.977 ± 0.398**
	CBCT -> CT	0.691 ± 0.090	0.912 ± 0.013	3.208 ± 1.130	2.246 ± 0.333*
Subject2	CBCT -> CBCT	**0.841 ± 0.031**	**0.916 ± 0.016**	**2.116 ± 0.564**	**2.754 ± 0.560**
	CBCT -> CT	0.807 ± 0.051*	0.810 ± 0.012*	3.026 ± 0.618*	5.909 ± 0.695*

Bold numbers denote the best measure in each column for each class across methods, and * indicates statistical significance under one-sided paired t-tests with $p < 0.05$ w.r.t. the best performance.

Table 3.3 reports the inference results where a test image is fed into the trained axial-view two-stream model, and inference is read out either from the path corresponding to its own

domain or the other one. It can be observed that CBCT achieves the better performance when following the path of its own domain, showing the utility of decoupled later inference.

The illustrative example in Fig. 3.5 shows that (1) the high-quality CT is easier to segment for both paths compared to CBCT, and (2) there is significant decoding difference between the two paths, and CBCT is segmented the best by following its tailored path.

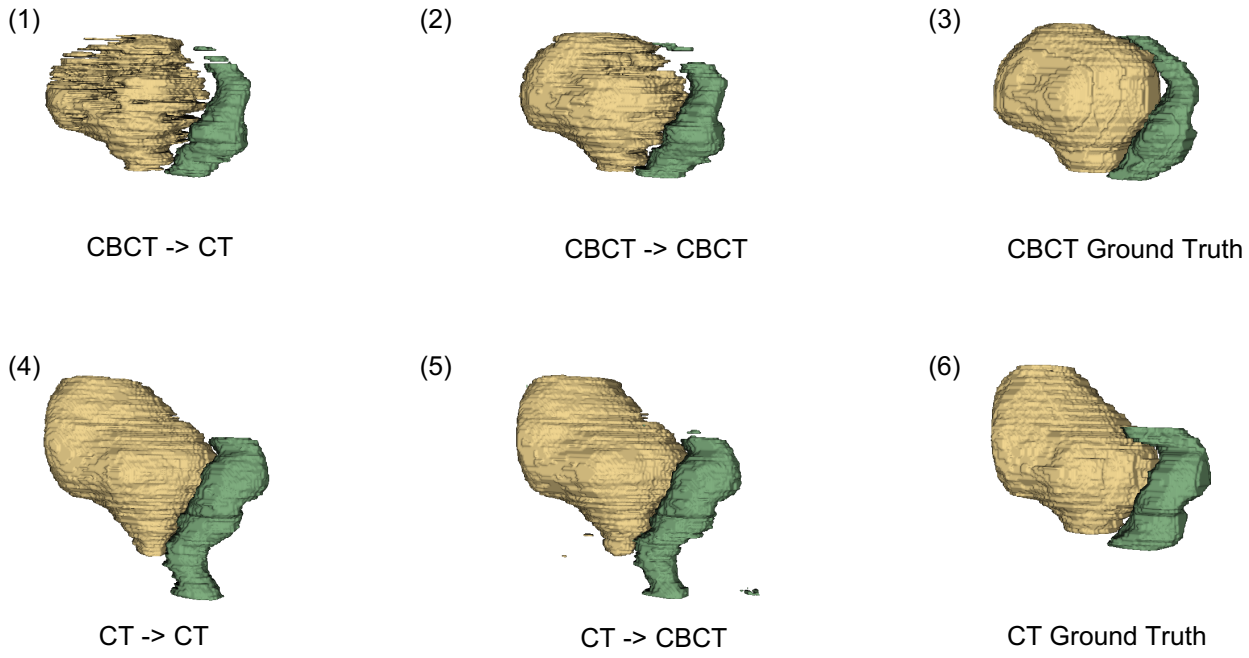


Figure 3.5: Results of each path of the proposed axial-view two-stream model: (1) CBCT data enter CT path, (2) CBCT data enter CBCT path, (3) CBCT ground truth, (4) CT data enter CT path, (5) CT data enter CBCT path, and (6) CT ground truth.

3.3.4.2 Contour Regularization

The role and impact of the proposed tensor-regularized view-ensemble is qualitatively presented in Fig. 3.6. It can be observed that even with the benefit of CT augmentation, single-view based segmentation still suffers from geometric "noise", which can be alleviated to a certain extent by averaging the soft memberships from different views and taking the

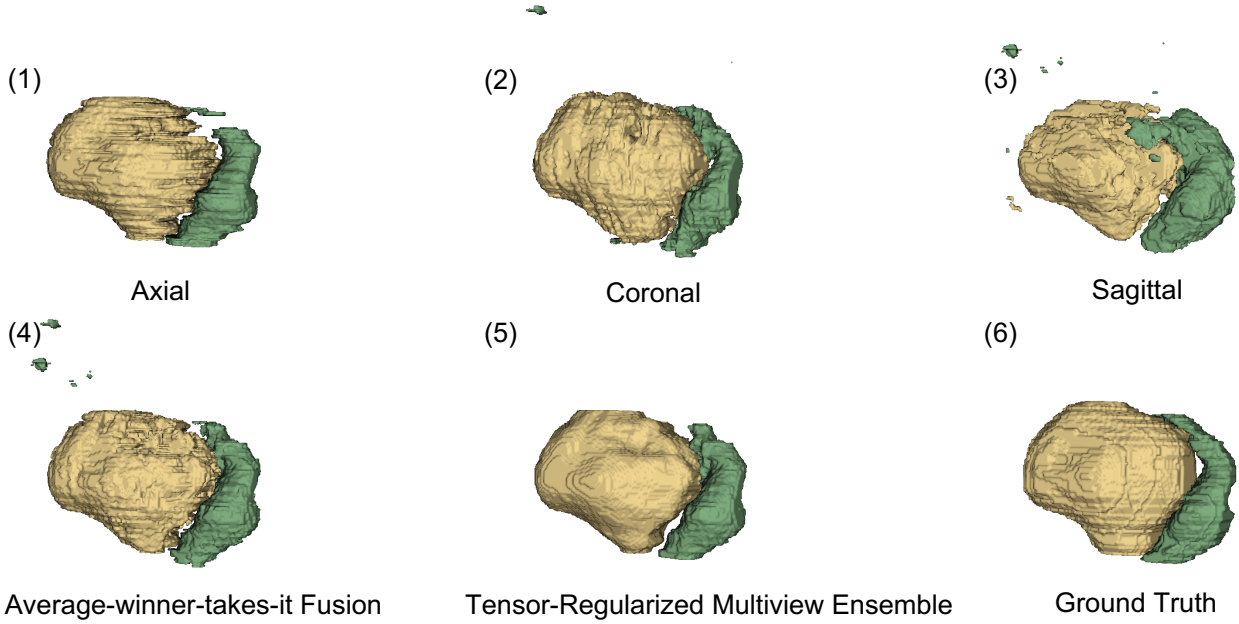


Figure 3.6: Rectum and bladder contour predictions of one CBCT fraction of one testing patient in 3D illustration. The results are by the proposed two-stream model trained from different views: (1) axial, (2) coronal, (3) sagittal, (4) the "average-winner-takes-it" fusion, (5) the proposed tensor-regularized ensemble, and (6) the ground truth. Green: rectum, yellow: bladder.

class associated with the highest averaged membership pixel-wise. The proposed method achieved the highest geometric integrity overall. Table 3.4 reports the impacts quantitatively with the four-fold cross validation, and shows that the proposed approach generally achieved either the best or a performance comparable to the best (without statistical significant inferiority).

While slice-wise morphological operations may provide ad-hoc correction to the contours, no good scheme exists for choosing a consistent parameter to achieve good performances across different slices. Fig. 3.7 shows that larger kernels may successfully fill the holes inside of an organ but can over-fill the boundaries. In contrast, dilation with smaller kernels can miss larger holes inside an organ. The table in Fig. 3.7 reports the measures across all

Table 3.4: Single-view vs. Multi-view Ensemble

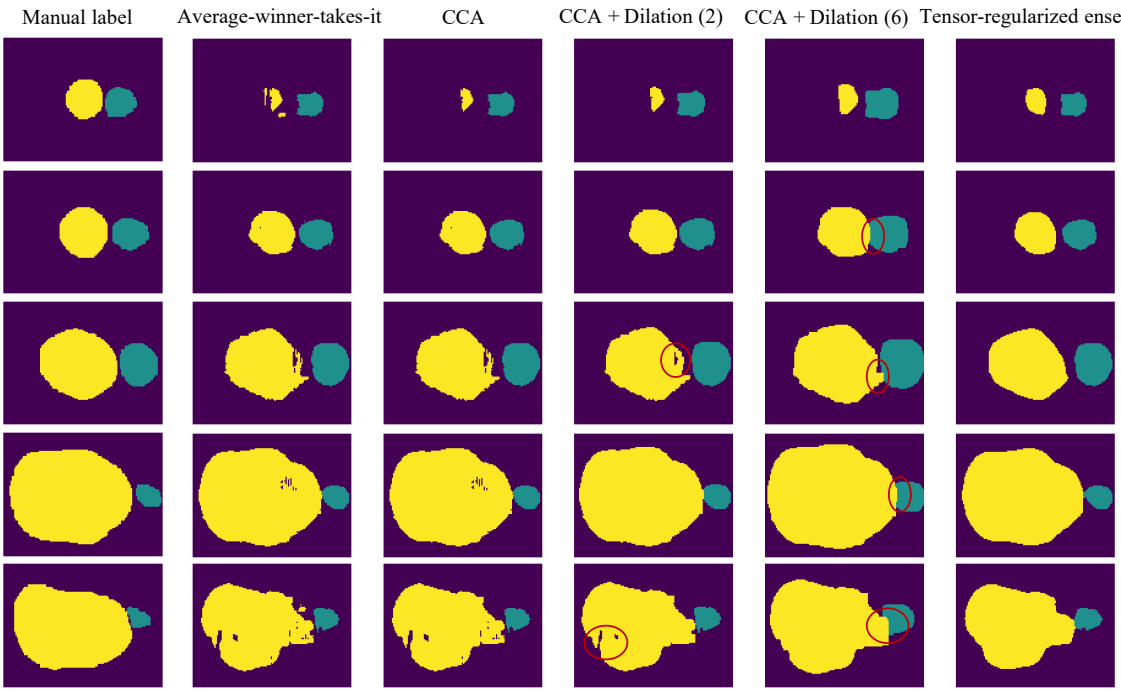
View/Metric	DSC		MSD (mm)	
	Rectum	Bladder	Rectum	Bladder
Axial view	**0.785 ± 0.075**	0.912 ± 0.053	**2.543 ± 1.204**	2.305 ± 1.571*
Coronal view	0.713 ± 0.079*	0.884 ± 0.067*	3.363 ± 1.501*	2.640 ± 1.765*
Sagittal view	0.707 ± 0.082*	0.862 ± 0.108*	3.655 ± 1.371*	2.815 ± 2.040*
Average-winner-takes-it	0.778 ± 0.069	0.915 ± 0.055	2.761 ± 1.433	1.843 ± 1.434*
Tensor-regularized ensemble	0.779 ± 0.069	**0.915 ± 0.055**	2.895 ± 1.496*	**1.675 ± 1.311**

Bold numbers denote the best measure in each column for each class across methods, and * indicates statistical significance under one-sided paired t-tests with $p < 0.01$ w.r.t. the best performance.

the axial slices for an example volume. The proposed method achieved significant better morphological measures than CCA combined with dilation, without compromising the DSC and MSD.

3.3.4.3 Hyperparameter Selection

Fig. 3.8 reports rectum and bladder DSC under different parameter setting, and it shows that the proposed ensemble scheme was robust against parameter selection. Good ranges for λ_1 and λ_2 are found to be $[0.1, 0.6]$ and $[0.1, 2]$, respectively. Our selection of ($\lambda_1 = 0.5$, $\lambda_2 = 2$) manages to achieve improved spatial smoothness without sacrificing DSC, and works well across different volume cases.



Rectum DSC	**0.842 ± 0.131**	0.843 ± 0.131	0.830 ± 0.111*	0.737 ± 0.085*	0.832 ± 0.137*
Bladder DSC	0.896 ± 0.123*	0.892 ± 0.132*	**0.904 ± 0.116**	0.901 ± 0.064	0.904 ± 0.102
Rectum MSD	**1.463 ± 0.738**	1.457 ± 0.771	1.709 ± 0.696*	3.374 ± 0.825*	1.605 ± 0.844*
Bladder MSD	3.284 ± 2.049*	3.456 ± 2.358*	2.946 ± 2.120*	3.307 ± 1.040*	**2.664 ± 1.850**
#Hole pixels	24.79 ± 41.27*	24.52 ± 41.32*	12.53 ± 28.08*	**1.34 ± 6.13**	2.93 ± 17.61
TV	439.79 ± 108.56	426.04 ± 103.54*	396.05 ± 88.32*	369.88 ± 70.21*	**344.145 ± 74.84**

Figure 3.7: Qualitative and quantitative comparison between the proposed tensor-regularized ensemble method and slice-wise morphological operations. Red circles point out the discrepancy introduced to the contour between two different dilation sizes. #Hole pixels denote the number of hole pixels inside organs, and the TV is 2D total variation. Bold numbers denote the best measure in each column for each class across methods, and * indicates statistical significance under one-sided paired t-tests with $p < 0.05$ w.r.t. the best performance.

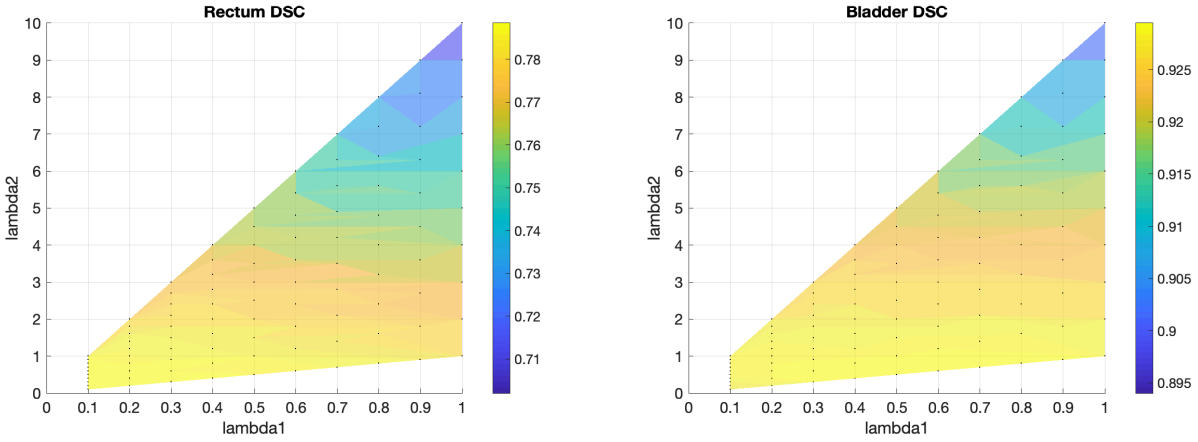


Figure 3.8: Rectum and bladder DSC under different TV regularization parameters

3.4 AIR BUBBLE-INDUCED PERFORMANCE DEGRADATION

3.4.1 Introduction to CBCT-specific Feldkamp Artifacts

While common domain embedding may help to enhance CBCT image quality by improving its signal-to-noise ratio (SNR) [44], CBCT-specific Feldkamp artifacts, induced by bony structures or air bubbles and exhibit as high-frequency streak-like patterns, are much harder to handle with their spatial and orientational dependency [62]. The Feldkamp artifacts can affect structure visibility in CBCT [63], and negatively impact the dose calculation accuracy in radiotherapy to the pelvis [64].

Unfortunately, there exists very limited work where such artifacts are addressed: they are handled by either artificial deletion or empirical correction. With the deletion approach, the volumes or slices (in 2D setting) with significant air bubbles are detected and excluded from the training process [65]. This could result in biased training and also failure to cope with air bubble presence at inference time. With the correction line, the detected air bubble regions are either interpolating with intensities from surrounding regions [66], or assigning a uniform value of typical soft tissues [67]. Fig. 3.9 illustrates such artifacts and their variability in appearance and spatial support.

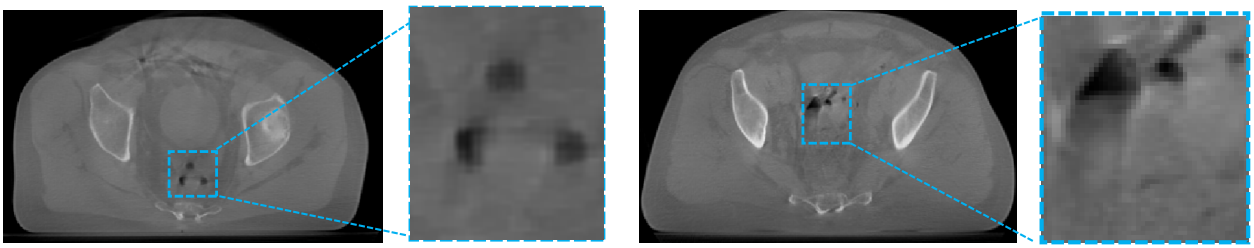


Figure 3.9: Illustration of air bubble-induced artifacts in CBCT imaging

While automatic segmentation using deep networks has demonstrated promise, it has shown inferior performance based on CBCT than planning CT with similar networks [42, 44, 48]. In addition, we have observed rectum segmentation from post-prostatectomy patients

are generally more challenging than from patients without surgery, even with comparable rectum morphology statistics. We hypothesize that this may be related to the higher proportion of post-surgical patients who experience abnormal bowel movements and constipation, resulting in much more frequent and severe air bubble presence in the rectum [68].

3.4.2 Establishment of Relationship between Air-bubbles and Segmentation Performance

In this work, we systematically examine the relation between air bubble severity and the task-specific endpoint of deep learning-based automatic rectum segmentation.

3.4.2.1 Air Bubble Characterization

Air bubbles were detected by thresholding the Hounsfield units at -500 and superimposed with either the manually contoured or the network-estimated mask, to differentiate its occurrence within the rectum from outside the rectum.

From each of the axial, sagittal, and coronal views, the portion of slices containing air bubbles was calculated with respect to the corresponding rectum range.

$$\text{Air}_v = \frac{\#\text{rectum slices with air}}{\#\text{total rectum slices}}, v = \text{axial, sagittal, coronal.} \tag{3.8}$$

3.4.2.2 Gaussian Mixture Model for Correlation Identification

It is reasonable to expect that 1) there exist factors other than air bubble severity that would also affect the segmentation performance, and 2) the influence of air bubble severity dominating the segmentation performance could be case dependent. Therefore, it is reasonable to consider a mixture model. To analyze the regression relationship, let $X = (\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n})$ be a collection of n independent observations, and each $\boldsymbol{x_i}, i = 1, 2, ..., n$ is a two-dimensional vector with air severity as its first coordinate and the corresponding Dice similarity coef-

ficient (DSC) in 3D as the second coordinate, i.e., $[\mathrm{Air}_v, \mathrm{DSC}_v]^{\mathrm{T}}$, for v = axial, sagittal, coronal.

By testing against stability to initial conditions and overfitting behavior, a two-component mixture model is established and a set of Bernoulli-distributed latent variables $\boldsymbol{z} = (z_1, z_2, ..., z_n)$ are introduced, such that:

$$P(z_i = 0) = p, \text{and } P(z_i = 1) = 1 - p, \ i = 1, 2, ...n,$$

and the associated conditional distribution:

$$P(\boldsymbol{x_i}|z_i = 0) \sim \mathcal{N}(\boldsymbol{\mu_0}, \Sigma_0), \text{and } P(\boldsymbol{x_i}|z_i = 1) \sim \mathcal{N}(\boldsymbol{\mu_1}, \Sigma_1), \ i = 1, 2, ..., n,$$

where $\mathcal{N}(\cdot)$ is a Gaussian probability density function. To estimate the parameters - the mixture weight p, mean vectors $\boldsymbol{\mu_0}$ and $\boldsymbol{\mu_1}$, and covariance matrices Σ_0 and Σ_1 in $\theta = (p, \boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \Sigma_0, \Sigma_1)$, the likelihood function

$$\mathcal{L}(\theta; X, \boldsymbol{z}) = P(X, \boldsymbol{z}|\theta) = P(X|\boldsymbol{z}, \theta)P(\boldsymbol{z}|\theta) = p\mathcal{N}(X, |\boldsymbol{\mu_0}, \Sigma_0) + (1 - p)\mathcal{N}(X|\boldsymbol{\mu_1}, \Sigma_1), \quad (3.9)$$

is maximized using the expectation-maximization (EM) algorithm [69].

Let the eigen-decomposition (ED) of covariance matrix $\Sigma_c, \ c = 0, 1$ be:

$$\mathrm{ED}(\Sigma_c) = Q_c \Lambda_c Q_c^{-1},$$

and let $\boldsymbol{q_c}$ be the eigenvector corresponding to the smaller eigenvalue, and then a local linear approximation for cluster c can be established by:

$$< \boldsymbol{x} - \boldsymbol{\mu_c}, \ \boldsymbol{q_c} > = \ 0, \ c = 0, 1, \quad (3.10)$$

where $< \cdot >$ denotes the inner product.

3.4.3 Relationship Results

The same four-fold cross-validation process was applied to ensure each sample had been treated as a test sample exactly once and the performance can be properly assessed. For

each view, 17×5 air bubble severity levels with the corresponding DSC metrics in 3D were calculated.

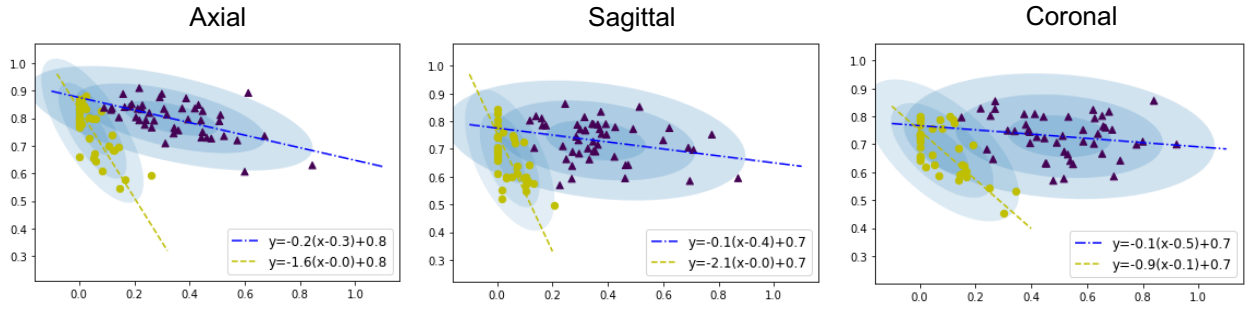


Figure 3.10: Air bubble severity in rectum (horizontal axis) vs. rectum DSC in 3D (vertical axis) for each view: clusters of purple triangles and yellow dots represent two separate clusters. The $-\cdot-$ style blue line and the $--$ style yellow line denote the fitted linear lines for the corresponding cluster. We name the cluster with the more negative regression slope as cluster 1 (yellow dots) and the other cluster as cluster 0 (purple triangle).

