

# Image Segmentation for Highly Variable Anatomy: Applications to Congenital Heart Disease

by

Danielle Frances Pace

B.Cmp.H., Queen's University (2007)

M.E.Sc., The University of Western Ontario (2010)

Submitted to the Department of Electrical Engineering and Computer  
Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

© Massachusetts Institute of Technology 2020. All rights reserved.

Author .....  
Department of Electrical Engineering and Computer Science  
June 5, 2020

Certified by.....  
Polina Golland  
Professor of Electrical Engineering and Computer Science  
Thesis Supervisor

Accepted by .....  
Leslie A. Kolodziejcki  
Professor of Electrical Engineering and Computer Science  
Chair, Department Committee on Graduate Students



# Image Segmentation for Highly Variable Anatomy: Applications to Congenital Heart Disease

by

Danielle Frances Pace

Submitted to the Department of Electrical Engineering and Computer Science  
on June 5, 2020, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

## Abstract

Automated segmentation of medical images can facilitate clinical tasks in diagnosis, patient monitoring, and surgical planning. However, current methods either rely on explicit correspondence detection, or use machine learning techniques that require a large collection of fully annotated and representative images. Neither of these approaches are suitable when anatomical variability is high and labeled data is limited. In this thesis, we formulate new interactive segmentation methods and evaluate their applicability to congenital heart disease, which involves a wide range of cardiac malformations and topological changes and for which few image analysis methods have been previously developed. We begin by describing the new imaging datasets that we have created to support our research in congenital heart disease. Next, we show that image patches can be used to exploit manual segmentations made on a small set of slice planes in order to automatically segment the rest of an image, and investigate the potential of active learning to automatically solicit user input. Third, we develop an iterative segmentation model that can be accurately learned from small datasets which do not necessarily include the same pathologies as a new image to be segmented, and demonstrate that our model better generalizes to patients with the most severe heart malformations. Ultimately, the methods developed here take a step towards bringing the benefits of medical image analysis to challenging clinical applications involving large anatomical variability and small datasets.

Thesis Supervisor: Polina Golland

Title: Professor of Electrical Engineering and Computer Science



## Acknowledgments

First, I would like to thank my advisor, Polina Golland. I didn't really get to know Polina until after I started at MIT in September, and in fact we couldn't meet at the EECS Visit Days for newly admitted students because she was still on sabbatical in Australia. So, the first piece of advice I received from Polina was actually via a slideshow shown at the event, which had a single written suggestion for new graduate students from each professor in the department. Polina's was to "be stubborn" (and even more so than your advisor). I think I've taken that to heart throughout this Ph.D.! Polina is a true advocate for her students and a superb mentor. I've benefited greatly from her immense knowledge and excellent guidance, both technically and personally, and her strategic thinking helped me learn how to better break down problems and focus my efforts whenever I became lost in all of the options. Through her example, Polina taught me to be strong in my convictions, compassionate towards others, and to strive for excellence. She inspires the courage to tackle hard problems, and gives a wealth of support when things are difficult. I truly could not have asked for a better advisor.

I greatly appreciate the assistance of Mehdi Hedjazi Moghari, my longtime collaborator at Boston Children's Hospital. Mehdi is always willing to help whenever I ask, and ensured that we kept the clinical utility of our projects in the forefront. He is always extremely clear and patient, whether teaching me how to read a cardiac MR scan or discussing directions for algorithm development, and it was a pleasure running our HVSMR challenge together.

The other two members of my committee, Phillip Isola and Juan Eugenio Iglesias, provided excellent insights and suggestions that undoubtedly improved the quality of this thesis. Thank you Phillip for encouraging me to search for a deeper understanding of my methods. I thoroughly appreciate Eugenio's broad and deep knowledge and his critical eye, both of which helped me identify gaps in my thinking and clarify my writing.

I have been extremely lucky to have had many wonderful technical and clinical

collaborators throughout graduate school. Adrian Dalca has been an exceptional collaborator who tirelessly gives 100% in whatever he does, whether debating potential algorithmic approaches, giving comments on a paper draft, or critiquing a presentation. Adrian has given me superb advice innumerable times, whether I was stuck on a derivation or trying to improve my algorithms. I wish to express my deepest gratitude to Jürgen Weese for offering his wide perspective and uncanny ability to get to the root of a problem during our progress meetings, and for hosting me for a memorable summer at Philips Research in Hamburg, Germany. Thanks to Tom Brosch for sharing his deep learning expertise and for his welcoming presence while I visited Hamburg. I also wish to thank Andrew Powell, Tal Geva and Sunil Ghelani at Boston Children’s Hospital for taking the time to teach me about all kinds of heart defects, for sharing their insights into what kinds of heart models and algorithmic tools would be impactful, and for always being so approachable and friendly. I am also grateful to Matthew Jolley, Christian Herz, Andras Lasso, Steve Pieper, Tina Kapur and Ron Kikinis for their help and suggestions.

The new manually segmented datasets described in this thesis would not have been possible without the tireless dedication of Hannah Contreras, Patricia Gao, Shruti Ghelani, Imon Rahaman, and Yue (Jerry) Zhang. The Pacers (they named themselves, I swear!) are all extremely motivated, bright and intellectually curious individuals who will all go far in their chosen paths.

Numerous others have shaped my experience at MIT. Being part of the Medical Vision Group has been a privilege and a joy. I have many fond memories of group lunches, spirited debates, shared conferences, and dinners at fantastic restaurants with Mazdak Abulnaga, Kayhan Batmanghelich, Polina Binder, George Chen, Adrian Dalca, Bernhard Egger, Courtney Guo, Georg Langs, Ruizhi Liao, Razvan Marinescu, Daniel Moyer, Nalini Singh, Ramesh Sridharan, Archana Venkataraman, Christian Wachinger, Clinton Wang, Peiqi Wang, Sandy Wells, Larry Zhang and Miaomiao Zhang. I very much appreciate the time we all spent helping each other with our projects. I would also like to thank everyone at the Medical Vision Reading Group, which for me has been a fantastic learning experience that greatly improved my

critical thinking, my knowledge of subjects not immediately related to my research topic, and my whiteboard presentation skills. More broadly, I would like to thank everyone in the Vision Graphics Neighborhood and at CSAIL for creating such a friendly and intellectually stimulating environment, and everyone at TIG for their help in maintaining our group's computing infrastructure.

My time at MIT would not have been the same without the friends I've made here, including Affi, Alin, Anna, Ari, Colm, Curtis, Deborah, Eva, Gautam, Guha, Gus, Guy, James, Jonas, Joy, Julian, Katie, Lindsay, Michel, Mike, Pavel, Sara, Twan and Zoya. I've spent many memorable evenings with Mandy, Mengfei, Ramya, Amy and Jennifer. I've also very much enjoyed working with and getting to know everyone in the MIT Canadians Club and the MICCAI Student Board.

I wish to acknowledge my parents, Mary Pace and Joe Pace, and my sister Emily. They have always provided unwavering encouragement, regardless of what I wanted to achieve. Thank you to my parents for teaching me the value of commitment and hard work, and for always being there to listen.

Finally, I would like to express my deepest gratitude to my fiancé Marek. His broad outlook has helped me achieve a greater balance than I ever thought was possible in graduate school, and his support, honesty and compassion has helped me work through the challenges it brought along the way. You have my heart.





# Contents

<b>Abstract</b>	<b>3</b>
<b>Acknowledgements</b>	<b>5</b>
<b>List of Figures</b>	<b>13</b>
<b>List of Tables</b>	<b>21</b>
<b>1 Introduction</b>	<b>23</b>
1.1 Problem Overview . . . . .	23
1.2 Clinical Motivation . . . . .	23
1.2.1 Congenital Heart Disease . . . . .	24
1.2.2 3D Heart Models for Congenital Heart Disease . . . . .	25
1.3 Problem Definition and Challenges . . . . .	30
1.4 Related Work . . . . .	33
1.5 Contributions . . . . .	36
<b>2 Datasets and Open Science for Congenital Heart Disease</b>	<b>39</b>
2.1 Background . . . . .	39
2.2 The HVSMR Dataset and Challenge . . . . .	40
2.3 The HVSMR+ and HVSMR++ Datasets . . . . .	43
2.4 Segmentation Evaluation . . . . .	48
2.5 Summary . . . . .	51

<b>3</b>	<b>Patch-Based Interactive Segmentation with Active Learning</b>	<b>53</b>
3.1	Background . . . . .	53
3.2	Patch-based Interactive Segmentation . . . . .	57
3.3	Empirical Study: Active Learning for Reference Selection . . . . .	62
3.4	Evaluation . . . . .	63
3.4.1	Data . . . . .	63
3.4.2	Parameter Selection . . . . .	64
3.4.3	Results . . . . .	64
3.5	Discussion . . . . .	68
<b>4</b>	<b>Learning Iterative Segmentation from Limited Data</b>	<b>73</b>
4.1	Background . . . . .	73
4.2	Iterative Segmentation Model . . . . .	79
4.2.1	Probabilistic Model . . . . .	79
4.2.2	Transition Probability Model . . . . .	80
4.2.3	Learning . . . . .	81
4.2.4	Inference . . . . .	84
4.3	Recurrent Neural Network . . . . .	85
4.3.1	RNN Architecture . . . . .	85
4.3.2	Training Data Generation . . . . .	87
4.3.3	Data Augmentation . . . . .	89
4.4	Evaluation . . . . .	90
4.4.1	Data . . . . .	91
4.4.2	Experimental Setup . . . . .	92
4.4.3	Qualitative Results . . . . .	93
4.4.4	Train on HVSMR+, Test on HVSMR++ . . . . .	95
4.4.5	Train and Test on Subsets of HVSMR++ . . . . .	98
4.4.6	Summary of Results . . . . .	101
4.4.7	User Interaction Mechanisms . . . . .	101
4.5	Discussion . . . . .	104

<b>5</b>	<b>Discussion and Conclusions</b>	<b>109</b>
5.1	Technical Directions . . . . .	109
5.1.1	User Interaction for Whole Heart Segmentation . . . . .	109
5.1.2	Dynamic Heart Models . . . . .	110
5.1.3	Weak Supervision . . . . .	111
5.1.4	Evaluation of Segmentation Accuracy . . . . .	112
5.2	Clinical Outlook . . . . .	113
5.2.1	Visualization and Surgical Planning . . . . .	114
5.2.2	Cardiac Function and Simulation . . . . .	114
5.3	Conclusions . . . . .	115
<b>A</b>	<b>Learning Iterative Segmentation: Supplemental Material</b>	<b>117</b>
A.0.1	Network Architectures . . . . .	117
A.0.2	Training Data Generation . . . . .	118
A.0.3	Data Augmentation . . . . .	119
A.0.4	Learning . . . . .	122
A.0.5	Additional Figures . . . . .	123
	<b>Bibliography</b>	<b>125</b>



# List of Figures

1-1	Internal anatomy of the heart. Adapted with permission from [5]. . .	24
1-2	3D heart models for a patient with DORV. 3D heart models can visualize the entire intracardiac blood pool, the shell around it (consisting of the thick muscle surrounding the ventricles and the thin walls surrounding the atria and great vessels), or each individual cardiac chamber and great vessel. . . . .	28
1-3	Image segmentation is required to create a 3D heart model from a patient’s MRI scan. . . . .	30
1-4	Example challenges related to image appearance in whole heart segmentation from cardiac MRI for CHD. (a) Lack of contrast at boundaries. (b) Different objects can appear locally similar, e.g., the aorta, pulmonary artery and left SVC pointed to by the blue arrows, and the left and right ventricles pointed to by the yellow arrows. (c) Inferior part of the ventricles outside the field of view. (d) MR inhomogeneity artifacts surrounding stents. . . . .	31
2-1	Ground truth HVSMR++ segmentations were created using a pipeline that merged manual annotations with model outputs. . . . .	46
2-2	Optional zones in the ground truth vessel segmentations. . . . .	47

3-1	<p>(a) Our patch-based interactive segmentation algorithm uses manual segmentations on limited image domains (“reference” slices or regions of interest) to segment the rest of the image (i.e., each “target”) slice. Note that the image has been rotated into a short-axis orientation, so that the apex of the heart points down and the bottom slices show cross-sections of the left and right ventricles. (b) An important consideration is where the user should provide input. A simple option is to uniformly distribute full short-axis slices. However, our interactive patch-based algorithm is very flexible, and annotations could be made on any slice or short-axis region. . . . .</p>	55
3-2	<p>An active learning loop for image segmentation comprises three steps: (1) uncertainty sampling to decide where the user should provide input; (2) a batch query in which the user manually labels many voxels; and (3) re-running the segmentation algorithm using the user’s new inputs.</p>	56
3-3	<p>The steps of patch-based segmentation, illustrated for a single target patch in which two entire reference slices are available. We use multipoint estimation, so each reference patch carries one label per voxel, and not a single label for the central voxel in the patch. Therefore, the eventual label for each voxel depends on contributions from all overlapping patches. . . . .</p>	58
3-4	<p>Example setup of our patch-based interactive segmentation, in which a target slice is segmented using two entire short-axis reference slices and a smaller region of interest between them. We also illustrate the exponential curves for the spatially adaptive in-plane and out-of-plane position weights <math>\delta_{\parallel}(x^i, \Omega_r)</math> and <math>\delta_{\perp}(x^i, \mathcal{R}_t)</math>, respectively . . . . .</p>	59

3-5	<p>Example 3D heart models (cut in half to visualize the interior) and segmentation results for a subject with DORV, from our patch-based interactive segmentation method with 3, 8 and 14 uniformly distributed reference slices. The interactive segmentation results are shown in yellow and the ground truth segmentation in red. Arrows indicate segmentation errors that are corrected by including more reference slices. The red line superimposed onto the 3D heart models indicates the position of the 2D image slice visualized below. . . . .</p>	65
3-6	<p>Accuracy of patch-based interactive segmentation as a function of the number of uniformly distributed reference slices. Segmentation accuracy was high, and increased with the number of manually segmented slices. . . . .</p>	66
3-7	<p>Segmentation accuracy of alternative reference selection methods. These are reported as the improvement over <b>uniform slice</b> selection, such that negative values indicate that <b>uniform slice</b> selection outperforms the method. The <b>oracle ROI</b> method is the most promising active learning approach. For experiments using four images (<b>random slice</b> and <b>optimal greedy slice</b>), thin lines represent each subject and the thick line corresponds to the mean. For experiments using twenty images (<b>oracle slice</b> and <b>oracle ROI</b>), we show the mean and standard deviation. <b>Random slice</b> selection scores are averages over five trials per subject. For <b>oracle ROI</b> active learning, the Dice improvement is reported as a function of the cumulative area that the user must segment. . . . .</p>	67
3-8	<p>Segmentation error (in green) overlaid onto an uncertainty map computed as the margin between the first- and second-place labels (white indicates high uncertainty, black indicates low uncertainty). Note that the intensity distributions in our image made it exceedingly rare for there to be votes for all three classes at an individual voxel, so measuring this margin is appropriate. This uncertainty measure did not correlate well with segmentation error. . . . .</p>	70

4-1	Example recurrent neural network (RNN) architectures. (a) An input sequence is processed to produce a single output, with recurrent connections at the hidden layer, e.g. for video classification. Note that the learned network parameters (here, $U$ , $W$ and $V$ ) are the same at each step). (b) An RNN trained to map an input sequence to an output sequence, e.g for language translation. (c) In this architecture, the recurrent connections directly use the output of each iteration as an input at the next step. (d) The architecture from (c) modified for a single input: this is the architecture of the RNN developed in this chapter. (Figure adapted from [130]). . . . .	75
4-2	Example results of an RNN for image segmentation that is trained with a loss function that considers the final output alone. The segmentation improves over time, but the evolution pattern would be difficult for a user to interact with (reproduced from [153], Copyright © 2018, IEEE).	76
4-3	Example results from our iterative segmentation model, which evolves segmentations in a predictable way. (a) Example vessel segmentation (SVC) for a heart with normal anatomy. (b) Example chamber segmentation (RV) for a heart with severe malformations and several previous surgeries. . . . .	77
4-4	Simplified schematic of our RNN architecture for iterative segmentation.	78
4-5	Probabilistic model for the proposed iterative segmentation. Given an image $\mathbf{x}$ , we assume that pairs of segmentations and stopping indicators $\{\mathbf{y}_t, s_t\}$ follow a first order Markov chain. Shaded nodes indicate observed variables. . . . .	80
4-6	Schematic of the RNN trained to jointly evolve the segmentation and predict the stopping indicator. The main block is an augmented U-Net architecture. (Here, the number of feature maps in the latent representation $\mathbf{h}_t$ is $C = 24$ ). . . . .	86



4-7	For each training image $\mathbf{x}$ , input partial segmentations, output partial segmentations and stopping indicators $(\mathbf{y}_{in}, \mathbf{y}_{out}, s)$ are generated on-the-fly during training from the complete ground truth segmentation $\mathbf{y}$ and seed $\mathbf{y}_0$ . . . . .	88
4-8	(a) Representative cardiac MR images showing a stent artifact and bright background areas showing surrounding fluid and vasculature. (b) We use data augmentation to create corrupted input segmentations $\mathbf{y}_{in}$ and uncorrupted output segmentations $\mathbf{y}_{out}$ so that the trained RNN is robust to errors in its intermediate results. . . . .	90
4-9	The five segmentation approaches to be compared, alongside representative inputs and outputs for left ventricle segmentation. . . . .	92
4-10	Examples from subjects with severe heart malformations where iterative segmentation with automatic stopping ( <b>Iter-A</b> ) has high accuracy. Each row depicts the segmentation propagation (as a 3D model overlaid onto a 2D image slice), ending with the automatically detected stopping point. The final column shows an overlay of the ground truth segmentation (dark colour) and the <b>Iter-A</b> result (lighter colour). . .	94
4-11	Results on 40 held-out HVSMR++ subjects after training using 20 HVSMR+ subjects: Summary statistics (Dice score). (Top) For mild and moderate subjects, all methods except <b>U-Net</b> had comparable performance. (Bottom) For severe subjects, the iterative segmentation methods ( <b>Iter-A</b> and <b>Iter-U</b> ) were superior, especially iterative segmentation with user stopping. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold. . . . .	96
4-12	<b>Iter-U</b> outperformed <b>U-Net+S</b> on held-out subjects with severe heart malformations when training with the very small HVSMR+ dataset. Points above the dotted line (white zone) indicate where <b>Iter-U</b> is better. . . . .	97

4-13	<p><b>Iter-A</b> versus <b>U-Net+S</b> for held-out subjects with severe heart malformations when training with the very small HVSMR+ dataset. Points above the dotted line (white zone) indicate where <b>Iter-A</b> is better. For each structure, <b>Iter-A</b> had a better average Dice score than <b>U-Net+S</b>, but there were some cases where the iterative model needed user stopping to achieve the best performance. . . . .</p>	97
4-14	<p>HVSMR++ cross-validation summary statistics (Dice score). (Top) For mild and moderate subjects, all five methods except <b>Iter-A</b> had comparable performance. (Bottom) For severe subjects, <b>Iter-U</b> had the best overall Dice score, with better or similar accuracy compared to <b>U-Net+S</b> for all structures except the SVC. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold. . . . .</p>	98
4-15	<p>Results on 12 held-out HVSMR++ test subjects after training using 48 HVSMR++ subjects: Summary statistics (Dice score). There are 2 mild and moderate subjects and 10 severe subjects. <b>Iter-U</b> had the best overall Dice score, with better or similar accuracy compared to <b>U-Net+S</b> for all eight structures except the SVC. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold. . . . .</p>	100
4-16	<p>Results on 12 held-out HVSMR++ test subjects after training using 20 HVSMR+ subjects: Summary statistics (Dice score). There are 2 mild and moderate subjects and 10 severe subjects. The iterative segmentation methods (<b>Iter-A</b> and <b>Iter-U</b>) were superior, especially iterative segmentation with user stopping. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold. . . . .</p>	100

4-17	Examples from subjects with severe heart malformations in which a user can choose a better segmentation from the output sequence than that chosen via automatic stopping. Ground truth segmentations are dark, <b>Iter-A</b> (top) and <b>Iter-U</b> (bottom) segmentations are lighter. . . . .	101
4-18	The number of iterations executed by automatic stopping (used by <b>Iter-A</b> ) is typically close to the ideal number of iterations (used by <b>Iter-U</b> ). These results are for 40 held-out HVSMR++ subjects after training using 20 HVSMR+ subjects. The median value for each structure is shown at the top of the graph. . . . .	102
4-19	Number of iterations and runtime required for iterative segmentation with user-directed stopping. These results are for 40 held-out HVSMR++ subjects after training using 20 HVSMR+ subjects. The median value for each structure is shown at the top of the graph. . . . .	102
4-20	Iterative segmentation with user stopping can have low accuracy in especially challenging cases. Ground truth segmentations are dark, <b>Iter-U</b> segmentations are lighter. . . . .	103
5-1	Comparing image segmentations of the heart's two ventricles illustrates the deficiencies of the Dice score and average surface distance, and the advantages of a spectral shape similarity measure (nWESD [191,192]). The segmentations in the top and bottom rows have the same number of incorrect pixels, even though those on the bottom are intuitively much better. For each column, bold scores indicate higher similarity to the original segmentation on the left. . . . .	113

A-1 HVSMR+ cross-validation summary statistics (Dice score). (Top) For mild and moderate subjects, all five methods had comparable performance. (Bottom) For severe subjects, the iterative segmentation methods (**Iter-A** and **Iter-U**) were superior, especially iterative segmentation with user stopping. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold. . . . . 123

# List of Tables

1.1	Variants of cardiac anatomy in CHD. All hearts are shown in the anterior view (patient faces towards the camera) or posterior view (patient faces away). Images are axial slices. The $\leftrightarrow$ symbol indicates that two structures are connected. . . . .	26
1.2	Variants of cardiac anatomy in CHD (continued). . . . .	27
2.1	Heart defects and diagnoses in the HVSMR+ and HVSMR++ datasets. Subjects can have multiple diagnoses, and are categorized as mild, moderate or severe according to the most serious defect in the image (e.g., a repaired VSD does not count as a VSD). Coincident variants are not used to categorize subjects. A dilated chamber or vessel is only listed if it is the sole diagnosis. S/P = Status post. . . . .	45
2.2	Ground truth definitions of each cardiac structure and their optional zones. . . . .	49
2.3	Seed points in the HVSMR+ and HVSMR++ datasets. References to “10 axial slices” assumes that the image’s height is 180 slices. The actual number of slices used is proportional to the image’s actual height.	50
4.1	Seed points to be clicked by the user. For more details, see Chapter 2.	89
A.1	Network architectures . . . . .	117
A.2	Step sizes $d_s$ . . . . .	118
A.3	Scale factor $s$ bounds for affine transformations . . . . .	119
A.4	Probability of creating a random intensity-shifted blob . . . . .	120

A.5 Cube size bounds for foreground blobs (0 indicates no foreground blobs) 122  
A.6 Segmentation boundary weights  $\omega_0$  . . . . . 122

# Chapter 1

## Introduction

### 1.1 Problem Overview

The main focus of this thesis is medical image analysis for tasks in which the anatomical variability in a population is significant. In particular, we focus on image segmentation, which is the problem of classifying each pixel in an image according to its anatomical label. Previous image segmentation methods rely on shape models, atlases or a set of exemplary pairs of images and segmentations, and have problems generalizing to inconsistent anatomy when labeled examples are scarce. This is because they require either explicit correspondence detection or a very large dataset of labeled images that illustrates all possible anatomical variants. Our aim is to develop interactive segmentation methods that input an image on which some annotations have been made by a user and output a highly accurate segmentation, and that (1) do not require a large amount of user interaction, (2) do not require a large dataset of segmented images, and (3) can handle changes in the location, shape, topology, number, and presence or absence of each anatomical structure to be segmented.

### 1.2 Clinical Motivation

Our work is driven by problems in analyzing cardiac magnetic resonance images (MRI) from patients with congenital heart disease (CHD). Specifically, we aim to develop

interactive image segmentation algorithms that can efficiently and accurately outline the cardiac structures of the whole heart. Our clinical goal is to enable more routine clinical use of patient-specific 3D heart models for surgical planning.

### 1.2.1 Congenital Heart Disease

Congenital heart disease includes all heart defects existing at birth, encompassing a wide array of potential cardiac malformations and topological changes [1]. Congenital heart disease is the leading cause of birth defect related deaths [2], and affects approximately 1% of births in the USA, about 25% of which is critical CHD for which surgery or other interventions are necessary [3]. Moreover, the life expectancy for CHD patients is improving, leading to an increasing population of adults with CHD that has not been previously seen [4].