Fig. 3.10 shows the relationship between the air bubble severity and the rectum segmentation performance, and visualizes the local Gaussian component with the corresponding linear presentation in (3.10). It clearly shows two clusters with distinct linear regression patterns. One shows a strong dependency of segmentation performance on the rectum air bubbles, and the other one suggests influences from alternative factors. The axial view shows a steeper slope than the other two views, indicating that the axial view is affected the most by the air bubble-induced Feldkamp artifacts.

Additionally, the consistency of the clustering structure amongst different views was assessed by examining the cluster membership triplets. $[0, 0, 0]$ and $[1, 1, 1]$ mean a volume is either consistently heavily influenced by the air bubble severity or not across all views, whereas a triplet of $[1, 1, 0]$ or $[1, 0, 0]$ indicates one out of the three views behaves differently.

Fig. 3.11 illustrates the clustering consistency amongst the three views with sums of the binary triplets. It shows that most samples (in purple triangles and yellow dots) are

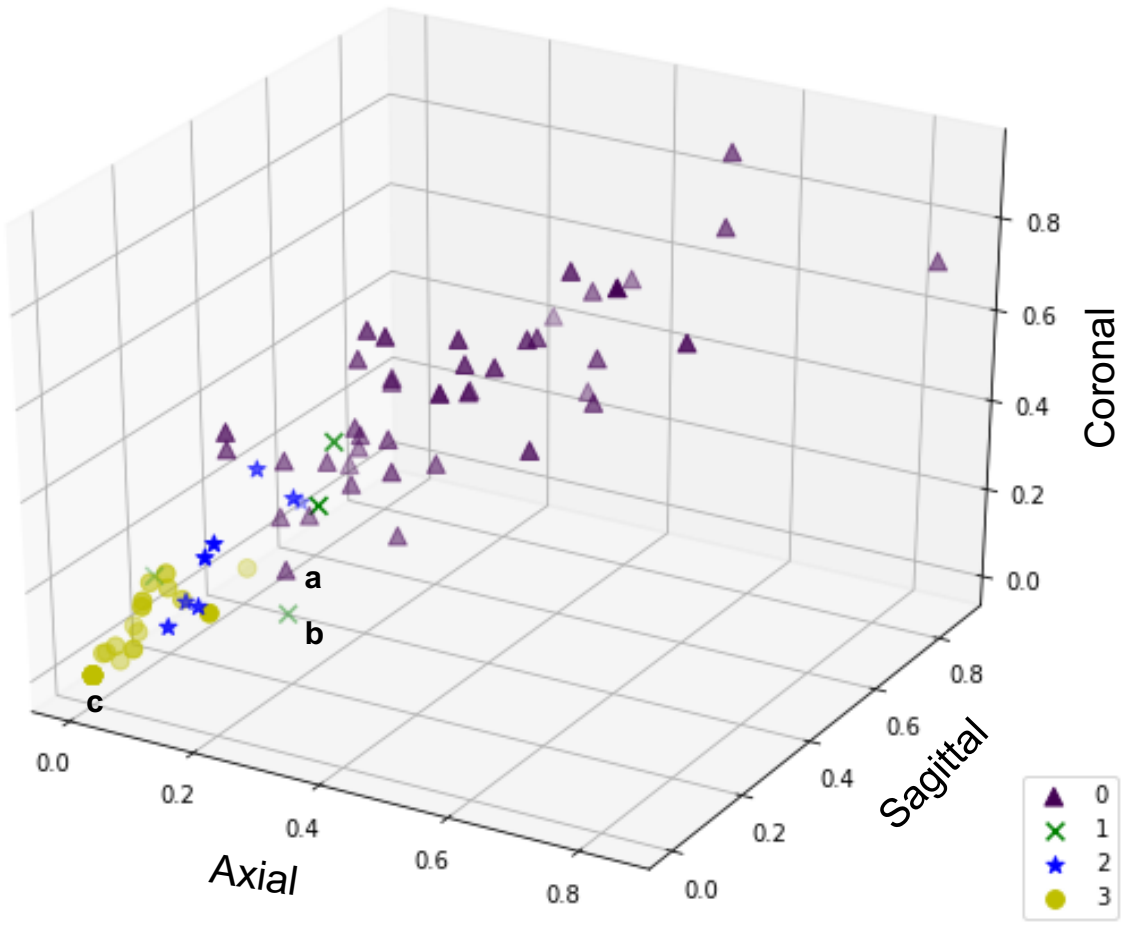
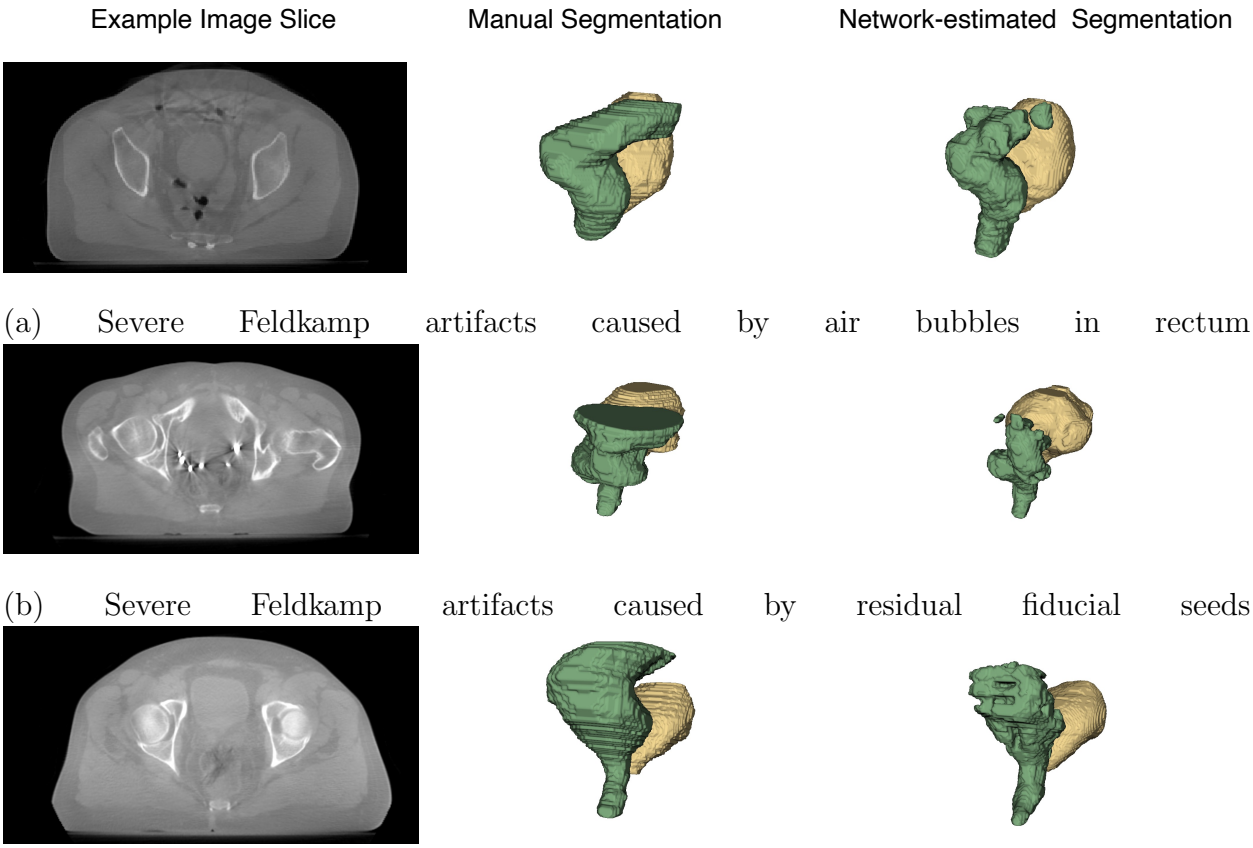


Figure 3.11: Cluster membership consistency across views: the yellow dots and purple triangles have a consistent cluster membership across the axial, sagittal, and coronal views, and the blue stars and green crossings have membership discrepancy. Symbols with colors denote the membership summation of the three views for each sample. Cases (**a**)-(**c**) are shown in Fig. 2.4.

consistent in the cluster membership, and only a few samples (10 out of 85, blue stars and green crossings) have discrepant cluster membership across views. While the two-cluster behavior in each view supports a mixture impact model, the high agreement of membership across different views indicates that the relative role of such attribute is highly consistent.



(a) Severe Feldkamp artifacts caused by air bubbles in rectum

(b) Severe Feldkamp artifacts caused by residual fiducial seeds

(c) Atypical rectum morphology

Figure 3.12: Example cases with the lowest DSC. (a) is from the cluster with strong dependency on the air bubbles, (b) and (c) are two samples from the cluster with lower performance dependency.

We investigated differences between the two clusters for the low-DSC points. Fig. 3.12(a),(b),(c) present the air bubble severity in descending order. Fig. 3.12(a) illustrates severe air bubbles in the rectum which had a strong impact on segmentation performance. In the cluster with weaker dependency on air bubble severity, most volumes had inconsistent

appearance of fiducial seeds, exemplified in Fig. 3.12(b), and some had abnormally thick rectum at the superior portion likely due to constipation, as shown in Fig. 3.12(c).

3.5 DISCUSSION AND CONCLUSIONS

3.5.1 Segmentation Network Structure Comparison

Table 3.2 shows moderate improvement of the two-stream UNet model over a 2.5D UNet that takes a brute-force inclusion of CT and CBCT input alluding to high relevance between CT and CBCT, in particular as the physicians relied heavily on transferring CT-derived contours to delineate CBCT.

A major challenge associated with using a 3D network is the requirement of data volume amount in supervised setting. Each labor-intensive CBCT contour set only provides a single data sample for 3D UNet - with the current set of 65×2 training volumes, the validation loss exhibits significant fluctuations, risking unstable convergence. A different but related consequence for such sample handling is more sensitive to out-of-distribution test, as shown in Fig. 3.4. In contrast, 2.5D models use slice stacks that can augment samples and offer robustness, stability, and low data-demand with the lower-dimensional model. The model also uses cross-validation assessment to prevent overfitting. Each model performance vs. its complexity is summarized in Table 3.5.

3.5.2 Segmentation Result Comparison with Literature

We have reported lower DSC values, particularly in the rectum, than the reported state-of-the-art CBCT-based pelvic segmentation results as 0.874 ± 0.096 and 0.814 ± 0.055 for the bladder and rectum in [42], and 0.916 ± 0.005 and 0.872 ± 0.201, respectively, in [44]. This is expected with our post-prostectomy dataset in comparison with the typical pelvic images with intact prostates - with prostate removed there is more pelvic cavity. We have

Table 3.5: Model Performance and Complexity Comparison

Model/Metric	DSC (Test Subject 2)		Training Time	Inference Time	#Parameters
	Rectum	Bladder			
2.5D UNet	0.827 ± 0.038	0.900 ± 0.009	2.3h	0.6 ± 0.2s	15M
3D UNet	0.781 ± 0.037*	0.686 ± 0.119*	5h	0.6 ± 0.0s	34M
2.5D two-stream UNet	0.841 ± 0.031	0.916 ± 0.016	2.5h	0.8 ± 0.2s	17M
Tensor-regularized ensemble[a]	**0.844 ± 0.048**	**0.921 ± 0.019**	2.5h	255.7 ± 28.9s	17M

[a]the tensor-regularized ensemble is a post-processing module of a 2.5D two-stream UNet; bold numbers denote the best measure for each class across methods, and * indicates statistical significance under one-sided paired t-tests with $p < 0.05$ w.r.t. the best performance.

observed pronounced intra- and inter-subject variation of the filling status of the rectum, as reported in Table 3.1: in our dataset of 17 patients and 85 fractions, the volumes of the rectum and bladder are 83 ± 39 cm^3 and 278 ± 162 cm^3 respectively, compared to 50 ± 10 cm^3 and 401 ± 130 cm^3 across 30 patients without surgery [70].

Restricted by the low CBCT image quality, and aiming to provide radiation coverage to the clinically defined prostate bed, the clinician's contours exhibit large uncertainty and tend to be less rigorous and consistent in the top and bottom of the rectums, as illustrated in Fig. 3.13. In fact, in a large scale prostate cancer study with ten contours of five physicians, the mean and std Dice of the contours on CBCT images were 0.9 ± 0.1 for the bladder and 0.7 ± 0.1 for the rectum [37]. Our reported values in Table 3.4 are consistent with human uncertainty.

3.5.3 Air-bubble in Rectum vs. Segmentation Performance

Our data exhibit air pockets with much higher frequency and severity compared to typical male pelvic imaging. This is possibly caused by post-surgical side-effects such as abdominal bloating, diarrhea, or constipation in post-prostatectomy patients [68], and post-surgery abnormal bowel movement makes it harder to comply to imaging protocols. Our observation

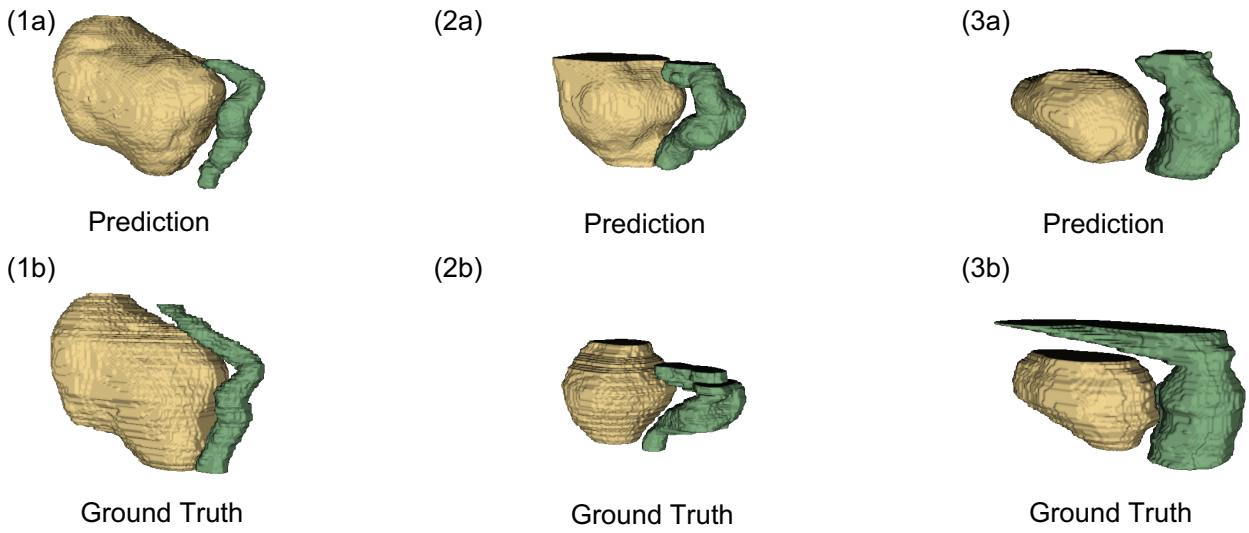


Figure 3.13: Example cases with low rectum/bladder DSC: (1) 0.672/0.915, (2) 0.652/0.818 (3) 0.656/0.910. The predictions maintain geometric integrity and clinical usability, while (1) and (3) miss some upper parts of the rectum compared to the physician contoured ground truth, and (2) predicts a larger bladder than the manual label.

agrees with the reported assessment for post-prostatectomy patients that the daily variations of the rectum and bladder filling status are larger than the patients with intact prostate [70, 71, 72]. The composition with artifacts caused by severe air bubbles further challenges rectum segmentation [73], and also leads to high uncertainty in human contours used as training labels.

3.5.4 Air-bubble Severity and Segmentation Performance Quantification

Air bubbles occurring closer to the peripheral versus those within the interior of the rectum impact segmentation differently. The current counting measure (3.8) for quantifying the air bubble severity does not capture spatial distribution, and a further improvement is being investigated to incorporate such information.

The study focuses on identifying and analyzing the cause of performance degradation and desires a normalized and stable performance measure. We think DSC is the best choice with its [0,1] range and robustness to sparse outliers, compared to other metrics such as Hausdorff distance. Either manual delineation or estimated rectum masks can be used to differentiate air bubbles located within or outside the rectum, despite the imperfection of the estimated rectum segmentation.

3.5.5 Conclusions and Future Development

We have proposed a novel method to segment pelvic organs from CBCT for post-prostatectomy patients. Our key contribution is a consistent ensemble logic, as reflected in both ensemble YOLO localizer and a tensor-regularized aggregation scheme to combine estimates from multiple views. The proposed method is shown to achieve comparable performance metric to manual confidence on pelvic CBCT, but on the much challenging post-prostatectomy cases with large spatial variation and severe artifacts.

By fitting a two-component mixture model, we have shown that severe air bubbles and

fiducial seeds induced Feldkamp artifacts are the primary contributors to performance degradation in deep-learning based rectum segmentation. Other contributing factors include out-of-distribution morphology at testing time.

The labeling uncertainty can be measured using our proposed agreement index between the constructed similarity graphs corresponding to the input image pairs and input label pairs. A probablistic UNet can also be adopted [74], which models a segmentation latent space accounting for all the segmentation variants caused by inter- and intra-observer variations. Plausible segmentation variants for an image input can be generated by multiple sampling processes, and the manual labeling uncertainty can then be derived by calculating the contour variations.

One pursuit direction for improving segmentation performance is the suppression or removal of air-pocket induced artifacts. A preliminary experiment with the threshold-and-fill approach did not improve segmentation performance [66], as severe air volumes lead to streaking behavior that is beyond local support, and interpolation aggressively may compromise the already low image contrast. The residual high-contrast fiducials cause similar Feldkamp artifacts. With deep learning approaches where abstract textures and context information are critical to support good learning-inference behavior, brute-force interpolation or bulk assignment approaches no longer meet the need to generate consistent feature values for the rectum. As a result, more sophisticated correction schemes either as preprocessing or integrated into the deep learning process are necessary to enhance segmentation performance. We are actively investigating alternative correction schemes, such as projection domain methods [75, 76].

Though not the emphasis of the current work, our segmentation method is compatible with extensions to more flexible auxiliary dataset sizes as well as more image modalities such as MRI. Using unbalanced streams would mitigate the need for labeled CBCT data [77], and the use of supplementary modalities may further complement either contour confidence or latent inference, similar to common domain embedding methods [51]. The auxiliary MR

for CBCT synthesis and planning CT can also be used to benefit a more comprehensive evaluation of the model generalizability.

CHAPTER 4

Intracranial Vessel wall Segmentation for Atherosclerotic Plaque Quantification

4.1 BACKGROUND

Stroke is a leading cause of morbidity and mortality in the US and worldwide [78, 79]. Intracranial atherosclerosis disease (ICAD), characterized by lipid deposition, inflammation, and remodeling in the artery vessel wall, remains a major risk factor for stroke occurrence. Magnetic resonance (MR) vessel wall imaging (VWI) is an emerging non-invasive technology to assist in ICAD evaluation, thanks to its high spatial resolution and superior dark-blood contrast [80, 81, 82]. Quantitative assessment of atherosclerotic lesions based on MR-VWI may provide valuable insights into the severity of ICAD [81]. Several morphological measurements, such as normalized wall index (NWI), arterial wall remodeling ratio (RR), and plaque-to-wall contrast ratio (CR), illustrated in Fig. 4.1, have been shown to be useful imaging surrogates for plaque burden quantification [83, 84, 85, 86, 87]. These measurements rely on accurate contouring of the vessel wall in a cross-sectional view.

Vessel wall contouring is typically performed manually and is subject to high inter- and intra- observer variations. These variations can induce high uncertainty on subsequent quantitative analysis on the small intracranial arteries. Moreover, with the advent of 3D VWI with large spatial coverage [88, 89], the presence of multiple ICAD lesions in a patient may incur intensive labor cost and exacerbate human errors. These limitations and concerns call for an automated method to improve segmentation accuracy, consistency, and efficiency.

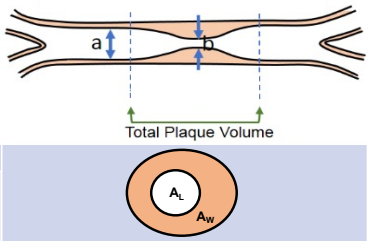
Clinical Quantitative Features	Definition	Illustration
1. Diameter Stenosis (DS)	(a-b) / a × 100%	
2. Normalized Wall Index (NWI)	A_w / (A_w + A_L)	
3. Remodeling Ratio (RR)	$(A_W+A_L)_{plaque}$ / $(A_W+A_L)_{reference}$	
4. Plaque-Wall Contrast Ratio (CR)	$Signal_{plaque}$ / $Signal_{reference}$	

Figure 4.1: Illustration of clinical stenosis quantification features: a and b are the diameter of the reference image slice and the most stenotic slice in a selected vessel segment, respectively. A_W and A_L are the area of the vessel wall and the lumen, respectively; and $\text{Signal}_{\text{plaque}}$ and $\text{Signal}_{\text{reference}}$ denote the vessel wall region image signals of the most stenotic VWI slice and the reference slice across a segment, respectively.

An end-to-end plaque analytical pipeline should consist of five modules (Fig. 4.2): 1) "image registration" between MR angiography (MRA) and VWI, 2) "centerline tracking" on MRA, 3) "vessel straightening and slicing" along the extracted centerline on VWI, 4) "vessel wall and lumen segmentation" on cross-sectional VWI, and 5) "plaque quantification". The end-to-end pipeline would ultimately provide the physician with the quantitative indices of the degree of plaque, which can be compared longitudinally for monitering the stages and development of stenosis to prevent stroke occurrence. In this study, we focus on the development of automated vessel wall segmentation module based on deep learning [90, 91]. More accurate segmentation would enable more accurate quantitative feature computation based on the contour areas, etc.