#### Anatomy of the Normal Heart:

Fig. 1-1 visualizes the anatomy of the heart, and includes almost all of the anatomical terms referred to in this thesis. Briefly, deoxygenated and oxygenated blood remain separate in the normal heart. Deoxygenated blood arrives via the superior vena cava (SVC) and inferior vena cava (IVC) into the right atrium (RA), and is pumped into

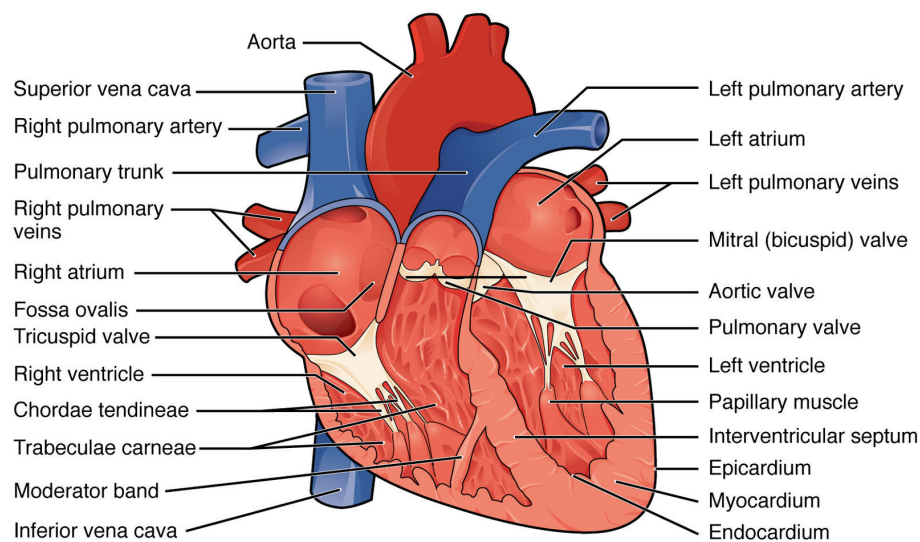


Figure 1-1: Internal anatomy of the heart. Adapted with permission from [5].



the right ventricle (RV) and then through the pulmonary arteries (PA) towards the lungs. Oxygenated blood returns from the lungs via the pulmonary veins, arriving into the left atrium (LA) and traveling through the left ventricle (LV), which pumps it through the aorta (AO) to the body.

### **Congenital Heart Defects:**

CHD can involve a wide range of heart defects, as shown in Tables 1.1 and 1.2. These include shape changes within a vessel or chamber, abnormal connectivity between cardiac structures (e.g., VSD, ASD, DORV, TGA, Glenn surgery, Fontan surgery), abnormal structure locations (e.g., inverted ventricles, inverted atria, dextrocardia, mesocardia), duplicated structures (e.g., bilateral SVC), and/or missing structures (e.g., common atrium, single ventricle). These problems often occur in combination.

Severe CHD requires multiple surgeries throughout infancy, childhood and adult life. For example, in DORV, surgeons may have to decide whether to (1) place a surgical baffle to connect the aorta to the left ventricle through the existing ventricular septal defect, or (2) detach and reattach the aorta and pulmonary artery to reestablish normal blood flow [6, 7]. Importantly, the heart of each CHD patient is different, exhibiting a unique combination of original heart defects, new atypical connections and implants from any prior surgeries, and shape changes from long-term cardiac remodeling [8].

### **1.2.2 3D Heart Models for Congenital Heart Disease**

To choose the best surgical approach and refine the preoperative plan, clinicians must understand each patient’s highly individual heart anatomy, evaluating the size and location of defects and determining their relationships with other cardiac structures.

Cardiac MRI is an attractive modality for preoperative imaging [9,10]. It captures cardiac anatomy and function in high resolution 3D (or 4D) images, while avoiding ionizing radiation (which is particularly important for children). However, visualizing the 3D block of image data remains a challenge. Clinicians often view one 2D

Table 1.1: Variants of cardiac anatomy in CHD. All hearts are shown in the anterior view (patient faces towards the camera) or posterior view (patient faces away). Images are axial slices. The  $\leftrightarrow$  symbol indicates that two structures are connected.

● LV ● RV ● LA ● RA ● AO ● PA ● SVC ● IVC


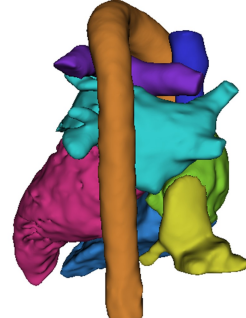
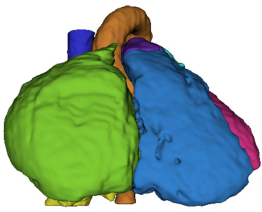
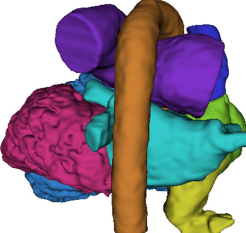
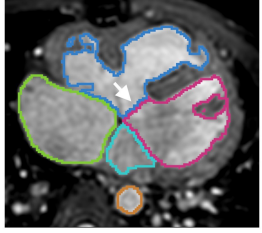
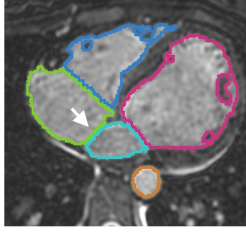
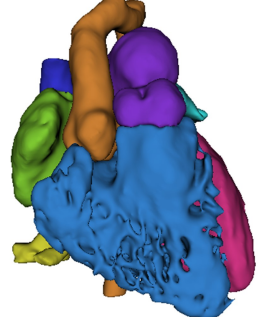
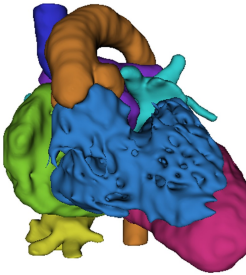
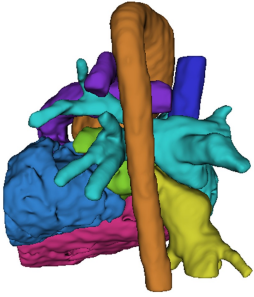


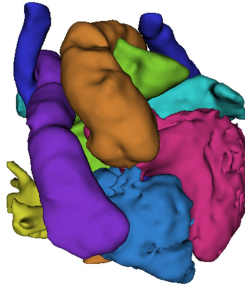
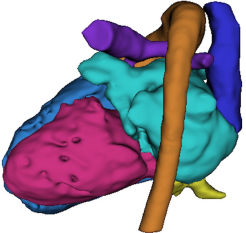
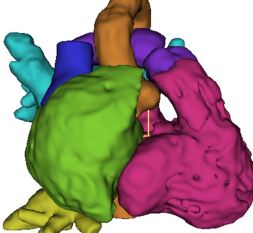
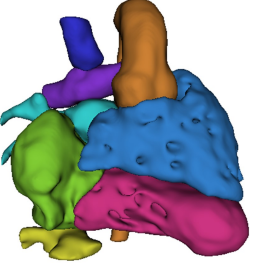
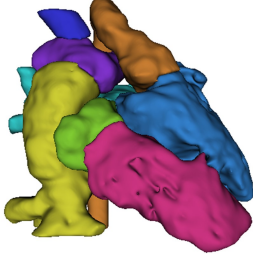
 <p><b>Normal Heart (Anterior View)</b></p> <ul style="list-style-type: none"> <li>• AO <math>\leftrightarrow</math> LV</li> <li>• PA <math>\leftrightarrow</math> RV</li> </ul>	 <p><b>Normal Heart (Posterior View)</b></p> <ul style="list-style-type: none"> <li>• Heart points to the left</li> <li>• SVC <math>\leftrightarrow</math> RA</li> <li>• IVC <math>\leftrightarrow</math> RA</li> </ul>
 <p><b>Severely Dilated Chamber</b></p> <ul style="list-style-type: none"> <li>• E.g., severely dilated RA</li> <li>• Rest of anatomy is normal in this patient</li> </ul>	 <p><b>Severely Dilated Vessel</b></p> <ul style="list-style-type: none"> <li>• E.g., severely dilated PA</li> <li>• Other defects were surgically repaired</li> </ul>
 <p><b>VSD</b></p> <ul style="list-style-type: none"> <li>• Ventricular Septal Defect</li> <li>• Hole in the wall between the two ventricles, i.e., LV <math>\leftrightarrow</math> RV</li> </ul>	 <p><b>ASD</b></p> <ul style="list-style-type: none"> <li>• Atrial Septal Defect</li> <li>• Hole in the wall between the two atria, i.e., LA <math>\leftrightarrow</math> RA</li> </ul>
 <p><b>DORV</b></p> <ul style="list-style-type: none"> <li>• Double Outlet Right Ventricle</li> <li>• AO <math>\leftrightarrow</math> RV</li> <li>• PA <math>\leftrightarrow</math> RV</li> <li>• Always has VSD</li> <li>• This patient has also undergone <b>PA Banding</b></li> </ul>	 <p><b>TGA</b></p> <ul style="list-style-type: none"> <li>• Transposition of the Great Arteries</li> <li>• AO <math>\leftrightarrow</math> RV</li> <li>• PA <math>\leftrightarrow</math> LV</li> <li>• Often has VSD or ASD</li> </ul>

Table 1.2: Variants of cardiac anatomy in CHD (continued).

 <p><b>Inverted Ventricles or Atria</b></p> <ul style="list-style-type: none"> <li>• Right ventricle and/or atrium is on left side</li> <li>• This patient has both</li> </ul>	<p><b>Dextrocardia</b></p> <ul style="list-style-type: none"> <li>• Heart points to the right</li> <li>• Often has <b>Left IVC, Left SVC</b> and/or inverted ventricles</li> </ul> 
 <p><b>Mesocardia + Pulmonary Atresia</b></p> <ul style="list-style-type: none"> <li>• Heart points center</li> <li>• PA valve did not form properly</li> <li>• This patient also has inverted ventricles, an ASD and DORV</li> </ul>	<p><b>TGA Surgery + Bilateral SVC</b></p> <ul style="list-style-type: none"> <li>• Rastelli procedure to restore normal connectivity</li> <li>• Two SVCs (one left, one right)</li> <li>• This patient also has an ASD</li> </ul> 
 <p><b>Common Atrium</b></p> <ul style="list-style-type: none"> <li>• Only one atrium</li> <li>• This patient also has DORV</li> </ul>	<p><b>Single Ventricle</b></p> <ul style="list-style-type: none"> <li>• Only one ventricle</li> <li>• This patient also has an ASD and PA banding</li> </ul> 
 <p><b>Glenn Surgery</b></p> <ul style="list-style-type: none"> <li>• SVC ↔ PA (bypass heart)</li> <li>• IVC ↔ RA</li> <li>• This patient also has DORV and <b>Superoinferior Ventricles</b></li> </ul>	<p><b>Fontan Surgery</b></p> <ul style="list-style-type: none"> <li>• Follows Glenn (SVC ↔ PA)</li> <li>• IVC ↔ PA (surgical baffle)</li> <li>• This patient also has inverted ventricles, an ASD and DORV</li> </ul> 

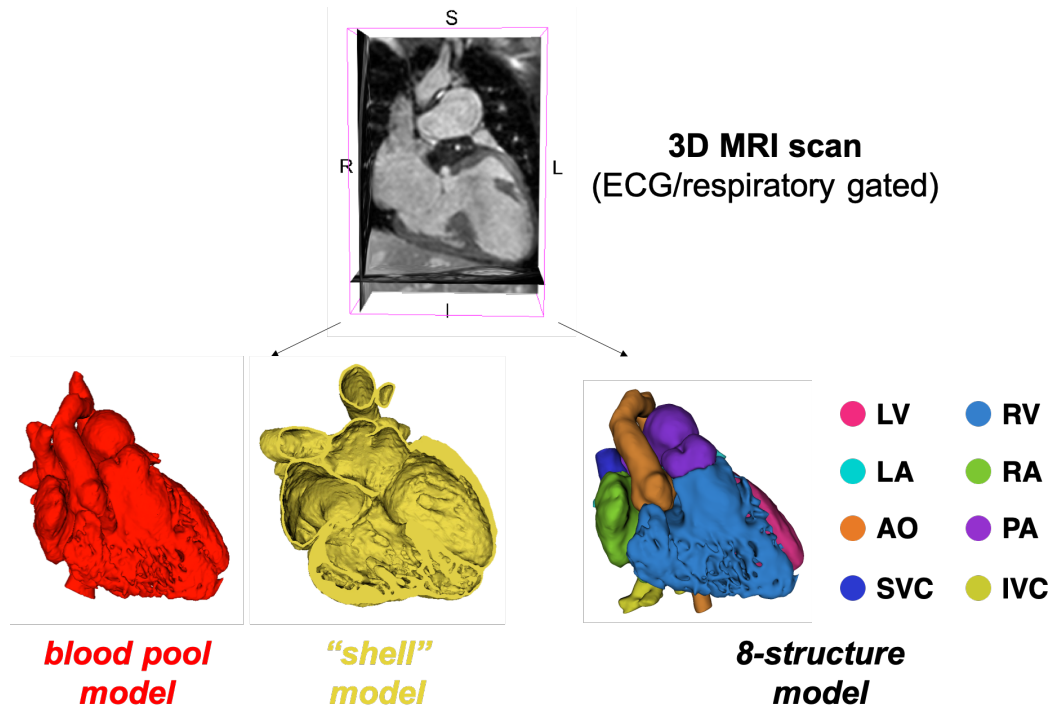


Figure 1-2: 3D heart models for a patient with DORV. 3D heart models can visualize the entire intracardiac blood pool, the shell around it (consisting of the thick muscle surrounding the ventricles and the thin walls surrounding the atria and great vessels), or each individual cardiac chamber and great vessel.

image slice at a time, mentally integrating them to form an impression of the 3D heart anatomy, or examine very coarse 3D blood pool models obtained through intensity thresholding, which can be inaccurate, dramatically obscured by surrounding vasculature, and inadequate for viewing anatomy deep within the heart [11]. Further analyzing the intracardiac anatomy during surgery is also difficult, due to blood in the field of view, restricted viewing portals through valves and incisions, and the flaccid heart.

Patient-specific 3D heart models hold great potential to enhance surgical planning for CHD, whether they are rendered on a computer screen or 3D-printed. In this thesis, we consider generating several different types of 3D heart models from a cardiac MRI scan (Fig. 1-2). The most simple 3D heart model reveals the intracardiac blood pool, while a “shell” model can be cut in half to better visualize the interior. Separately labeling each structure produces heart models that may be more intuitive, and aids

segmentation by reducing the shape variability of each labeled piece of the anatomy.

Studies have shown that using a 3D heart model may lead to a greater appreciation of the true locations and sizes of intracardiac structures, aid decision making and consensus, and even alter the surgical plan from that originally based on imaging [11–15]. In particular, 3D-printed heart models provide an anatomically faithful and tactile experience [16], can be used as physical phantoms for surgical practice [17], and may reduce exposure to anesthesia and cardiopulmonary bypass via decreased intraoperative times (again, important for children) [18, 19]. 3D heart models also have applications in medical education [20].

Automatic segmentation would also facilitate the computation of several quantitative metrics of cardiac function, such as chamber volumes, ejection fraction, myocardial mass and thickening, aortic dimensions, and ventricular motion analyses [21, 22]. In current clinical practice, such indices are typically derived from 2D or 2D cine MRI. For CHD patients whose anatomy does not match what is expected by commercial software, these images must be contoured manually or the automatic results must be heavily adjusted. Segmenting 3D images avoids cross-referencing 2D images from multiple views, and may be more accurate than estimating inherently 3D measurements from sparse 2D data. Looking forward, separately delineating each cardiac structure in 4D MRI (3D + time) data promises to enable future research into simulating post-surgical hemodynamics, assessing joint atrio-ventricular function, and quantifying vessel wall stiffness.

Building a 3D heart model requires image segmentation (Fig. 1-3). Delineating all of the cardiac structures in a patient’s MRI scan is known as “whole heart segmentation” [23–25]. In a 2016 review of segmentation techniques used in the medical literature for 3D printing heart models in congenital heart disease, Byrne *et al.* [26] found that most clinical studies used either simple image segmentation methods (e.g. thresholding, region growing and manual editing) or a specific software (Mimics, Materialise, Leuven, Belgium, proprietary algorithm). The lack of accurate, robust whole heart segmentation for congenital heart disease is the bottleneck that currently precludes widespread adoption of 3D heart models for surgical planning,

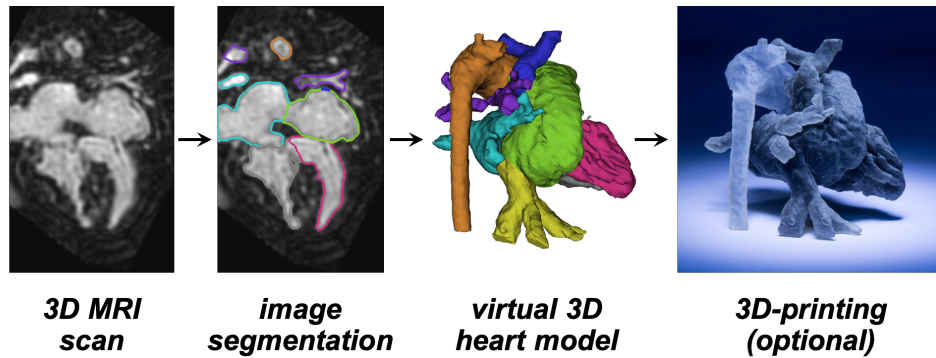


Figure 1-3: Image segmentation is required to create a 3D heart model from a patient’s MRI scan.

since segmentation currently involves extensive manual effort requiring many hours per subject [12, 26–28].

### 1.3 Problem Definition and Challenges

The input to our methods is a 3D cardiac MRI scan, acquired with electrocardiography (ECG) and respiratory gating to capture the heart at a single point in the cardiac cycle without motion artifacts. This type of 3D cardiac MRI is widely available in clinics. A useful 3D heart model can be created after segmenting the intracardiac blood pool, heart walls (including the thick ventricular myocardium and thin walls surrounding the atria and great vessels), and background. However, a more complete whole heart segmentation involves separately outlining the left ventricle, right ventricle, left atrium (including the pulmonary veins), right atrium, aorta, pulmonary artery, superior vena cava and inferior vena cava.

#### **Extreme Anatomical Variability:**

CHD can affect the size, shape, location, connectivity, background appearance, number and existence of cardiac structures. There are additional variations due to age, because a larger extent of the the head and torso is imaged for babies and young children. Cropping the image around the heart standardizes the field of view, but residual differences remain, e.g., only children have a large visible thymus lying in

front of the heart. The anatomical variability of congenital heart disease is at or beyond the limits of what has been previously attempted in automated medical image analysis. Strong anatomical priors cannot be enforced, and relating information across subjects with dramatically different heart configurations is very difficult.

### Image Appearance:

Additional difficulties related to cardiac MRI are illustrated in Fig. 1-4. The valves and thin walls that separate neighboring structures are often beyond the imaging resolution (typically around  $1\text{ mm}^3$ ), and hence there is no contrast at object boundaries. Different chambers and great vessels often appear very similar locally: in the normal heart their identity can be resolved using global context, but this is not straightforward for CHD due to heart malformations. In addition, the tips of the ventricles are sometimes placed out of the field of view during image acquisition to reduce scan time. Finally, cardiac MR is subject to a number of artifacts. Specifically, steady-state free precession (SSFP) images are subject to B0 inhomogeneities, and the pulmonary veins in particular can be poorly visible because deoxygenated blood has a lower T2 and their position near the lungs induces off-resonance artifacts. Finally, very dark regions surround metal implants such as stents.

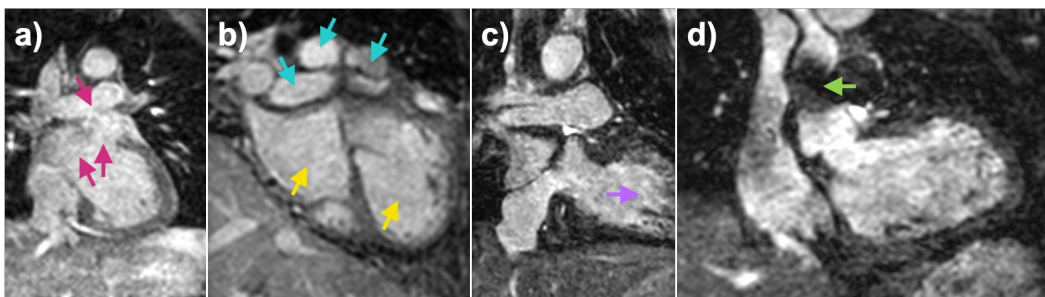


Figure 1-4: Example challenges related to image appearance in whole heart segmentation from cardiac MRI for CHD. (a) Lack of contrast at boundaries. (b) Different objects can appear locally similar, e.g., the aorta, pulmonary artery and left SVC pointed to by the blue arrows, and the left and right ventricles pointed to by the yellow arrows. (c) Inferior part of the ventricles outside the field of view. (d) MR inhomogeneity artifacts surrounding stents.

### **Limited Training Data:**

Modern machine learning can yield excellent performance given a large annotated dataset with limited domain shift. However, as in other fields within science and engineering [29–31], limited training data is a persistent issue in medical image analysis. This is especially true for new applications of medical imaging that are outside of the standard clinical routine, including ours. In part, this is due to the effort and medical training required to manually label large 3D images, which only grows more arduous as the analysis task is more difficult. Scarce training data precludes attempts to model anatomical subtypes separately in an attempt to reduce anatomical variability, as in [32, 33]. Moreover, patients with unique combinations of defects and prior surgeries defy categorization anyway. In our datasets, specific heart abnormalities are often represented by a single sample, and will be unseen during training if that sample is assigned to the validation or test set. In order to be useful, any approach based on machine learning must be able to learn from images that represent only a subset of the possible anatomy, and generalize well to previously unseen anatomical configurations.

### **Approach:**

As described above, any methods that we develop must be able to handle extraordinary anatomical variability while generalizing well from small, imbalanced datasets. Experiments performed in this thesis demonstrate that state-of-the-art fully automatic methods fail to do so. We therefore focus on developing efficient interactive segmentation methods.

In interactive segmentation, a user works with the computer to segment an image by providing limited manual inputs and/or correcting errors [34–44]. For example, the user can provide inputs by identifying anatomical landmarks, painting a few brushstrokes of tissue labels, or segmenting a few 2D slices in a 3D image. Interactive segmentation enables accurate image segmentation for very difficult problems, by aiding object localization and/or providing information on local appearance and



shape. In many cases, the human user can apply their medical knowledge and logical reasoning, while the segmentation algorithm may be able to contour precise boundaries faster, more accurately, and more reproducibly. Finally, research in interactive segmentation is motivated by the fact that some interaction is unavoidable, since clinicians must validate any segmentation used for decision making and correct the errors that are inevitable in automatic segmentation.

## 1.4 Related Work

In this section, we review prior work in whole heart segmentation and image analysis for congenital heart disease. Additional technical background relevant to the methods developed in this thesis will be reviewed at the beginning of each subsequent technical chapter.

### **Whole Heart Segmentation:**

Whole heart segmentation has a rich history, especially for hearts that do not exhibit substantial deviations from normal anatomy. When this doctoral work began, state-of-the-art methods were based on adapting a canonical mesh model [45, 46] or used multi-atlas segmentation [47]. Both of these approaches rely on finding correspondences between the image to be segmented and some representation of the expected anatomy. The first approach warps a generic surface mesh of the heart towards potential boundaries in the image [48–51], while multi-atlas segmentation uses a dataset of segmented atlas images to label a target image via dense deformable image registration and label fusion [52–56]. However, the substantial changes in heart geometry and topology in CHD makes shape modeling, image registration, and correspondence detection extremely difficult. The required correspondences between the model or atlas and the image to be segmented would be very complex, and may not even exist. In particular, fitting a heart model imposes a strong shape prior, since the amount of allowed deformation is limited according to the variability in the training dataset, and is therefore impractical for segmenting pathological cases for which it has not been

trained. On the other hand, deformable image registration is required for multi-atlas segmentation but typically fails when the input images are very different. Hence, multi-atlas segmentation would require a very large number of previously segmented scans and very sophisticated atlas selection. In fact Zuluaga *et al.* [57] actually exploit the widespread mis-registrations to perform computer-aided diagnosis by predicting whether subjects had normal anatomy or whether they had undergone an arterial switch or atrial switch operation to repair TGA.

More specialized approaches have been proposed to address different modes of variation in the heart. Prior work has addressed the physical distortions and appearance changes that arise when segmenting the LV and RV in patients with adult-onset heart diseases, such as left ventricular hypertrophy, dilated left ventricle, pulmonary hypertension, heart failure or myocardial infarction [58–61]. However, these methods still rely on probabilistic atlases or point distribution models that are built using data from normal subjects, again requiring one to find a spatial transformation between the abnormal heart and the normal model via anatomical landmarks or image registration. The concept is viable when one considers cardiac ventricles that can change in shape, size and wall thickness, but such methods are unlikely to perform well for whole heart segmentation in CHD.

Other groups have addressed how to handle topological changes in the anatomy in the context of segmenting the LA, since the number of pulmonary veins and their connectivity to the left atrial chamber naturally varies between individuals. Multi-atlas segmentation can be improved by more appropriately weighting the contribution from each atlas image [62]. One can also create a separate model for each expected anatomical variant, apply each one, and then automatically choose amongst the results based on the quality of each model’s local fit as estimated by region growing or machine learning [32, 33]. However, modeling each subtype of CHD would be infeasible without a prohibitively large database of segmented images, since the number of defects and their potential combinations is huge. Others have proposed part-based models for LA segmentation, which separately label the left atrial chamber and each pulmonary vein and later enforce the consistency between them [63]. Such an approach

would be difficult for our application, considering the large number of structures to be segmented and their uncertain connectivity.

More recently, state-of-the-art methods in medical image segmentation train a convolutional neural network (CNN) to segment images [64]. The U-Net architecture (encoder-decoder with skip connections) is especially popular [65]. CNNs have been extensively applied to whole heart segmentation, often using a dataset from the Multi-Modality Whole Heart Segmentation (MM-WHS) challenge held at MICCAI 2017 [24]. This dataset consists of 60 computed tomography (CT) images and 60 MRI scans from a variety of patients, including 16 images from CHD patients (7 training, 9 testing). In this context, researchers have investigated the use of two-step network cascades (heart localization or coarse segmentation followed by a segmentation refinement network) [66,67], deep supervision [68], multi-planar CNNs [69,70], losses based on the Dice score [68,71], and integration of statistical shape priors [69]. To date, the MM-WHS challenge results confirm the difficulty of our task: the MR images were more difficult to segment than the CT images (likely due to lower contrast, signal-to-noise ratio and spatial resolution, as well as imaging artifacts), and accuracy in the CHD patient subset was lower due to shape changes.