4.2 MR-VWI DESCRIPTION

T1-weighted MR VWI from 80 patients diagnosed with ICAD were involved for this study. The images were acquired with a whole-brain MR VWI protocol [92, 89], using a 3-Tesla

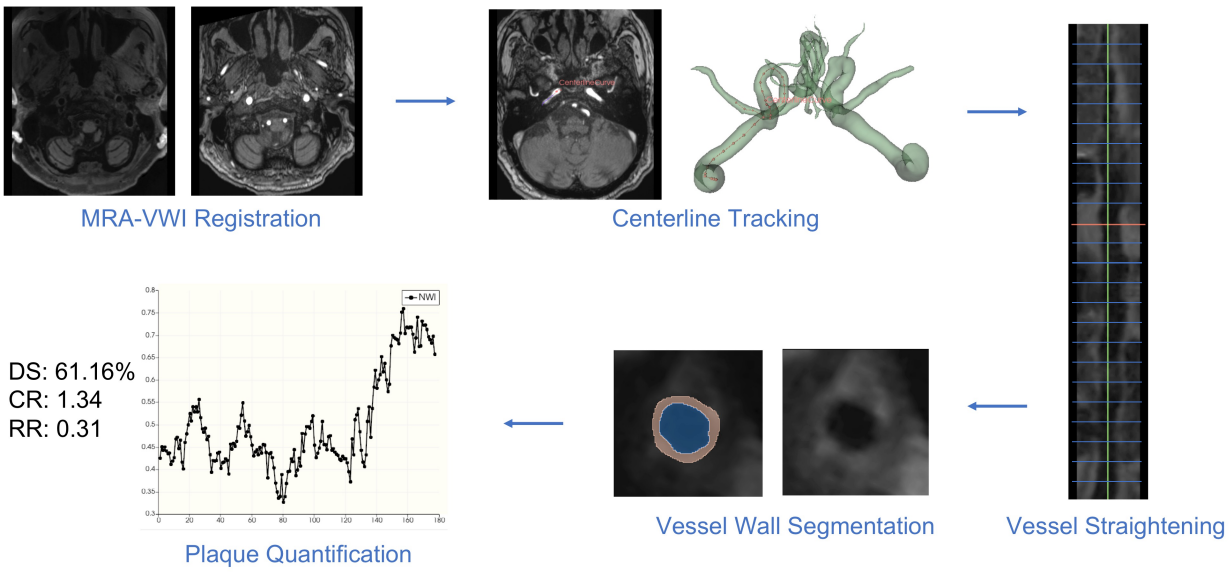


Figure 4.2: Overall automatic pipeline schema of an end-to-end plaque analytic

whole-body system (MAGNETOM Prisma; Siemens Healthcare, Erlangen, Germany) and a 64-channel head/neck coil (Siemens Healthcare). The images were acquired at an isotropic spatial resolution of 0.55 mm. The following four arterial segments including the one that involved the identified plaque were used for segmentation sample preparation: the intracranial internal carotid artery, the middle cerebral artery, the intracranial vertebral artery, and the basilar artery. 3D Slicer (version 4.11.0) was used to generate 30 contiguous 2D cross-sectional slices with 0.55 mm slice thickness and 0.1 mm in-plane resolution from each segment [93]. The ground truth lumen and vessel wall were labeled by an experienced radiologist using ITK-SNAP (version 3.8.0) [94].

4.3 RELATED VESSEL SEGMENTATION METHODS

Conventional automated or semi-automated vessel wall segmentation methods applied to MR-VWI images are usually based on explicit model fitting. For example, the shape of a whole carotid vessel was approximated as elliptic, and was translated, deformed, and

rotated iteratively to fit the outer vessel wall boundary [95]. In each iteration, the similarity of the ellipse to the outer wall boundary was evaluated with the average intensity gradient magnitude along the ellipse. The ellipse with the highest intensity gradient average was obtained as the final outer wall boundary. In addition to the 2D model, a 3D model has been investigated by deforming a 3D cylindrical non-uniform rational B-spline surface to fit the inner and the outer vessel wall boundary of a carotid artery [96]. A tube model was initialized by rings with pre-specified diameters and numbers of control points, and the control point locations were adjusted iteratively with signal intensities. The major disadvantages of these methods are the long computation time for iterative model fitting and the potential model misfit when the shape assumptions are violated.

As an alternative to the parametric approaches, level-set based methods can perform numerical computations of curves and surfaces on a fixed Cartesian grid and handle varying topology with ease [97]. Level-set based active contour approaches have been investigated to extract the lumen and outer wall boundaries by minimizing an energy function with fidelity force to align boundary with high gradients, and regularization for smoothness on in-plane contour shape and consistency across adjacent slices [98]. Typical ordinary differential equation (ODE)-based level-set methods usually require long computation time.

Recent research has been utilizing deep neural networks to perform automated vessel wall segmentation, using either a multi-class or multi-label setting. The multi-class methods predict multiple mutually exclusive classes by the same number of output channels, usually with softmax activation in the last network layer. With this setting, Shi *et al.* proposed a 2D UNet to segment the intracranial vessel and reported Dice similarity coefficients (DSC) of 0.89 and 0.77 for the lumen and vessel wall, respectively [99]. In contrast, the semantic segmentation to predict in a multi-label setting can be overlapped, and each pixel can have multiple class memberships. The multi-label setting usually has sigmoid activation in the last network layer, where binary prediction is performed for each class. With this setting, a 2.5D UNet was developed to segment the lumen, whole vessel, and background for the carotid

arteries, and achieved DSC of 0.96 and 0.97 for the lumen and whole vessel, respectively [100].

4.4 UNET++ WITH DICE + HAUSDORFF DISTANCE LOSS

4.4.1 2.5D UNet++ Model and Loss Function

As a preliminary study, we used a 2.5D UNet++ model structure and adopted a loss function composed of both a soft Dice coefficient loss and a distance-transform approximated Hausdorff distance (HD) loss [10, 101]. The UNet++ has dense connections among different semantic scales in the network, which offers structure adaptation. The addition of HD loss encourages the geometric conformality of the segmentation to the manual labels. The modified segmentation network yielded better performances across metrics compared to the benchmark 2D UNet model.

4.4.1.1 UNet++

UNet++ was proposed for 1) efficiently training an ensemble of UNets with multiple depths, and 2) establishing dense and more effective skip connections among varying semantic scales of the network [10]. UNet++ model structure is illustrated in Fig. 4.3. We utilized the deep supervision option in our design, where each of the D sub-decoders, i.e., $X^{0,i}, i \in [1, D]$ outputs a prediction and contributes to the total loss. We minimized the soft Dice coefficient (DC) loss for each sub-decoder:

$$L_{dc}\left(p,\ y^{i}\right) = \sum_{c=1}^{C}\left(1 - \frac{1}{N}\sum_{n=1}^{N}\frac{2p_{n,c}\ \cdot\ y_{n,c}^{i}}{{p_{n,c}}^{2}\ +\ {y_{n,c}^{i}}^{2}}\right) \tag{4.1}$$

where $y_{n,c}$ and $p_{n,c}$ is the prediction of pixel n for class c and the ground truth, respectively; N is the total number of pixels in a batch, and C is the number of classes.

Specifically, we assigned equal weights to the loss from each sub-decoder leading to the

overall soft DC loss function:

$$\mathcal{L}_{DC} = \sum_{i=1}^{D} \mathcal{L}_{dc}\left(p,\ y^i\right). \tag{4.2}$$

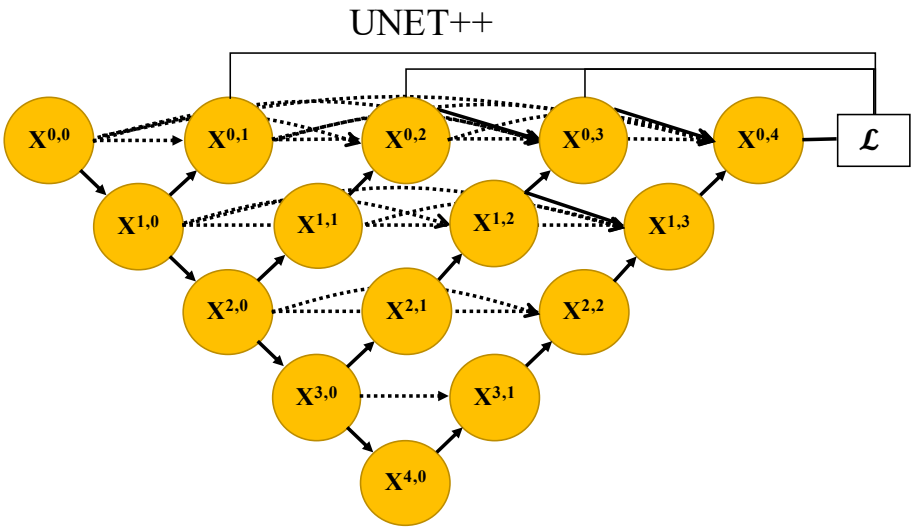


Figure 4.3: UNet++ structure for intracranial vessel wall segmentation: each node is a convolution block, downward arrows are down-sampling, upward arrows are up-sampling, and dot arrows are skip connections

4.4.1.2 Hausdorff Distance Loss

HD is computed between the predicted segmentation boundary and the ground truth, which indicates the biggest point-wise matching discrepancy. The bidirectional HD between the ground truth set P and the predicted set Y is:

$$HD\left(P, Y\right) = max\left(hd(P, Y)\,,\ hd(Y, P)\right), \tag{4.3}$$

where

$$hd(P,\ Y) = max_{p \in P} min_{y \in Y} ||p - y||_2, \tag{4.4}$$

$$hd(Y,\ P) = max_{y \in Y} min_{p \in P} ||p - y||_2. \tag{4.5}$$

Karimi et al. proposed three methods for approximating HD and incorporated the estimated HD loss in the overall loss function, which enabled a direct minimization of the HD [101]. We adopted the HD loss approximated by the distance transform (DT), due to its effective implementation. For a 2D binary image $X[i, j]$, with 0 representing the background and 1 the foreground, the DT of X is:

$$DT_X[i, j] = min_{k,l} d([i, j], [k, l]). \tag{4.6}$$

where $[k, l]$ is the indices of the foreground pixels, and d here denotes the Euclidean distance in our case. The HD loss is:

$$\mathcal{L}_{HD}(p, y) = \frac{1}{N} \sum_{i,c,n} (p_{n,c} - \ y^i_{n,c})^2 \cdot (dt^2_{p_{n,c}} + dt^2_{y^i_{n,c}\prime}). \tag{4.7}$$

Here c is the foreground classes, and dt_p and $dt_{y\prime}$ denote the DT of the ground truth p and the predicted binary segmentation mask $y\prime$, respectively.

4.4.1.3 Overall Loss Function

Combining both the soft DC loss and the HD loss with a hyperparameter λ, we formulated our overall loss function as:

$$\mathcal{L}_{all} = \lambda \mathcal{L}_{DC} + \ \mathcal{L}_{HD}. \tag{4.8}$$

4.4.2 Assessment and Network Specifications

For the network development, we split a subset of total patient cohort that contained 30 patients into 24:3:3 patients for training, validation, and testing, respectively. The performance of our model was evaluated on the testing set with the following four criteria in 2D, i.e., 1) Dice similarity coefficient (DSC), 2) 95 percentile HD (HD_95), 3) mean surface distance (MSD) from the prediction to the ground truth, and 4) the mean absolute error of NWI (MAE_NWI). While the DSC measures the integrative area discrepancy, the HD and MSD are good indicators for discrepancies at the segmentation boundaries. NWI is a clinical

morphological feature defined as: $\frac{|VW|}{|VW \cup Lumen|}$, where $|\cdot|$ here indicates area, VW denotes vessel wall. NWI ranges from 0 to 1, with a higher value indicating a heavier plaque burden.

The training data were augmented by randomly flipping vertically and horizontally and isotropic and anisotropic zooming-in images. The 2.5D network took three consecutive 128 $\times$ 128 slices as the triple-input to the network to estimate the middle slice label. The classification contained three classes: background, the lumen, and the vessel wall. The network specifically classified background, lumen, and the outer vessel boundary with a sigmoid activation, and subtracted lumen from the outer boundary as the vessel wall. The soft DC loss of the vessel wall was also added to the total soft DC loss [100].

Similar to the model structure in our previous study, the basic convolution block consists of a 2D convolution layer, a batch-normalization layer, a Parametric Rectified Linear Unit (PReLU) layer, and another 2D convolution layer [99]. All UNet and UNet++ models have a depth of four. The base number of channels is 32, and is doubled or halved at the down- or up-sampling processes in both UNet and UNet++. The hyperparameter λ in Eq. (4.8) was tuned with respect to the validating DC loss as 1. We trained the model for 100 epochs with an Adam optimizer and a learning rate of 10^{-5}. A batch size of 16 was used for all models.

A voting scheme was adopted for UNet++ models, which weighted-averaged the predictions among all sub-decoders. We assigned a weighting of $[0, 0.1, 0.8, 0.1]$ to node $X^{0,i}$, $i \in [1, 4]$, respectively, based on the validation performance.

4.4.3 Segmentation Results

Table 4.1 reports the segment-wise mean and standard deviation of the testing data. Fig. 4.4 shows typical segmentation performance with two examples. Fig. 4.5 presents the comparison of NWI curves by each model, and a 3D illustration of an example vessel segment of the ground truth and by the proposed model. It can be observed that the proposed 2.5D

UNet++ model with DC + HD loss generally achieves the best performance across various quantitative measures, and also visually best resembles the ground truth.

Table 4.1: Quantitative Comparison of 2D and 2.5D UNet and UNet++ Models

Model	**Loss Function**	**Class**	**Metric**			
			DSC	**HD_95 (mm)**	**MSD (mm)**	**MAE_NWI**
2D UNet	DC	Lumen	0.9163 ± 0.0522	0.3467 ± 0.5173	0.1034 ± 0.0787	0.0732 ± 0.0294
		Vessel Wall	0.7452 ± 0.1046	0.6146 ± 0.7147	0.1764 ± 0.1270	
2.5D UNet	DC	Lumen	0.9080 ± 0.0641	0.3641 ± 0.5674	0.1047 ± 0.0765	0.0975 ± 0.0425
		Vessel Wall	0.7521 ± 0.1006	0.5360 ± 0.5686	0.1657 ± 0.0971	
2.5D UNet	DC + HD	Lumen	0.9039 ± 0.0665	0.3339 ± 0.4772	0.0998 ± 0.0444	0.0836 ± 0.0422
		Vessel Wall	0.7615 ± 0.0969	**0.4784 ± 0.4994**	0.1423 ± 0.0640	
2.5D UNet++	DC	Lumen	0.9116 ± 0.0723	0.3771 ± 0.6205	0.1061 ± 0.1034	0.0811 ± 0.0404
		Vessel Wall	0.7758 ± 0.0957	0.5281 ± 0.5886	0.1494 ± 0.0882	
2.5D UNet++	DC + HD	Lumen	**0.9172 ± 0.0598**	**0.3252 ± 0.5071**	**0.0940 ± 0.0781**	**0.0725 ± 0.0333**
		Vessel Wall	**0.7833 ± 0.0867**	0.4914 ± 0.5743	**0.1408 ± 0.0917**	

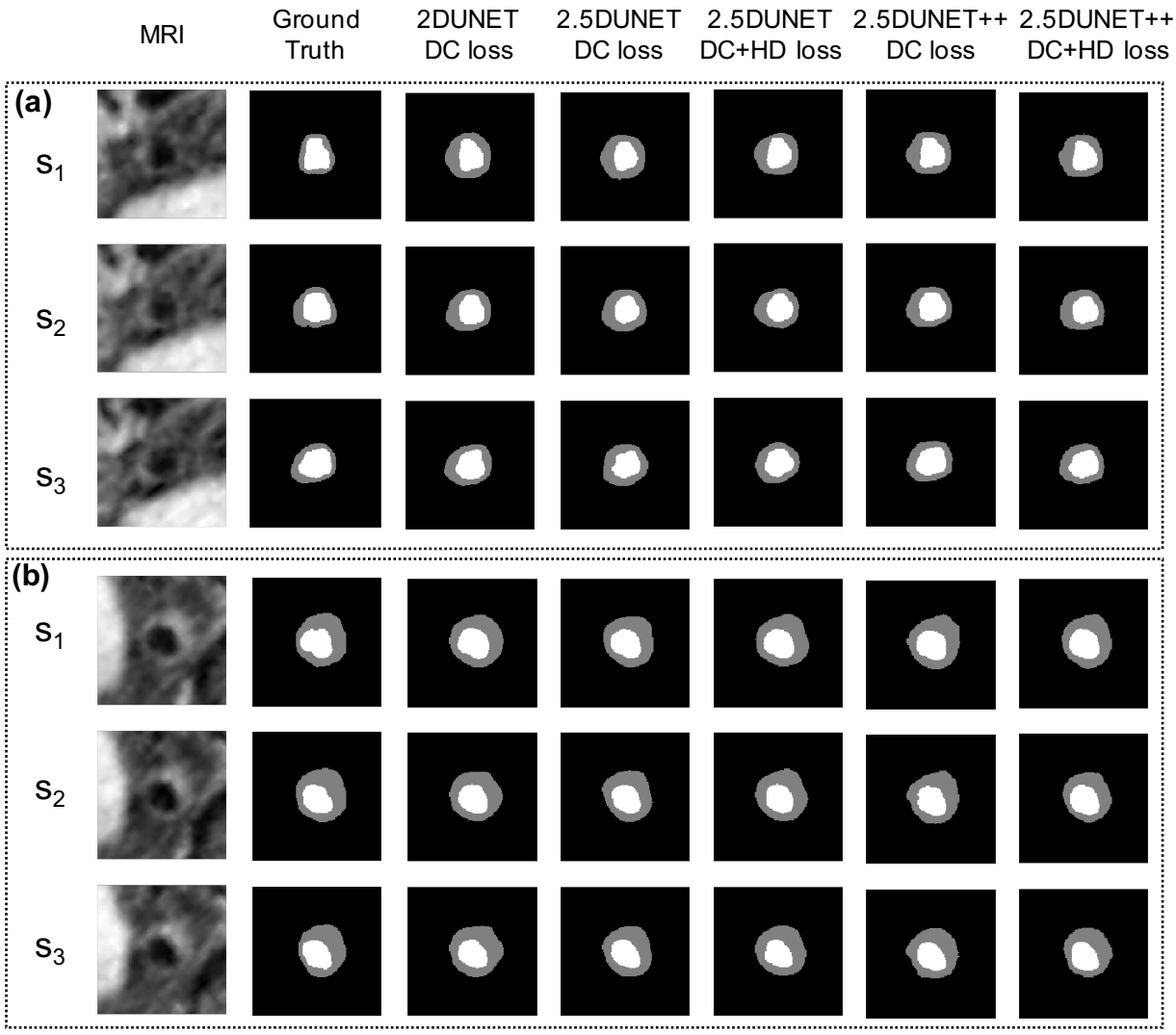


Figure 4.4: Visualization of vessel wall segmentation performance by 2D and 2.5D UNet and UNet++ models: dashed block (a) and (b) are two 3-slice examples from two vessel segments. The 1st column is the original consecutive MRI slices (s1, s2, and s3), the 2nd to the last columns show the ground truth and estimated segmentation from each model of the corresponding MRI slice, respectively. Black is the background, grey is the vessel wall, and white is the lumen.

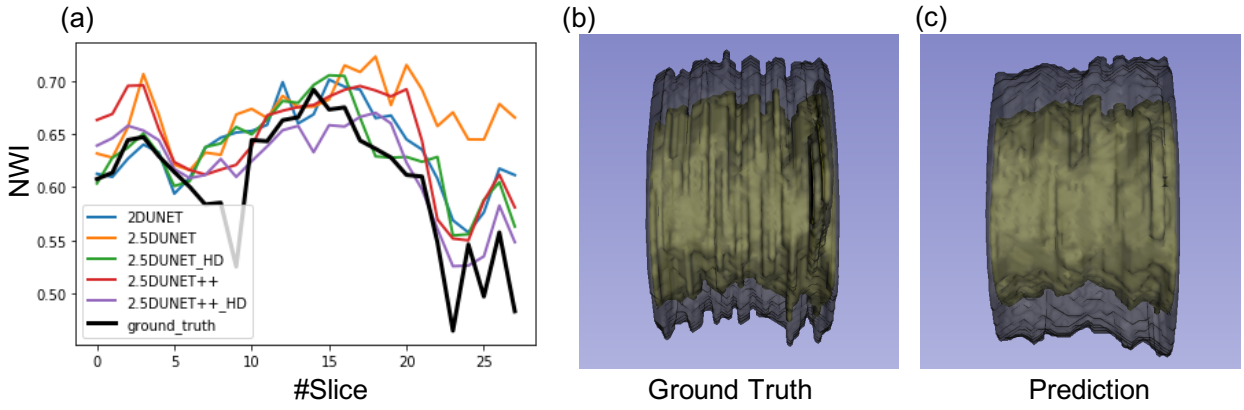


Figure 4.5: Comparisons of NWI curves by each 2D and 2.5D UNet and UNet++ model (a); the ground truth segmentation (b); and the predicted segmentation by the proposed model (c) in 3D illustration of an example vessel segment (30 consecutive slices). Inner yellow is the lumen, and outer grey is the vessel wall.

4.5 TIERED SEGMENTATION INCORPORATING CLASS INCLUSION

4.5.1 Morphological Problems of Multi-channel Segmentation Results

While the reported DSC values were reasonably high, they only indicate a good overlap between the labeled and predicted class memberships, by treating the lumen and the whole vessel (or vessel wall) as separate classes without considering the intrinsic coupling that the lumen resides inside the entire vessel. In other words, the set of lumen pixels should be a proper subset of the whole vessel pixel set, and we use the word "inclusion" hereafter to indicate this concept. This is particularly a concern for deep learning methods as there is little control once the network is trained. In the intracranial arteries, the contrast of the outer vessel boundary may be low, and vessel shapes are more irregular due to tortuosity and frequent branching. As a result, existing networks have been observed to generate morphologically infeasible solutions, such as lumen pixels outside of the vessel, isolated pixel sets, and highly oscillatory pattern or peaky singularity on the boundary, as illustrated in Fig. 4.6.

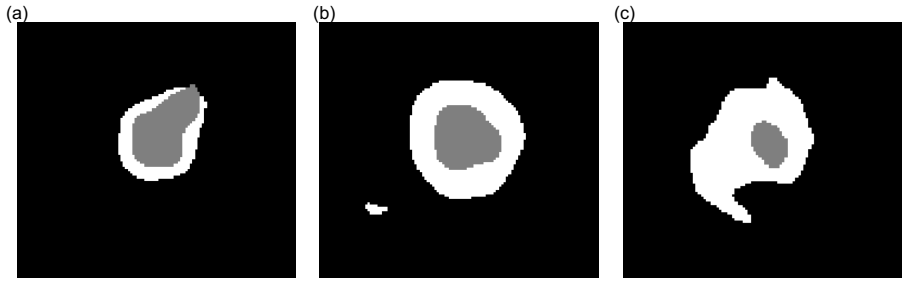


Figure 4.6: Morphologically infeasible examples of vessel wall segmentation generated by a naïve multi-label 2.5D UNet model: (a) lumen pixels outside of the vessel, (b) isolated pixel sets, and (c) highly oscillatory boundary.