### **Image Analysis for Congenital Heart Disease:**

Algorithm development regarding automated image analysis for congenital heart disease is very limited, and has mostly been aimed towards computer-aided diagnosis [72–74] or heart function quantification [75–79]. Although image segmentation is often required as part of the analysis pipeline, most often it was not the main focus. Only two of these papers have a significant segmentation component, both of which focus on Tetralogy of Fallot. Zhang *et al.* [72] segment the LV and RV in MR time series data using a hybrid active shape and active appearance model approach, while Mansi *et al.* [75] fit a geometric model to the image using marginal space learning, a probabilistic boosting tree and steerable features. However, all of these prior works have two deficiencies. First, they consider only the two ventricles [72,73,77,79], the right ventricle alone [75,76], or the aorta [74,78], rather than the whole heart. Sec-

ond, most limit their focus to a single subtype of CHD [72–75, 78, 79], while those that consider multiple subtypes required extensive manual inputs [76, 77].

## 1.5 Contributions

To the best of our knowledge, we have developed the first datasets and algorithms for whole heart segmentation in cardiac MRI for patients with diverse subtypes of congenital heart disease. This includes the first method to segment the blood pool and heart walls, and the first method to separately outline individual cardiac chambers and great vessels.

### 1. Datasets and Open Science for Congenital Heart Disease

We have created the first public dataset for whole heart segmentation in congenital heart disease patients, which consists of twenty 3D cardiac MRI scans with ground truth segmentations of the blood pool and heart walls [80]. The aim is to foster increased research in medical image analysis for the understudied CHD population through open data. This dataset has been continually expanded and refined, and has since grown to contain sixty images with ground truth segmentations of the LV, RV, LA, RA, AO, PA, SVC and IVC. The datasets and our efforts in open science are described in Chapter 2.

### 2. Patch-Based Interactive Segmentation with Active Learning

We present a new interactive algorithm to segment the cardiac blood pool, the ventricular myocardium and the thin walls surrounding the atria and great vessels from cardiac MRI for patients with congenital heart disease. In Chapter 3, we describe a new interactive segmentation method that exploits expert segmentations on a small set of short-axis slice regions, and automatically delineates the remaining volume using patch-based segmentation [81]. We also investigate the potential of active learning to automatically solicit user input in areas where segmentation error is likely to be high. Validation is performed on twenty CHD subjects with a variety of congenital heart defects. We show

that active learning strategies that ask the user to manually segment uncertain regions of interest within short-axis slices yield higher accuracy with less user input than approaches that query entire short-axis slices.

### 3. Learning Iterative Segmentation from Limited Data

This work addresses the need for whole heart segmentation to individually label each cardiac chamber and great vessel for patients with congenital heart disease. State-of-the-art image segmentation methods use a convolutional neural network (CNN) to directly segment an image in one step, requiring a large collection of manually annotated images to capture the anatomical variability in a cohort. In Chapter 4, we propose a novel iterative segmentation model, implemented as a recurrent neural network (RNN), which can be accurately learned from a small dataset [82]. The user provides a single landmark per structure, and a segmentation is evolved over multiple steps until reaching a stopping point that can be automatically determined or user-defined. The model grows segmentations in a predictable way that is defined via the training data, and we show that a loss function that evaluates the entire sequence of output segmentations can be optimized using training images alongside input-output pairs of partial segmentations. Our experiments demonstrate that, compared to conventional models that segment an image in one step, our iterative segmentation offers better generalization to patients with the most severe heart malformations, especially when training data is very limited.

These advances represent significant contributions towards our clinical goal of enabling 3D heart models to improve surgical planning for patients with congenital heart disease. A discussion of future technical directions and clinical outlooks is provided in Chapter 5.



# Chapter 2

## Datasets and Open Science for Congenital Heart Disease

In this chapter, we describe the novel datasets for whole heart segmentation in congenital heart disease that we have continuously developed and expanded over the course of this doctoral work. We will detail three versions, called HVSMR, HVSMR+ and HVSMR++. HVSMR stands for “Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease”. The original HVSMR dataset contains images from twenty subjects with segmentations of the blood pool and ventricular myocardium. This dataset was publicly released in a Challenge held at MICCAI 2016, and was the first open dataset of its kind [80]. Later, the HVSMR+ and HVSMR++ datasets were created, which have more detailed segmentations of individual cardiac chambers and great vessels, and contain progressively more images.

### 2.1 Background

Open source software (e.g., [83–89]) has had a critical impact on medical image analysis over many years. Open data is increasingly recognized as being just as important, especially since the proliferation of machine learning in radiology [90, 91].

Curating a new medical image dataset involves careful subject selection, tedious ground truth annotation, attention to the unstructured text in clinical reports, con-

sideration of patient privacy, and questions of data ownership and control [90–93]. If the aim is to automatically reproduce an existing clinical workflow, then ground truth classifications, measurements or image segmentations may be available in the hospital PACS (Picture Archiving and Communication System). However, for new applications of medical imaging, ground truth annotations must usually be generated, typically manually. This requires extensive training that cannot in general be crowd-sourced. In addition, the many slices that make up a 3D medical image makes annotation tasks like landmark detection and image segmentation extremely time consuming, especially for complex tasks and when consistency across slices must be maintained.

For these reasons, large annotated datasets are less prevalent in medical imaging than in mainstream computer vision. There exist only a few initiatives involving thousands of subjects, and these may not provide any annotations. Examples include ADNI [94], the UK Biobank [95], and the Rotterdam Study [96] (see [92] for a few more datasets). At present, most public medical image datasets are relatively small [97], and many methodological developments rely on private data. More concretely, a (non-peer reviewed) study found that the median size of the MR and CT datasets used in MICCAI 2018 articles was around 70 [98], and a recent report found that more than half of the papers published in MICCAI 2014-2018 used only private data [99].

When this thesis work began, no dataset was available for whole heart segmentation in CHD patients, and, perhaps consequently, very little research had been undertaken in this area. We grew our own dataset with considerable effort over several years. After we released the HVSMR dataset in October 2016, research on image segmentation for CHD immediately increased, demonstrating the importance of data in our research community and the major impact of our contribution.

## 2.2 The HVSMR Dataset and Challenge

The original HVSMR dataset contains twenty 3D cardiac MR images from patients with a variety of congenital defects, plus ground truth segmentations of the cardiac



blood pool and ventricular myocardium. The images were acquired during clinical practice at Boston Children’s Hospital (Boston, MA, USA), retrieved retrospectively from the hospital PACS, and manually segmented.

**Subject Selection:** Cases with high image quality were chosen on a rolling basis, by considering the signal homogeneity of the blood and myocardium and the strength of off-resonance artifacts. Only images in which the LV and RV were completely visible were selected.

**Images:** High resolution images of the entire heart were acquired in an axial view on a clinical 1.5T scanner (Philips Achieva). Acquisitions were performed using the Heart-NAV technique [100], which uses a free-breathing steady-state free precession (SSFP) pulse sequence, with ECG and respiratory navigator gating used to freeze cardiac and respiratory motion (TR=3.4ms, TE=1.7ms,  $\alpha=60^\circ$ ). Intravenous gadolinium-based contrast agent (Ablavar (gadofosveset) or Gadovist) was used in some patients. The proportion of imaging studies that use contrast has been reduced over time at Boston Children’s Hospital to around 20-30% currently. Each image had a different size ( $\sim 390 \times 390 \times 165$ ) and near-isotropic resolution ( $\sim 0.9 \times 0.9 \times 0.85$  mm).

**Intensity Normalization:** Since intensity distributions vary across cardiac MR scans, intensity normalization is required. We created an intensity normalization scheme based on estimating the mean blood pool and lung intensities in each image. Each image was rescaled to a range  $\approx [-0.1, 3.3]$  by fitting a linear transfer function that mapped estimates of the typical blood pool and lung intensities to 0.8 and 0.07, respectively. For each image, we estimated the blood pool intensity by automatically extracting a slab of the cropped images that typically contains the ventricles, and used the peak of the intensity histogram corresponding to the blood pool using the Mean Shift algorithm [101]. Similarly, we estimated the typical lung intensity in the image by extracting a slab in the upper portion of the cropped image that typically contains the lungs only, and used the mode of the resulting intensity histogram.

**Segmentations:** The blood pool and myocardium in each image was segmented by a trained rater using 3D Slicer<sup>1</sup> [85], which was reviewed by two hospital experts who advised the rater on any required corrections. Segmentations were done in an approximate short-axis view and then transformed back into the original image space. Manual segmentation considered all three viewing planes, but segmentation quality in the short-axis view was the deciding factor during review.

The blood pool class includes the cardiac chambers and great vessels taken together. All of the great vessels except the aorta were extended only a few centimeters past their origin, since very long vessels can be visually disruptive when 3D heart models are used for surgical planning. The ventricular myocardium class also included the coronaries if they traveled within it.

**The HVSMR 2016 Challenge:** The first HVSMR Challenge was held at MIC-CAI 2016<sup>2</sup>, with full conference papers published in Lecture Notes in Computer Science [80]. The submission system and leaderboards remain open. The challenge was organized by releasing 10 training images (with segmentations) and 10 test images (segmentations not publicly available). Scoring is done through an automated submission system where participants can submit segmentation results, and considers a weighted average of the Dice score, average boundary distance, and Hausdorff distance for the blood pool and myocardium classes (see Section 2.4 for definitions).

In publicly releasing this dataset, we have sparked a broader research effort towards medical image analysis for congenital heart disease. There have been approximately 100 submissions to the HVSMR leaderboard since the Challenge began, and about 35 medical imaging conference and journal papers have been published using the dataset (as of May 2020).

---

<sup>1</sup><http://www.slicer.org>

<sup>2</sup><http://segchd.csail.mit.edu>

## 2.3 The HVSMR+ and HVSMR++ Datasets

While a segmentation of the cardiac blood pool and myocardium is sufficient to display or 3D-print a patient-specific 3D heart model, there are many advantages to delineating each cardiac structure separately. On the methodological side, aiming for a more detailed segmentation may make the segmentation task easier, since the shape of each cardiac structure is less variable than that of the entire blood pool. Hence, feasible shapes could possibly be learned during training, or explicit shape priors could be incorporated. Clinically, it would allow for more detailed analyses of cardiac structure and function, whether for an individual patient or across an entire population. A second goal of ours was to increase the size of the dataset, to more comprehensively evaluate our methods and, for those methods based on machine learning, to assess their dependence on the size of the training dataset.

The HVSMR+ and HVSMR++ datasets consist of 3D cardiac MR images with ground truth segmentations that separately delineate the LV, RV, LA, RA, AO, PA, SVC and IVC. The myocardium was deemed less important, as it would likely be simple to derive once these eight structures are segmented. The HVSMR+ dataset has the same 20 cases as HVSMR. The HVSMR++ dataset is larger, comprising 60 scans that include the 20 HVSMR images.

**Subject Selection:** 3606 cases were identified by searching the written radiology reports at Boston Children’s Hospital (dating back to January 2012) for keywords indicating that a 3D MR scan was acquired. In addition to free-text descriptions, each report contains standardized “cardiology codes” (local to Boston Children’s Hospital) that enumerate hundreds of different patient diagnoses, abnormalities in cardiac anatomy or function, and prior interventions. Since clinicians manually choose the most relevant options as they write reports, many findings are not codified. In addition, the codes often describe initial diagnoses that have already been surgically corrected and are no longer applicable to the scan. However, the codes do provide a simple way to find patients with certain conditions.

We identified codes pertaining to important congenital heart defects and corrective

surgeries, and manually selected 40 high quality images to add to the HVSMR dataset. We aimed to create a balanced dataset that samples the different conditions and their combinations as uniformly as possible, while recognizing that some imbalance is inevitable since some defects are much more common than others. A trained rater reviewed each image to verify that its list of diagnoses was correct and complete .

**Subject Categorization:** After assessing each heart’s anatomical malformations (and not the patient’s prognosis), the cases were classified as mild, moderate or severe under the advise of a cardiologist. The relevant conditions and their prevalence in the HVSMR+ and HVSMR++ datasets are summarized in Table 2.1. To summarize, mild hearts had roughly normal anatomy<sup>3</sup>, prior CHD surgery with restoration of normal anatomy, and/or a mildly or moderately dilated chamber or vessel. Moderate hearts had abnormal connectivity, holes in the interior heart wall, bilateral SVC, a severely dilated chamber or vessel, and/or a congenital connective tissue disorder causing extremely curvy vessels. Severe hearts exhibited global heart malpositions or situs inversus (mirror image of normal anatomy), common atrium, single ventricle, and/or major prior reconstructive surgery resulting in highly abnormal anatomy.

**Images:** The vast majority of images were acquired using the Heart-NAV technique described above; very few were acquired using other protocols. All images were manually cropped tightly around the heart and intensities were normalized as described above.

---

<sup>3</sup>For example,

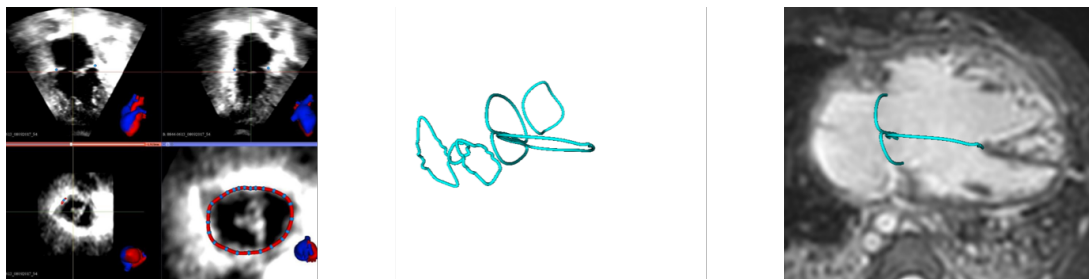
- patent foramen ovale – small hole in the wall between the two atria (present in all fetuses but normally closes after birth),
- vessel stenosis – narrowing in a blood vessel,
- hypertrophic cardiomyopathy – thickened ventricular myocardium,
- crossed PAs – abnormal origin and bend of the PA branches as they arise from the main PA trunk,
- coronary aneurysm – dilated coronary artery (artery supplying blood to the heart muscle itself),
- hypertension – high blood pressure,
- Marfan syndrome – systemic connective tissue disorder with many symptoms including heart problems, very thin chest often visible on cardiac MRI.

Table 2.1: Heart defects and diagnoses in the HVSMR+ and HVSMR++ datasets. Subjects can have multiple diagnoses, and are categorized as mild, moderate or severe according to the most serious defect in the image (e.g., a repaired VSD does not count as a VSD). Coincident variants are not used to categorize subjects. A dilated chamber or vessel is only listed if it is the sole diagnosis. S/P = Status post.

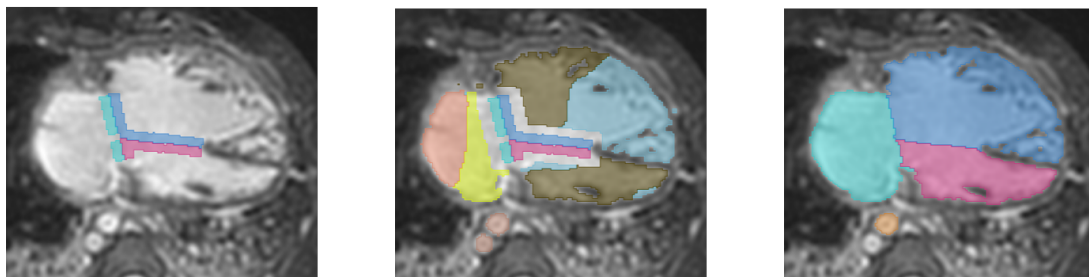
	<b>HVSMR+</b>		<b>HVSMR++</b>	
	<b>20 Subjects</b>		<b>60 Subjects</b>	
Age	18.1 ± 11.2		11.6 ± 11.7	
<b>Mild</b>	<b>10</b>	<b>50%</b>	<b>12</b>	<b>20%</b>
~Normal	5	25%	6	10%
Mild/Moderate Dilation	5	25%	6	10%
<b>Moderate</b>	<b>6</b>	<b>30%</b>	<b>11</b>	<b>18%</b>
VSD	6	30%	30	50%
ASD	3	15%	23	38%
DORV	3	15%	19	32%
D-Loop TGA	1	5%	5	8%
S/P Arterial Switch	–	–	1	2%
Bilateral SVC	2	10%	9	15%
Severe Dilation	2	10%	4	7%
Tortuous Vessels	–	–	2	3%
<b>Severe</b>	<b>4</b>	<b>20%</b>	<b>37</b>	<b>62%</b>
Heterotaxy	1	5%	14	23%
Dextrocardia	2	10%	10	17%
Mesocardia	1	5%	5	8%
Inverted Ventricles	3	15%	16	27%
Inverted Atria	–	–	7	12%
Left/Central IVC	1	5%	15	25%
Left/Central SVC	1	5%	6	10%
L-Loop TGA	1	5%	5	8%
S/P Atrial Switch	–	–	1	2%
S/P Rastelli Procedure	1	5%	2	3%
Common Atrium	1	5%	10	17%
Single Ventricle	–	–	10	17%
S/P Glenn Procedure	1	5%	24	40%
S/P Fontan Procedure	1	5%	9	15%
<b>Coincident Variants</b>				
Superoinferior Ventricles	–	–	2	3%
Double IVC	–	–	1	2%
PA Atresia or MPA Stump	1	5%	8	13%
S/P PA Banding	1	5%	7	12%
Aorta-PA Anastomosis	–	–	4	7%
Marfan Syndrome	2	10%	3	5%
MR Artifact (Aorta)	1	5%	9	15%
MR Artifact (PA)	2	10%	13	22%

**Segmentations:** As described above, the 20 images in the HVSMR dataset already contained ground truth segmentations of the blood pool and ventricular myocardium. A trained rater manually divided each blood pool model into its constituent parts by dropping landmarks at the relevant interfaces and fitting a local separating plane, thus creating the HVSMR+ dataset.

The 40 new HVSMR++ images were segmented using a pipeline that leveraged the existing 20 HVSMR+ images and segmentations (Fig. 2-1). A tool originally designed for valve contouring in ultrasound [102, 103] was used to quickly annotate roughly planar interfaces between the different heart structures, and these contours were superimposed onto the image grid. A 3D U-Net convolutional neural network trained on the HVSMR+ dataset was applied to each new image. This network did not perform well after it was trained using a such a small dataset, and its segmentation of the eight heart chambers and great vessels was often very inaccurate (although the algorithms that we had already developed at this point in time were more accurate,



Custom module to manually contour all valves and septal defects (“holes” in heart)



Contours onto image grid

Combine with output from automatic 3D CNN

Manually reassign islands and cleanup further

Figure 2-1: Ground truth HVSMR++ segmentations were created using a pipeline that merged manual annotations with model outputs.

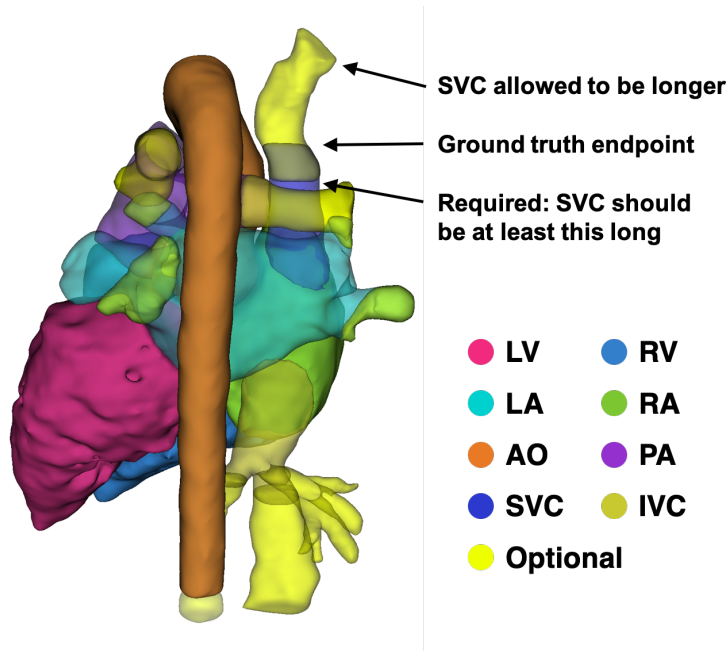


Figure 2-2: Optional zones in the ground truth vessel segmentations.

we did not want to bias the ground truth towards our methods). Nevertheless, once combined with the manually contoured interfaces, segmentation cleanup via island relabeling and further painting or erasing was faster than manual segmentation from scratch. The boundaries were carefully reviewed and adjusted to avoid bias towards the neural network output as much as possible, and the resulting segmentations were validated by hospital expert when the trained rater deemed the correct segmentation to be ambiguous.

Note that this pipeline for ground truth annotation might be bootstrapped in future to create bigger and bigger labeled datasets: once more images have been segmented, they can be used to train a better model whose outputs require less manual editing, and so on.

Results from a previous whole heart segmentation challenge noted that fair evaluation can be problematic when vessel lengths are not standardized in the ground truth [24]. One may not want to penalize algorithms that produce vessel segmentations that are slightly too short or too long, as this does not make the segmentations any less clinically useful [24,49,50]. To this end, for all sixty images we created ground truth segmentations with consistent endpoints that were based on cardiac landmarks

for model training, but also defined “optional zones” that defined a minimum required vessel length and a permitted continuation (Fig. 2-2). More details are provided in Table 2.2.

**Seed Points:** Seed points were used to simulate a user click within each structure. These were manually created for the 20 images in HVSMR+ and automatically defined for the sixty images in HVSMR++. More details are provided in Table 2.3.

**Future Work:** The HVSMR++ dataset could be made more balanced by adding more D-Loop TGA, L-Loop TGA, arterial switch and atrial switch patients. We chose to reduce the complexity of the dataset by not including patients with Tetralogy of Fallot, hypoplastic left heart syndrome or truncus arteriosus, but these conditions could be added in future. Finally, the ground truth PA branch length was defined with respect to a landmark that can be easily and reproducibly localized (the truncus anterior on the right PA branch). This leads to PA branches whose lengths vary across subjects. Another potential option is to use a constant branch length instead.

## 2.4 Segmentation Evaluation

### Dice Score:

We use the Dice score throughout this thesis to quantify segmentation accuracy.

Given a ground truth segmentation  $\mathbf{y}$  of an object, the Dice score for a given segmentation  $\hat{\mathbf{y}}$  is

$$DSC(\mathbf{y}, \hat{\mathbf{y}}) = 100 \cdot \frac{2|\mathbf{y} \cap \hat{\mathbf{y}}|}{|\mathbf{y}| + |\hat{\mathbf{y}}|}, \quad (2.1)$$

which is a volume overlap score where  $DSC(\mathbf{y}, \hat{\mathbf{y}}) = 100$  indicates perfect overlap and  $DSC(\mathbf{y}, \hat{\mathbf{y}}) = 0$  indicates no overlap.

Note that the Dice score is more sensitive to segmentation errors for small or thin structures (e.g., the myocardium and vessel walls, IVC, and SVC) than in larger structures (e.g., the entire blood pool or the four cardiac chambers).

If one does not want a vessel segmentation to be penalized for being slightly too



Table 2.2: Ground truth definitions of each cardiac structure and their optional zones.

<b>LV</b>	Typically bordered by the mitral valve, aortic valve, and/or VSD (if present). If there is a single ventricle, it is labeled as LV. As advised by cardiologists, the papillary muscles are not included in order to create realistic 3D heart models.
<b>RV</b>	Typically bordered by the tricuspid valve, pulmonary valve, and/or VSD (if present). The LV and RV are differentiated by considering chamber shape, wall thickness, presence/absence of trabeculations, and the radiology report. As advised by cardiologists, the trabeculations are not included in order to create realistic 3D heart models.
<b>LA</b>	Typically bordered by the mitral valve and/or ASD (if present). If there is a common atrium, it is labeled as LA. The ground truth includes the pulmonary veins (PVs) until they branch. The PVs can be optionally shorter, this was defined by manually cutting each PV to require its stump only.
<b>RA</b>	Typically bordered by the tricuspid valve, SVC insertion, IVC insertion, and/or ASD (if present).
<b>AO</b>	From the aortic valve through the ascending and descending aorta, until the most inferior level of the LV/RV/LA/RA/PA. Includes two ascending aorta branches for AO-PA anastomosis. Can optionally continue to the bottom of the image.
<b>PA</b>	Typically includes the main PA trunk from the pulmonary valve to the bifurcation point, plus the left and right branches with equal length (defined by the distance from the bifurcation point to behind the truncus anterior on the right hand side). For Glenn/Fontan patients, there is no main PA, only two branches. Effort was made to track through MR inhomogeneity artifacts; if this was impossible then any disconnected segments were labeled as optional. The distal $\sim 25\%$ of each branch in the ground truth is optional. Each branch can optionally continue until it splits at its lower lobe anterior basal segmental branch.
<b>SVC</b>	From the axial slice at the level of its bifurcation into the brachiocephalic veins, down to its insertion into the attached atrium (angled according to atrium curvature) or the PA branches (Glenn/Fontan patients). A second SVC may also be present (bilateral SVC). The superior $\sim 25\%$ in the ground truth is optional. The right/left SVC can optionally continue higher, through the right/left brachiocephalic vein and right/left internal jugular vein, respectively.
<b>IVC</b>	From its insertion into the attached atrium (angled according to atrium curvature) or inferior level of the PA branches (Fontan patients, baffle included), down through the hepatic segment and subsequent branching. The ground truth was defined by identifying the level of the first bifurcation, counting down by 5% of the image height, and then dilating the pre-bifurcation segment to this level (so that branches are cut at an angle). The non-optional segment was defined by repeating this using the lowest axial slice in which the IVC appeared round (i.e., above any branching) and counting down by $2/3 \cdot 5\%$ of the image height. Can optionally continue branching to the bottom of the image.