Realizing the importance to account for inclusion morphology, Chen *et al.* proposed a carotid artery segmentation network in the polar coordinate system [102]. By converting

MR-VWI images into a polar coordinate system using an estimated lumen center as the reference origin, the segmentation problem became a regression task where the distances from the lumen center to the lumen boundary and the whole vessel boundary along each radial direction were predicted. The segmentation convolutional neural network (CNN) contained a fully connected (FCN) layer to predict the polar coordinates for the lumen boundary and whole vessel boundary in t sampled polar directions.

DSCs of 0.961 and 0.860 were reported for the lumen and the vessel wall, respectively, compared to 0.922 and 0.774 from the conventional Cartesian coordinate system. It was claimed that performing segmentation in the polar system has the advantages of: 1) ensuring contour continuity when enforcing the distance from the lumen center to the predicted whole vessel boundary to be larger than that of the lumen boundary, and 2) easily differentiating adjacent arteries from the artery to be segmented. A prerequisite for segmenting in the polar coordinate is a reliable definition of the centerline. A tracklet refinement algorithm was proposed for lumen center localization and centerline tracking to meet this requirement [102].

While this combination of centerline tracking and polar analysis for vessel wall may work well for the large carotid arteries, it is a lot more challenging to ensure a good automatic centerline for the much smaller intracranial vessels, whose signal and contrast strength could be low or disruptive even in angiography. In this study, we propose and develop a novel method to address the demand to account for topology inclusion with much relaxed requirement on "centerline" or origin definition.

4.5.2 Propose Vessel Inner-Outer Boundary Inclusion

We propose to account for the inclusion morphology with coupled level-set functions and using a deep neural network approach as the overall structure. In particular, we develop a network with a *single* output channel to infer the soft "tiered" memberships of the lumen, whole vessel, and background simultaneously, in sharp contrast to the typical multi-channel

predictions in multi-class or multi-label settings. Fidelity is defined based on class agreement between the "ground truth" labels and the prediction derived from the level-set as in Eq. (4.14). The training cost is further regularized with penalty Eq. (4.15) and Eq. (4.17) to encourage smoothness of the network predicted value function and the vessel wall boundaries, respectively. Fig. 4.7. illustrates the general schema of the proposed tiered method. We deploy a 2.5D UNet structure with ResNet backbone in our implementation.

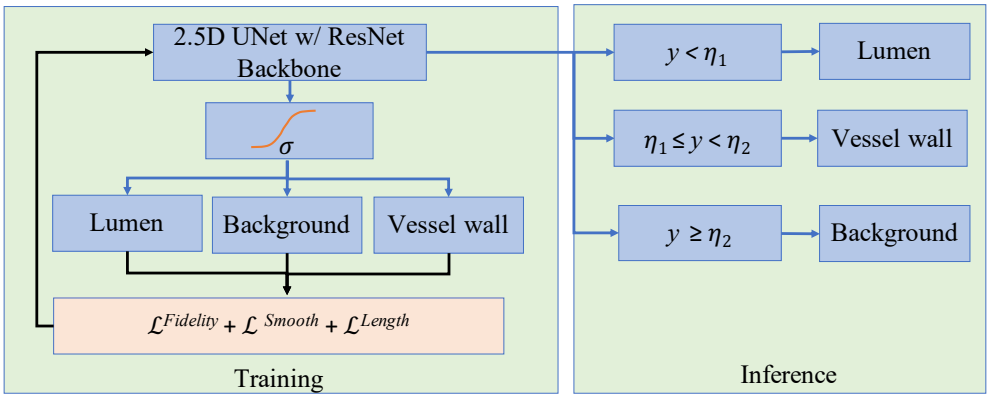


Figure 4.7: Schema of the proposed method to account for the inclusion between lumen and the whole vessel: the training objective is the weighted sum of three loss terms: the fidelity on soft Dice $\mathcal{L}^{\text{Fidelity}}$ as Eq. (4.14), the $l2$-norm of the network predicted value function gradient $\mathcal{L}^{\text{Smooth}}$ as Eq. (4.15), and the total variation-based length penalty $\mathcal{L}^{\text{Length}}$ as Eq. (4.17) on the inner and outer vessel wall boundaries; the inference process simply maps the network output y into the predicted classes according to its values in the tier system.

4.5.2.1 Level-set Formulation

To encode inclusion, we consider the ordinal relations among various level-sets with respect to a single level-set function. Under a 2D setting, let $\phi(x) : R^2 \rightarrow R$ be a level-set function, the lumen and the whole vessel pixels are associated with:

$$\begin{aligned} &\Omega_{\text{lumen}}\{x : \phi(x) < \eta_1\}, \\ &\Omega_{\text{whole\_vessel}}\{x : \phi(x) < \eta_2\}. \end{aligned} \tag{4.9}$$

We take advantage of the simple relation that for $\eta_1 < \eta_2$, $\Omega_{\text{lumen}} \subset \Omega_{\text{whole\_vessel}}$ reflects the inclusion relationship. In this specific application, we may define the background as the complement of the larger set $\Omega_{\text{background}} = D - \Omega_{\text{whole\_vessel}}$, where $D \in R^2$ denotes the entire segmentation domain. Without loss of generality, we may set $\eta_1 = \frac{1}{3}$, $\eta_2 = \frac{2}{3}$. Fig. 4.8. illustrates the level-set idea.

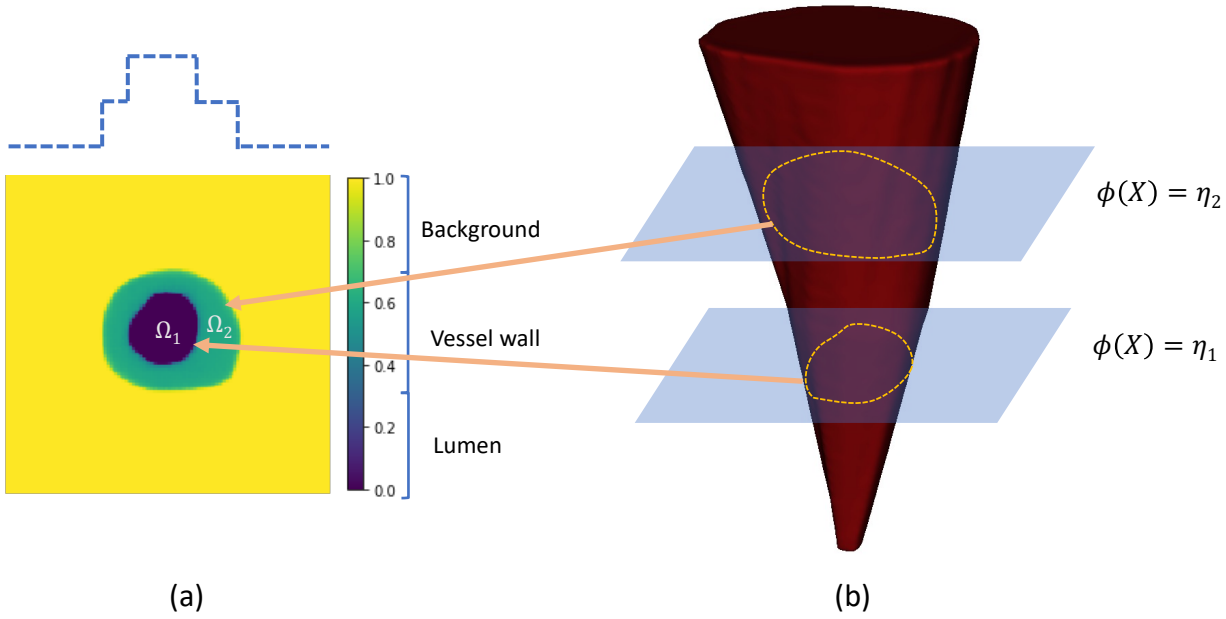


Figure 4.8: Illustration of level-set scheme: (a) is the output level-set map from the proposed segmentation neural network with class inclusion: Ω_1 denotes the lumen, Ω_2 is the vessel wall, and $D - (\Omega_1 \cup \Omega_2)$ is the background. The dashed blue line illustrates the change of level-set function height with a ray starts from the background and encounters the vessel wall and lumen subsequently and goes back to the background. (b) is the illustration of the level-set function of the whole vessel and the lumen.

The membership of a pixel $x \in R^2$ is obtained by taking the level-set function through a

Heaviside function H:

$$\begin{cases} H(\eta_1 - \phi(x)) = 1, \ x \in \Omega_{\text{lumen}} \\ H(\eta_2 - \phi(x)) \cdot H(\phi(x) - \eta_1) = 1, \ x \in \Omega_{\text{vessel\_wall}} \\ H(\phi(x) - \eta_2) = 1, \ x \in \Omega_{\text{background}}, \end{cases} \tag{4.10}$$

where

$$H(x) = \begin{cases} 1, \ x \geq 0 \\ 0, \ x < 0; \end{cases} \tag{4.11}$$

which is then relaxed to a continuous differential sigmoid function $S(x) = \frac{1}{1+e^{-x}}$ to generate a "soft membership" for each class given in Eq. (4.12).

The corresponding continuous probability-like relaxation of the network predicted value function y to y' is given by

$$\begin{cases} y'_{\text{lumen}} = S(\eta_1 - y), \\ y'_{\text{vessel\_wall}} = S(\eta_2 - y) \cdot S(y - \eta_1), \\ y'_{\text{background}} = S(y - \eta_2). \end{cases} \tag{4.12}$$

4.5.2.2 UNet with ResNet Backbone Structure

A 2.5D UNet model with ResNet backbone is used for level-set inference [9, 13], as demonstrated in Fig. 4.9. The convolution blocks in the UNet model each consists of one convolution layer followed by batch normalization, and another convolution layer. With a ResNet backbone, a skip-connection is inserted after the input of each convolution block and is passed through a 1×1 convolution to add the feature of the previous layer to the last layer of a convolution block. The network has a single channel output via a 1×1 convolution layer with sigmoid activation. This single-channel prediction maps each pixel's value to its corresponding class membership.

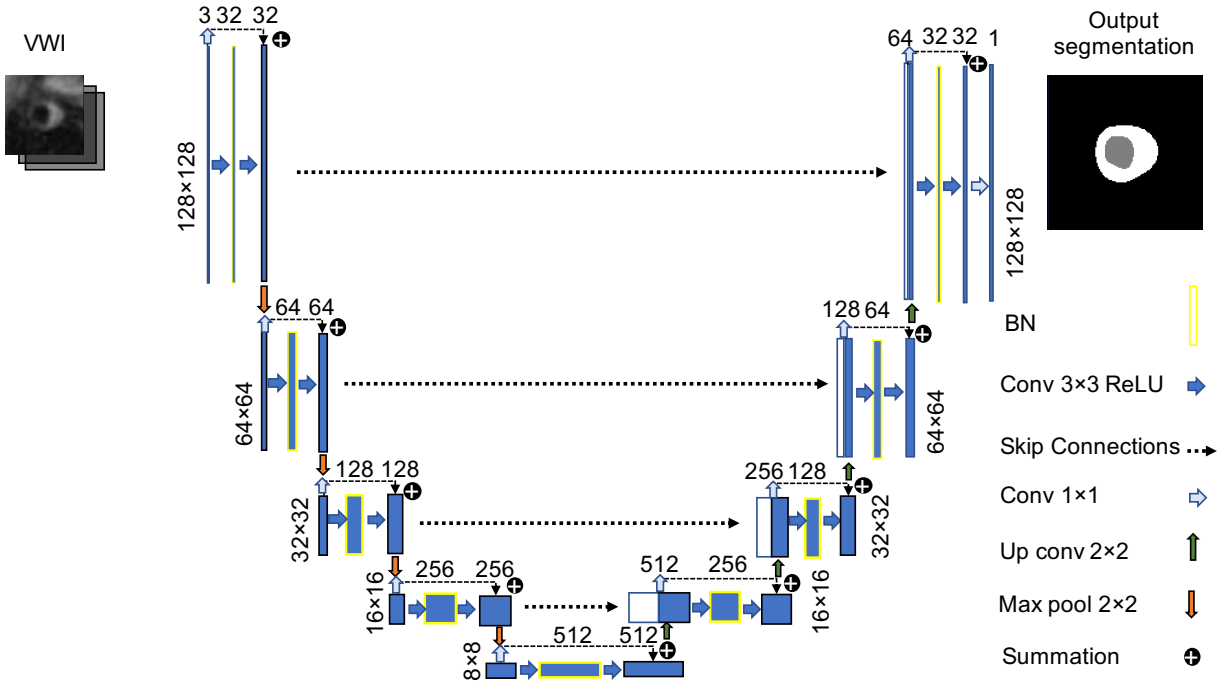


Figure 4.9: Illustration of the proposed tiered segmentation network structure: a skip–connection is inserted in each convolution block. Consecutive VWI slices are input to the network and a single-channel prediction of the background (black), lumen (gray), and the vessel wall (white) is output via sigmoid activation for the middle slice.

4.5.2.3 Training Objective

The deep neural network is trained to minimize an objective function consisting of three terms: a fidelity term to match the derived level-sets with the training labels and two regularization terms to encourage smoothness for the network predicted value function and the class boundaries, respectively.

The overall loss function is a summation of the three terms weighted by regularization hyperparameters λ and γ:

$$\mathcal{L} = \mathcal{L}^{\text{Fidelity}} + \lambda \mathcal{L}^{\text{Smooth}} + \gamma \mathcal{L}^{\text{Length}} \tag{4.13}$$

The fidelity term defines the agreement between the predicted and the given labels using soft Dice criterion for the lumen, vessel wall, and background classes.

$$\mathcal{L}^{\text{Fidelity}} = \sum_{c} (1 - \frac{1}{N} \sum_{n=1,\ldots,N} \frac{2 p_{n,c} y'_{n,c}}{p_{n,c}^2 + y_{n,c}'^2}), \tag{4.14}$$

where $y'_{n,c}$ and $p_{n,c}$ are the soft prediction from Eq. (4.12) and the "ground truth" labels for the nth pixel of class c = lumen, vessel wall, background, respectively. N is the total number of pixels in a batch.

To encourage clear and robust differentiation between the adjacent classes, i.e., lumen vs. vessel wall, and vessel wall vs. background, we introduce an l_2 norm to the gradient of the network output y to prevent oscillation and promote stable region-wise homogeneous membership.

$$\mathcal{L}^{\text{Smooth}} = \frac{1}{N} \sum ||\nabla y||^2, \tag{4.15}$$

where ∇ is the spatial differential operator:

$$\nabla y = (y_{i+1,j} - y_{i,j}, y_{i,j+1} - y_{i,j}), \tag{4.16}$$

with i and j indexing over the horizontal and vertical axes in 2D images.

Penalizing the magnitude of the gradient encourages smooth transitions in y, and has two important consequences (1) it leads to congruent connected labeled regions upon inference,

and (2) for any ray starting from the lumen, smooth membership transition ensures a good chance of encountering a decent-sized vessel wall class before entering the background class, as illustrated by the example profile view in Fig. 4.8 (a).

As in active contour approaches, we further impose a length penalty based on total variation (TV) on the vessel wall class to reduce the roughness of the inner and outer boundaries of the vessel wall [103, 104]:

$$\mathcal{L}^{\text{Length}} = \frac{1}{N} \sum ||\nabla y'_{\text{vessel\_wall}}||. \tag{4.17}$$

4.5.3 Assessment Criteria

4.5.3.1 Conventional measure

The primary goal of segmentation is label agreement and it is typical to measure segmentation performance by calculating DSC, HD_95, and MSD. DSC measures the globally overlapping degree, while HD and MSD measure the biggest and the averaged point-wise matching discrepancy, between the prediction and the ground truth, respectively.

4.5.3.2 Clinical measure

The clinically relevant quantification feature - the lumen and vessel wall area as well as NWI are also adopted as measures to match a common clinical practice. We report the MAE of the NWI, lumen area (A_{lumen}), and vessel wall area ($A_{\text{vessel\_wall}}$), where area is measured in pixels.

4.5.3.3 Geometric measure

To assess segmentation quality, the MAE of the inner and outer boundary length (L) of the vessel wall, as well as the mean error (ME) of the lumen area and vessel wall area are reported in pixels. We further propose two metrics to measure the geometric integrity. To

quantify the existence of isolated pixels as in Fig. 2.1.(b), connected component analysis is applied and the summed area of small islands (denoted as N_{Iso}) is reported in pixels. The numbers of violation of inclusion as in Fig. 2.1.(a) is measured using membership gradient:

$$N_{\text{V}} = \sum \mathbf{I}\{\nabla y^d > 1\}, \tag{4.18}$$

where $\mathbf{I}$ denotes the indicator function, and y^d is the categorized membership which has the value of 2 for lumen pixels, 1 for vessel wall, and 0 for background. N_V counts the amount of lumen pixels that directly connect to the background pixels.

The mean and the standard deviation of each measure above are reported for a test set. One-sided paired t-tests with $p < 0.05$ are applied between the measure achieved by each method in comparison and the best measure, for each class, where applicable.

4.5.4 Network Specifications and Comparison with Benchmark Methods

We randomly split the recorded 80 patients into 74 : 3 : 3 for training, validation, and testing, respectively. Again, each patient was associated with four segments, and each segment had 30 2D cross-sectional image slices. Metrics and statistics were calculated slice-wise. In our implementation of UNet with ResNet backbone, the depth of the UNet model was four, and the base number of channels was 32. The network took three consecutive slices as input, and output the class prediction for the middle slice. The learning rate for all segmentation models was 10^{-4} for a total of 50 epochs, with Adam optimizer and a batch size of 64. The regularization hyperparameter λ was 0.1 and γ was 0.5, all tuned with respect to the validation performance.

The proposed method was compared with the conventional multi-label segmentation [100], and the polar-coordinated segmentation methods [104], both qualitatively and quantitatively.

The benchmark multi-label method utilized the same 2.5D UNet structure with ResNet backbone as the proposed method, with three output channels representing the prediction of

the lumen, whole vessel, and background, respectively. The training objective was the sum of soft Dice across these three independent classes, regularized with length penalty for the lumen and the whole vessel classes, as in Eq. (4.17).

To compare with segmentation in polar coordination system [104], the images were first resampled to 256×256 from 128×128 with nearest neighbor interpolation before polar conversion, and the model predictions were eventually converted back to the Cartesian coordinates. The same segmentation network structure as the proposed method with $128 \times 128 \times 3$ input size was used, and an FCN with $2t = 256$ nodes was attached to the last layer of the UNet. Specifically, the prediction of the multi-label and the proposed tiered models was upsampled to 256×256 to maintain the same dimension as the results by the polar method and also to achieve a smoother segmentation.

The polar intersection over union (IoU) loss function was used for network training for the polar method as in Eq. (4.19) [105]. Manually extracted centerlines were used as the polar origin instead of the iteratively refined centerline as in [102] to alleviate the challenge of fully automated centerline tracking for small intracranial vessels. The samples whose labels cannot be polar converted were removed from the training set, and two examples of such samples are illustrated in the Results Section Fig. 4.11.

$$\text{Polar IoU Loss} = \log \frac{\sum_{i=1}^{t} \min(d, d')}{\sum_{i=1}^{t} \max(d, d')}, \tag{4.19}$$

where d and d' are the ground truth and the predicted coordinates in the polar system, respectively, along each of the t directions.

Ablation study was performed to assess the contribution of each component in the overall loss function of the proposed tiered method, where the weighting hyperparameters were the same as the proposed method.

4.5.5 Segmentation Results

4.5.5.1 Method Comparisons

Fig. 4.10. illustrates the qualitative results of the conventional multi-label method, the polar segmentation method, and the proposed tiered method. It can be observed that the proposed tiered method achieved smoother boundaries and better resemblance to the ground truth manual segmentation, compared to other methods. The method helped to alleviate the overestimation of the vessel wall area compared to the conventional multi-label segmentation method, and achieved a better preservation of morphology than the polar method when the segmented shapes deviated further from regular circles, shown from the column (d) where the segmentation resembled a union of two circles, despite an over-regulated vessel wall shaping compared to the conventional multi-label method.

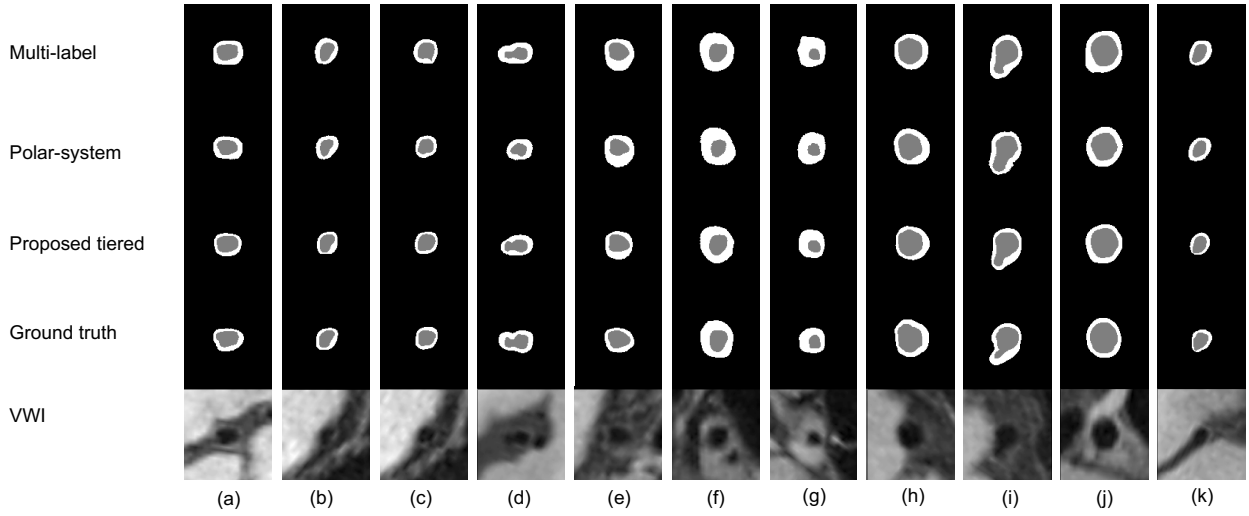


Figure 4.10: Qualitative visualization of the segmentation results by the proposed tiered segmentation method with class inclusion: each column is an example slice, and each row on the top panel corresponds to a different segmentation method corresponding to the cross–sectional vessel wall image on the bottom. The colors gray, white, and black indicate the lumen, vessel wall, and background, respectively.