Table 2.3: Seed points in the HVSMR+ and HVSMR++ datasets. References to “10 axial slices” assumes that the image’s height is 180 slices. The actual number of slices used is proportional to the image’s actual height.

	<b>Summary</b>	<b>Automatic Localization (HVSMR++)</b>
<b>LV</b>	Center region	Centroid of its midaxial slice.
<b>RV</b>	Center region	Centroid of its midaxial slice.
<b>LA</b>	Center region	Centroid of its midaxial slice, after morphological erosion to remove PVs.
<b>RA</b>	Center region	Centroid of its midaxial slice.
<b>AO</b>	Bottom of descending aorta	Centroid of the segment extracted from its bottom 10 axial slices.
<b>PA</b>	Bottom of main PA trunk, or midpoint between the two PA branches (Glenn/Fontan patients without a main PA stump)	Centroid of the segment extracted from the bottom 10 slices of the main PA trunk (via dilation of attached LV/RV, or taking the bottom 10 axial slices if the main PA is unattached), or centroid of the segment at the intersection between the two PA branches via manual dilation of the PA optional zones (Glenn/Fontan patients without a main PA stump).
<b>SVC</b>	At its superior end	Centroid of the segment extracted from its top 10 axial slices.
<b>IVC</b>	Center of hepatic segment, or below IVC-PA connection (Fontan patients)	Centroid of the segment extracted above any branching, or centroid of the segment extracted from its top 10 axial slices (Fontan patients).

short or too long compared to the ground truth segmentation, the optional zone is subtracted from the ground truth segmentation and the predicted segmentation before computing the Dice score. In this way, only the “required” regions are compared.

#### **Average Boundary Distance and Hausdorff Distance:**

In addition to the Dice score, we also used the average boundary distance and the Hausdorff distance to evaluate submissions to the HVSMR 2016 Challenge. Both of these evaluate physical distances between spatial coordinates on the ground truth

segmentation boundary  $\partial\mathbf{y}$  and the given segmentation boundary  $\partial\hat{\mathbf{y}}$ . In the following equations,  $d(\mathbf{v}, \hat{\mathbf{v}})$  is the Euclidian distance between two spatial coordinates  $\mathbf{v}$  and  $\hat{\mathbf{v}}$ .

The average boundary distance is

$$ABD(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2} \left( \frac{1}{|\partial\mathbf{y}|} \sum_{\mathbf{v} \in \partial\mathbf{y}} \min_{\hat{\mathbf{v}} \in \partial\hat{\mathbf{y}}} d(\mathbf{v}, \hat{\mathbf{v}}) + \frac{1}{|\partial\hat{\mathbf{y}}|} \sum_{\hat{\mathbf{v}} \in \partial\hat{\mathbf{y}}} \min_{\mathbf{v} \in \partial\mathbf{y}} d(\mathbf{v}, \hat{\mathbf{v}}) \right), \quad (2.2)$$

where  $ABD(\mathbf{y}, \hat{\mathbf{y}}) = 0$  indicates perfect overlap and a large  $ABD(\mathbf{y}, \hat{\mathbf{y}})$  indicates that points on the two segmentation boundaries are far apart on average.

The Hausdorff distance is

$$HD(\mathbf{y}, \hat{\mathbf{y}}) = \max \left\{ \max_{\mathbf{v} \in \partial\mathbf{y}} \min_{\hat{\mathbf{v}} \in \partial\hat{\mathbf{y}}} d(\mathbf{v}, \hat{\mathbf{v}}), \max_{\hat{\mathbf{v}} \in \partial\hat{\mathbf{y}}} \min_{\mathbf{v} \in \partial\mathbf{y}} d(\mathbf{v}, \hat{\mathbf{v}}) \right\}, \quad (2.3)$$

where  $HD(\mathbf{y}, \hat{\mathbf{y}}) = 0$  indicates perfect overlap and a large  $HD(\mathbf{y}, \hat{\mathbf{y}})$  indicates that there exists at least one point on one segmentation that is far from all of the points on the other segmentation.

## 2.5 Summary

In this chapter, we have described three imaging datasets that we have built to expose the vast range of heart defects in congenital heart disease. Our efforts in making these datasets public have been instrumental in opening new developments in medical image analysis for congenital heart disease. For our purposes, they have supported the development and evaluation of novel methods for whole heart segmentation, which are described in the next chapters.



# Chapter 3

## Patch-Based Interactive Segmentation with Active Learning

In this chapter, we focus on segmenting the blood pool, myocardium and background in cardiac MRI. To the best of our knowledge, this chapter represents one of the first demonstrations towards clinically practical segmentation for patients with CHD in order to enable routine use of 3D heart models for surgical planning [81].

### 3.1 Background

Image patches are small blocks of pixels (e.g.,  $3 \times 3$ ,  $5 \times 5$ , or  $7 \times 7$ ). Patches are often used for image analysis because they provide more context than individual pixel intensities while remaining relatively low dimensional, and because many patches can be extracted from few images [104–108]. Patch-based segmentation is an established technique in medical image analysis to transfer segmentation labels across subjects [107, 108]. An advantage of patch-based segmentation methods is that they are not limited by the accuracy of an inter-subject deformable registration step, which is especially error prone when there is large motion. In fact, they do not require any explicit spatial correspondence detection. Instead, the label propagation is much more flexible: following a rough image alignment (e.g., via affine registration), each small patch of the new image is segmented by finding the most similar patches from

the previously segmented atlas images and then fusing their labels.

Patch-based segmentation can be seen as an instantiation of the k-nearest neighbors algorithm, in which the training data  $(p_1, l_1), (p_2, l_2), \dots, (p_n, l_n)$  consists of  $d \times d$  intensity patches  $p_i \in \mathbb{R}^{d^2}$  with associated labels  $l_i \in \{1, \dots, L\}$  (classification of the central voxel only) or  $l_i \in \{1, \dots, L\}^{d^2}$  (“multi-point” classification of all voxels). One must define the patch similarity measure that is used to identify the most similar training patches. This can consider for example patch intensities, gradients and/or location, typically after spatially constraining the search. Atlas selection can also be used to select the best training images or regions to consider [55, 56]. One must also choose a label fusion strategy, e.g., weighted or unweighted majority voting, which can consider local and/or global features [47, 109].

Patch-based segmentation has been shown to perform very well in multiple application domains, including for cardiac MRI [110–113]. Previous research has investigated different patch similarity measures, e.g., by incorporating anatomical landmarks [112] or spectral features [110], and label fusion strategies, e.g., by learning local support vector machine classifiers that use gradient and context features in addition to patch intensities [111]. Efficient implementations have also been demonstrated [114, 115].

Traditional patch-based segmentation can be more attractive than atlas-based segmentation when anatomical variability is high, but still requires images to be affinely registered and a search window established, so that patches from widely disparate portions of the anatomy are not matched. This is not feasible for congenital heart disease, unless perhaps one had a very large database of segmented images (and a very robust atlas selection procedure) so that information would be transferred only from the relevant atlases.

In this chapter, we propose to use patch-based segmentation *within* a given 3D image volume, presenting the first interactive patch-based segmentation method and demonstrating that it provides accurate whole heart segmentation in CHD. The method uses a small set of manually labeled slices within the 3D image. These regions provide patient-specific information on the heart’s shape and on the local ap-

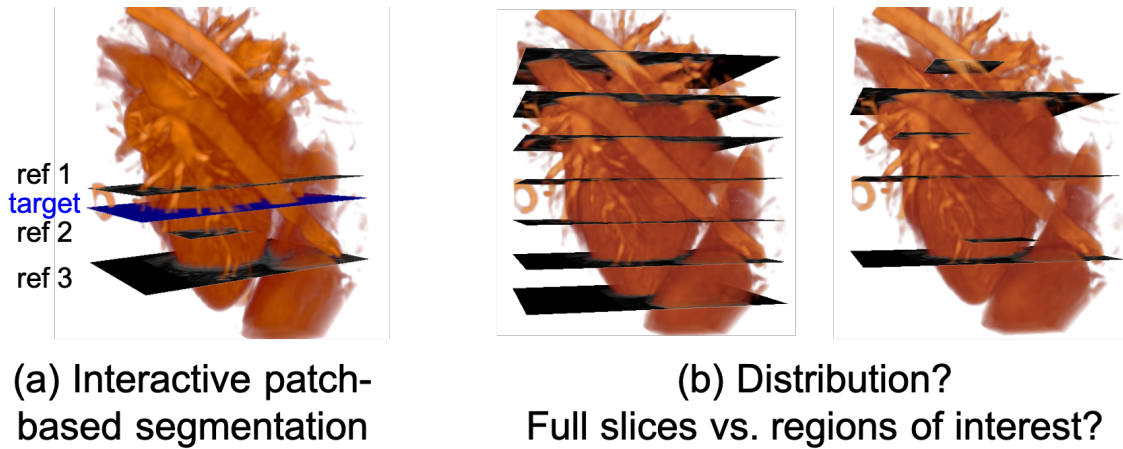


Figure 3-1: (a) Our patch-based interactive segmentation algorithm uses manual segmentations on limited image domains (“reference” slices or regions of interest) to segment the rest of the image (i.e., each “target”) slice. Note that the image has been rotated into a short-axis orientation, so that the apex of the heart points down and the bottom slices show cross-sections of the left and right ventricles. (b) An important consideration is where the user should provide input. A simple option is to uniformly distribute full short-axis slices. However, our interactive patch-based algorithm is very flexible, and annotations could be made on any slice or short-axis region.

pearance of the blood pool, myocardium and surrounding organs, which is exploited by the algorithm to infer labels in the remaining parts of the image (Fig. 3-1a). This approach can adapt to complicated shapes (e.g., the entire blood pool), and using patches allows fine details in the segmentation to be maintained (e.g., the thin walls and valves that separate the vessels and atria).

We decided to work with short-axis slices because clinicians are already accustomed to segmenting short-axis views for making cardiac function measurements such as ejection fraction [116]. The short-axis orientation is a standard radiological view in cardiac imaging, and is defined by displaying the slices whose normals are parallel to the “long-axis” line that intersects the center of the mitral valve (between the LV and LA) and the heart’s apex (the tip at the bottom of the heart). In addition, the overall mass of the heart’s shape is fairly well aligned with the long-axis, making it easier to propagate labels up and down in short-axis slices than in the axial slices in which the data is acquired.

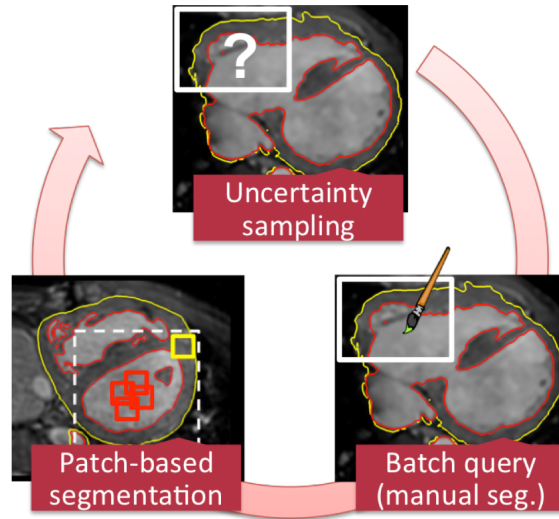


Figure 3-2: An active learning loop for image segmentation comprises three steps: (1) uncertainty sampling to decide where the user should provide input; (2) a batch query in which the user manually labels many voxels; and (3) re-running the segmentation algorithm using the user’s new inputs.

Moreover, we examine active learning methods to further reduce the number of interactions. Specifically, as depicted in Fig. 3-1b, where should the user provide manual annotations for optimal performance?

At each step of an active learning session, the algorithm directs the user to manually label part of the data deemed most informative [117]. These methods aim to achieve the same accuracy with fewer user interactions compared to systems in which the user decides where to provide input. Most active learning methods for interactive medical image segmentation rely on uncertainty sampling with a batch selection query strategy [118–123]. This active learning loop is illustrated in Figure 3-2. In the uncertainty sampling step, the algorithm selects the voxels in which it is least confident. Confidence can be measured using image-based metrics [120, 123], label probability maps [119, 122], ensemble methods that assess the disagreement among votes [118], or SVM classifiers that choose the data to query based on the distance to the margin [121]. A batch query then asks the user to label multiple voxels in each interaction step. A query can involve annotating sets of the most informative voxels [118, 119, 121], segmenting entire slices [120, 123] or deciding whether or not to include an entire hypothesized object [122]. Finally, the segmentation algorithm is



re-run once the new annotations are added to the training dataset.

Within our patch-based interactive segmentation framework for high-quality segmentation in CHD, we explore the potential benefits of active learning with batch queries based on uncertainty sampling. We show that methods that select entire slices for manual delineation fail to perform significantly better than a simple strategy based on a uniform distribution of the input slices. In contrast, active learning queries that asks the user to segment regions of interest (ROIs) within short-axis planes are more accurate with less user interaction.

## 3.2 Patch-based Interactive Segmentation

In this section we describe our patch-based interactive segmentation algorithm that incorporates user annotations. The method also provides a baseline for our study of active learning strategies for cardiac MRI segmentation.

Given image  $I : \Omega_I \rightarrow \mathbb{R}$ , where the image domain is  $\Omega_I \in \mathbb{R}^3$ , we seek a label map  $L : \Omega_I \rightarrow \{bp, myo, bg\}$  that parcellates image  $I$  into blood pool, myocardium and background. For the purpose of creating 3D heart models, the myocardium class includes the thick ventricular myocardium, the papillary muscles, and the walls surrounding the atria and great vessels, i.e., the heart model’s “shell” that can be 3D-printed or displayed.

At each step of the interactive segmentation procedure, the user is presented with a 2D slice or region of interest, and asked to manually segment it. We denote the set  $\mathcal{R}_I$  of manually segmented reference regions  $r_i$  containing voxel intensities with associated labels as

$$\begin{aligned}
 r_i : \Omega_i &\rightarrow (\mathbb{R}, \{bp, myo, bg\}), \\
 \text{where } \Omega_i &\in \mathbb{R}^2 \text{ and } \Omega_i \in \Omega_I, \\
 \mathcal{R}_I &= \{r_i\}.
 \end{aligned}
 \tag{3.1}$$

Each reference domain  $\Omega_i$  is defined on a short-axis plane. In the simplest case,  $\Omega_i$  is an entire short-axis slice plane, but it may represent a smaller region within one. In our baseline algorithm, the expert segments entire short-axis slices that are uniformly

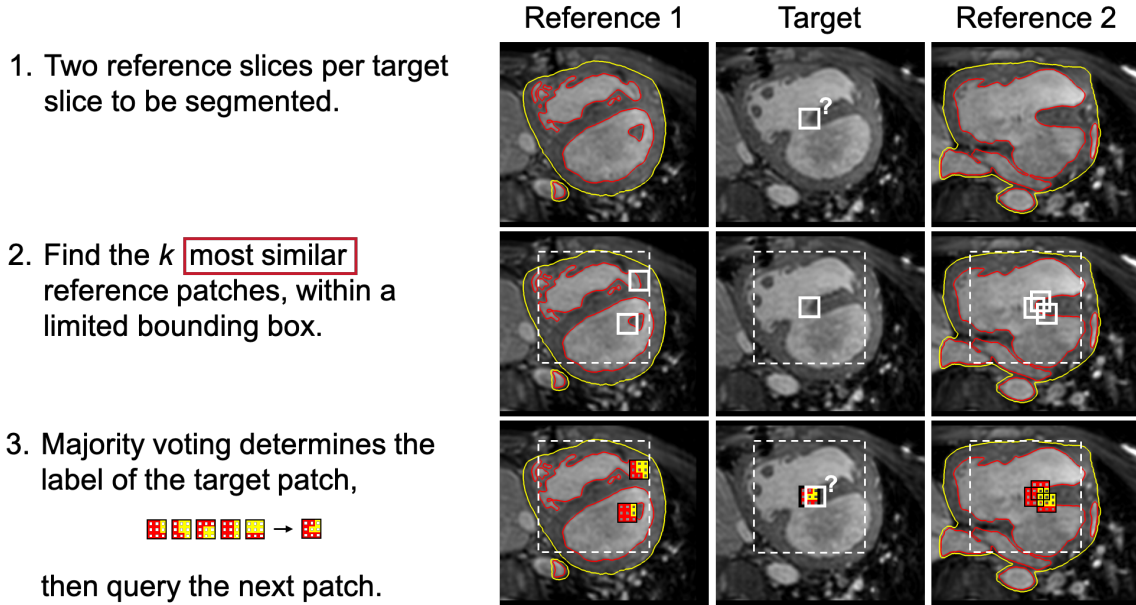


Figure 3-3: The steps of patch-based segmentation, illustrated for a single target patch in which two entire reference slices are available. We use multipoint estimation, so each reference patch carries one label per voxel, and not a single label for the central voxel in the patch. Therefore, the eventual label for each voxel depends on contributions from all overlapping patches.

distributed in the MRI volume.

Once the user annotations are provided, a patch-based method is used to update the segmentation volume [107, 108]. An overview of the algorithm is shown in Fig. 3-3. The search region, patch similarity measure, and label fusion are as follows.

### Search Region:

Not all manually segmented reference regions are useful when segmenting a given target slice  $t$ , since the physically closest references are most informative. In Fig. 3-4, we illustrate each target's set of relevant reference regions  $\mathcal{R}_t \in \mathcal{R}_I$ , from which a library of intensity patches with corresponding labels will be constructed.

If all of the reference domains  $\Omega_i$  are entire short-axis slices, each remaining target slice in the volume is segmented using the two closest reference slices, one above and one below. If there are smaller ROIs, each target slice is segmented using patches from the two closest entire reference slices plus all of the ROIs between them. An

ROI segmentation “shadows” the region behind it.

Once the set  $\mathcal{R}_t$  of relevant reference regions is found, the search window for each target patch is further limited to a 2D in-plane bounding box centered around it.

### Patch Similarity Measure:

To segment patch  $p_t(x^i)$  centered at voxel position  $x^i$  in target slice  $t$ , we must find the  $K$  most similar labeled patches in  $\mathcal{R}_t$ . Here, we describe a patch similarity measure that we have developed for patch-based interactive segmentation.

We use  $x = [x_1, x_2, x_3]$  to denote the three coordinates of position  $x$ , where  $x_1$  and  $x_2$  are in-plane (short-axis) coordinates and  $x_3$  is the out-of-plane coordinate.

Given a patch  $p_r(x^j)$  centered at voxel position  $x^j$  in a reference  $r \in \mathcal{R}_t$  with

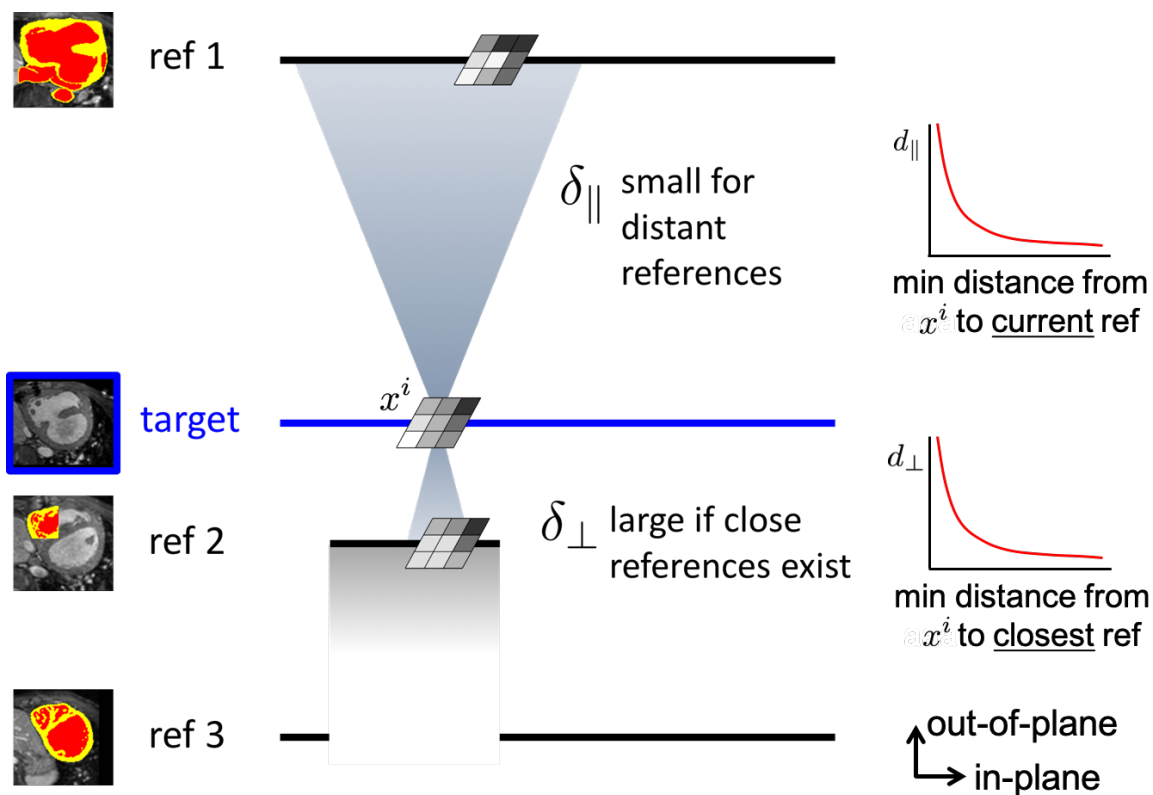


Figure 3-4: Example setup of our patch-based interactive segmentation, in which a target slice is segmented using two entire short-axis reference slices and a smaller region of interest between them. We also illustrate the exponential curves for the spatially adaptive in-plane and out-of-plane position weights  $\delta_{\parallel}(x^i, \Omega_r)$  and  $\delta_{\perp}(x^i, \mathcal{R}_t)$ , respectively

domain  $\Omega_r$ , the distance between patch  $p_t(x^i)$  and patch  $p_r(x^j)$  depends on the patch intensities, gradients and positions:

$$d(p_t(x^i), p_r(x^j)) = \alpha \|p_t(x^i) - p_r(x^j)\|^2 + \beta \|\nabla p_t(x^i) - \nabla p_r(x^j)\|^2 + \delta_{\parallel}(x^i, \Omega_r) \left[ (x_1^i - x_1^j)^2 + (x_2^i - x_2^j)^2 \right] + \delta_{\perp}(x^i, \mathcal{R}_t) (x_3^i - x_3^j)^2. \quad (3.2)$$

Here,  $\alpha$  and  $\beta$  are weights on the relative importance of the intensity and gradient terms, respectively. The position weights  $\delta_{\parallel}(x^i, \Omega_r)$  and  $\delta_{\perp}(x^i, \mathcal{R}_t)$  are for the in-plane and out-of-plane components of the positions, respectively. Making them spatially adaptive means that they can adjust to the geometry of each situation, as illustrated in Fig. 3-4. Essentially, the out-of-plane position weight does a soft selection amongst the relevant reference regions, while the in-plane position weight determines an effective search window tailored to each reference.

First, we want the out-of-plane position weight  $\delta_{\perp}(x^i, \mathcal{R}_t)$  to be high when the target patch is physically close to any one of its reference regions, to encourage matching to that close reference since it likely contains the same structures. To this end, the out-of-plane position weight is defined as a function of the distance from  $x^i$  to the closest point within *any* of its references in  $\mathcal{R}_t$ :

$$\delta_{\perp}(x^i, \mathcal{R}_t) = \gamma_1 \exp(-\gamma_2 \cdot D_{\perp}(x^i, \mathcal{R}_t)) + \gamma_3, \quad (3.3)$$

where  $D_{\perp}(x^i, \mathcal{R}_t) = \min_{r \in \mathcal{R}_t} \min_{x^j \in \Omega_r} \|x^i - x^j\|$ .