Table 4.2: Proposed Tiered Method vs. Benchmarks in Conventional and Clinical Measures

Input/Class/Metric		DSC	HD_95 (mm)	MSD (mm)	A_MAE	NWI_MAE
Conventional multi-label segmentation	Lumen	0.924 ± 0.047	0.298 ± 0.477	0.087 ± 0.056	273.0 ± 357.5	0.065 ± 0.028*
	Vessel Wall	**0.794 ± 0.082**	0.394 ± 0.431*	0.119 ± 0.059*	600.5 ± 476.5*	
Polar-system segmentation	Lumen	0.893 ± 0.053*	0.643 ± 0.703*	0.233 ± 0.103*	466.8 ± 400.9*	0.077 ± 0.035*
	Vessel Wall	0.781 ± 0.079*	0.698 ± 0.609*	0.233 ± 0.070*	554.7 ± 460.7*	
Proposed tiered segmentation	Lumen	**0.925 ± 0.048**	**0.286 ± 0.436**	**0.083 ± 0.037**	**257.6 ± 325.9**	0.050 ± 0.015
	Vessel Wall	0.786 ± 0.084*	**0.345 ± 0.419**	**0.103 ± 0.032**	**490.6 ± 387.5**	

*significant under one-sided t-test with $p < 0.05$, and bold numbers denote the best measure for each class across methods. The image size is 256×256.

Table 4.2 reports the quantitative performance of the above methods for comparison. The tiered method generally achieved the best measure across metrics among all the methods. The significant reduction in the MAE of the NWI and areas indicates that the morphological improvement offered by the tiered approach has manifested favorably into quantitative clinical endpoints.

Table 4.3 reports the geometric integrity across the testing set. The results show that none of the method had the problem of lumen pixels directly connecting to the background as a violation of inclusion for our specific randomly selected test-set. However, the proposed tiered method achieved significantly less isolated pixels of vessel wall compared to the conventional multi-label method. The proposed method also achieved the smallest MAE of boundary lengths and alleviated the under-estimation of lumen area and the over-estimation of vessel wall area of the other two methods.

Fig. 4.11. shows two examples where the polar conversion encountered problems. These types of samples were removed from the training set, and their inference results were illustrated. Despite maintaining a good geometric integrity, the results were not very close in morphology to the "ground truth" segmentation and image cues, as the relatively complex morphology with tortuous boundaries was not seen during training.

Table 4.3: Proposed Tiered Method vs. Benchmarks in Geometric Measures

Input/Class/Metric		A	A_ME	L	L_MAE	N_{Iso}	N_{V}
Conventional multi-label segmentation	Lumen	2810 ± 1357	-138.2 ± 428.0	227.4 ± 55.13	13.15 ± 23.83*	4.015 ± 40.82*	0
	Vessel Wall	3361 ± 1098	454.5 ± 617.3	335.4 ± 59.99	18.96 ± 22.44		
Polar-system segmentation	Lumen	2490 ± 1263	-458.1 ± 410.8	233.8 ± 60.67	13.34 ±18.61*	**0**	0
	Vessel Wall	3274 ± 1169	367.5 ± 620.4	350.1 ± 67.37	29.18 ± 20.69*		
Proposed tiered segmentation	Lumen	2933 ± 1397	-15.45 ± 415.1	231.5 ± 55.17	**12.37 ± 23.26**	0.050 ± 0.873	0
	Vessel Wall	2627 ± 1035	-279.1 ± 559.4	318.8 ± 62.34	**15.32 ± 22.54**		

*significant under one-sided t-test with $p < 0.05$, and bold numbers denote the best measure for each class across methods. For the length measure L, lumen denotes the inner boundary, and vessel wall denotes the outer boundary for simplicity. The image size is 256×256.

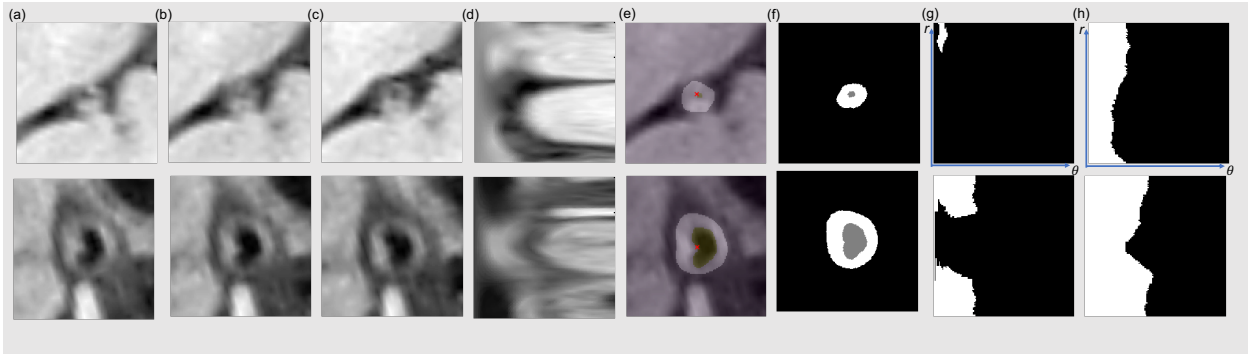


Figure 4.11: Two example slices in two rows where polar conversion is not applicable. (a),(b),(c): three consecutive VWI slices; (d): polar conversion of the middle slice (b); (e): ground truth lumen (yellow) and vessel wall (white) of (b), the red crossing shows the location of the image center (or polar origin); (f): the predicted labels by the polar method; (g): polar-converted ground truth lumen segmentation of (b); (h): polar-converted ground truth whole vessel segmentation of (b). The first example shows that when the lumen area is too small and the pre-detected lumen center (image center) is outside of the lumen area, the polar method encounters multiple intersections with the vertical axis. The second example shows that a non-convex shape leads to problems in polar conversion, as a line radiates from a detected lumen center can encounter multiple points on the segmentation boundary.

4.5.5.2 Ablation Studies

We compared the results obtained by using the loss function of 1) only the soft Dice loss $\mathcal{L}^{\text{Fidelity}}$, 2) soft Dice as the fidelity and the smooth loss term $\mathcal{L}^{\text{Smooth}}$, 3) soft Dice and the length penalty term $\mathcal{L}^{\text{Length}}$, and 4) the proposed soft Dice together with the smooth loss and length penalty. Fig. 4.12. illustrates the qualitative results of the ablation studies. The results show that the $\mathcal{L}^{\text{Smooth}}$ was essential for reducing holes in the segmentation and regularizing morphology. The $\mathcal{L}^{\text{Length}}$ term further regulated the morphology, and helped smooth out the segmentation boundaries and reduce small isolated pixel sets occur in the background. The quantitative results in Table 4.4 show that the proposed method generally achieved the best performance, and each term was critical for the method formulation.

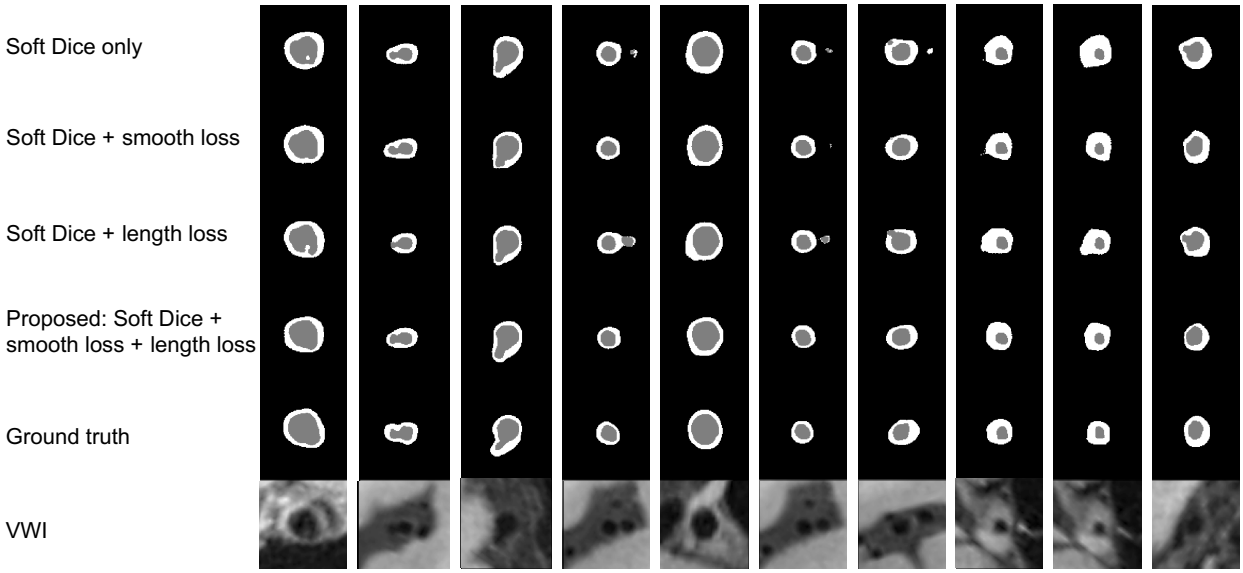


Figure 4.12: Ablation studies for the proposed tiered loss function: each column is an example slice, and each of the first four rows is a different loss function. The proposed method achieves the best and smoothest shaping compared to with other objective alternatives, and thus each term is critical to the proposed loss function. Gray is the lumen, white the vessel wall, and black the background.

Table 4.4: Proposed Tiered Method Ablation Studies

Input/Class/Metric		DSC	HD_95 (mm)	MSD (mm)	NWI_MAE
Soft Dice only	Lumen	0.924 ± 0.048	0.332 ± 0.539*	0.091 ± 0.062*	0.0651 ± 0.0260*
	Vessel Wall	0.791 ± 0.085	0.402 ± 0.454*	0.123 ± 0.062*	
Soft Dice + smooth loss	Lumen	0.925 ± 0.046	0.289 ± 0.440	0.085 ± 0.040	0.0504 ± 0.0163
	Vessel Wall	**0.795 ± 0.080**	0.359 ± 0.429*	0.106 ± 0.038*	
Soft Dice + length loss	Lumen	0.925 ± 0.047	0.327 ± 0.506*	0.096 ± 0.090*	0.0503 ± 0.0148
	Vessel Wall	0.793 ± 0.078	0.370 ± 0.426	0.109 ± 0.054*	
Soft Dice + smooth loss + length loss	Lumen	**0.925 ± 0.048**	**0.286 ± 0.436**	**0.083 ± 0.037**	**0.0498 ± 0.0146**
	Vessel Wall	0.786 ± 0.084*	**0.345 ± 0.419**	**0.103 ± 0.032**	

*significant under one-sided t-test with $p < 0.05$, and bold numbers denote the best measure for each class across all methods.

4.6 DISCUSSION AND CONCLUSIONS

4.6.1 UNet++ and HD Loss

UNet++ structure achieves a better performance than the UNet, with dense- and skip-connections offering more flexibility in the scale adaptation. Moreover, a weighted sum of the outputs from the deep supervisions further enhances the performance.

The addition of the HD loss term further helps reshape the class boundaries to better conform to the ground truth segmentation. It also appears to be a good surrogate objective to boost the NWI estimate. The selection of the trade-off hyperparameter λ needs to be handled with care, as the scale of the DC loss and HD loss is different. For simplicity, we assigned a fixed λ instead of a ratio between HD and DC loss, as suggested by Karimi et al., and managed to maintain the desired advantage [101].

4.6.2 NWI Oscillations and Estimation Discrepancy

UNet++ with soft DC and HD loss segmentation model is used for assessing NWI oscillations and estimation discrepancy. The NWI oscillations shown in Fig. 4.5 (a) and the zigzags of

the manual labeled vessel considered as ground truth segmentation in cross-sectional slices in Fig. 4.5 (b) suggest possible room for improvement in accuracy and consistency of the ground truth labels. The artificial oscillations may explain the moderate improvement of the 2.5D model over our previous 2D model, as the benefit of longitudinal smoothness from the former may not be properly represented in the current manual labels. This observation shows that another independent label or review is warranted and a spatial filter may be applied to improve the quality of the true labels.

The signed error between the prediction and the ground truth NWI corresponds to a 95% confidence interval of $[-0.03, 0.18]$, as shown in Fig. 4.13 (left). A one-sided paired t-test with $p = 0.05$ rejects the null hypothesis 11 out of the 12 testing vessel segments. The systematic over-estimation of the NWI is more prominent for relatively normal vessel segments, as shown in Fig. 4.13 (right). This was caused by a bias of the NWI distribution in the training set, as cases with extremely big and small NWI values are rarely found and incorporated in the training set. This may be alleviated by incorporating more "rare cases" in the training set or performing data augmentation. If needed, one may consider a bias correction scheme on the NWI during post-processing.

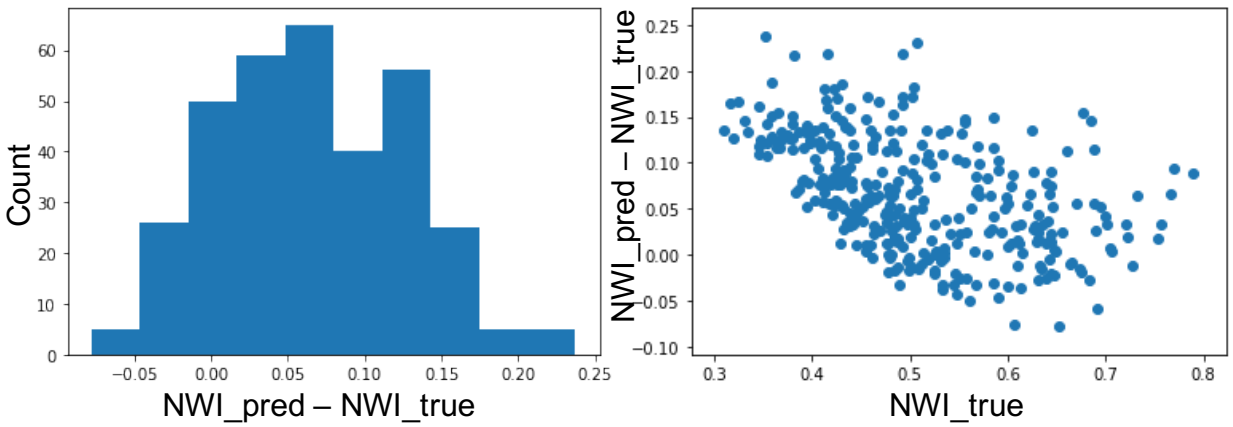


Figure 4.13: Histogram of NWI signed error (left) and signed error distribution (right) by using 2.5D UNet++ with soft DC and HD loss model

4.6.3 Proposed Tiered Method vs. Benchmark Method

With two regularization terms to encourage the smoothness of the membership transition and segmentation boundaries, the proposed tiered method achieved significantly better morphological feasibility than the conventional multi-label method without much compromising typical segmentation performance in DSC, HD, and MSD. The clinically relevant plaque assessment - NWI, A_{lumen}, and $A_{\text{vessel\_wall}}$ also enjoyed significant improvements, promising advantages in downstream clinical tasks.

In comparison, compromised segmentation performance was observed with the polar-system method [102], possibly caused by replacing the soft Dice loss with the polar IoU as the objective to maintain geometric integrity. Originally proposed to segment carotid arteries with larger sizes and more regular and circular/elliptical vessel shapes, the polar method was challenged with the finer and more torturous intracranial cases, which was the focus in our study. A related observation was that the finer structure in intracranial vessels also demanded more numerical stability during coordinate conversions: the polar method required meticulous definition of centerline or image center to maintain ray-wise convexity. While the training samples with out-of-lumen center or non-convex vessel shapes may be removed with additional adjudication and manual examination, it is impractical to remove such samples at inference time, which eventually gives rise to system breakdown or erroneous results. A derived benefit of our method's robustness in being compatible with all cases is the avoidance of selective removal so that the network can receive a broad exposure without artificial bias.

4.6.4 Class Inclusion with Level Set

It is worth noting that the geometric inclusion is not a consequence of the native level-set representation, where a multi-phase one uses either $n-1$ or $\log_2(n)$ level-set functions to represent n phases and allows each region to evolve [106]. When applied to vessel segmenta-

tion problems, these methods handle the inner and outer vessel wall boundaries separately without accounting for their relative placement [98, 107]. The logic addition and subtraction used in the composite multiphase may provide some adjacency constraint but is insufficient to reflect the enclosing vessel "ring" on lumen [106], unlike the proposed tiered level-set derived from a single value function.

In addition to the morphological benefits with tiered level-sets, the proposed method inherits efficiency advantage from the deep learning with fast inference, compared to typical level-set methods solved with iterations.

4.6.5 Conclusions and Future Development

As a preliminary study, a 2.5D UNet++ structure with a loss function composed of both soft Dice coefficient loss and Hausdorff distance loss is proposed, which yields universal improvements in various metrics for intracranial vessel segmentation. To particularly preserve the inclusion relationship between the lumen and the whole vessel, a novel and effective segmentation method based on deep neural networks is further proposed. The proposed method relates the classes intrinsically with a function whose value provides an ordinal indication for the tiered class membership, which has achieved better segmentation accuracy and morphology both qualitatively and quantitatively compared to benchmark methods. The proposed tiered method can be adopted to any applications that have similar inclusive settings between classes to generate morphological feasible segmentation solutions, and the improved morphology promises better evaluation support.

Further assessment with normalized wall index indicates that quantitative clinical endpoints may misalign with the common segmentation metrics despite their close association. Future work includes manual contour quality assurance and potential calibration schemes to use NWI quantitatively. The current tiered method requires tuning two hyperparameters for balancing regularization weights. We are actively investigating alternative regularization schemes to either simplify the design or learn the hyperparameters [108]. Furthermore, we

are working on extending the proposed tiered method to segment the entire vessel structure within the brain to further take advantage of the level-set's flexibility in handing topology transitions and coping with bifurcations.

CHAPTER 5

Deep Learning-guided Iterative Refinement to Improve Data Quality and Label Consistency

5.1 INTRODUCTION

Artificial intelligence has demonstrated great success in biomedical applications. In particular, when clinical insight is challenging to be described with a quantitative objective, deep learning techniques with supervision, either full, partial, or weak, have been developed to characterize and infer the intrinsic input-output relationship from training data. This is a common practice for a large set of problems, including but not limited to segmentation, detection, localization, recognition, classification and so forth. Manual labels are typically considered as the ground-truth during both model training and performance evaluation, resulting in a strong dependency of the overall accuracy and stability on the label quality. On the other hand, it is widely realized that the label quality itself could be subject to uncertainty.

Such uncertainty can be roughly categorized into two types. The first type is directly related to the quality and "accuracy" of the label, due to operator experience and training. We put the term "accuracy" in quotation mark because there is no absolute truth, which relates to the second category, and is termed as style uncertainty by us. In style uncertainty, the labels are all biomedically feasible and meet clinical needs. They can be considered perturbational samples due to personal style (inter-observer) or instantiation style (intra-observer), but all of them are correct and legitimate, from the quality perspective.

Realizing the important role of data, there has been extensive discussion about selectivity on big data, but mostly focusing on data reduction for relevance, compression, dimension reduction, and redundancy elimination in the general domain [109, 110]. In the centric development of biomedical applications, it is typically the case that data volume does not impose as big an infrastructure burden, but data quality is of the primary interest. Along this line, efforts have been made to refine labels using iterative generative adversarial networks (GAN) [111, 112], with a common idea to use the network for outlier detection or sample quality assessment, and then adjust the sample contribution in further network training or model development. This rationale is quite analogous to the noted simultaneous truth and performance level estimation (STAPLE) line of work [113], where the truth estimation module maps to the GAN development and the weight adjustment resembles that of performance level estimation.

In this work, we work on a quite different perspective by assuming that all input labels are of acceptable quality, as in the cases with most adjudicated clinical dataset, and that there is room of perturbation that would make such label reside within the range of feasibility or clinical acceptance. Instead of passively adjusting the contributing weight of certain subset of samples, we take a proactive approach to engage the physicians to truly incorporate clinical insight and maintain a tight connection between the adjustment scheme and clinical conformality.

We demonstrate our design with a use case of intracranial vessel wall segmentation task. It is a semantic segmentation task that aims to label the pixel-wise membership of the lumen and the vessel wall of the intracranial arteries based on magnetic resonance (MR) vessel wall images [80, 81]. The segmentation is highly challenged by the small size of the intracranial vessels. The limited signal contrast and conspicuity at the lumen and outer wall boundaries can greatly contribute to labeling variation between similar input images. Furthermore, radiologists typically use a software platform to generate such segmentation labels in a slice-by-slice manner, based on cross-sectional 2D MR vessel wall image slices, and the relations

between contiguous slices are only implicitly incorporated in the clinical contouring process.

Fig. 5.1 shows the variation of labeling on two adjacent slices from the same board-certified experienced radiologist, at two different labeling instances. While all these contours are clinically acceptable (even upon independent review by another radiologist), they present significant differences. In particular, the bottom row exhibits a smoother cross-slice behavior than the top row. These differences may have an impact on a semantic segmentation network learned from such data, and ultimately manifest into clinical quantifications, e.g., in terms of normalized vessel wall index that characterizes the plaque burden and correlates to stenosis detection or identification.