Second, we focus on the in-plane position weight  $\delta_{\parallel}(x^i, \Omega_r)$  for each reference region  $r$ . If a given reference is physically close to the target slice, then the matching structure for each target patch is probably located at a similar in-plane position within the reference, and  $\delta_{\parallel}(x^i, \Omega_r)$  should be high. In contrast, we want to be able to search more widely within distant reference slices, i.e.,  $\delta_{\parallel}(x^i, \Omega_r)$  should be low. This enables matching of structures that might change shape substantially across neighboring slices, and is especially useful when few references are available. The in-plane position weight is therefore different for *each* reference, and is defined as the

distance from  $x^i$  to the closest point in the reference domain  $\Omega_r$ :

$$\begin{aligned} \delta_{\parallel}(x^i, \Omega_r) &= \lambda_1 \exp(-\lambda_2 \cdot D_{\parallel}(x^i, \Omega_r)) + \lambda_3, \\ \text{where } D_{\parallel}(x^i, \Omega_r) &= \min_{x^j \in \Omega_r} \|x^i - x^j\|. \end{aligned} \quad (3.4)$$

### Label Fusion:

Once the  $K$  nearest neighbor patches from the reference regions are found, the labels of the target patch are determined through majority voting with multipoint estimation. For target patch  $p_t(x^i)$ , we use  $l_t(x^i) \in \{bp, myo, bg\}^{d^2}$  to denote its inferred labels; this is a patch of labels centered at voxel position  $x^i$  in target slice  $t$ . Similarly, we use  $l_{r_k}(x^k) \in \{bp, myo, bg\}^{d^2}$  to denote the labels of the  $k$ -th nearest neighbor patch  $p_{r_k}(x^k)$  centered at voxel position  $x^k$  in a reference region  $r_k$ . Majority voting on the  $K$ -nearest neighbor patches maximizes the voxel-wise label posterior probabilities

$$p(l_t(x^i) = l \mid p_t(x^i), \mathcal{R}_I) = \frac{1}{K} \sum_{k=1}^K \mathbb{1}[l_{r_k}(x^k) = l], \quad (3.5)$$

$$l_t(x^i) = \operatorname{argmax}_{l \in \{bp, myo, bg\}} p(l_t(x^i) = l \mid p_t(x^i), \mathcal{R}_I), \quad (3.6)$$

where  $\mathbb{1}[\cdot]$  is the indicator function and all operations operate element-wise on  $d \times d$  patches. In practice, patches are extracted in a sliding window and overlap each other. This means that the labeling for each voxel is influenced by all of the patches that overlap it, and there are actually  $K \cdot d^2$  votes at each voxel. Multipoint estimation results in a smoother segmentation, which removes the need for additional smoothness constraints that could potentially eliminate small walls inside the heart and surrounding the great vessels.

We also investigated weighted majority voting for label fusion,

$$\begin{aligned} w(p_t(x^i), p_{r_k}(x^k)) &= \exp\left(-\theta \cdot d(p_t(x^i), p_{r_k}(x^k))^2\right), \\ p(l_t(x^i) = l \mid p_t(x^i), \mathcal{R}_I) &= \frac{\sum_{k=1}^K w(p_t(x^i), p_{r_k}(x^k)) \cdot \mathbb{1}[l_{r_k}(x^k) = l]}{\sum_{k=1}^K w(p_t(x^i), p_{r_k}(x^k))}, \end{aligned} \quad (3.7)$$

for a variety of values  $\theta$ , which gives more influence to reference patches that are more similar to the target patch (simple majority voting is a special case where  $\theta = 0$ ), and an adaptive weighted majority voting scheme [107] in which the weight factor  $\theta$  is set at each voxel according to the distance to the first nearest neighbor:

$$\theta = \frac{1}{d(p_t(x^i), p_{r_{k=1}}(x^{k=1})) + \epsilon}, \quad (3.8)$$

with  $\epsilon = 1\text{e} - 4$  a small positive constant. If this distance is small, then  $\theta$  is large and only similar patches have significant votes. If this distance is large, then none of the nearest neighbors are similar to the target patch and they contribute more equally.

### 3.3 Empirical Study: Active Learning for Reference Selection

Here, we investigate different batch query strategies for automatically choosing the reference subdomains  $\{\Omega_i\}$  to be segmented by the user. Our baseline algorithm is called **uniform slice**, and uniformly distributes entire short-axis slices. All other methods are initialized by segmenting three uniformly distributed short-axis slices. Another baseline is **random slice** selection, which is a common baseline in active learning [120, 123]. Neither the **uniform slice** nor the **random slice** strategies require an iterative back-and-forth with the user.

We first evaluate two active learning workflows that rely on the local uncertainty to select either entire short-axis slices or smaller regions of interest. To decouple the effect of the reference domain size from that of a specific uncertainty sampling method, we use a ground truth manual segmentation to identify the next region to be segmented. At each step, the slice or region with the highest cumulative segmentation error (evaluated over its in-plane domain and  $\pm h$  slices) is selected for manual input. We refer to these two iterative approaches as **oracle slice** and **oracle ROI** selection, since the uncertainty estimation is perfect by construction. We emphasize that our goal is to investigate the effect of the interaction mechanism, as this approach is clearly

infeasible for segmentation of novel images. In practice, uncertainty can be measured using metrics that locally estimate segmentation accuracy by assessing the entropy of the patch vote distributions, the alignment of label boundaries with image gradients, or the intensity homogeneity within small regions assigned the same label [120, 123].

We also implemented an iterative **optimal greedy slice** selection that after each step, aims to maximize the overall improvement in segmentation accuracy throughout the entire image volume. In each step, this strategy exhaustively tries each possible new reference slice, evaluates the global segmentation accuracy after its inclusion using the ground truth segmentation, and keeps the slice that maximally reduces error.

## 3.4 Evaluation

### 3.4.1 Data

Validation was performed using a precursor to the HVSMR dataset described in Chapter 2. The dataset includes twenty cardiac MRI scans from patients with a variety of heart defects plus corresponding ground truth segmentations of the blood pool and myocardium, in which the myocardium class included the thick muscle surrounding the two ventricles, the septum between them, the thin walls surrounding the atria and great vessels, and the cardiac valves if visible. In contrast, the HVSMR dataset is a subsequent refinement that includes the ventricular myocardium only.

Each high resolution 3D MR image was manually cropped to a tight region around the heart and rotated into an approximate short-axis orientation (note that there is residual rotation around the long axis that is not consistent across subjects). The final image size after cropping was different for each scan (the average image size was  $\approx 120 \times 150 \times 200$ ). Each image was smoothed slightly using anisotropic diffusion.

The ground truth segmentations of the blood pool and myocardium were used to simulate the user input provided in interactive segmentation.

### 3.4.2 Parameter Selection

For patch-based interactive segmentation, we used  $5 \times 5$  patches and limited the nearest neighbor search to a  $101 \times 101$  in-plane bounding box. Testing found that retrieving  $k = 5$  or  $k = 10$  nearest neighbor patches for each target patch led to the best segmentation accuracy; results are presented for  $k = 10$  here. The performance of weighted majority voting varied little so long as  $\theta < 0.4$  and adaptive weighted majority voting had no appreciable improvement. Here, we present results for simple majority voting label fusion.

The weights governing the relative influence of the intensity, gradient and position terms in the patch similarity measure from eqns. (3.2) - (3.4) were determined empirically using four images. The best values were found to be  $\alpha = 1$ ,  $\beta = 1$ ,  $\gamma = [8.49, 0.02, 0.0375]$  and  $\lambda = [1.62, 0.2, 1.25]$ . To determine the six parameters for the spatially adaptive position weights ( $\gamma$  and  $\lambda$ ), we ran the proposed intra-image patch-based segmentation algorithm multiple times, where each trial used a different number of uniformly distributed short-axis slices and different constant position weights  $\delta_{\perp}$  and  $\delta_{\parallel}$ . We plotted the best performing position weights for each target slice as a function of the distance to the closest reference slice, fit a decaying exponential curve, and verified that using the exponential function outperformed all of the constant position weights  $\delta_{\perp}$  and  $\delta_{\parallel}$  (i.e., using spatially adaptive position weights yielded an improvement). In practice, the in-plane position weights  $\delta_{\parallel}(x^i, \Omega_r)$  were rounded to the nearest member of the set  $\{0.5, 1, 2, 3, 4, 5\}$  in order to reuse results from the  $K$ -nearest neighbor patch lookups across all of our experiments.

For both of the oracle selection strategies (**oracle slice** and **oracle ROI**), we evaluated the local segmentation error on each slice (i.e.,  $h = 0$ ). For **oracle ROI**, we used ROIs of size  $39 \times 39$ .

### 3.4.3 Results

Fig. 3-5 shows example heart models and segmentations created using our patch-based interactive segmentation, after instantiation with 3, 8 and 14 uniformly dis-



tributed reference slices, respectively. The improvement in accuracy when more input is provided is clear. A high quality model could be created using only 14 reference short-axis segmentations out of  $\sim 200$  slices. Even the heart model instantiated with only three reference slices showed a roughly correct global structure, although the underlying segmentation has some errors.

Fig. 3-6 plots segmentation accuracy using **uniform slice** selection for the blood pool and myocardium. The patch-based segmentation method achieved good accuracy using relatively few segmented slices, especially considering the difficulty of whole heart segmentation in CHD. As expected, the error decreased as the simulated user provided more manual input. In particular, manually segmenting 14 of approximately 200 short-axis slices provided a good trade-off between manual effort and segmentation accuracy, achieving a Dice score of  $96.6 \pm 0.6$  for the blood pool and  $86.7 \pm 2.0$  for the myocardium.

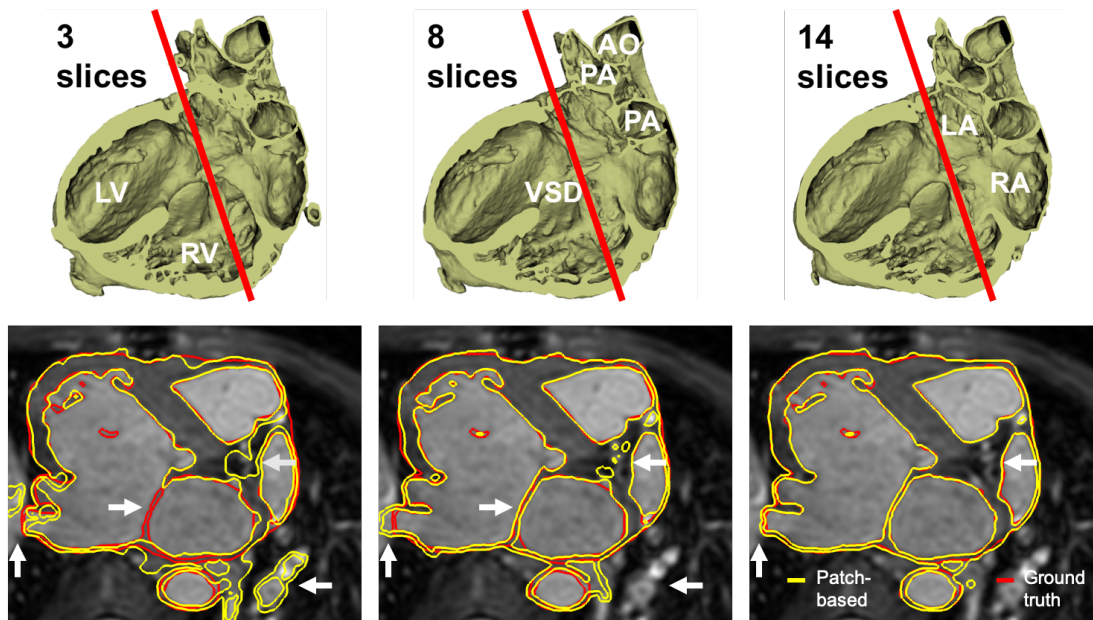


Figure 3-5: Example 3D heart models (cut in half to visualize the interior) and segmentation results for a subject with DORV, from our patch-based interactive segmentation method with 3, 8 and 14 uniformly distributed reference slices. The interactive segmentation results are shown in yellow and the ground truth segmentation in red. Arrows indicate segmentation errors that are corrected by including more reference slices. The red line superimposed onto the 3D heart models indicates the position of the 2D image slice visualized below.

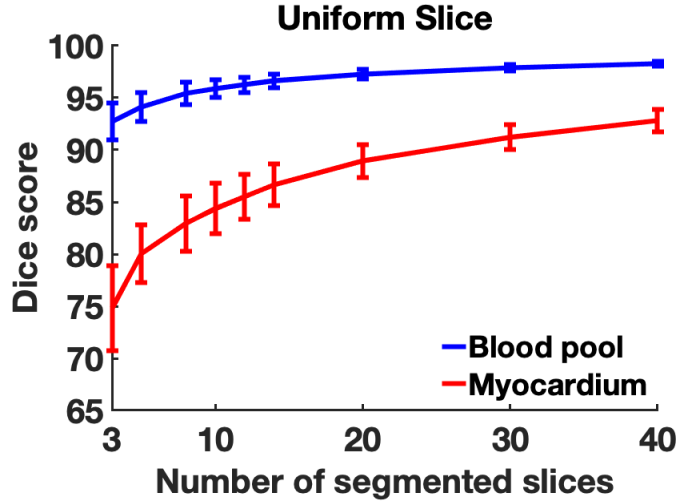


Figure 3-6: Accuracy of patch-based interactive segmentation as a function of the number of uniformly distributed reference slices. Segmentation accuracy was high, and increased with the number of manually segmented slices.

We observed that the slices selected by active learning were sampled more densely near the base of the heart (near the atria and many great vessels) and less so near the apex (near the two ventricles), which correlates with the relative difficulty of segmenting these areas.

Experimental results for active learning are reported in Fig. 3-7. All methods substantially outperformed **random slice** selection, including **uniform slice** selection. **Random** slice selection can leave large gaps between annotated slices, especially when the number of segmented slices is small, which makes it difficult to propagate manual segmentations through the rest of the image. This suggests that random selection is not always the most appropriate baseline when evaluating new active learning methods, although it is widely used.

The **oracle slice** active learning strategy, which selects entire short-axis slices according to their local error, did not achieve a meaningful improvement compared to uniform slice selection. Even the **optimal greedy slice** strategy, which directly measures the impact of adding a slice to the set of manually segmented slices, only showed a modest improvement compared to **uniform slice** distribution. This suggests that there is not much scope for improvement for active learning methods that iteratively choose entire short-axis slices.

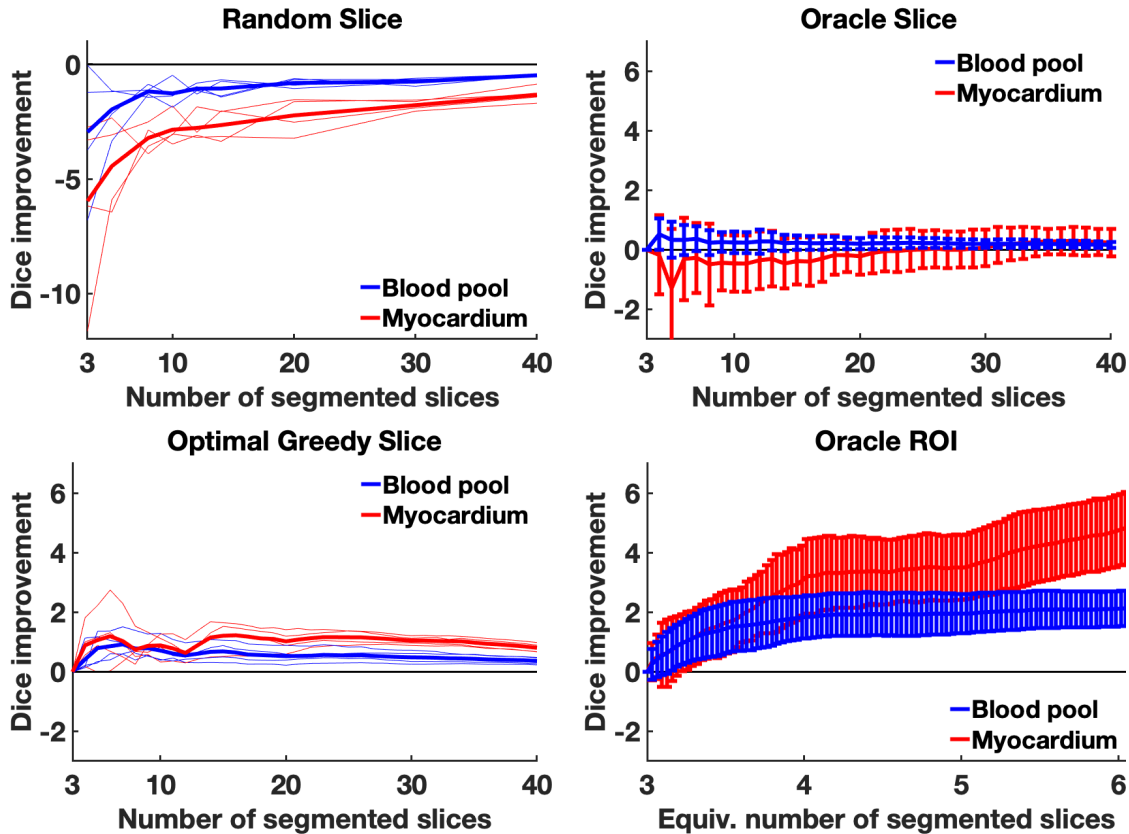


Figure 3-7: Segmentation accuracy of alternative reference selection methods. These are reported as the improvement over **uniform slice** selection, such that negative values indicate that **uniform slice** selection outperforms the method. The **oracle ROI** method is the most promising active learning approach. For experiments using four images (**random slice** and **optimal greedy slice**), thin lines represent each subject and the thick line corresponds to the mean. For experiments using twenty images (**oracle slice** and **oracle ROI**), we show the mean and standard deviation. **Random slice** selection scores are averages over five trials per subject. For **oracle ROI** active learning, the Dice improvement is reported as a function of the cumulative area that the user must segment.

The **oracle ROI** active learning strategy did show a substantial improvement in segmentation accuracy given the same amount of user input ( $\sim 5$  Dice score improvement for the myocardium and  $\sim 2$  Dice score improvement for the blood pool). Compared to asking the user to segment entire slices, having the user segment smaller regions of interest better targets areas with concentrated errors, and leads to more efficient interactive segmentation.

Manual delineation of  $\sim 14$  short-axis slices requires less than one hour of an

expert’s time, versus 8 or more hours for the entire volume. The runtime of our current implementation of patch-based segmentation was roughly one hour per scan. The computation time associated with adding a new reference region is proportional to its size and the number of affected target slices.

### 3.5 Discussion

In this chapter, we have presented a new patch-based interactive segmentation method. We demonstrated that it has high accuracy for whole heart segmentation for patients with congenital heart disease, even when the user manual segments a small proportion of slices that are uniformly distributed in the image volume, e.g.,  $\sim 14/200$ . Our experiments also showed that active learning approaches, in which uncertain regions of interest are identified and labeled by the user, have potential to reduce segmentation time. To the best of our knowledge, this is one of the first works tackling image segmentation for congenital heart disease and aiming to make creating 3D heart models for surgical planning more efficient.

Our results add support to the idea that active learning can benefit interactive segmentation of medical images. Two works that proposed active learning for interactive segmentation workflows in which the user provides dense manual segmentations on arbitrarily oriented slice planes also (1) showed that slice plane selection based on active learning outperformed choosing planes at random, and (2) demonstrated high accuracy (e.g., Dice score  $\geq 90$ ) after relatively few planes were annotated (e.g., 2-8 planes) [120, 123]. These investigations focused on segmenting structures with relatively simple shapes (bones, liver, brain ventricles and putamen, thigh muscles, and the hepatic vein). We found a similar trend for more complex anatomical structures (the cardiac blood pool and our “shell” myocardium class), which our patch-based interactive segmentation method could segment with high accuracy.

Our experiments raised an interesting question of what should be used as a surrogate measure for the amount of user interaction. The results reported in Fig. 3-7 use the total area that is segmented as such a measure. This is similar to previous

studies that count the number of slice planes segmented, the number of landmarks clicked, etc. None of these consider the mental load and user time needed to provide the requested annotations. Several alternative metrics for user effort are plausible, and may depend on whether the user will densely paint labels on all voxels or draw curves delineating different tissue subjects. To this end, we also examined accuracy as a function of the number of edge voxels in the chosen reference regions, which is one proxy of how arduous the region is to segment. When evaluated this way, the differences between **oracle slice** and **oracle ROI** active learning disappeared, and simply uniformly distributing the reference slices may be the best choice. In this case, even the **optimal greedy slice** strategy, which tends to select slice planes with very intricate details, could perform worse than uniform slice selection. A user study evaluating the time required to manually segment slices versus ROIs would be the best way to determine the most appropriate approximation of interaction time. It would also allow us to evaluate the robustness of our approach to user error, since here we simulate an optimal user that provides manual segmentations drawn from the ground truth.

There are several potential future directions for both patch-based interactive segmentation and active learning. In patch-based interactive segmentation, having the user segment slices with different orientations would be useful, whether they come from standard orthogonal axes or have completely arbitrary orientations. Currently, it can be difficult to identify object boundaries that lie roughly in the short-axis plane, since this interface is not reflected in nearby short-axis slices on which the user can provide input. Also, we found that our patch-based algorithm produced rather smooth segmentations, even without an explicit smoothness constraint. This is likely because neighboring target patches tend to match to neighboring reference patches. More explicitly promoting this behavior [105, 124] may be helpful, or it may prove to be too brittle when the number of manually segmented slices is small and the target and reference slices show very different parts of the anatomy. Future work could focus on reducing the algorithm runtime, e.g., through the use of approximate nearest neighbors. In the clinic, patch correspondences could potentially be precom-

puted as well, in the time between image acquisition and when a technician works on the segmentation.

Regarding active learning, recall that model uncertainty for a given data point can be used as a surrogate for how useful labeling that data point might be. Now that we've determined that active learning based on ROIs can improve performance under idealized conditions in which the most "uncertain" voxels are the incorrectly labeled ones, an effective uncertainty measure must be formulated. We found that directly computing an uncertainty map from the patch votes (specifically, by comparing the number of votes for the most popular label versus the second most popular label) did not correlate well with segmentation error (Fig. 3-8). First, some incorrectly segmented regions are not flagged as uncertain. For a given target patch, the  $k$ -nearest neighbor reference patches can be very close spatially with very similar labels, in which case all of them vote identically and uncertainty is low even if the segmentation is incorrect. Second, the label of almost every border voxel is uncertain. This is because the ground truth segmentations, which were created manually, do not segment each boundary in a completely identical way. Even if a target patch is matched appropriately to reference patches that depict the correct edge orientation,

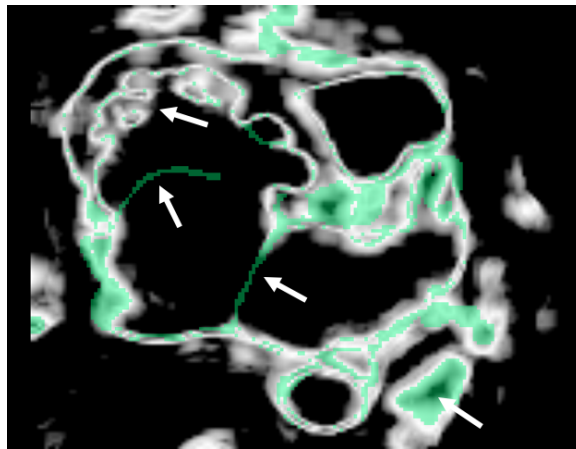


Figure 3-8: Segmentation error (in green) overlaid onto an uncertainty map computed as the margin between the first- and second-place labels (white indicates high uncertainty, black indicates low uncertainty). Note that the intensity distributions in our image made it exceedingly rare for there to be votes for all three classes at an individual voxel, so measuring this margin is appropriate. This uncertainty measure did not correlate well with segmentation error.

the labels of those reference patches may not overlap exactly. Besides finding a better uncertainty measure, working with a more sophisticated active learning model that probabilistically models the expected error reduction given a candidate ROI could also improve segmentation accuracy with minimal user effort [117].

During the course of this doctoral work, features learned from data have steadily superseded handcrafted features. One option for future work would be to learn the patch similarity measure that underpins methods such as the one developed in this Chapter [125, 126]. Alternatively, deep neural networks promise fast inference and high accuracy without the need for extensive feature engineering, and have been successfully applied to many computer vision and medical image analysis problems. In this Chapter, we have provided a method that reduces the amount of user effort for whole heart segmentation in CHD patients, from fully manual segmentation, to dense annotations on relatively few 2D slices, and potentially to segmentations on even smaller regions of interest. In our effort to further decrease the user interaction to a few clicks, we also chose to leverage deep learning. This was partially inspired by the results of the HVS MR Challenge that we held at MICCAI 2016 (see Section 2.2), in which the top performing automatic methods used deep neural networks [127, 128]. However, learning an accurate deep neural network model typically requires a large training dataset when anatomical variability is high. In Chapter 4, we propose a neural network model that can be accurately learned from a very small dataset, and demonstrate improved generalization to subjects with severe heart malformations.





# Chapter 4

## Learning Iterative Segmentation from Limited Data

To the best of our knowledge, this work represents the first whole heart segmentation for CHD that segments the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), aorta (AO), pulmonary artery (PA), superior vena cava (SVC) and inferior vena cava (IVC). It is an extension of our previous work on segmenting the AO and LV using the smaller HVSMR+ dataset [82].

### 4.1 Background

State-of-the-art segmentation methods train a deep convolutional neural network (CNN) to segment an image in one step [64, 65, 129]. These methods aim to optimize the parameters of a model that inputs an image and outputs a probability map, which at each voxel represents the estimated probability of each anatomical label. CNNs [130] use convolution operations followed by a nonlinear activation functions to generate successive sets of learned feature maps that form useful representations for image segmentation. The low-dimensional, shared weights of the convolution layers reduce the number of model parameters, based on the assumption that their output feature maps are computed based on local information that disregards explicit spatial coordinates, i.e., segmentations should be equivariant to translation. CNNs gener-

ally include pooling layers to make the intermediate feature maps smaller, reduce the number of convolutions required to expand the model’s receptive field, and introduce invariance to small translations. Upconvolution layers can be used to create feature maps with finer resolution, and skip connections can be added to combine results from different parts of the network, for example to merge features that are derived from fine details in the input image with features that consider a wider context [65]. In a supervised setting, the CNN parameters are optimized using a training dataset of images with ground truth segmentations. This is typically done by maximizing the empirical log likelihood of the ground truth segmentations under the model parameters. Deep neural networks offer very fast inference with an accuracy that is often similar or better than traditional image analysis methods, but typically require a large training dataset to achieve good performance when the variability in the inputs is high.

An alternative is to learn a model that iteratively segments an image over multiple steps, at each step conditioning on a previous partial solution to make progress towards the final answer. One simple approach that follows this general strategy is to use a cascade of neural networks, e.g., following a network that outputs a coarse segmentation or a region of interest with a fine-grained segmentation network [66, 131–136]. Such a two-step cascade can be iterated further, repeatedly using the estimated segmentation to re-crop the image and repeating until convergence [136, 137]. An iterative strategy is also a natural choice for instance segmentation, which requires separately locating and segmenting all of the objects of each class label in an image. Several approaches have been proposed to sequentially output a segmentation of each object, using an internal memory or an attention mechanism [138–140].

Recurrent neural networks (RNNs) are popular methods for analyzing data sequentially [130]. RNNs model the repeated application of a recursive function, using the same learned parameters at each step. This parameter sharing reduces the model’s size and enables inference on sequences of arbitrary length. At each step, the network inputs information from the previous step via recurrent connections. As illustrated in Fig. 4-1, the global architecture can be constructed in a variety of ways. For

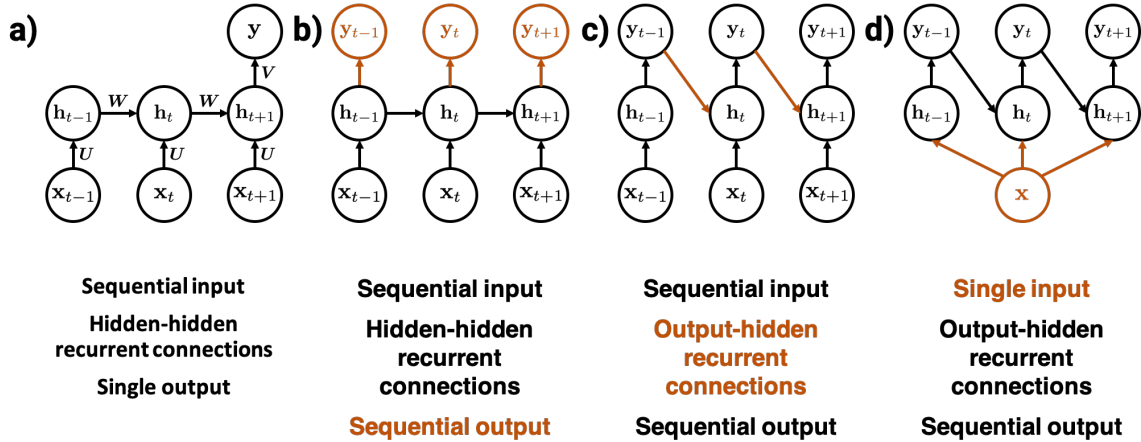


Figure 4-1: Example recurrent neural network (RNN) architectures. (a) An input sequence is processed to produce a single output, with recurrent connections at the hidden layer, e.g. for video classification. Note that the learned network parameters (here,  $U$ ,  $W$  and  $V$ ) are the same at each step). (b) An RNN trained to map an input sequence to an output sequence, e.g. for language translation. (c) In this architecture, the recurrent connections directly use the output of each iteration as an input at the next step. (d) The architecture from (c) modified for a single input: this is the architecture of the RNN developed in this chapter. (Figure adapted from [130]).

example, the input or output can be a sequence or a single image, and the recurrent connections can link the analogous hidden layers of consecutive iterations or connect outputs to hidden units. RNNs have been used to learn dependencies across time (e.g., object tracking [141]), space (e.g., image segmentation [142, 143]) or both time and space (e.g., cardiac cine MRI analysis [144]). They have been widely applied to image generation [145, 146], object recognition [147], human pose estimation [148, 149] and image captioning [150].

This chapter develops an iterative segmentation model, and its RNN implementation, that starts from a user-supplied seed and progressively outlines the entire structure via a sequence of output segmentations. Moreover, we develop a novel loss to learn the model’s parameters so that it grows segmentations in a predictable, task-specific manner. We investigate two variants: one in which the user decides when the segmentation should stop growing, and another in which this stopping point is automatically predicted. Our model is reminiscent of traditional active contours, level

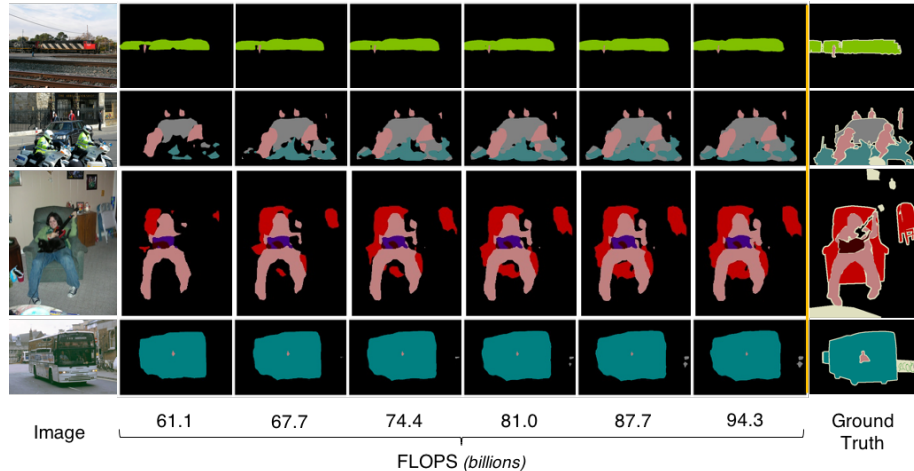


Figure 4-2: Example results of an RNN for image segmentation that is trained with a loss function that considers the final output alone. The segmentation improves over time, but the evolution pattern would be difficult for a user to interact with (reproduced from [153], Copyright © 2018, IEEE).

sets and particle filters [151,152], but leverages the powerful feature learning afforded by modern machine learning. Our aim is to develop an effective segmentation scheme initialized by a single click per structure, while keeping in mind further potential user interaction as it may be needed in our challenging application.

Most previously proposed RNNs for image segmentation use a loss function that evaluates the final output alone or encourages every intermediate segmentation in the output sequence to match the complete ground truth segmentation as much as possible [153,154]. Alternative approaches model level sets [155–157] or sequentially segment small areas pulled from an internal list of potential regions of interest [158]. In practice, this results in output sequences that progressively refine an initial coarse segmentation of the entire object, and/or produce unpredictable growth patterns (Fig. 4-2). Consider a user in-the-loop who aims to monitor the segmentation process and correct errors. For these unpredictable segmentation patterns, it remains difficult for a user to quickly scan for errors (especially considering 3D segmentations) and inject knowledge at intermediate stages - what should the user fix at an intermediate phase to guide the model to the correct answer?

In contrast, our iterative model is trained to produce *predictable* increments in

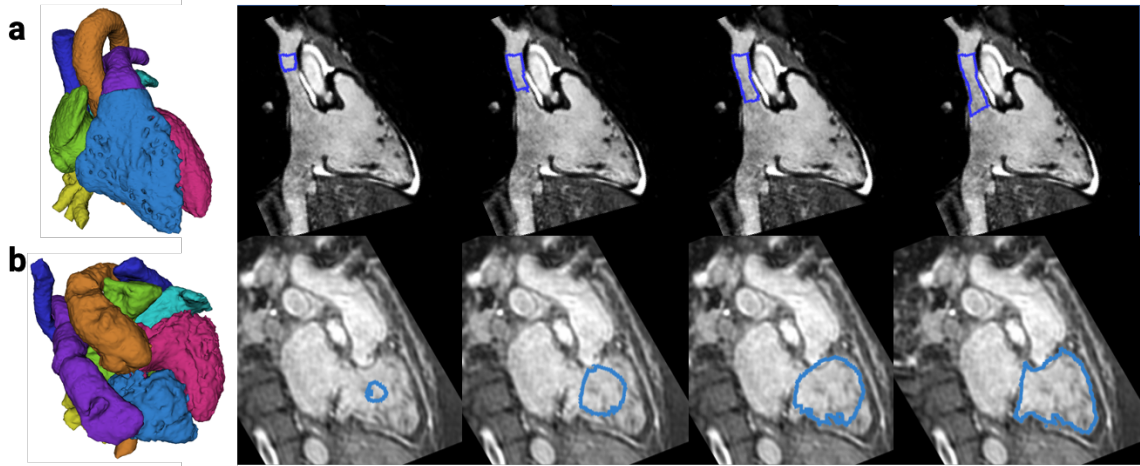


Figure 4-3: Example results from our iterative segmentation model, which evolves segmentations in a predictable way. (a) Example vessel segmentation (SVC) for a heart with normal anatomy. (b) Example chamber segmentation (RV) for a heart with severe malformations and several previous surgeries.

the segmentation. This segmentation pattern is defined by training data as described below. Fig. 4-3 shows some example segmentations in the heart, which are trained to grow along vessel centerlines and spherically outwards towards chamber borders. The user can (1) easily monitor progress, because the region in which growth is expected is spatially limited, (2) easily fix mistakes at intermediate steps, e.g., if a vessel segmentation begins to grow in an incorrect direction, and (3) choose between multiple feasible solutions in the output sequence, e.g., choosing an intermediate segmentation result if the segmentation grows too large or asking for more iterations if it is too small.

To accomplish this, we propose a novel loss that evaluates the entire *sequence* of output segmentations against a desired segmentation trajectory. Maximizing the likelihood of entire observed sequences is known as teacher forcing [130,159]. We show how our loss can be factored into a sum over decoupled time steps, avoiding back-propagation through time and allowing us to learn model parameters from images alongside input-output pairs of partial segmentations. We derive these pairs on-the-fly during training from complete ground truth segmentations. However, our model

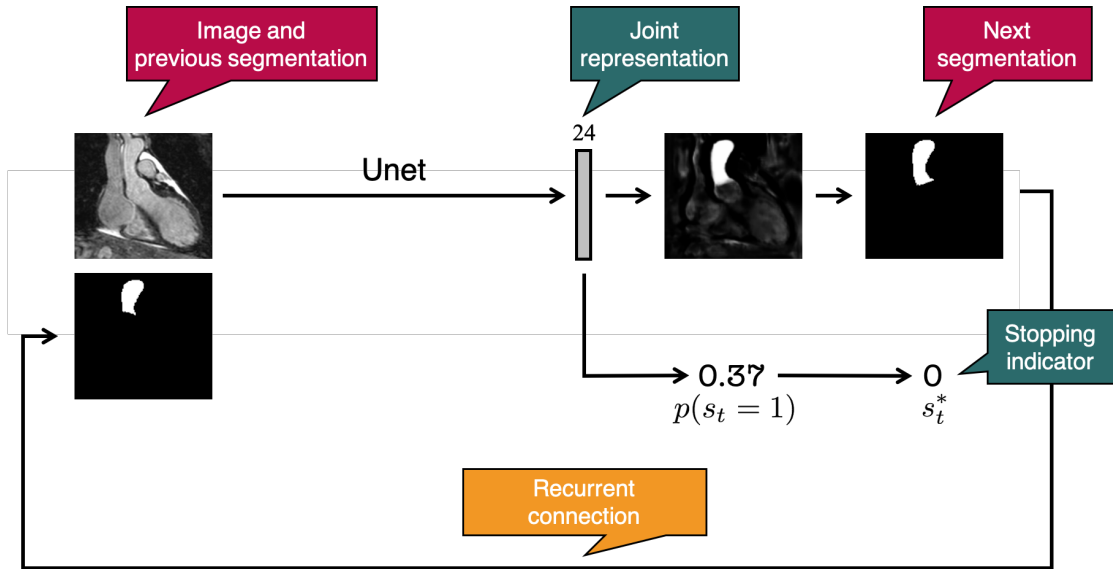


Figure 4-4: Simplified schematic of our RNN architecture for iterative segmentation.

can be trained to follow any desired evolution pattern, which is entirely specified by the training data, because it operates directly on the 3D image grid, unlike approaches that learn to progressively trace a contour [160, 161] or perform slice-by-slice analysis [162, 163].

Our iterative segmentation has several benefits. Iterative segmentation models can learn both local structure and long-range dependencies in the output domain, because inference at each pixel also inputs label estimates for each of the pixels in its receptive field [132, 154]. This has potential to more effectively propagate information from distant landmarks. Another consequence is that the model’s field of view is implicitly expanded without increasing the number of parameters. Empirically, our model better maintains the connectivity of each anatomical structure compared to direct segmentation methods. We also show that it can be learned from very small datasets that do not necessarily include the same pathology present in the test image.

Fig. 4-4 shows a schematic of our RNN architecture. At each step, the image to be segmented and a partial segmentation form the input to a modified U-Net [65] that is trained to predict the next segmentation in the sequence. This segmentation becomes the input segmentation at the next step, via recurrent connections between outputs and the first hidden layer. The U-Net contains a learned representation that

is used to jointly learn a binary stopping indicator that denotes whether the output segmentation is complete.

We validate the proposed iterative segmentation model using a dataset of 3D cardiac MRI scans from patients with a diversity of CHD types. We compare it to “direct” segmentation methods that we have developed for this problem which, like all feedforward neural networks, segment the image in a single step. We show better generalization in the context of learning with limited training data, by demonstrating improved segmentation accuracy in subjects with the most severe cardiac malformations when learning using small datasets.

## 4.2 Iterative Segmentation Model

Given an image  $\mathbf{x} : \Omega \rightarrow \mathbb{R}$  and an initial segmentation  $\mathbf{y}_0 : \Omega \rightarrow \{0, \dots, L - 1\}$ , we seek a segmentation label map  $\mathbf{y} : \Omega \rightarrow \{0, \dots, L - 1\}$  that assigns one of  $L$  anatomical labels to each voxel in image  $\mathbf{x}$ . Although we focus on binary segmentation of each object in this chapter (i.e.,  $L = 2$ ), the model is easily extended to jointly evolve the segmentations of multiple objects. In practice, the initial segmentation  $\mathbf{y}_0$  for each anatomical structure is created by centering a small sphere around a seed point placed by the user.

### 4.2.1 Probabilistic Model

We model the segmentation label map  $\mathbf{y}$  as the endpoint of a sequence of segmentations  $\mathbf{y}_0, \dots, \mathbf{y}_T$  that captures a growing and evolving portion of the anatomy of interest. In particular,  $\mathbf{y}_t : \Omega \rightarrow \{0, \dots, L - 1\}$  for time steps  $t = 0, \dots, T$ .

The number of iterations required to segment a given image depends on the shape and size of the object. To capture this, we introduce a sequence of stopping indicators  $s_0, \dots, s_T$ , where  $s_t \in \{0, 1\}$  and  $s_t = 1$  indicates that the segmentations should finish evolving at  $\mathbf{y}_t$ . Note that  $s_0 = 0$  always.

Given an image  $\mathbf{x}$ , we assume pairs of segmentations and stopping indicators

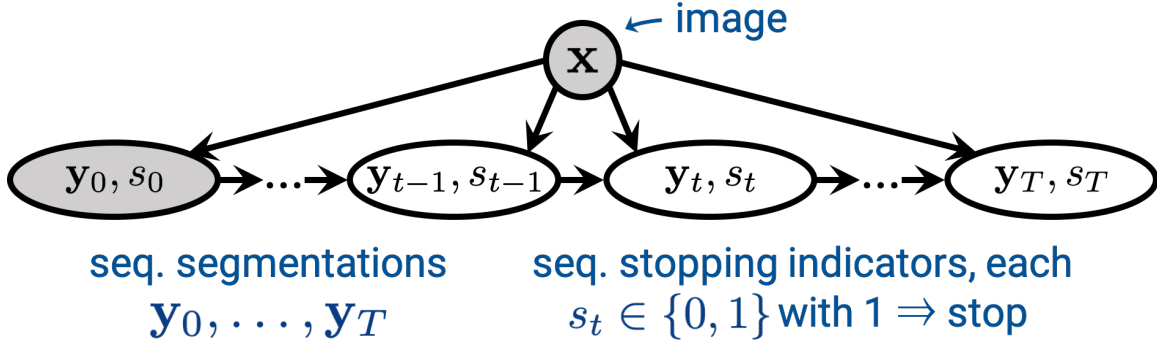


Figure 4-5: Probabilistic model for the proposed iterative segmentation. Given an image  $\mathbf{x}$ , we assume that pairs of segmentations and stopping indicators  $\{\mathbf{y}_t, s_t\}$  follow a first order Markov chain. Shaded nodes indicate observed variables.

$\{\mathbf{y}_t, s_t\}$  follow a first order Markov chain (Fig. 4-5), i.e.,

$$p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_0, \dots, \mathbf{y}_{t-1}, s_0, \dots, s_{t-1}) = p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1}), \quad (4.1)$$

which leads to the recursion

$$p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_0, s_0) = \sum_{\mathbf{y}_{t-1}} \sum_{s_{t-1}} \underbrace{p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1})}_{\text{transition probability}} \cdot \underbrace{p(\mathbf{y}_{t-1}, s_{t-1} | \mathbf{x}, \mathbf{y}_0, s_0)}_{\text{recursive definition}}. \quad (4.2)$$

## 4.2.2 Transition Probability Model

To complete the recursion in eqn. (4.2), the transition probability  $p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1})$  must be defined. We do this by separately considering the cases  $s_{t-1} = 1$  and  $s_{t-1} = 0$ .

When  $s_{t-1} = 1$ , the segmentation  $\mathbf{y}_{t-1}$  is the finished segmentation, and the transition model is trivial:

$$p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 1) = \mathbb{1}[\mathbf{y}_t = \mathbf{y}_{t-1}] \cdot \mathbb{1}(s_t = 1), \quad (4.3)$$

where  $\mathbb{1}[\cdot]$  denotes the indicator function.

When  $s_{t-1} = 0$ , the segmentation's evolution is not yet finished. We introduce a



deterministic latent representation

$$\mathbf{h}_t = h(\mathbf{x}, \mathbf{y}_{t-1}) \tag{4.4}$$

that captures all of the necessary information from the given image  $\mathbf{x}$  and previous segmentation  $\mathbf{y}_{t-1}$ , i.e.,

$$p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 0) = p(\mathbf{y}_t, s_t | \mathbf{h}_t). \tag{4.5}$$

We model the segmentation  $\mathbf{y}_t$  and stopping indicator  $s_t$  as conditionally independent given the latent representation  $\mathbf{h}_t$ :

$$p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 0) = p(\mathbf{y}_t | \mathbf{h}_t) \cdot p(s_t | \mathbf{h}_t). \tag{4.6}$$

This independence assumption is justified by the fact that deciding whether  $\mathbf{y}_t$  is the final segmentation is equivalent to deciding whether  $\mathbf{y}_{t-1}$  is one step from completion, due to the predictable segmentation evolution (see Section 4.2.3). Hence, to estimate the stopping indicator  $s_t$ , knowledge of  $\mathbf{y}_t$  is not informative once we are given  $\mathbf{h}_t$ , which contains information about the image  $\mathbf{x}$  and previous segmentation  $\mathbf{y}_{t-1}$ .

Finally, we model  $h(\mathbf{x}, \mathbf{y}_{t-1})$ ,  $p(\mathbf{y}_t | \mathbf{h}_t)$  and  $p(s_t | \mathbf{h}_t)$  as stationary functions, i.e., they do not depend on the time step  $t$ .

### 4.2.3 Learning

We aim to learn the parameters of a model for the required transition probability  $p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 0)$ . In this section, we develop a loss that considers the segmentation’s evolution in addition to the final prediction.

We begin by assuming access to a training dataset  $\mathcal{D}$  containing images  $\{\mathbf{x}\}$  and variable-length ground truth sequences of segmentations  $\{\mathbf{y}_0, \dots, \mathbf{y}_{T(\mathbf{x})-1}, \mathbf{y}_{T(\mathbf{x})}\}$  and stopping indicators  $\{s_0, \dots, s_{T(\mathbf{x})-1}, s_{T(\mathbf{x})}\} = \{0, \dots, 0, 1\}$ , such that the final segmentation is the sole complete segmentation. This dataset encapsulates the preferred

segmentation growth dynamics.

Following the concept of teacher forcing [130, 159], we seek the parameter values that minimize the expected negative log-likelihood over entire *sequences* of segmentations and stopping indicators, conditioned on the image and initial conditions. In particular, we seek parameters  $\boldsymbol{\theta}^*$  that minimize our loss  $\mathcal{L}(\boldsymbol{\theta})$ , i.e.,

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\theta}), \quad (4.7)$$

where

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}, \mathbf{y}_0, \dots, \mathbf{y}_{T(\mathbf{x})}, s_0, \dots, s_{T(\mathbf{x})} \sim \mathcal{D}} \left[ -\log p(\mathbf{y}_1, \dots, \mathbf{y}_{T(\mathbf{x})}, s_1, \dots, s_{T(\mathbf{x})} | \mathbf{x}, \mathbf{y}_0, s_0; \boldsymbol{\theta}) \right]. \quad (4.8)$$

Expanding the loss term yields

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{x}, \mathbf{y}_0, \dots, \mathbf{y}_{T(\mathbf{x})}, s_0, \dots, s_{T(\mathbf{x})} \sim \mathcal{D}} \left[ \sum_{t=1}^{T(\mathbf{x})} -\log p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1}; \boldsymbol{\theta}) \right], \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}_0, \dots, \mathbf{y}_{T(\mathbf{x})}, s_0, \dots, s_{T(\mathbf{x})} \sim \mathcal{D}} \left[ \sum_{t=1}^{T(\mathbf{x})} -\log p(\mathbf{y}_t | h(\mathbf{x}, \mathbf{y}_{t-1}); \boldsymbol{\theta}) - \log p(s_t | h(\mathbf{x}, \mathbf{y}_{t-1}); \boldsymbol{\theta}) \right]. \end{aligned} \quad (4.9)$$

In the first step, we see that teacher forcing leads to a sum over decoupled time steps, due to the Markov property in eqn. (4.1). This greatly simplifies training, because back-propagation through time is no longer required. In the second step, we see that eqn. 4.9 is an expectation over a segmentation loss and a stopping indicator loss.

Since the loss is a sum over decoupled time steps, training data that predefines entire output sequences is actually unnecessary. The loss can be equivalently minimized using a simplified dataset  $\mathcal{D}'$  consisting of tuples  $\{\mathbf{x}, \mathbf{y}_{in}, \mathbf{y}_{out}, s\}$ , where the segmentations  $\mathbf{y}_{in}$  and  $\mathbf{y}_{out}$  correspond to consecutive time steps and  $s$  denotes whether  $\mathbf{y}_{out}$  is a complete segmentation:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}, \mathbf{y}_{in}, \mathbf{y}_{out}, s \sim \mathcal{D}'} \left[ -\log p(\mathbf{y}_{out} | h(\mathbf{x}, \mathbf{y}_{in}); \boldsymbol{\theta}) - \log p(s | h(\mathbf{x}, \mathbf{y}_{in}); \boldsymbol{\theta}) \right]. \quad (4.10)$$

These input-output pairs can be generated on-the-fly during training: more details on training data generation are provided in Section 4.3.2.

Again, eqn. (4.10) is an expectation over the sum of a segmentation loss and a stopping indicator loss, which are defined as follows. The segmentation  $\mathbf{y}_t$  and stopping indicator  $s_t$  are predicted jointly, and both of their losses influence the parameters for the latent representation  $\mathbf{h}_t$ . This multi-task approach often improves learning, and results in fewer trainable parameters compared to training two separate networks [164, 165].

### Segmentation Loss:

We assume that the label of each voxel in the segmentation  $\mathbf{y}_{out}$  is conditionally independent of all other voxels given  $h(\mathbf{x}, \mathbf{y}_{in})$ . Predicted segmentations can therefore be represented as probability maps, at each voxel storing the parameters of a categorical distribution over  $L$  labels. Given a ground truth output segmentation  $\mathbf{y}_{out}$  and a predicted voxel-wise segmentation probability map  $\hat{\mathbf{y}}_{out}$  that arises from model parameters  $\theta$ , the segmentation loss is therefore a voxel-wise cross-entropy loss, to which we add spatially varying weights:

$$\mathcal{L}_{seg}(\mathbf{y}_{out}, \hat{\mathbf{y}}_{out}) = \sum_{\mathbf{v} \in \Omega} \sum_{l=0}^{L-1} -\omega_{\mathbf{y}_{out}}^l(\mathbf{v}) \cdot \mathbf{y}_{out}^l(\mathbf{v}) \cdot \log \hat{\mathbf{y}}_{out}^l(\mathbf{v}). \quad (4.11)$$

We use spatially varying weights  $\omega_{\mathbf{y}_{out}}^l(\mathbf{v})$  with two goals in mind. The first is class rebalancing, hence the dependence on each label  $l$ . The second is to encourage segmentations to “snap” to the borders, by more strongly penalizing errors near ground truth segmentation boundaries [166], hence the dependence on the ground truth output segmentation  $\mathbf{y}_{out}$ .

First, we define the class rebalancing weights  $\omega^l$  as the inverse label frequencies in the segmentations from the training data, normalized to sum to one.