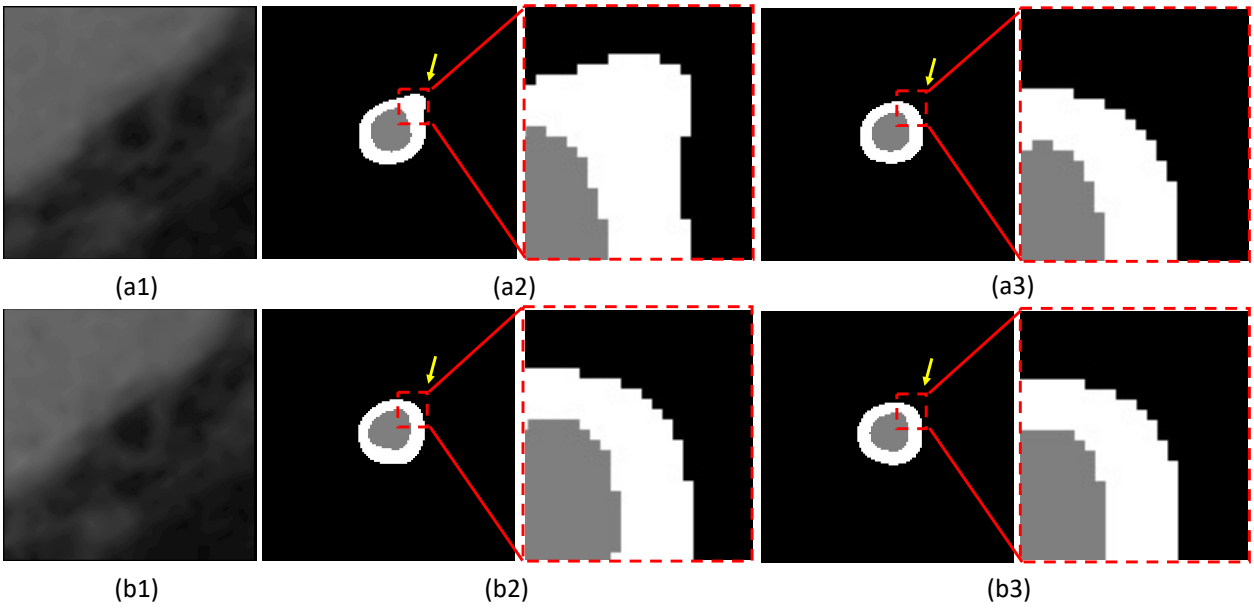


Figure 5.1: Manual labels of lumen (gray) and the vessel wall (white) of two cross-sectional vessel wall images. (a1) and (b1) are adjacent image slices with 0.55 mm in-plane distance; (a2) and (a3), (b2) and (b3) are two plausible contour solutions depicted by a radiologist across two labeling times of the corresponding images in the same row, with the associated zoom-in looks of the largest contour discrepancies.

Admitting the feasibility of all variations as "ground-truth", we hypothesize that prefer-

ence should be given to the specific realizations that support a lower dimensional model, as a cohort. This motivates the iterative procedure depicted in Fig. 5.2.

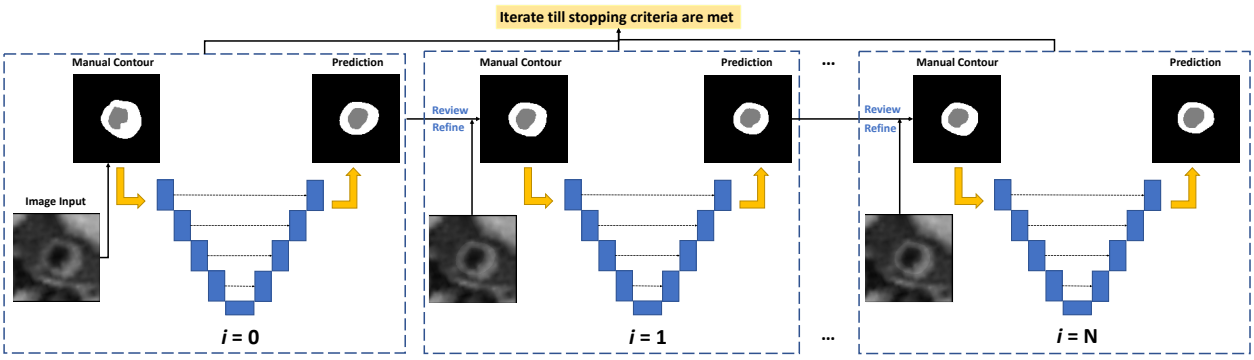


Figure 5.2: Iterative label refinement schema. For iteration i, the radiologist aligns the network prediction with the axial-view vessel image to review and refine the contours by accepting it or applying necessary modifications. The iteration ends when a proposed stopping criterion based on equivalence tests is met. The equivalence tests apply to the DSC calculated with the lumen and the vessel wall prediction compared to the corresponding round of "ground truth", between two neighboring rounds of refinement.

We utilize the guidance from deep neural networks to prompt and guide physician review and refinement. Upon such adjustment, a transfer learning scheme is used to quickly adapt the network parameters to the new label data. This process continues until a stopping criterion is met.

We show that the iterative refinement process improves the end-to-end consistency and the quality of the labeling practice by 1) enhancing the smoothness of contour boundaries along a vessel segment, 2) boosting the segmentation consistency between similar-intensity-distributed input images, and 3) increasing the agreement between human labels and network prediction, which would lead to a more stable down-stream clinical quantification based on the contours.

5.2 USE CASE DATA DESCRIPTION AND BACKGROUND

Under IRB approval, we obtained MR vessel wall imaging (VWI) from 80 patients with diagnosed intracranial atherosclerotic disease. All imaging data were acquired on a 3-Tesla system (MAGNETOM Prisma; Siemens Healthcare, Erlangen, Germany) equipped with a 64-channel head/neck coil using a whole-brain VWI protocol [92]. Four vessel segments with a high likelihood of plaque presence were included for each patient: the intracranial internal carotid artery, the middle cerebral artery, the intracranial vertebral artery, and the basilar artery. Each segment contained 30 contiguous 2D cross-sectional slices with 0.55 mm slice thickness and 0.10 mm in-plane resolution generated in 3D Slicer [93]. The manual lumen and vessel wall contours were labeled and refined by an experienced radiologist using ITK-SNAP [94].

In previous studies, various network structures and cost functions were proposed to perform the automatic vessel wall segmentation under supervised setting. Despite improved morphological feasibility, the segmentation accuracy measured in Dice similarity coefficient (DSC) and mean surface distance (MSD) between the network prediction and the provided labels only exhibited moderate improvement.

5.3 METHODS

Our logic of considering the uncertainty or non-uniqueness of the ground-truth was motivated in part by probabilistic networks where a distribution of output is generated instead of a point estimate. On the other hand, the integration into a clinical operation makes it desirable to have a definitive output. The statistical rationale and clinical pragmatism had driven us to consider a perturbational approach where preference is given to samples who are more likely to be of high quality. Since there is no absolute ground truth, we use consistency with an underlying parsimonious model as a surrogate for such quality measure.

Our overall procedure can be interpreted as a block-descent scheme for solving the joint optimization problem for both the set of labeling fields and the network model: the label estimate is encouraged as conformal as possible to a lower-dimensional network model, under the constraint of clinical acceptance criteria.

5.3.1 Low-complexity Deep Segmentation Network

A 2.5D UNet with ResNet backbone structure is adopted as the segmentation network [13]. The UNet structure has a contractive path that captures context, and a symmetric expanding path that enables localization [9]. Each convolution block consists of two convolution layers, where a batch normalization (BN) layer is inserted after the first convolution layer. In each block the feature learned by the first convolution layer is added to the feature of the last convolution layer before going to a Rectified Linear Unit (ReLU) activation, to incorporate the previous information by skip-connections. The network structure is illustrated in Fig. 5.3.

The 2.5D network takes a stack of consecutive VWI slices, and uses multi-label setting that outputs the odds of the background, lumen, and the whole vessel for the middle slice in three output channels with sigmoid activation. The vessel wall class is obtained by subtracting the whole vessel by the lumen during inference time. The training objective consists of the average soft Dice loss of the background, lumen, whole vessel, as well as the subtracted vessel wall class [100].

5.3.2 Iterative Refinement Process

The refinement process alternates between the network training based on the most recently modified label data, and the network-guided label review and refinement. This iterative process continues until a stopping criterion is met. In our particular use case, we defined the stopping criterion based on statistical equivalence between the network predictive per-

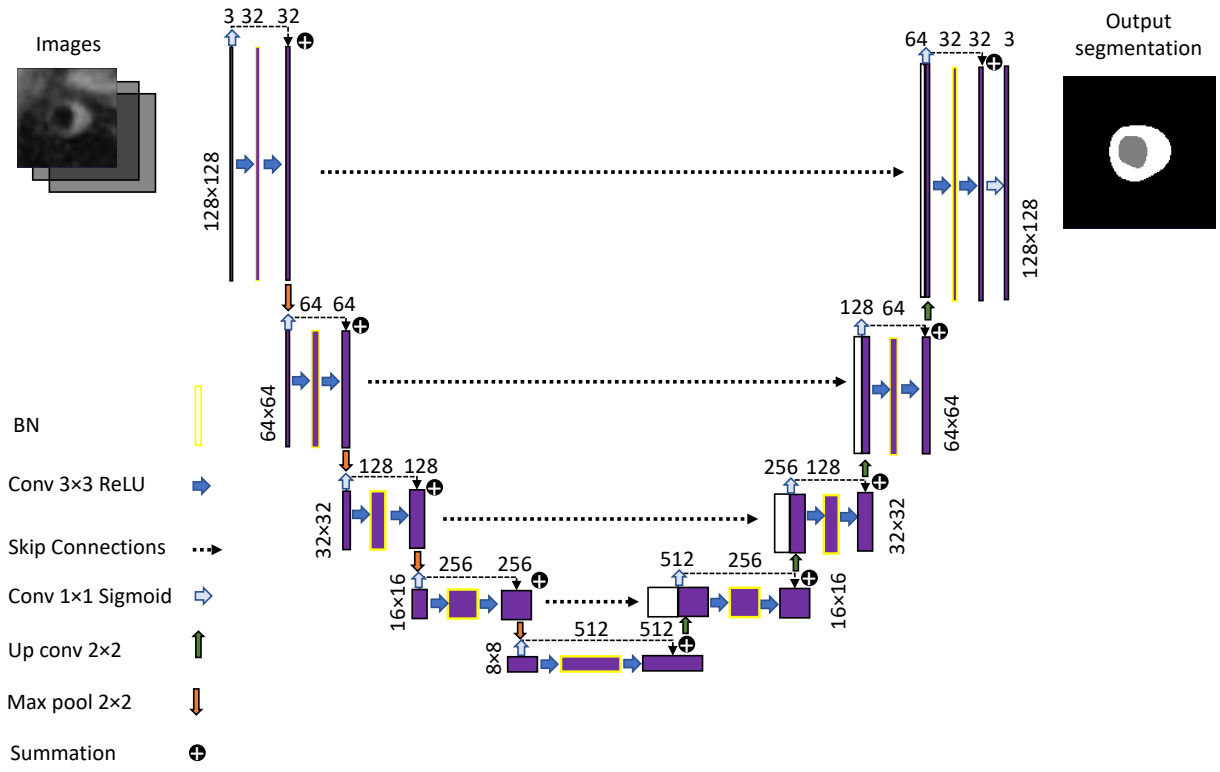


Figure 5.3: 2.5D UNet structure with ResNet backbone as the segmentation network for iterative refinement process realization. The input takes a stack of three consecutive image slices, and predicts the lumen (gray), whole vessel (white), and the background (black) in three output channels with sigmoid activation in the final layer.

formance in DSC between two consecutive refinement rounds.

Initially, the segmentation network is trained by the original contours labeled by a radiologist based on 2D cross-sectional VWI images only. In the subsequent each round of contour refinement, the network predicted labels are overlayed with the corresponding VWI slices. The predictions are reviewed by the radiologist by visual feedback for either accepting the contours or applying necessary contour adaptions where the network makes intolerable mistakes. The refined contours are then used for the next round of network training. In each round, the network prediction is assessed with the corresponding round of manual contours. The refinement process terminates when statistical equivalence is reached between the current and the previous iteration, suggesting insignificant differences between the two iterations.

5.3.3 Performance Evaluation Criteria

2D DSC and MSD from the prediction to the ground truth are adopted as measures for segmentation accuracy. A clinically relevant measure – normalized wall index (NWI) [83], computed as the vessel wall area divided by the whole vessel area, is further incorporated as a measure to reflect a typical down-stream clinical assessment. NWI ranges from 0 to 1, with a higher value indicating a more severe plaque burden.

5.3.3.1 Compliance with the lower dimensional network model

The compliance with lower dimensional network is measured by DSC and MSD. DSC measures the overlapping degree between the prediction and the ground truth, while MSD measures the average distance discrepancy from the prediction to the ground truth, in mm.

5.3.3.2 Piecewise smoothness along segment

The NWI, as a clinical measure of plaque level, is expected to be piecewise smooth, with abrupt changes upon transition from moderate to severe plaque burden. To this end, we use total variation (TV) index to quantify piecewise smoothness along segment as

$$TV = \sum_{i} |NWI_{i+1} - NWI_{i}|, \tag{5.1}$$

with i indexing over the axial slices of a segment.

5.3.3.3 Consistency across similar input samples

Cross-sample consistency indicates that more similar input should correspond to more similar output. To this end, we adopt a graph characterization approach, as illustrated in Fig. 5.4. To measure the degree of "similar" between two input images, mutual information (MI) is adopted, and the similarity within the output group is measured by the DSC of the predicted classes across the slices in a segment. Both MI and DSC range from 0 to 1, with 1 as the largest similarity. To evaluate the capability that only "similar" input images may yield "similar" output contours, we presented the input and output groups as undirected weighted graphs with the same node correspondence [114].

For a randomly selected test set with m samples, MI and DSC are calculated on each pair of input images and each pair of predicted contours, respectively, to form m×m adjacency matrices. MI is defined as:

$$MI\left(X,Y\right) = H\left(X,Y\right) - H\left(X|Y\right) - H\left(Y|X\right), \tag{5.2}$$

where

$$H(X) = -\sum_{i=1}^{n} P\left(x_i\right) logP\left(x_i\right) \tag{5.3}$$

is the entropy, given a discrete random variable X with possible outcomes $x_i, i = 1, 2, ..., n$, occurring with a corresponding probability $P(x_i)$. $H(X,Y)$ and $H(X|Y)$ are joint entropy

and conditional entropy of random variables X and Y, respectively. Under DSC and MI definitions, the adjacency matrices are both symmetric.

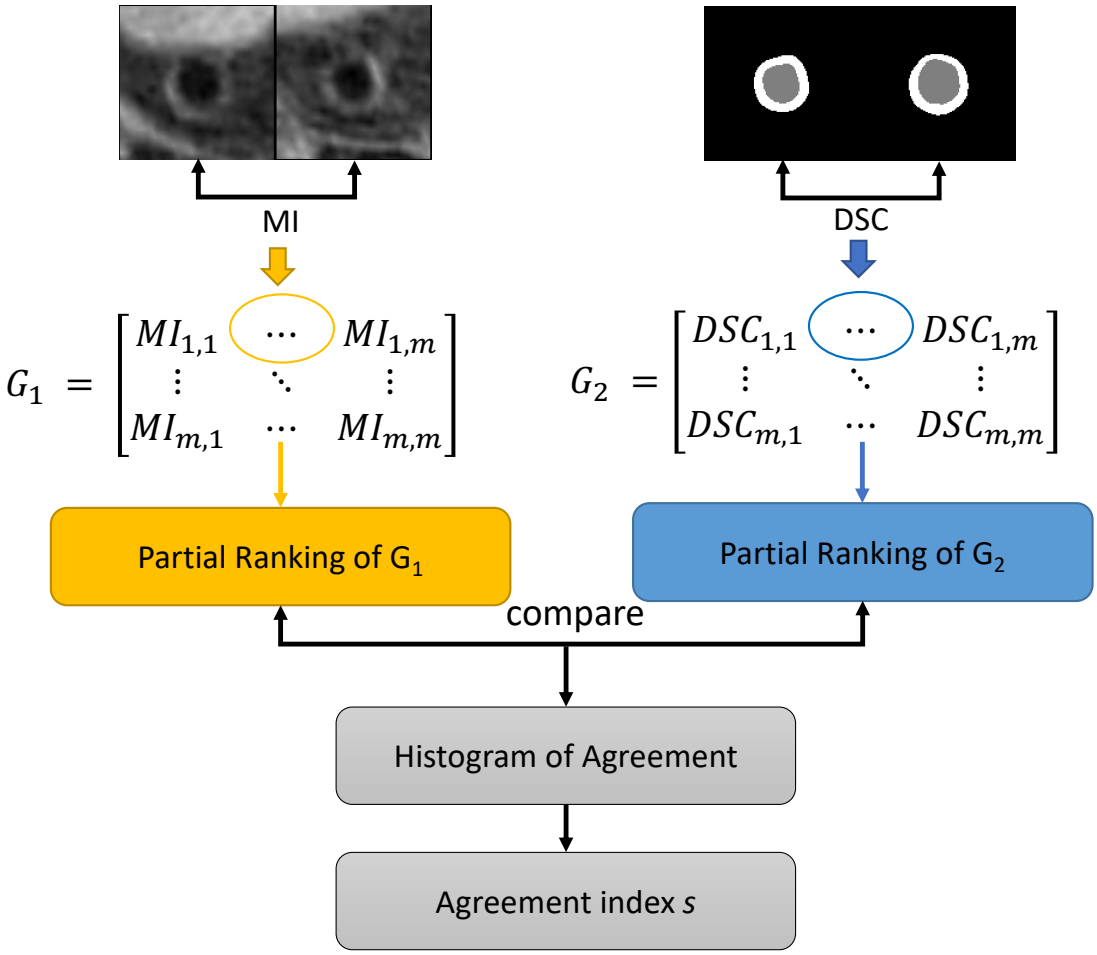


Figure 5.4: Evaluation of agreement between input images and contours for the proposed iterative refinement process

Graph similarity between the adjacency G_1 derived from intensity MI and G_2 derived from contour similarity is then measured by agreement index in Eq. (5.4),

$$s\left(G_1, G_2\right) = \frac{\sum_{i=1}^{m} \sum_{j=i+1}^{m} \mathbb{I}\left(G_{1i} > G_{1j},\ G_{2i} > G_{2j}\right) + \mathbb{I}\left(G_{1i} \leq G_{1j},\ G_{2i} \leq G_{2j}\right)}{\binom{m}{2}}, \tag{5.4}$$

where $\mathbb{I}$ is the indicator function, i and j are element indices of the adjacency matrices, and $\binom{m}{2} = \frac{m(m-1)}{2}$ is the total number of distinct pairs that are formed by the m test examples. The agreement index between two graphs increases with more consistent ranking.

5.4 METHOD SPECIFICATIONS

The 2.5D UNet took three consecutive image slices as the input stack and the size of the input images and output contours was 128×128. The base number of channels was 32, and the network depth was four. The network was trained with learning rate 0.0001 over 50 epochs on GPU GTX 1080 Ti, and the code implementation was with Tensorflow 2. Adam optimizer and a batch size of 32 were used.

The number of bins in calculating entropy in Eq. (5.2) was 50, and the number of samples in the test set for evaluating network agreement index in Eq. (5.4) was $m = 200$. We chose the number of stopping refinement iteration based on equivalence tests of lumen and vessel wall DSC with $(0, 0.03)$ equivalence interval.

To refine the contour and assess the network performance of each axial slice of the whole patient cohort, a five-fold cross-validation study was performed, where each fold included 64 subjects for training and 16 subjects for testing. Again, each subject was associated with four vessel segments, and each segment had 30 consecutive axial MR slices. Measures and statistics were calculated slice-wise. The network hyperparameters such as the number of training epochs, learning rate, and the network depth, etc. were optimized with respect to the validation results of our previous study in Chapter 4, and were kept fixed during the cross-validation process.

5.5 RESULTS

5.5.1 Round-specific Compliance with Low-dimensional Model

Fig. 5.5 reports the DSC and MSD of the lumen and the vessel wall classes vs. the number of refinement iterations. Predictive power of the same 2.5D network improved with both rounds. The increased DSC and reduced MSD indicate better agreement was reached between the manual and the predicted contours with iterative refinement process.

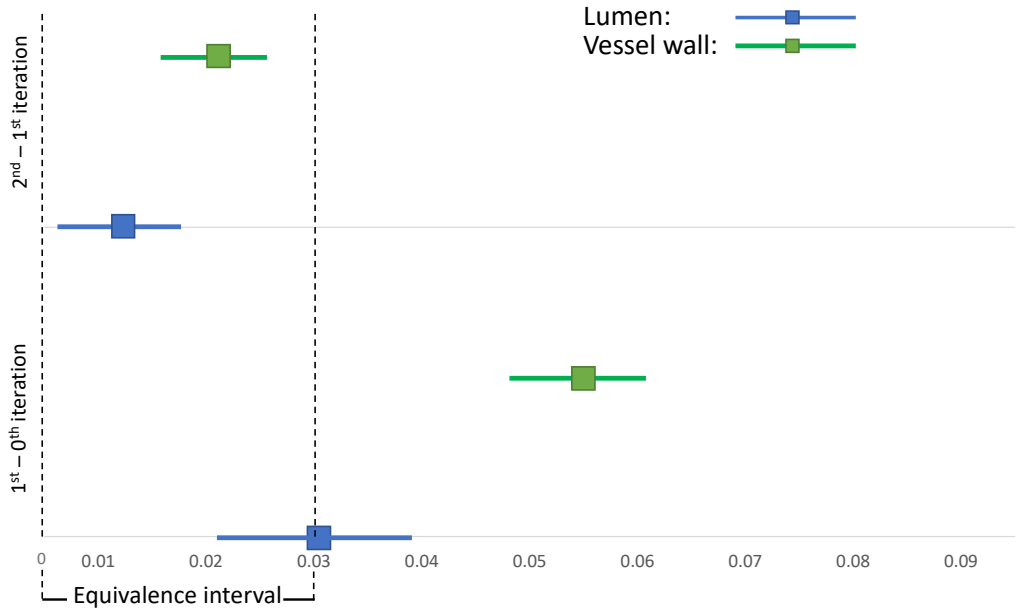


Figure 5.5: Lumen and vessel wall 2D DSC and MSD vs. the number of refinement iterations. 0 on the horizontal axis denotes the initial contour without refinement, 1 and 2 denote the first and second round of refinement, respectively.

Fig. 5.6 reports the tests of statistical difference and equivalence between the 0th and 1st refinement iteration and those between the 1st and 2nd iteration. Under significance level of 0.01, while statistically significant differences can be seen from both rounds, it can

be observed that the incremental improvement tends to diminish at latter rounds. With equivalence interval set at $(0, 0.03)$ for both the lumen and the vessel wall, we see that the change induced by the second round of refinement is not extreme enough to be of interest and it triggered our stopping criterion.