Second, we introduce a weight map  $\omega_{\mathbf{y}_{out}} : \Omega \rightarrow \{0, \omega_0\}$  that contains a constant boundary weight  $\omega_0 > 0$  for voxels whose minimum distance  $d_{\mathbf{y}_{out}}(\mathbf{v})$  to a boundary

in the ground truth output segmentation  $\mathbf{y}_{out}$  is less than a constant distance  $d_0$ :

$$\omega_{\mathbf{y}_{out}}(\mathbf{v}) = \begin{cases} \omega_0, & \text{if } d_{\mathbf{y}_{out}}(\mathbf{v}) < d_0, \\ 0, & \text{otherwise.} \end{cases} \quad (4.12)$$

Then, the weights  $\omega_{\mathbf{y}_{out}}^l(\mathbf{v})$  are

$$\omega_{\mathbf{y}_{out}}^l(\mathbf{v}) = \omega^l + \omega_{\mathbf{y}_{out}}(\mathbf{v}). \quad (4.13)$$

### Stopping Indicator Loss:

The distribution of the stopping indicator  $s$  is Bernoulli, so the stopping indicator loss is a binary cross-entropy loss, again weighted for class rebalancing. Given a ground truth stopping indicator  $s \in \{0, 1\}$  and a predicted probability of stopping  $\hat{s} \in [0, 1]$  from model parameters  $\boldsymbol{\theta}$ , we have

$$\mathcal{L}_{stop}(s, \hat{s}) = -(1 - \omega_s) \cdot s \log \hat{s} - \omega_s \cdot (1 - s) \log(1 - \hat{s}), \quad (4.14)$$

where the class rebalancing weight  $\omega_s$  is the proportion of training instances in which the stopping indicator equals one.

## 4.2.4 Inference

Since the recursion in eqn. (4.2) is computationally intractable due to the summation over all possible segmentations  $\mathbf{y}_{t-1}$ , given model parameters  $\boldsymbol{\theta}$  we infer  $\mathbf{y}_t$  and  $s_t$  using point estimates, directly from the most likely previous segmentation  $\mathbf{y}_{t-1}^*$  and stopping indicator  $s_{t-1}^*$ :

$$p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_0, s_0) \approx p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}^*, s_{t-1}^*; \boldsymbol{\theta}), \quad (4.15)$$

where  $\mathbf{y}_{t-1}^*, s_{t-1}^* = \underset{\mathbf{y}_{t-1}, s_{t-1}}{\operatorname{argmax}} p(\mathbf{y}_{t-1}, s_{t-1} | \mathbf{x}, \mathbf{y}_0, s_0; \boldsymbol{\theta})$ .

This recursion continues until  $p(s_t = 1 | \mathbf{x}, \mathbf{y}_0, s_0; \boldsymbol{\theta}) > 0.5$ , at which point the segmentation  $\mathbf{y}_t^*$  is deemed the final segmentation and iterative segmentation stops. A user can override this automatic stopping prediction by choosing an earlier segmentation or asking for more iterations.

## 4.3 Recurrent Neural Network

Our RNN has parameters  $\boldsymbol{\theta} = \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s\}$  and implements the recursion

$$\begin{aligned} \mathbf{h}_t &= h(\mathbf{x}, \mathbf{y}_{t-1}; \boldsymbol{\theta}_h), \\ \mathbf{y}_t^* &= \operatorname{argmax}_{\mathbf{y}_t} p(\mathbf{y}_t | \mathbf{h}_t; \boldsymbol{\theta}_y), \\ s_t^* &= \operatorname{argmax}_{s_t} p(s_t | \mathbf{h}_t; \boldsymbol{\theta}_s). \end{aligned} \tag{4.16}$$

Recurrent connections between the hidden layers of successive steps would break the Markov property in eqn. (4.1). Instead, we assume that an image, partial segmentation and stopping indicator capture everything about the current state that is needed to predict the next step [130].

### 4.3.1 RNN Architecture

Our RNN is depicted in Fig. 4-6. It is constructed by joining copies of a 3D U-Net architecture [65] that we augment to model  $p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 0)$ . At each time step  $t$ , the augmented U-Net has  $L + 1$  input channels containing the image to be segmented and a binary mask for each of the anatomical labels in the input segmentation  $\mathbf{y}_{t-1}^*$ , including the background. There are two outputs: the output segmentation  $\mathbf{y}_t^*$ , which becomes the input segmentation in the next time step, and the stopping indicator  $s_t^*$ , which indicates whether the segmentation process is finished.

Recall that in the U-Net architecture, a final bank of learned feature maps is used to produce the final segmentation probability map. In our RNN, these learned feature maps form the latent representation  $\mathbf{h}_t = h(\mathbf{x}, \mathbf{y}_{t-1}; \boldsymbol{\theta}_h)$  that includes all information from the image  $\mathbf{x}$  and input segmentation  $\mathbf{y}_{t-1}$  needed to jointly predict the output

segmentation  $\mathbf{y}_t$  and stopping indicator  $s_t$ . The size of  $\mathbf{h}_t$  equals the dimensions of image  $\mathbf{x}$  multiplied by the number of channels  $C$  (note that this is not a bottleneck layer).

The inputs first undergo a downsampling path, which consists of a series of  $3 \times 3 \times 3$  convolutions with ReLU activations, and maxpooling layers that increase the receptive field size while maintaining a reasonably sized model and imparting some translation invariance. This results in a set of learned low-resolution features aiming to capture global context. Second is an upsampling path, which consists of upconvolution layers and additional  $3 \times 3 \times 3$  convolutions with ReLU activations. This eventually recovers full spatial resolution, while skip connections at each resolution level concatenate features from the downsampling path, allowing successively finer details in the input to be considered. At each resolution level, the number of computed feature maps doubles. All model parameters up to this point form  $\theta_h$ , as the output of this part of the network architecture is the latent representation  $\mathbf{h}_t$ .

The output segmentation probability map  $p(\mathbf{y}_t | \mathbf{h}_t; \theta_y)$  is computed as in the standard U-Net. Each voxel stores the parameters of a categorical distribution over  $L$

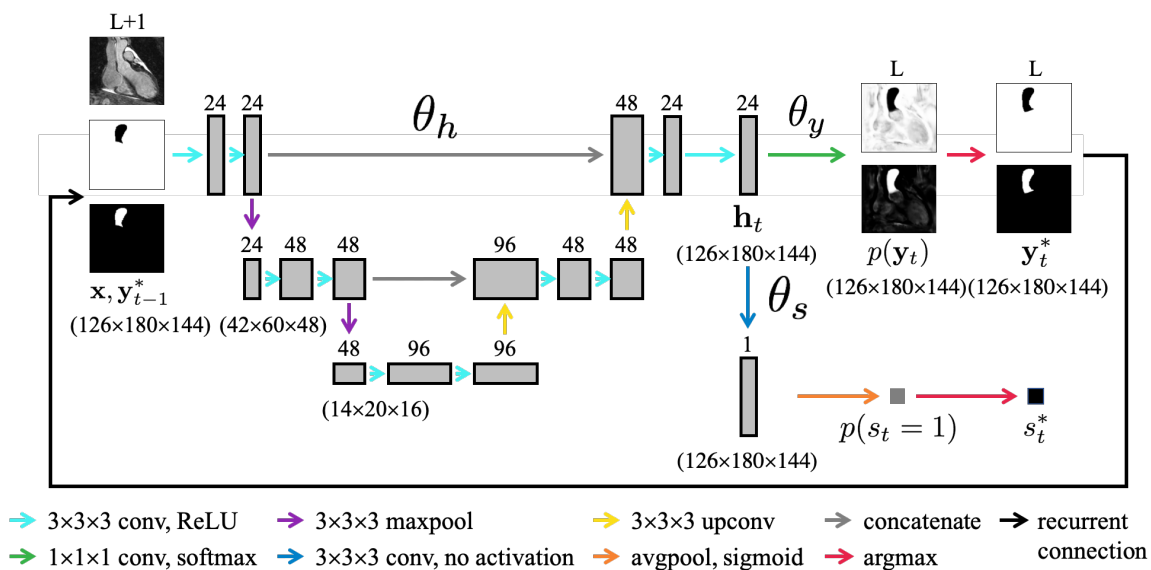


Figure 4-6: Schematic of the RNN trained to jointly evolve the segmentation and predict the stopping indicator. The main block is an augmented U-Net architecture. (Here, the number of feature maps in the latent representation  $\mathbf{h}_t$  is  $C = 24$ ).

labels. These are computed from the feature maps in  $\mathbf{h}_t$  via  $C \cdot L$   $1 \times 1 \times 1$  convolutions, whose parameters form  $\boldsymbol{\theta}_y$ , and a softmax activation. Finally, a voxel-wise argmax yields  $\mathbf{y}_t^*$ .

The scalar parameter  $p(s_t = 1|\mathbf{h}_t; \boldsymbol{\theta}_s)$  of the stopping indicator’s Bernoulli distribution is computed via  $C$   $3 \times 3 \times 3$  convolutions that reduce the latent representation  $\mathbf{h}_t$  to a single channel. The parameters of these convolutions form  $\boldsymbol{\theta}_s$ . A global average pooling with sigmoid activation yields a scalar in  $[0, 1]$  representing  $p(s_t = 1|\mathbf{h}_t; \boldsymbol{\theta}_s)$ . Finally, an argmax yields  $s_t$ .

### 4.3.2 Training Data Generation

Recall that the training data  $\mathcal{D}' = \{\mathbf{x}, \mathbf{y}_{in}, \mathbf{y}_{out}, s\}$  should illustrate the application-dependent segmentation evolution pattern that the RNN should learn. In our case, for every training image we have a ground truth segmentation  $\mathbf{y}$  for each anatomical structure and an example seed  $\mathbf{y}_0$  (for details on how these are generated, see Chapter 2). During each training epoch, we automatically generate one sample from  $\mathcal{D}'$  for each training image. The process is visualized in Figure 4-7. Note that we will train a neural network to evolve all segmentations natively in 3D, i.e., not via slice-by-slice processing.

#### Great Vessel Segmentations:

These are trained to grow along the vessel centerline at a constant rate. Before training, we use fast marching [167] to precompute a distance map that is zero in the background and for each foreground voxel, stores the distance of the shortest path to the seed point that remains within the ground truth segmentation. During training, this map can be thresholded to get arbitrary partial segmentations. Corresponding input-output segmentation pairs  $(\mathbf{y}_{in}, \mathbf{y}_{out})$  are created by first thresholding at a distance  $d_1$  chosen uniformly at random to form  $\mathbf{y}_{in}$ , and at  $d_2 = d_1 + d_s$  to form  $\mathbf{y}_{out}$ , where  $d_s$  is the desired step size.

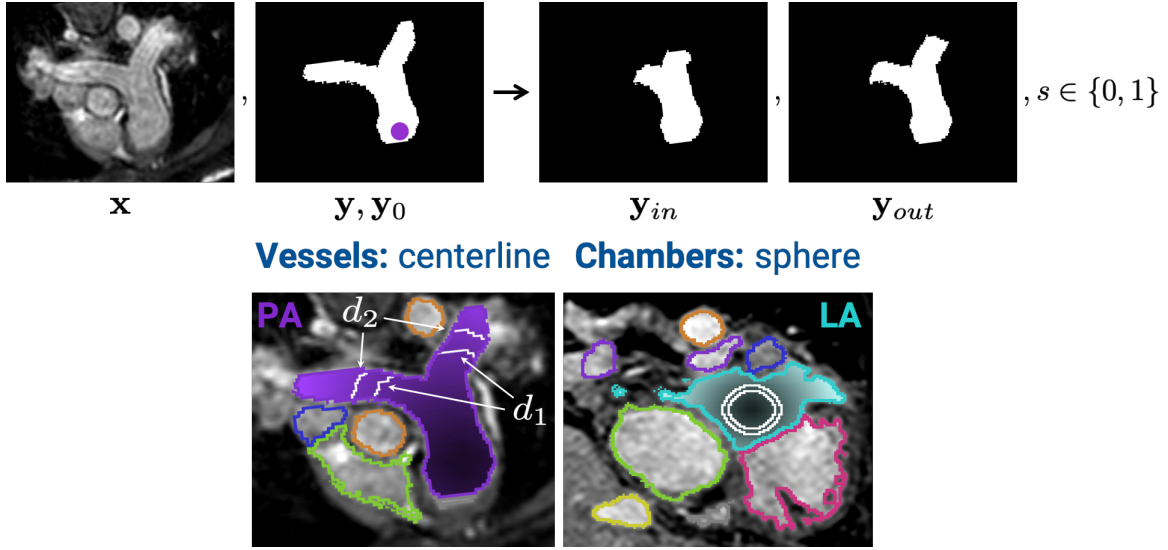


Figure 4-7: For each training image  $\mathbf{x}$ , input partial segmentations, output partial segmentations and stopping indicators  $(\mathbf{y}_{in}, \mathbf{y}_{out}, s)$  are generated on-the-fly during training from the complete ground truth segmentation  $\mathbf{y}$  and seed  $\mathbf{y}_0$ .

### Chamber Segmentations:

These are trained to grow outward at a constant rate. Since these are larger structures, during training we first randomly perturb the seed point by moving  $\mathbf{y}_0$  within the chamber center. We then generate two concentric spheres centered on it: the radius  $r_1$  of the smaller sphere is chosen uniformly at random, and the larger radius is  $r_2 = r_1 + r_s$ , where  $r_s$  is the desired step size. Both spheres are intersected with the ground truth complete segmentation to form  $(\mathbf{y}_{in}, \mathbf{y}_{out})$ .

### Stopping Indicators:

The ground truth stopping indicator  $s \in \{0, 1\}$  is computed by comparing the output segmentation  $\mathbf{y}_{out}$  with the complete ground truth segmentation  $\mathbf{y}$ . For vessels, we evaluate whether the distance threshold  $d_2$  used for the output segmentation is close to the maximum distance in the fast marching image. For chambers, we consider the proportion of voxels in the complete ground truth segmentation that are contained in the output segmentation.



### Seed Point Locations:

Seed points (Table 4.1) were chosen to maximize their potential for automatic detection in future. For example, the seed for the aorta could have been placed at the aortic valve, and segmentations grown away from the heart, which would be easy to complete accurately. However, the descending aorta is more salient, so we grow segmentations in the opposite direction, towards the aortic valve. For all but the PA, segmentations must grow toward one or more boundaries with another structure (typically a valve, ASD/VSD, or the connection of the IVC/SVC into the adjacent atrium or vessel). The lack of contrast at these borders, which separate the global blood pool, provides a challenging test case for our automatic stopping.

Table 4.1: Seed points to be clicked by the user. For more details, see Chapter 2.

<b>Chambers</b>		<b>Great Vessels</b>	
<b>LV</b>	Center region	<b>AO</b>	Bottom of descending aorta
<b>RV</b>	Center region	<b>PA</b>	Bottom of main PA trunk
<b>LA</b>	Center region	<b>SVC</b>	Superior end (two for bilateral SVC)
<b>RA</b>	Center region	<b>IVC</b>	Center of hepatic segment

### 4.3.3 Data Augmentation

We found data augmentation to be essential to learn from a small training dataset, and use it to mimic the diversity of heart shapes and sizes, global intensity changes (caused by variable acquisition settings and inhomogeneity artifacts) and noise (induced by elevated heart rates or arrhythmias). We apply random affine transformations (translation, rotation, scaling and shearing), nonlinear transformations, left-right and anterior-posterior flips (which are helpful due to dextrocardia and other cardiac malpositions in CHD), constant intensity shifts, and additive Gaussian noise.

Cardiac MRI exhibits both inhomogeneity artifacts from previously implanted stents and heterogeneous background appearance due to the surrounding vasculature (Fig. 4-8(a)). To simulate this variability, we perform additional data augmentation

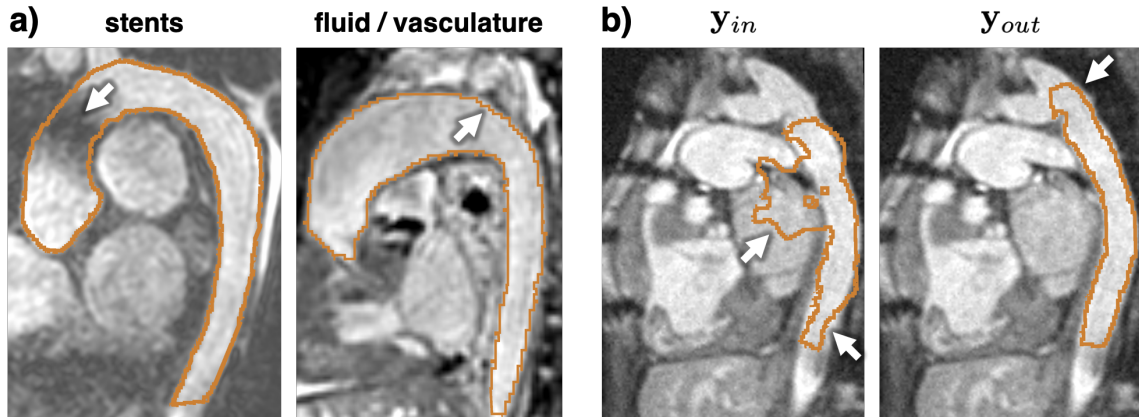


Figure 4-8: (a) Representative cardiac MR images showing a stent artifact and bright background areas showing surrounding fluid and vasculature. (b) We use data augmentation to create corrupted input segmentations  $\mathbf{y}_{in}$  and uncorrupted output segmentations  $\mathbf{y}_{out}$  so that the trained RNN is robust to errors in its intermediate results.

for the AO and PA by adding random dark regions inside the vessel and random dark or bright regions next to it.

Finally, if the augmented U-Net for  $p(\mathbf{y}_t, s_t | \mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 0)$  is trained solely using error-free input segmentations  $\mathbf{y}_{in}$ , then it may not operate well at test time when it must perform inference on its own imperfect outputs. As shown in Fig. 4-8(b), we address this by corrupting the input segmentations  $\mathbf{y}_{in}$  using random nonrigid deformations. We also add random foreground blobs in the segmentation that vary in number, location and size. The output segmentation  $\mathbf{y}_{out}$  remains unchanged. During training, the RNN must learn to remove input segmentation errors while simultaneously growing the segmentation appropriately. Hence, when the model recursively operates on its own results, it will be more robust to errors in its input and able to correct them.

## 4.4 Evaluation

In this section, we evaluate our iterative model’s performance for the challenging application of whole heart segmentation for congenital heart disease patients, and compare it to several automatic and interactive learning-based methods that directly segment an image in one step.

### 4.4.1 Data

We used the HVSMR+ and HVSMR++ datasets described in Chapter 2. (Recall that the 60 HVSMR++ cases include the 20 HVSMR+ cases). These datasets contain 3D MRI scans with ground truth segmentations of the LV, RV, LA, RA, AO, PA, SVC and IVC, a seed point for each structure, and a “mild”, “moderate” or “severe” label according to each heart’s cardiac malformations. Each image was cropped to a tight region around the heart and resized to  $\approx 128 \times 180 \times 144$  (depending on the network architecture used).

The datasets were used in two ways:

- **Train on HVSMR+, Test on HVSMR++:** These experiments test an algorithm’s ability to generalize from training on a very small dataset biased towards more normal anatomy (HVSMR+: 20 subjects, 10/20 mild, 6/20 moderate, 4/20 severe) to evaluation on a larger, but still relatively small, dataset that has more severe cases (HVSMR++: 60 subjects, 12/60 mild, 11/60 moderate, 37/60 severe).

We present results both for (1) cross-validation on the HVSMR+ dataset, which was used for training and model selection, and for (2) inference on the 40 additional subjects from HVSMR++, which act as held-out test subjects (40 subjects, 2/40 mild, 5/40 moderate, 33/40 severe).

- **Train and Test on Subsets of HVSMR++:** These experiments test an algorithm’s accuracy when a larger and more balanced dataset is available.

We present results both for (1) cross-validation on a subset of the HVSMR++ dataset (48 subjects, 11/48 mild, 10/48 moderate, 27/48 severe), which was used for training and model selection, and for (2) inference on 12 held-out test subjects from the HVSMR++ dataset (12 subjects, 1/12 mild, 1/12 moderate, 10/12 severe). To do this, the HVSMR++ dataset was split into a cross-validation group and a test group by manually choosing test images that encapsulate the entire range of potential congenital heart defects, based on each image’s cardiology codes. The 20 HVSMR+ cases all belong to the HVSMR++ cross-validation group, since they had already been used for algorithm development.

Each cross-validation dataset was split randomly into four folds (i.e., train on 15, test on 5 for the 20 HVSMR+ cases, and train on 36, test on 12 for the 48 HVSMR++ cases). Each fold had an approximately equal number of mild, moderate and severe cases. The resulting four models were then applied to the held-out test subjects. All results are presented by grouping the results from all four models together.

#### 4.4.2 Experimental Setup

We compared five segmentation approaches, which are depicted in Fig. 4-9. The first two are fully automatic and the remaining three are interactive methods.

- **U-Net-All**: Conventional U-Net for multiclass segmentation of all 8 structures;
- **U-Net**: Conventional U-Net for binary segmentation of each anatomical structure;
- **U-Net+S**: Conventional U-Net for binary segmentation, also inputs the Euclidean distance map to the user-specified seed [44];
- **Iter-A**: Iterative segmentation with automatic stopping; and
- **Iter-U**: The same model as **Iter-A**, but simulates a user who chooses the stopping point by keeping the best segmentation from the first 40 iterations.

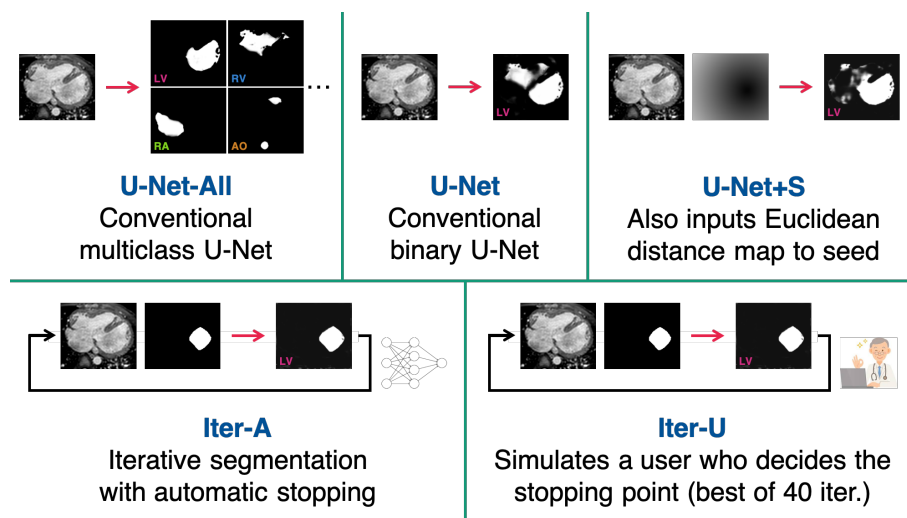


Figure 4-9: The five segmentation approaches to be compared, alongside representative inputs and outputs for left ventricle segmentation.

Experimentally, we found that the best **U-Net-All** model had 4 levels,  $C = 24$  learned channels at the first level,  $3 \times 3 \times 3$  maxpooling after the first level,  $2 \times 2 \times 2$  maxpooling after the second and third levels, and  $\approx 3,600,000$  parameters. The best architecture for all binary segmentations, whether **U-Net**, **U-Net+S** or iterative, consistently had 3 levels,  $C = 24$  learned channels at the first level,  $3 \times 3 \times 3$  maxpooling, and  $\approx 870,000$  parameters. Data augmentation varied slightly for each structure, as for example some structures are expected to vary in size more than others. Additional details are provided in Appendix A.

We implemented our method using Keras<sup>1</sup> [168] with a Tensorflow<sup>2</sup> [169] backend. Model parameters were optimized using adadelta [170] with the default Keras parameters, for 2000 epochs with a batch size of one.

All output segmentations were post-processed to keep only the island connected to the seed. If the segmentation did not contain the seed, or if no seed is available (**U-Net-All**, **U-Net**), the largest connected component was retained for each structure (or two largest connected components for cases with bilateral SVC or double IVC). When computing the Dice score for segmentation evaluation, vessel segmentations were not penalized for being slightly too long or too short compared to the ground truth segmentation, as described in Chapter 2.

### 4.4.3 Qualitative Results

Fig. 4-10 shows example successes of iterative segmentation with automatic stopping.

Visual inspection revealed the most challenging structures: the **PA**, whose difficulty is corroborated by the results from a recent whole heart segmentation challenge [24], **LA**, as the faint pulmonary veins were often missed, **RV**, as the pointy apex of this crescent-like structure were sometimes missed, **bilateral SVCs**, as these are quite different from normal despite being in the moderate category, and **IVC**, as the insertion of the IVC into an atrium is not defined by a valve, so this boundary is subject to lower inter- and intra-rater reliability in the ground truth segmentation.