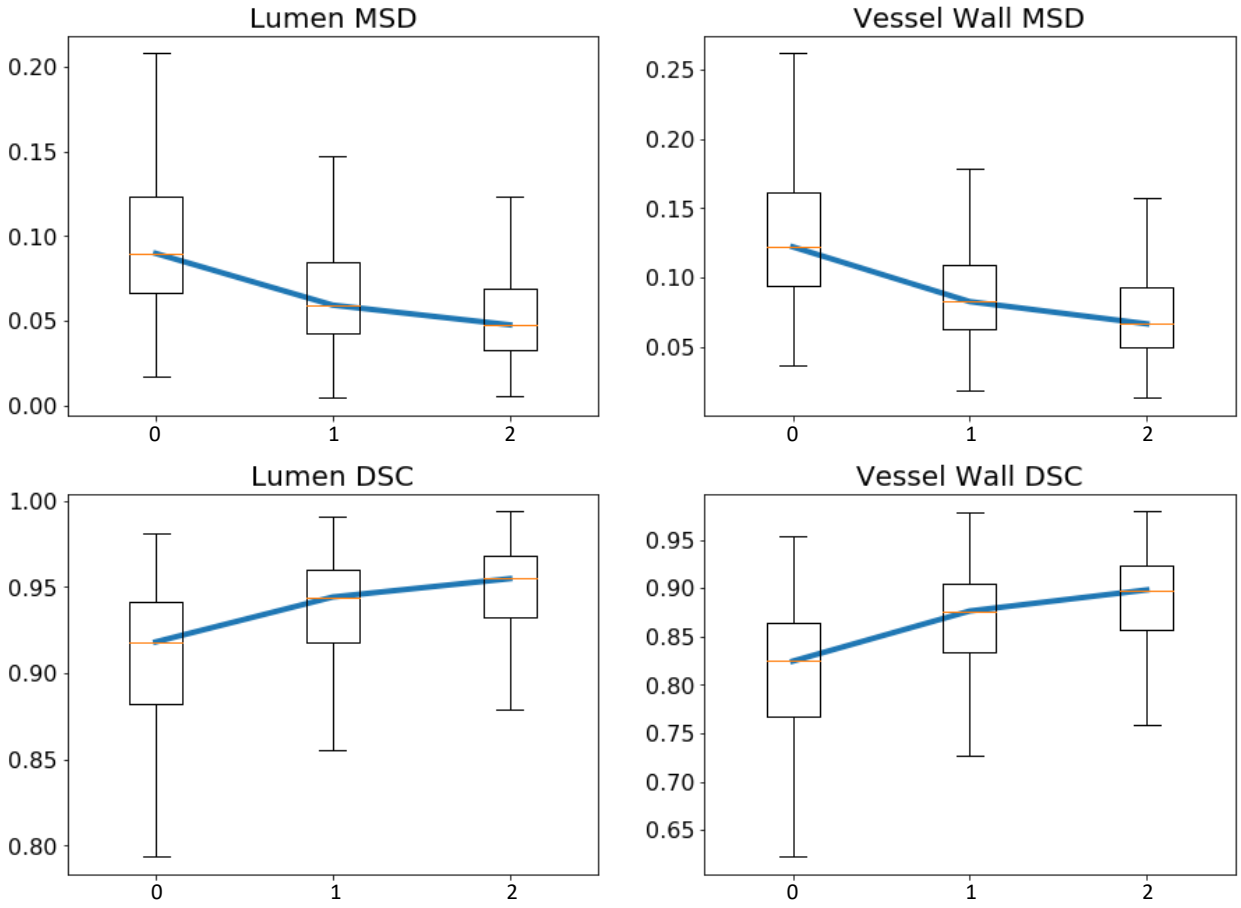


Figure 5.6: Lumen and vessel wall equivalence tests as iterative refinement stopping criteria. The solid lines denote 99% confidence interval, the blocks denote mean difference, and the dashed line is the equivalence interval.

Fig. 5.7 illustrates examples of the manual and network predicted segmentation as the iterative refinement proceeds. Columns (a) and (b) show examples where the problem of missing lumen was successfully fixed after the last round of refinement. This shows that as

the training data become more consistent and coherent, the same low-dimensional model may manage to allocate its representation power to the necessary variations. Columns (c) and (d) display two cases with vessel bifurcations, where the manual contours had significant variations from different instances/rounds. Column (c) is an uncertain transient bifurcation benefited from the review and refinement, where a contour of either elongated or circular shape was clinically acceptable. On the other hand, the network prediction in (d) persisted to disagree with the manual contours, but an improvement of vessel morphology in the last round was seen compared to the initial prediction. Columns (e)-(i) show that the manual contours and network predictions improve to agree with each other as the iterative refinement proceeds, which indirectly indicates that the labeling consistency had improved, and it resulted in better compatibility with a parsimonious model

5.5.2 Piece-wise Smoothness along Segment

Table 5.1 summarizes the quantitative results for each refinement iteration. From the table, TV of NWI of the manual and predicted contours measured in Eq. (5.1) was both significantly reduced from the 0th to the 2nd refinement iteration, under paired t-tests with $p < 0.01$.

Table 5.1: Quantitative Evaluation for Each Refinement Iteration

	DSC_Lumen	**DSC_VW**	**MSD_Lumen**	**MSD_VW**	**TVnwi_Manual**	**TVnwi_Pred**	***s*_Manual**	***s*_Pred**
0	0.893 ± 0.108	0.806 ± 0.086	0.104 ± 0.071	0.137 ± 0.067	1.022 ± 0.245	0.757 ± 0.181	0.519	0.523
1	0.924 ± 0.090*	0.860 ± 0.070*	0.073 ± 0.074*	0.095 ± 0.058*	0.839 ± 0.237*	0.643 ± 0.183*	0.521*	0.543*
2	0.938 ± 0.078*	0.879 ± 0.072*	0.058 ± 0.056*	0.080 ± 0.055*	0.763 ± 0.242*	0.586 ± 0.182*	0.522	0.556*

Each row is a refinement iteration. VW denotes the vessel wall, and s denotes the network agreement index measured in Eq. (5.4). Manual denotes the manual contour and pred is the predicted contour. *Results are significant with paired t-tests under $p < 0.01$ between 1st and 0th iteration, and between 2nd and 1st iteration.

An example visualization of the evolution of segmentation contours and the corresponding NWI is shown in Fig. 5.8 and Fig. 5.9. It demonstrates that both the predicted contours

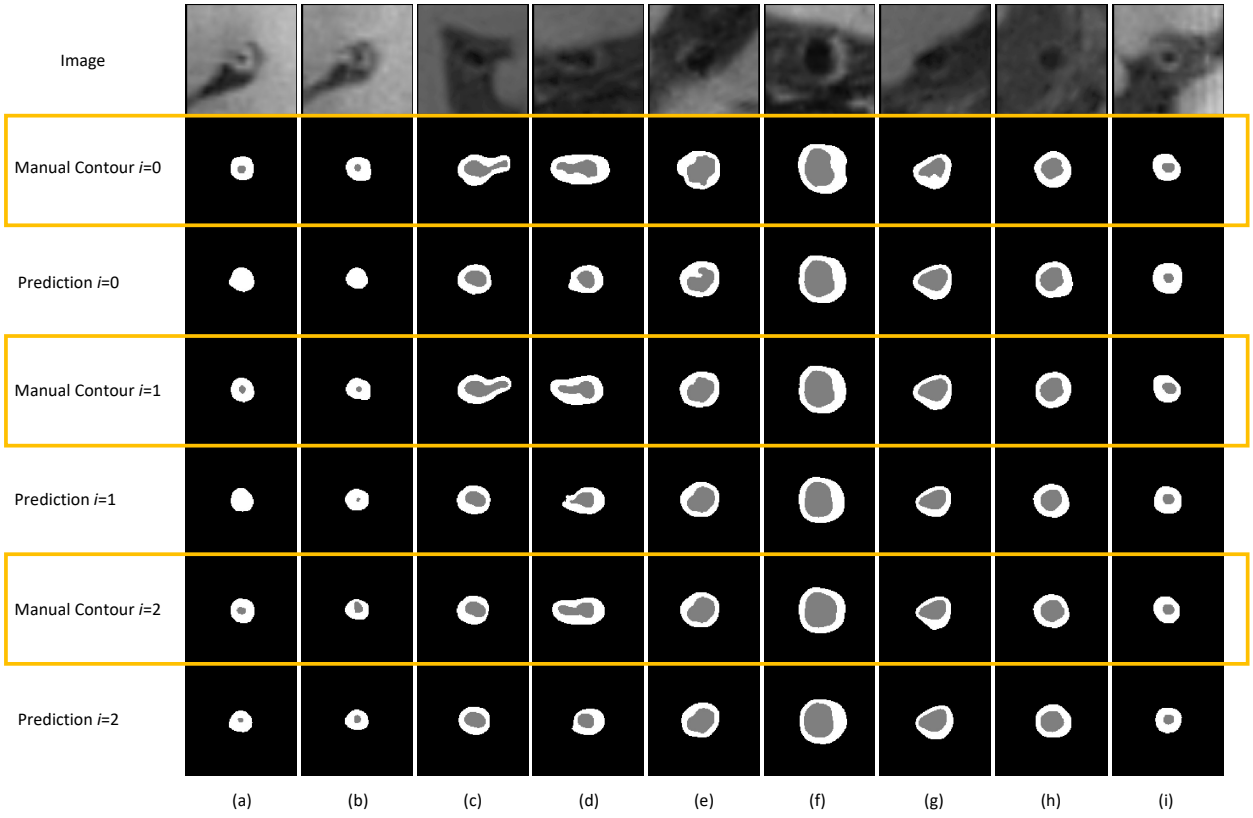


Figure 5.7: Illustration of the lumen (gray) and the vessel wall (white) contours of the manual and network predicted segmentation for each refinement iteration. Each column is an example slice. For visualization, the highlighted yellow boxes are the manual labels.

and the post-refinement manual contours exhibit higher level of smoothness than the manual contours at the initial round.

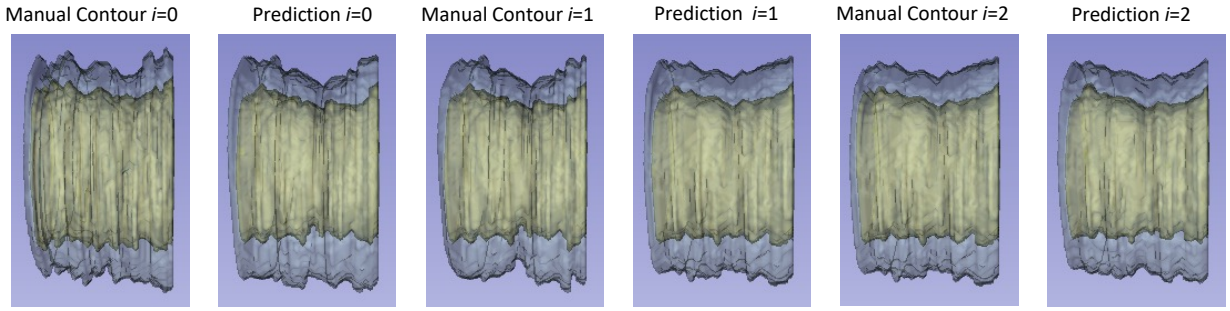


Figure 5.8: Contour smoothness illustration of the manual and predicted segmentation from each refinement iteration: lumen (inner yellow) and the vessel wall (outer transparent blue) are from a randomly selected vessel segment consisting of 28 slices.

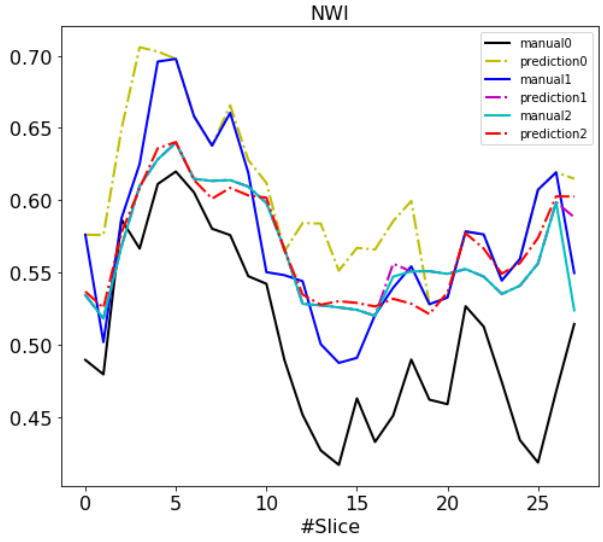


Figure 5.9: NWI of the manual (denoted as manual) and network predicted (denoted as prediction) labels of a randomly selected vessel segment consisting of 28 slices, for each refinement iteration

As a side observation, smoother contouring boundaries and NWI curves along the segment can correspond to more realistic and feasible morphology and may better support the

subsequent clinical quantification.

5.5.3 Consistency across Similar Input Samples

The agreement index s was 0.519, 0.521, 0.522 for the 0th, 1st, and 2nd round of refinement, respectively, for the human label; and 0.523, 0.543, and 0.556, respectively, for the network prediction. It shows an improving manner of the network input-output consistency with label refinement.

5.6 DISCUSSION AND CONCLUSIONS

5.6.1 Observations from Radiologist Review

Our radiologist reported 70% vs. 30% slices for accepting vs. revising in the 1st refinement round and 90% vs. 10% in the 2nd round, which took 60 hours and 30 hours, respectively, for a total of 10K 2D image slices. While complete reviewing is demanding and labor intensive, it serves our purpose to have a pilot study to maintain close faithfulness to clinical judgement and obtain insight from a nominal proportion to be examined and refined. In the real application scenario, one can restrict the attention to a subset of manual labels with high discrepancy from network prediction. With a low dimension network, there is minimal risk of overfit.

Some common mistakes made by the segmentation network were a) failure to identify the lumen for slices with extremely stenotic lumen, b) partial lesion identification in slices with exceptionally eccentric thickening of the vessel wall, c) limited representation power with sparse bifurcation cases, and d) erroneous inclusion of nearby brain parenchyma as part of the vessel wall when there is no clear cerebrospinal fluid (CSF) signal to provide separating contrast.

The scenarios (a-c) are results from sparse or imbalanced occurrence, as both are rare

events in training. Data sampling techniques can be incorporated to address this issue. The scenarios (d) may be resulted from the inconsistent human labeling based on low-contrast structure boundaries. Physicians could have a more comprehensive contouring and reasoning process based on auxiliary information such as patient-specific risk profile and stenosis condition, which are currently inaccessible to the neural network. Incorporation of shape prior and conditioning may be possible directions for improvements [115, 116].

5.6.2 Conclusions and Future Development

In this study, we have demonstrated that human label variation, even when small and clinically acceptable, may affect the prediction performance of a neural network and manifest into subsequent analytics. Using consistency to a parsimonious network model as the quality surrogate, we have proposed a novel and simple iterative refinement scheme to systematically perform guided perturbational refinement of ground truth labels. We have demonstrated that conformality to network model, consistency in adjacent slices, and consistency across different samples are significantly improved quantitatively and qualitatively. Further investigations are planned to improve rigor and efficiency of the stopping criterion, and to incorporate metric learning and attention to label consistency assessment.

While we are confident about the overall rationale and design, there are a few modules that warrant problem-specific tuning and investigation. Currently the stopping criterion is defined with respect to the equivalence task to address the question of "whether the improvement or modification is large enough to be clinically interesting". The equivalence interval selection $(0, 0.03)$ is quite arbitrary in the current stage and would desire more rigor. In addition, the stopping is claimed in a retrospective fashion once a non-interesting improvement has occurred. Given the amount of effort required to perform a refinement round, it would be highly desirable if such decision can be made in a prospective fashion. A performance predictor or an application-context specific hard stopping for the terminal round would be very helpful.

To measure network mapping consistency, the current approach assesses graph agreement after establishing two graphs, based on MI measures for images pairs and DSC measures for label pairs. While this is a reasonable first approach, it has much room for improvement. It is known that label adjacency and image adjacency are related non-trivially, and there are many options to establish or assess the existence of isomorphism for consistency. A metric learning approach could be used [117]. There is also a gap that exists between clinical assessment of relevance which has clear local attention, in contrast to the global full domain-based MI. The introduction of a self-attention scheme could be helpful [118].

CHAPTER 6

Discussions

Our current text processing work is limited to text de-identification. As a matter of fact, we have also devoted time to text understanding and feature visualization. By using RNN in a binary classification task, the textual features contributing the most to the classification would be the driving force of the problem formulation and warrants further investigation. Before performing classification by neural networks, the proposed preprocessing method can be applied to remove private information. Besides, text style transfer methods can also be applied to unify the writing styles, focusing the classification on nontrivial contents.

The pelvic organ segmentation project is currently limited to pelvic organs and we are in the process of pushing its application to the segmentation of the treatment target: the prostate bed. The segmentation of the prostate bed needs neighboring rectum and bladder organs as location support, as the structure itself is "invisible" and not based on image cues. Therefore, we take a sequential logic to segment the neighboring organs first and then the prostate bed. The contouring height and range of the prostate bed also depends on the existence of seminal vesicles structures and surgical clips, where the seminal vesicles should be localized by MR modality. Deep learning -based segmentation of the target volumes with injected prior would benefit in reducing the prominent observer inconsistency in labeling the virtual structure. Further investigation can be applied to associate and correlate the segmentation with the subsequent dose calculation and clinical treatment outcomes.

For the intracranial vessel wall segmentation task, a more definitive and clinical endpoint-driven goal such as the diagnosis of ICAD vessel segments would enable more substantial

task-specific analysis. Additionally, we have built an automated end-to-end plaque quantification pipeline in a free and open-sourced clinical software -the 3D Slicer, and we are currently drafting a manuscript of the pipeline including the steps from MRA-VWI registration to clinical feature quantification.

The current study serves as a precursor to a more comprehensive and combined text-image analysis. As text and images can provide complementary information, both modalities can be integrated into a single system, e.g., as common domain embeddings or one modality as a prior to be incorporated to the other, to boost the overall performance of a clinical endpoint such as diagnosis accuracy. Specifically, medical reports may provide information on the overall health condition and the risk profile of a patient, while the images provide the rich spatial differential content.

With the growing availability of text and images, as well as images with captions, the translation from one modality to the other and the joint training of text and images would help the understanding of feature representations and can be built into a robust integrated system, which are worthwhile directions to pursue.

REFERENCES

[1] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[2] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[4] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2017.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *CoRR*, vol. abs/1505.0, 2015.

[10] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," 2020.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 779–788, 2016.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, 2016.

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," 2014.

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020.

[16] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra, "Draw: A recurrent neural network for image generation," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1462–1471.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[18] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[19] "Methods for De-identification of PHI — HHS.gov. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. accessed: Dec 07, 2020," .

[20] Ishna Neamatullah, Margaret M. Douglass, Li Wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford, "Automated de-identification of free-text medical records," *BMC Medical Informatics and Decision Making*, vol. 8, pp. 1–17, 2008.

[21] Andrea C. Fernandes, Danielle Cloete, Matthew T.M. Broadbent, Richard D. Hayes, Chin Kuo Chang, Richard G. Jackson, Angus Roberts, Jason Tsang, Murat Soncul, Jennifer Liebscher, Robert Stewart, and Felicity Callard, "Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records," *BMC Medical Informatics and Decision Making*, vol. 13, no. 1, 2013.

[22] Bin He, Yi Guan, Jianyi Cheng, Keting Cen, and Wenlan Hua, "CRFs based de-identification of medical records," *Journal of Biomedical Informatics*, vol. 58, pp. S39–S46, 2015.

[23] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits, "De-identification of patient notes with recurrent neural networks," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 596–606, 2017.

[24] ©2020 IEEE. Reprinted with permission from Hanyue Zhou and Dan Ruan, "An embedding-based medical note de-identification approach with minimal annotation," *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 263–268, 2020.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pp. 1–12, 2013.

[26] Yifan Sun, Nikhil Rao, and Weicong Ding, "A simple approach to learn polysemous word embeddings," *CoRR*, vol. abs/1707.01793, 2017.

[27] Jenny Rose Finkel, Trond Grenager, and Christopher Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, , no. 1995, pp. 363–370, 2005.

[28] Tzvika Hartman, Michael D. Howell, Jeff Dean, Shlomo Hoory, Ronit Slyper, Itay Laish, Oren Gilon, Danny Vainstein, Greg Corrado, Katherine Chou, Ming Jack Po, Jutta Williams, Scott Ellis, Gavin Bee, Avinatan Hassidim, Rony Amira, Genady Beryozkin, Idan Szpektor, and Yossi Matias, "Customization scenarios for de-identification of clinical notes," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 14, 2020.

[29] Aron Henriksson, Hercules Dalianis, and Stewart Kowalski, "Generating features for named entity recognition by learning prototypes in semantic space: The case of de-identifying health records," *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*, pp. 450–457, 2014.

[30] Edward Loper Bird Steven and Ewan Klein, *Natural language processing with python*, O'Reilly Media Inc., 2009.

[31] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2016.

[32] Clark Alexander, "Inducing Syntactic Categories by Context Distribution Clustering," in *Proceedings of CoNLL-2000 and LLL-2000*, 2000, pp. 91–94.

[33] Cheng Peng, Ergun Ahunbay, Guangpei Chen, Savannah Anderson, Colleen Lawton, and X. Allen Li, "Characterizing interfraction variations and their dosimetric effects in prostate cancer radiotherapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 79, no. 3, pp. 909–914, 2011.

[34] Michael Wahl, Martina Descovich, Erin Shugard, Dilini Pinnaduwage, Atchar Sudhyadhom, Albert Chang, Mack Roach, Alexander Gottschalk, and Josephine Chen, "Interfraction anatomical variability can lead to significantly increased rectal dose for patients undergoing stereotactic body radiotherapy for prostate cancer," *Technology in Cancer Research & Treatment*, vol. 16, no. 2, pp. 178–187, 2017, PMID: 27199276.

[35] Gianluca Ingrosso, Roberto Miceli, Elisabetta Ponti, Andrea Lancia, Daniela Cristino, Francesco Pasquale, Pierluigi Bove, and Riccardo Santoni, "Interfraction prostate displacement during image-guided radiotherapy using intraprostatic fiducial markers and a cone-beam computed tomography system: A volumetric off-line analysis in relation to the variations of rectal and bladder volumes," *Journal of Cancer Research and Therapeutics*, vol. 15, 01 2018.

[36] Feng Liu, Ergun Ahunbay, Colleen Lawton, and X. Allen Li, "Assessment and management of interfractional variations in daily diagnostic-quality-ct guided prostate-bed irradiation after prostatectomy," *Medical Physics*, vol. 41, no. 3, pp. 031710, 2014.

[37] Jinkoo Kim, Sanath Kumar, Chang Liu, Hualiang Zhong, Deepak Pradhan, Mira Shah, Richard Cattaneo, Raphael Yechieli, Jared R Robbins, Mohamed A Elshaikh, and Indrin J Chetty, "A novel approach for establishing benchmark CBCT/CT deformable image registrations in prostate cancer radiotherapy," *Physics in Medicine and Biology*, vol. 58, no. 22, pp. 8077–8097, nov 2013.

[38] Claudio Fiorino, Franca Foppiano, Paola Franzone, Sara Broggi, Pietro Castellone, Michela Marcenaro, Riccardo Calandrino, and Giuseppe Sanguineti, "Rectal and bladder motion during conformal radiotherapy after radical prostatectomy," *Radiotherapy and Oncology*, vol. 74, no. 2, pp. 187–195, 2005.

[39] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, Eds., Cham, 2016, pp. 424–432, Springer International Publishing.

[40] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *CoRR*, vol. abs/1606.04797, 2016.

[41] Eliott Brion, Jean Léger, A. M. Barragán-Montero, Nicolas Meert, John A. Lee, and Benoit Macq, "Domain adversarial networks and intensity-based data augmentation for male pelvic organ segmentation in cone beam CT," *Computers in Biology and Medicine*, vol. 131, no. October 2020, 2021.

[42] Jean Léger, Eliott Brion, Paul Desbordes, Christophe De Vleeschouwer, John A. Lee, and Benoit Macq, "Cross-domain data augmentation for deep-learning-based male

pelvic organ segmentation in cone beam CT," *Applied Sciences (Switzerland)*, vol. 10, no. 3, pp. 1154, feb 2020.

[43] Xiaoqian Jia, Sicheng Wang, Xiao Liang, Anjali Balagopal, Dan Nguyen, Ming Yang, Zhangyang Wang, Jim Xiuquan Ji, Xiaoning Qian, and Steve Jiang, "Cone-Beam Computed Tomography (CBCT) Segmentation by Adversarial Learning Domain Adaptation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Dinggang Shen, Tianming Liu, Terry M Peters, Lawrence H Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, Eds., Cham, 2019, pp. 567–575, Springer International Publishing.

[44] Hongfei Sun, Rongbo Fan, Chunying Li, Zhengda Lu, Kai Xie, Xinye Ni, and Jianhua Yang, "Imaging Study of Pseudo-CT Synthesized From Cone-Beam CT Based on 3D CycleGAN in Radiotherapy," *Frontiers in Oncology*, vol. 11, no. March, pp. 1–16, 2021.

[45] Yang Lei, Tonghe Wang, Sibo Tian, Xue Dong, Ashesh B. Jani, David Schuster, Walter J. Curran, Pretesh Patel, Tian Liu, and Xiaofeng Yang, "Male pelvic multi-organ segmentation aided by CBCT-based synthetic MRI," *Physics in Medicine and Biology*, vol. 65, no. 3, 2020.

[46] Yabo Fu, Yang Lei, Tonghe Wang, Sibo Tian, Pretesh Patel, Ashesh B. Jani, Walter J. Curran, Tian Liu, and Xiaofeng Yang, "Pelvic multi-organ segmentation on cone-beam CT for prostate adaptive radiotherapy," *Medical Physics*, vol. 47, no. 8, pp. 3415–3422, 2020.

[47] Alessio Spantini, R. Baptista, and Y. Marzouk, "Coupling techniques for nonlinear ensemble filtering," *arXiv: Methodology*, 2019.

[48] Anjali Balagopal, Samaneh Kazemifar, Dan Nguyen, Mu Han Lin, Raquibul Hannan, Amir Owrangi, and Steve Jiang, "Fully automated organ segmentation in male pelvic CT images," *Physics in Medicine and Biology*, vol. 63, no. 24, pp. 245015, 2018.

[49] Anjali Balagopal, Dan Nguyen, Howard Morgan, Yaochung Weng, Michael Dohopolski, Mu Han Lin, Azar Sadeghnejad Barkousaraie, Yesenia Gonzalez, Aurelie Garant, Neil Desai, Raquibul Hannan, and Steve Jiang, "A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy," apr 2020.

[50] Amirkoushyar Ziabari, Derek C. Rose, Matthew R. Eicholtz, David J. Solecki, and Abbas Shirinifard, "A 2.5d Yolo-Based Fusion Algorithm for 3d Localization of Cells," in *Conference Record - Asilomar Conference on Signals, Systems and Computers*, 2019, vol. 2019-Novem, pp. 2185–2190.

[51] Vanya V. Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker, "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI," in *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, 2018, vol. 2018-Janua, pp. 547–556.

[52] Andreas Themelis and Panagiotis Patrinos, "Douglas-rachford splitting and admm for nonconvex optimization: Tight convergence results," *SIAM J. Optim.*, vol. 30, pp. 149–181, 2020.

[53] Guoyin Li and Ting Kei Pong, "Douglas–rachford splitting for nonconvex optimization with application to nonconvex feasibility problems," *Mathematical Programming*, vol. 159, pp. 371–401, 2016.

[54] Jan Lellmann, Jörg Kappes, Jing Yuan, Florian Becker, and Christoph Schnörr, "Convex Multi-class Image Labeling by Simplex-Constrained Total Variation," in *Scale Space and Variational Methods in Computer Vision*, Xue-Cheng Tai, Knut Mørken, Marius Lysaker, and Knut-Andreas Lie, Eds., Berlin, Heidelberg, 2009, pp. 150–162, Springer Berlin Heidelberg.

[55] Weiran Wang and Miguel Á. Carreira-Perpiñán, "Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application," *CoRR*, vol. abs/1309.1541, 2013.

[56] Álvaro Barbero and Suvrit Sra, "Modular proximal optimization for multidimensional total-variation regularization," *Journal of Machine Learning Research*, vol. 19, pp. 1–82, 2018.

[57] Jorge J. Moré and D. C. Sorensen, "Computing a trust region step," *SIAM J. Sci. Stat. Comput.*, vol. 4, no. 3, pp. 553–572, Sept. 1983.

[58] Zong Ben Xu, Hai Liang Guo, Yao Wang, and Hai Zhang, "Representative of L1/2 regularization among Lq (0<q⩽1) regularizations: an experimental study based on phase diagram," *Zidonghua Xuebao/Acta Automatica Sinica*, vol. 38, no. 7, pp. 1225–1228, 2012.

[59] Qihui Lyu, Daniel O'Connor, Tianye Niu, and Ke Sheng, "Image-domain multimaterial decomposition for dual-energy computed tomography with nonconvex sparsity regularization," *Journal of Medical Imaging*, vol. 6, no. 04, pp. 1, 2019.

[60] J P Mckelvey, "Simple transcendental expressions for the roots of cubic equations," *Citation: American Journal of Physics*, vol. 52, pp. 269, 1984.

[61] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg,

wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham, "ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," Apr. 2021.

[62] S. Mori, M. Endo, T. Obata, T. Tsunoo, K. Susumu, and S. Tanada, "Properties of the prototype 256-row (cone beam) ct scanner," *European Radiology*, vol. 16, pp. 2100–2108, 2006.

[63] Elisabeth Weiss, Jian Wu, William Sleeman, Joshua Bryant, Priya Mitra, Michael Myers, Tatjana Ivanova, Nitai Mukhopadhyay, Viswanathan Ramakrishnan, Martin Murphy, and Jeffrey Williamson, "Clinical evaluation of soft tissue organ boundary visualization on cone-beam computed tomographic imaging," *International Journal of Radiation Oncology*Biology*Physics*, vol. 78, no. 3, pp. 929–936, 2010.

[64] Leah N. McDermott, Markus Wendling, Jasper Nijkamp, Anton Mans, Jan-Jakob Sonke, Ben J. Mijnheer, and Marcel van Herk, "3d in vivo dose verification of entire hypo-fractionated imrt treatments using an epid and cone-beam ct," *Radiotherapy and Oncology*, vol. 86, no. 1, pp. 35–42, 2008.

[65] Annika Hänsch, Volker Dicken, Jan Klein, Tomasz Morgas, Benjamin Haas, and Horst K. Hahn, "Artifact-driven sampling schemes for robust female pelvis CBCT segmentation using deep learning," in *Medical Imaging 2019: Computer-Aided Diagnosis*, Kensaku Mori and Horst K. Hahn, Eds. International Society for Optics and Photonics, 2019, vol. 10950, pp. 212 – 219, SPIE.

[66] T E Marchant, C J Moore, C G Rowbottom, R I MacKay, and P C Williams, "Shading correction algorithm for improvement of cone-beam CT images in radiotherapy," *Physics in Medicine and Biology*, vol. 53, no. 20, pp. 5719–5733, oct 2008.

[67] Tianye Niu, Ahmad Al-Basheer, and Lei Zhu, "Quantitative cone-beam ct imaging in radiation therapy using planning ct as a prior: First patient studies," *Medical Physics*, vol. 39, no. 4, pp. 1991–2000, 2012.

[68] Peyman Adibi, Hamid Mazdak, Ali Darakhshandeh, and Ali Toghiani, "Change in functional bowel symptoms after prostatectomy: A case-control study," *Journal of research in medical sciences : the official journal of Isfahan University of Medical Sciences*, vol. 16, pp. 130–5, 02 2011.

[69] T.K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[70] Om Prakash Gurjar, Ramesh Arya, and Harsh Goyal, "A study on prostate movement and dosimetric variation because of bladder and rectum volumes changes during the course of image-guided radiotherapy in prostate cancer," *Prostate International*, vol. 8, no. 2, pp. 91–97, 2020.

[71] Takashi Hanada, Yutaka Shiraishi, Toshio Ohashi, Junichi Fukada, Tomoki Tanaka, Atsunori Yorozu, and Naoyuki Shigematsu, "Variations in Rectal Volumes and Dosimetry Values Including NTCP due to Interfractional Variability When Administering 2D-Based IG-IMRT for Prostate Cancer," *Journal of Radiotherapy*, vol. 2014, pp. 1–7, jul 2014.

[72] Hanne Tøndel, Arne Solberg, Stian Lydersen, Christer Andre Jensen, Stein Kaasa, and Jo-Asmund Lund, "Rectal volume variations and estimated rectal dose during 8 weeks of image-guided radical 3D conformal external beam radiotherapy for prostate cancer," *Clinical and Translational Radiation Oncology*, vol. 15, pp. 113, feb 2019.

[73] Zhi Chen, Zhaozhi Yang, Jiazhou Wang, and Weigang Hu, "Dosimetric impact of different bladder and rectum filling during prostate cancer radiotherapy," *Radiation Oncology*, vol. 11, 08 2016.

[74] Simon A.A. Kohl, Bernardino Romera-Paredes, Klaus H. Maier-Hein, Danilo Jimenez Rezende, S. M.Ali Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger, "A hierarchical probabilistic u-net for modeling multi-scale ambiguities," *arXiv*, pp. 1–25, 2019.

[75] Zheng Hui, Jie Li, Xiumei Wang, and Xinbo Gao, "Image fine-grained inpainting," 2020.

[76] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu, "Occlusion robust face recognition based on mask learning with pairwisedifferential siamese network," 2019.

[77] Roger Bermudez-Chacon, Pablo Marquez-Neila, Mathieu Salzmann, and Pascal Fua, "A domain-adaptive two-stream U-Net for electron microscopy image segmentation," in *Proceedings - International Symposium on Biomedical Imaging*. may 2018, vol. 2018-April, pp. 400–404, IEEE Computer Society.

[78] Farida B. Ahmad and Robert N. Anderson, "The Leading Causes of Death in the US for 2020," *JAMA*, vol. 325, no. 18, pp. 1829–1830, 05 2021.

[79] Stephen JX. Murphy and David J. Werring, "Stroke: causes and clinical features," *Medicine*, vol. 48, no. 9, pp. 561–566, 2020.

[80] D. M. Mandell, M. Mossa-Basha, Y. Qiao, C. P. Hess, F. Hui, C. Matouk, M. H. Johnson, M. J.A.P. Daemen, A. Vossough, M. Edjlali, D. Saloner, S A Ansari, B A Wasserman, and D. J. Mikulis, "Intracranial vessel wall MRI: Principles and expert consensus recommendations of the American society of neuroradiology," 2017.

[81] Jae W. Song, Athanasios Pavlou, Jiayu Xiao, Scott E. Kasner, Zhaoyang Fan, and Steven R. Messé, "Vessel Wall Magnetic Resonance Imaging Biomarkers of Symptomatic Intracranial Atherosclerosis: A Meta-Analysis," *Stroke*, vol. 52, no. 1, pp. 193–202, 2020.

[82] Jeffrey D. Bodle, Edward Feldmann, Richard H. Swartz, Zoran Rumboldt, Truman Brown, and Tanya N. Turan, "High-resolution magnetic resonance imaging," *Stroke*, vol. 44, no. 1, pp. 287–292, 2013.

[83] Jiayu Xiao, Matthew M. Padrick, Tao Jiang, Shuang Xia, Fang Wu, Yu Guo, Nestor R. Gonzalez, Shujuan Li, Konrad H. Schlick, Oana M. Dumitrascu, Marcel M. Maya, Marcio A. Diniz, Shlee S. Song, Patrick D. Lyden, Debiao Li, Qi Yang, and Zhaoyang Fan, "Acute ischemic stroke versus transient ischemic attack: Differential plaque morphological features in symptomatic intracranial atherosclerotic lesions," *Atherosclerosis*, vol. 319, pp. 72–78, jan 2021.

[84] Ye Qiao, Zeeshan Anwar, Jarunee Intrapiromkul, Li Liu, Steven R. Zeiler, Richard Leigh, Yiyi Zhang, Eliseo Guallar, and Bruce A. Wasserman, "Patterns and implications of intracranial arterial remodeling in stroke patients," *Stroke*, vol. 47, no. 2, pp. 434–440, 2016.

[85] Ye Qiao, Eliseo Guallar, Fareed K. Suri, Li Liu, Yiyi Zhang, Zeeshan Anwar, Saeedeh Mirbagheri, Yuan Yuan Joyce Xie, Nariman Nezami, Jarunee Intrapiromkul, Shuqian Zhang, Alvaro Alonso, Haitao Chu, David Couper, and Bruce A. Wasserman, "MR imaging measures of intracranial atherosclerosis in a population-based study," *Radiology*, vol. 280, no. 3, pp. 860–868, 2016.

[86] Jiayu Xiao, Shlee Song, Konrad Schlick, Shuang Xia, Tao Jiang, Tong Han, Robert Jackson, Márcio Diniz, Oana Dumitrascu, Patrick Lyden, Debiao Li, qi Yang, and Zhaoyang Fan, "Disparate trends of atherosclerotic plaque evolution in stroke patients under 18-month follow-up: a 3d whole-brain magnetic resonance vessel wall imaging study," *The Neuroradiology Journal*, p. 197140092110269, 06 2021.

[87] Fang Wu, Qingfeng Ma, Haiqing Song, Xiuhai Guo, Marcio A. Diniz, Shlee S. Song, Nestor R. Gonzalez, Xiaoming Bi, Xunming Ji, Debiao Li, Qi Yang, and Zhaoyang Fan, "Differential features of culprit intracranial atherosclerotic lesions: A whole-brain vessel wall imaging study in patients with acute ischemic stroke," *Journal of the American Heart Association*, vol. 7, no. 15, aug 2018.

[88] Ye Qiao, David A. Steinman, Qin Qin, Maryam Etesami, Michael Schär, Brad C. Astor, and Bruce A. Wasserman, "Intracranial arterial wall imaging using three-dimensional high isotropic resolution black blood mri at 3.0 tesla," *Journal of Magnetic Resonance Imaging*, vol. 34, no. 1, pp. 22–30, 2011.

[89] Zhaoyang Fan, Qi Yang, Zixin Deng, Yuxia Li, Xiaoming Bi, Shlee Song, and Debiao Li, "Whole-brain intracranial vessel wall imaging at 3 Tesla using cerebrospinal fluid–attenuated T1-weighted 3D turbo spin echo," *Magnetic Resonance in Medicine*, vol. 77, no. 3, pp. 1142–1150, mar 2017.

[90] ©2021 IEEE. Reprinted with permission from Hanyue Zhou, Jiayu Xiao, Zhaoyang Fan, and Dan Ruan, "Intracranial vessel wall segmentation for atherosclerotic plaque quantification," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1416–1419.

[91] ©2022 IEEE. Reprinted with permission from Hanyue Zhou, Jiayu Xiao, Zhaoyang Fan, and Dan Ruan, "Intracranial vessel wall segmentation with deep learning using a novel tiered loss function to incorporate class inclusion," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022.

[92] Qi Yang, Zixin Deng, Xiaoming Bi, Shlee S. Song, Konrad H. Schlick, Nestor R. Gonzalez, Debiao Li, and Zhaoyang Fan, "Whole-brain vessel wall MRI: A parameter tune-up solution to improve the scan efficiency of three-dimensional variable flip-angle turbo spin-echo," *Journal of Magnetic Resonance Imaging*, vol. 46, no. 3, pp. 751–757, sep 2017.

[93] Ron Kikinis, Steve D. Pieper, and Kirby G. Vosburgh, "3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support," in *Intraoperative Imaging and Image-Guided Therapy*, pp. 277–289. Springer New York, 2014.

[94] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, jul 2006.

[95] Isabel M. Adame, Rob J. van der Geest, Bruce A. Wasserman, Mona Mohamed, Johan Hans C. Reiber, and Boudewijn P. F. Lelieveldt, "Automatic plaque characterization and vessel wall segmentation in magnetic resonance images of atherosclerotic carotid arteries," in *Medical Imaging 2004: Image Processing*, J. Michael Fitzpatrick and Milan Sonka, Eds. International Society for Optics and Photonics, 2004, vol. 5370, pp. 265 – 273, SPIE.

[96] Ronald van 't Klooster, Patrick J.H. de Koning, Reza Alizadeh Dehnavi, Jouke T. Tamsma, Albert de Roos, Johan H.C. Reiber, and Rob J. van der Geest, "Automatic lumen and outer wall segmentation of the carotid artery using deformable three-dimensional models in mr angiography and vessel wall images," *Journal of Magnetic Resonance Imaging*, vol. 35, no. 1, pp. 156–165, 2012.

[97] Stanley Osher and Ronald P Fedkiw, "Level Set Methods: An Overview and Some Recent Results," *Journal of Computational Physics*, vol. 169, no. 2, pp. 463–502, 2001.

[98] Yan Wang, Florent Seguro, Evan Kao, Yue Zhang, Farshid Faraji, Chengcheng Zhu, Henrik Haraldsson, Michael Hope, David Saloner, and Jing Liu, "Segmentation of lumen and outer wall of abdominal aortic aneurysms from 3d black-blood mri with a registration based geodesic active contour model," *Medical Image Analysis*, vol. 40, pp. 1–10, 2017.

[99] Feng Shi, Qi Yang, Xiuhai Guo, Touseef Ahmad Qureshi, Zixiao Tian, Huijuan Miao, Damini Dey, Debiao Li, and Zhaoyang Fan, "Intracranial Vessel Wall Segmentation Using Convolutional Neural Networks," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2840–2847, 2019.

[100] Jiayi Wu, Jingmin Xin, Xiaofeng Yang, Jie Sun, Dongxiang Xu, Nanning Zheng, and Chun Yuan, "Deep morphology aided diagnosis network for segmentation of carotid artery vessel wall and diagnosis of carotid atherosclerosis on black-blood vessel wall MRI," *Medical Physics*, vol. 46, no. 12, pp. 5544–5561, 2019.

[101] Davood Karimi and Septimiu E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," 2019.

[102] Li Chen, Jie Sun, Gador Canton, Niranjan Balu, Daniel S Hippe, Xihai Zhao, Rui Li, Thomas S Hatsukami, Jenq-Neng Hwang, and Chun Yuan, "Automated Artery Localization and Vessel Wall Segmentation Using Tracklet Refinement and Polar Conversion," *IEEE Access*, vol. 8, pp. 217603–217614, 2020.

[103] C. R. Vogel and M. E. Oman, "Iterative methods for total variation denoising," *SIAM Journal of Scientific Computing*, vol. 17, no. 1, pp. 227–238, 1996.

[104] Xu Chen, Bryan M. Williams, Srinivasa R. Vallabhaneni, Gabriela Czanner, Rachel Williams, and Yalin Zheng, "Learning active contour models for medical image segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 11624–11632, 2019.

[105] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo, "Polarmask: Single shot instance segmentation with polar representation," *CoRR*, vol. abs/1909.13226, 2019.

[106] Luminita A Vese and Tony F Chan, "A Multiphase Level Set Framework for Image Segmentation Using the Mumford and Shah Model *," *International Journal of Computer Vision*, vol. 50, no. 3, pp. 271–293, 2002.

[107] E. Ukwatta, J. Awad, A. D. Ward, D. Buchanan, J. Samarabandu, G. Parraga, and A. Fenster, "Three-dimensional ultrasound of carotid atherosclerosis: Semiautomated segmentation using a level set-based method," *Medical Physics*, vol. 38, no. 5, pp. 2479–2493, 2011.

[108] A Helen Victoria and · G Maragatham, "Automatic tuning of hyperparameters using Bayesian optimization," *Evolving Systems*, vol. 12, pp. 217–223, 2021.

[109] Muhammad Habib ur Rehman, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah, and Samee U. Khan, "Big Data Reduction Methods: A Survey," *Data Science and Engineering*, vol. 1, no. 4, pp. 265–284, dec 2016.

[110] Maciej Beręsewicz, Risto Lehtonen, Fernando Reis, Loredana Di Consiglio, and Martin Karlberg, *An overview of methods for treating selectivity in big data sources*, European Commission, 2018.

[111] Simon K. Warfield, Kelly H. Zou, and William M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[112] Yunqiao Yang, Zhiwei Wang, Jingen Liu, Kwang-Ting Cheng, and Xin Yang, "Label Refinement with an Iterative Generative Adversarial Network for Boosting Retinal Vessel Segmentation," 2019.

[113] Guohua Cheng, Hongli Ji, and Yan Tian, "Walking on two legs: Learning image segmentation with noisy labels," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, UAI 2020*, 2020, pp. 330–339.

[114] Mattia Tantardini, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi, "Comparing methods for comparing networks," *Scientific Reports*, vol. 9, no. 1, pp. 1–19, 2019.

[115] Chen Chen, Carlo Biffi, Giacomo Tarroni, Steffen Petersen, Wenjia Bai, and Daniel Rueckert, "Learning Shape Priors for Robust Cardiac MR Segmentation from Multi-view Images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11765 LNCS, pp. 523–531.

[116] Qian Yue, Xinzhe Luo, Qing Ye, Lingchao Xu, and Xiahai Zhuang, "Cardiac Segmentation from LGE MRI Using Deep Neural Network Incorporating Shape and Spatial Priors," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11765 LNCS, pp. 559–567.

[117] Tingting Zhao and Dan Ruan, "Learning image based surrogate relevance criterion for atlas selection in segmentation," *Physics in Medicine and Biology*, vol. 61, no. 11, pp. 4223–4234, may 2016.

[118] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei Shi Zheng, Jonathan Li, and Alexander Wong, "Squeeze-And-Attention networks for semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13062–13071.