---

<sup>1</sup><http://keras.io>

<sup>2</sup><http://www.tensorflow.org>

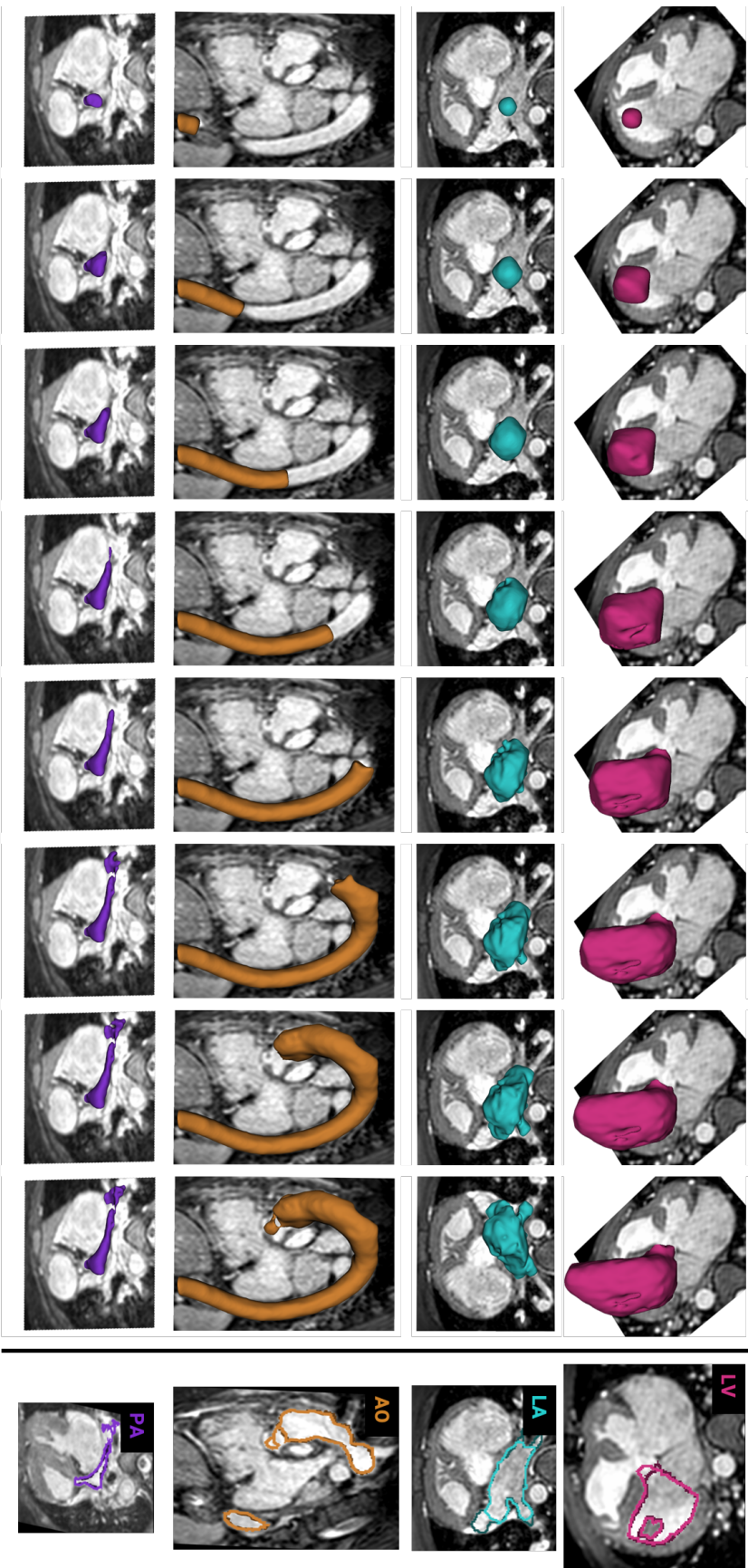


Figure 4-10: Examples from subjects with severe heart malformations where iterative segmentation with automatic stopping (**Iter-A**) has high accuracy. Each row depicts the segmentation propagation (as a 3D model overlaid onto a 2D image slice), ending with the automatically detected stopping point. The final column shows an overlay of the ground truth segmentation (dark colour) and the **Iter-A** result (lighter colour).

Our main experimental results are described below. Results from the mild and moderate cases are combined because they were largely similar, especially compared to those from the severe cases.

#### 4.4.4 Train on HVSMR<sub>+</sub>, Test on HVSMR<sub>++</sub>

Fig. 4-11 summarizes segmentation accuracy when models trained using the 20 HVSMR<sub>+</sub> cases are applied to 40 held-out HVSMR<sub>++</sub> cases. Cross-validation on the HVSMR<sub>+</sub> dataset follows similar trends, and is shown as Fig. A-1 in Appendix A.

##### Mild and Moderate Cases:

On average all methods performed well for mild and moderate cases, as shown in Fig. 4-11 (Top). Iterative segmentation with user stopping had the best overall score, albeit by a small margin. These anatomies are very well represented in the training data and do not exhibit major cardiac malformations, and so good generalization is possible despite the very small HVSMR<sub>+</sub> dataset used for training.

##### Severe Cases:

User input was much more important for severe cases due to extreme anatomical variations, as shown in Fig. 4-11 (Bottom). Considering the direct methods, **U-Net+S** outperformed **U-Net-All** and **U-Net**, which is expected since the object localization provided by the user seed yields an important signal for segmentation. However, all three direct segmentation methods had much lower accuracy than in mild and moderate subjects.

In severe subjects, iterative segmentation with a user-determined stopping point (**Iter-U**) had the best mean segmentation accuracy for all eight structures, and achieved an overall Dice score of  $81.0 \pm 15.1$ . The superiority of **Iter-U** over **U-Net+S** is clearly seen in Fig. 4-12, which compares **Iter-U** and **U-Net+S** for all 253 data points (i.e., 8 structures  $\times$  33 severe subjects, plus/minus some structures for subjects with common atrium, single ventricle, bilateral SVC or double IVC). Visually,

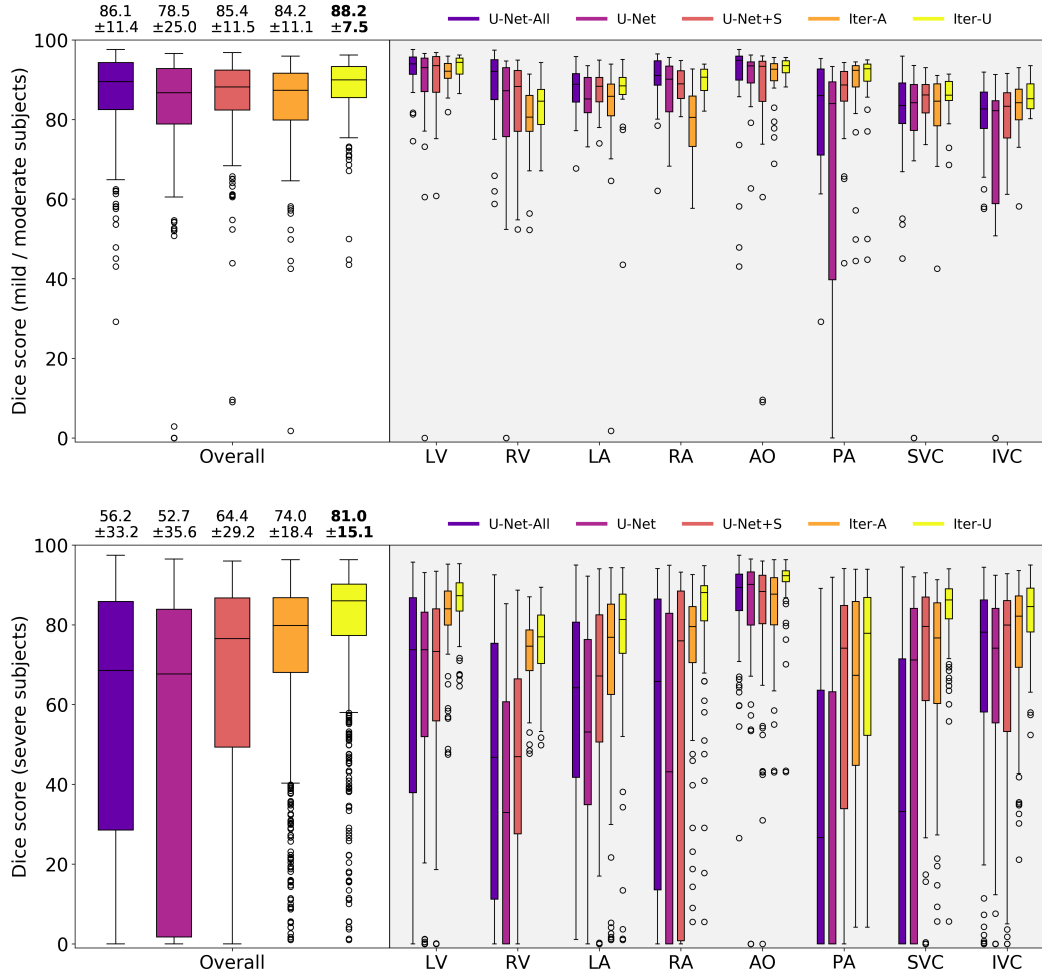


Figure 4-11: Results on 40 held-out HVSMR++ subjects after training using 20 HVSMR+ subjects: Summary statistics (Dice score). (Top) For mild and moderate subjects, all methods except **U-Net** had comparable performance. (Bottom) For severe subjects, the iterative segmentation methods (**Iter-A** and **Iter-U**) were superior, especially iterative segmentation with user stopping. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold.

**Iter-U** had lower accuracy than **U-Net+S** in only  $8/253 = 3\%$  of structures, most of which are PAs (the PA also had the lowest accuracy of all structures).

**Iter-A** requires the same level of user input as **U-Net+S**, namely a single click per structure, but was more accurate on average (overall Dice score:  $74.0 \pm 18.4$  for **Iter-A** versus  $64.4 \pm 29.2$  for **U-Net+S**). Moreover, **Iter-A** had a better mean segmentation accuracy than **U-Net+S** for all eight structures. See Fig. 4-13 for a direct comparison of **Iter-A** and **U-Net+S**.



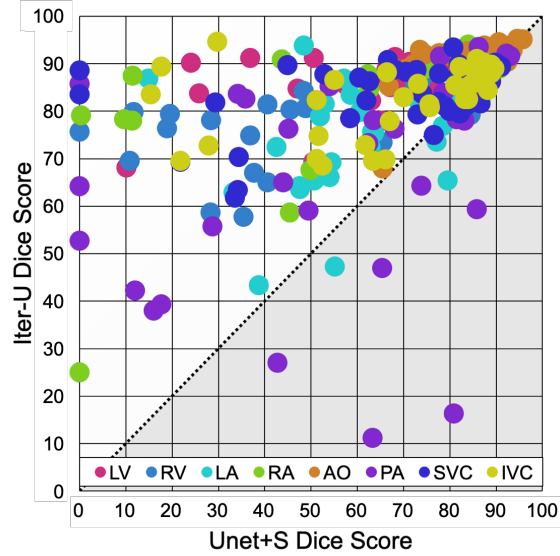


Figure 4-12: **Iter-U** outperformed **U-Net+S** on held-out subjects with severe heart malformations when training with the very small HVSMR+ dataset. Points above the dotted line (white zone) indicate where **Iter-U** is better.

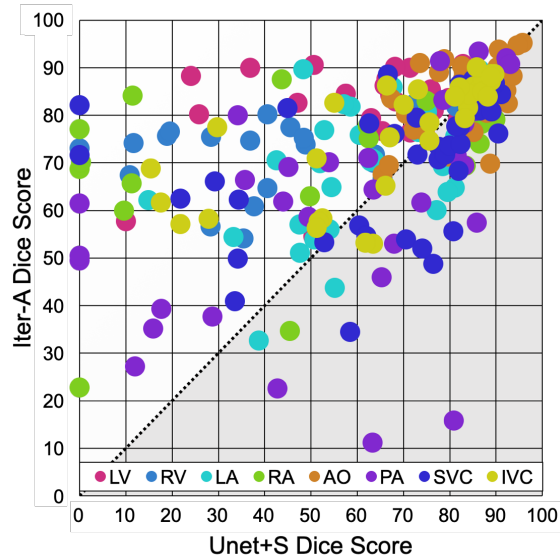


Figure 4-13: **Iter-A** versus **U-Net+S** for held-out subjects with severe heart malformations when training with the very small HVSMR+ dataset. Points above the dotted line (white zone) indicate where **Iter-A** is better. For each structure, **Iter-A** had a better average Dice score than **U-Net+S**, but there were some cases where the iterative model needed user stopping to achieve the best performance.

### 4.4.5 Train and Test on Subsets of HVSMR++

Next, we discuss the results from models trained using the larger HVSMR++ dataset (60 subjects total), and compare to those from models trained using the smaller HVSMR+ dataset (20 subjects total).

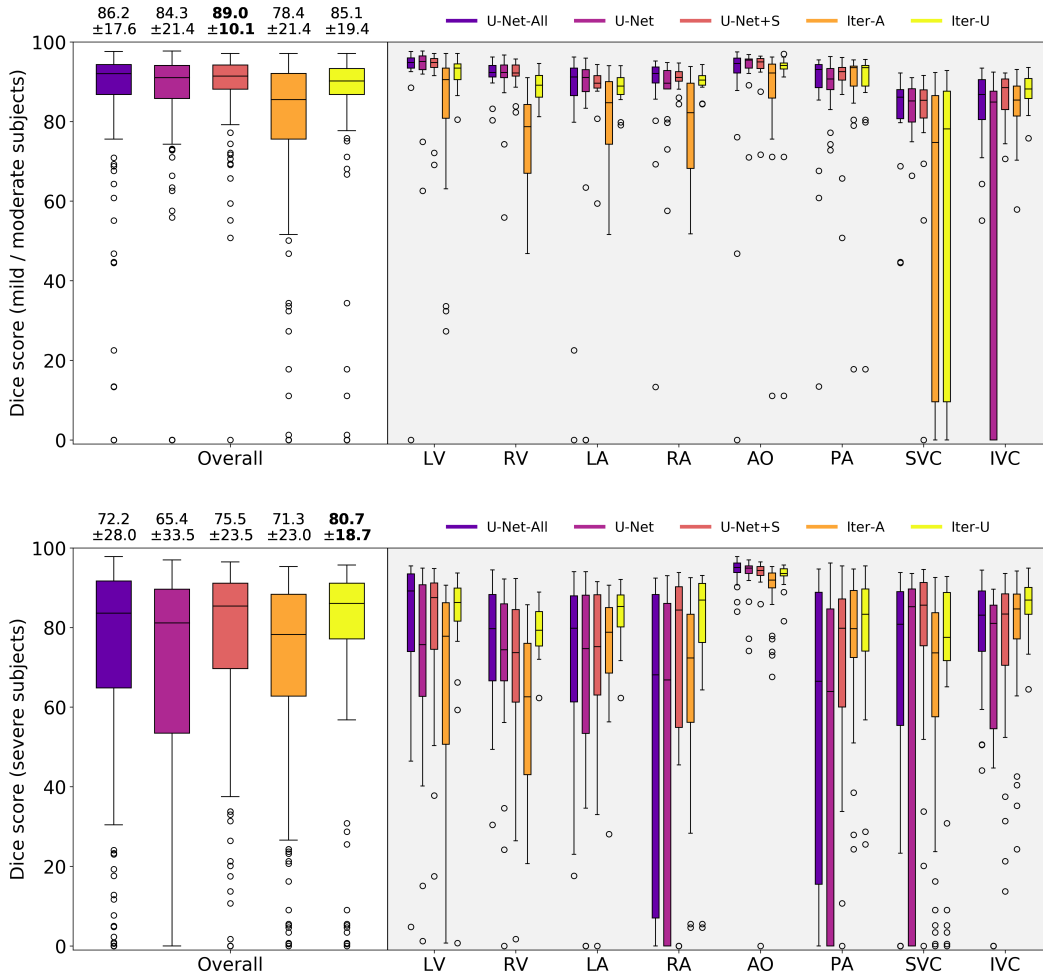


Figure 4-14: HVSMR++ cross-validation summary statistics (Dice score). (Top) For mild and moderate subjects, all five methods except **Iter-A** had comparable performance. (Bottom) For severe subjects, **Iter-U** had the best overall Dice score, with better or similar accuracy compared to **U-Net+S** for all structures except the SVC. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold.

### Mild and Moderate Cases:

For training using the larger HVSMR++ dataset, segmentation accuracy for mild and moderate subjects cannot be assessed using the HVSMR++ test subjects because there are too few of them (2/12 subjects). Therefore, we present the results from the HVSMR++ cross-validation group (48 subjects). Mild and moderate cases still form the largest cluster of subjects used for training ( $21/48 = 44\%$ ), but they are not as predominant as they were in HVSMR+ ( $16/20 = 80\%$ ). Note that although there are now more severe cases ( $27/48 = 56\%$ ), they still form a very heterogeneous group that contains different types of heart defects and their combinations.

The cross-validation results for mild and moderate subjects are shown in Fig. 4-14 (Top). As we've seen before, the three direct segmentation methods performed well for mild and moderate subjects. For HVSMR++ cross-validation, **U-Net+S** had the best overall Dice score. **Iter-U** had relatively similar performance as **U-Net+S** for all structures except the SVC and RV, which reduced its overall score, and **Iter-A** had lower accuracy than **U-Net+S** for all structures.

### Severe Cases:

First, we consider the results when models trained using the 48 HVSMR++ cases were applied to the 12 held-out HVSMR++ test subjects, which showed similar trends to the results for severe subjects for HVSMR++ cross-validation from Fig. 4-14 (Bottom). Most of the test subjects are categorized as severe (10/12 subjects). The results are shown in Fig. 4-15. In both cross-validation and held-out testing, **Iter-U** again had the best overall Dice score. **Iter-U** was better or comparable to **U-Net+S** for all cardiac structures except for the SVC (the SVC also had the lowest accuracy of all structures).

Finally, we can directly compare the results on the 12 HVSMR++ test subjects for models trained using 48 HVSMR++ subjects (Fig. 4-15) versus 20 HVSMR+ subjects (Fig. 4-16; these results are very similar to those in Fig. 4-11 (Bottom), in which the same models were evaluated on even more severe subjects).

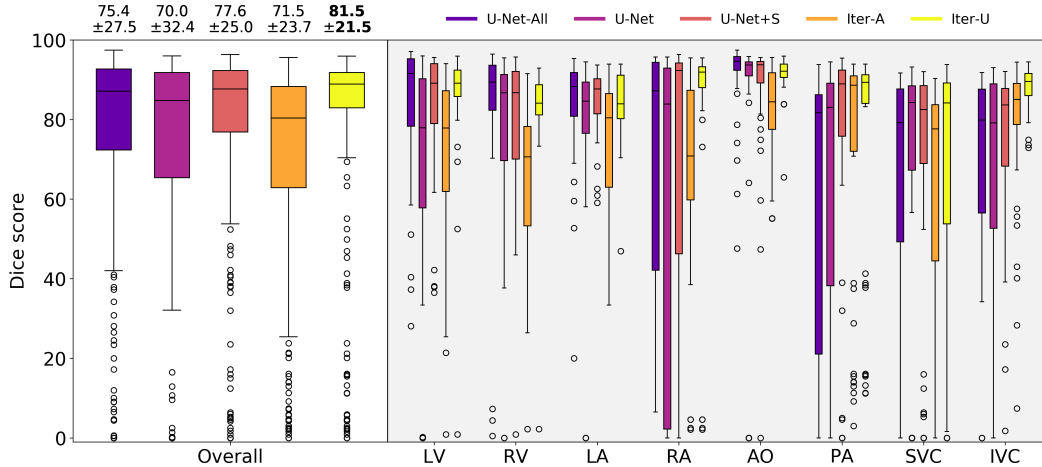


Figure 4-15: Results on 12 held-out HVSMR++ test subjects after training using 48 HVSMR++ subjects: Summary statistics (Dice score). There are 2 mild and moderate subjects and 10 severe subjects. **Iter-U** had the best overall Dice score, with better or similar accuracy compared to **U-Net+S** for all eight structures except the SVC. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold.

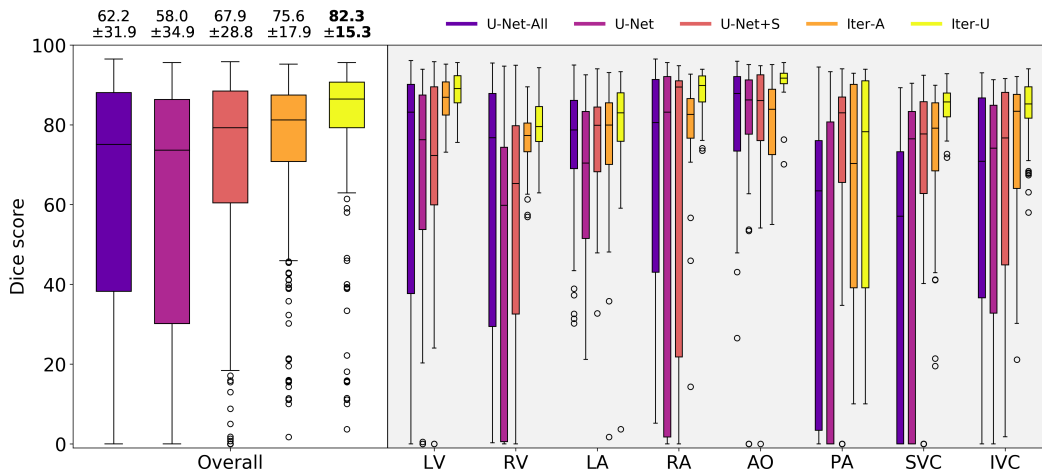


Figure 4-16: Results on 12 held-out HVSMR++ test subjects after training using 20 HVSMR+ subjects: Summary statistics (Dice score). There are 2 mild and moderate subjects and 10 severe subjects. The iterative segmentation methods (**Iter-A** and **Iter-U**) were superior, especially iterative segmentation with user stopping. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold.

#### 4.4.6 Summary of Results

The main messages are as follows. Regardless of whether a smaller or larger dataset is used, **Iter-U** had the best performance for severe subjects. Both **Iter-U** and **Iter-A** outperformed **U-Net+S** for the smallest dataset. Overall, the two iterative methods performed equally well regardless of whether training was done using a small dataset biased towards more normal anatomy (HVSMR+) or a larger dataset with more severe subjects (HVSMR++). In contrast, the direct segmentation methods required more training data to achieve better performance, and even then did not reach the accuracy of **Iter-U**.

#### 4.4.7 User Interaction Mechanisms

Some examples in which iterative segmentation with user stopping outperformed automatic stopping are shown in Fig. 4-17. These illustrate that in our proposed setup, a user can ask for more iterations if the object is under-segmented, or go back in the

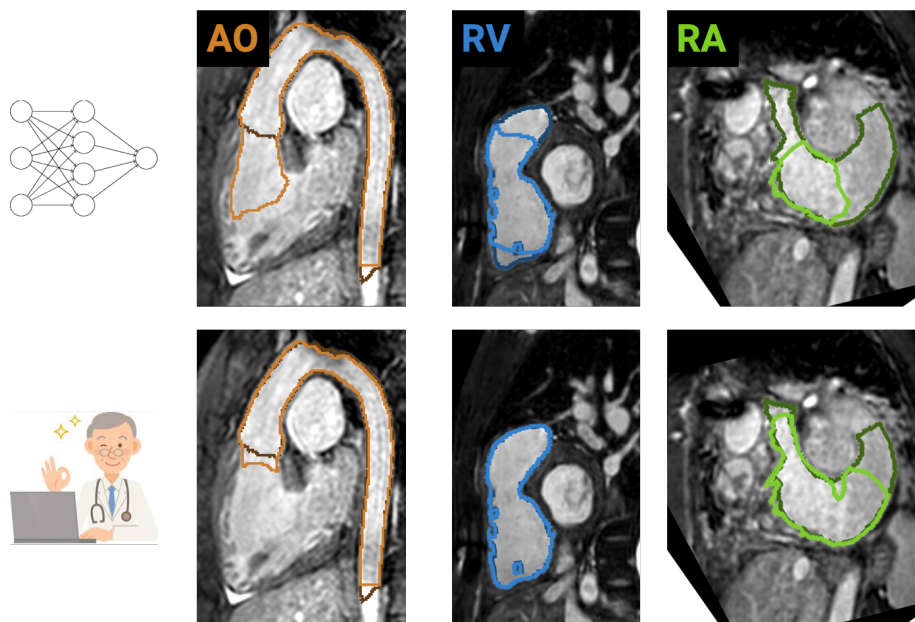


Figure 4-17: Examples from subjects with severe heart malformations in which a user can choose a better segmentation from the output sequence than that chosen via automatic stopping. Ground truth segmentations are dark, **Iter-A** (top) and **Iter-U** (bottom) segmentations are lighter.

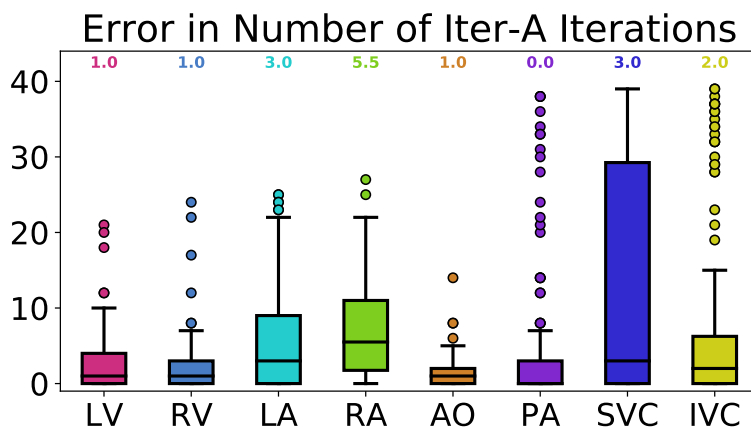


Figure 4-18: The number of iterations executed by automatic stopping (used by **Iter-A**) is typically close to the ideal number of iterations (used by **Iter-U**). These results are for 40 held-out HVSMR++ subjects after training using 20 HVSMR+ subjects. The median value for each structure is shown at the top of the graph.

sequence and pick an earlier segmentation if the segmentation has grown too much.

The error in the number of iterations selected by automatic stopping directly causes any decrease in segmentation accuracy for **Iter-A** compared to **Iter-U**. However, Fig. 4-18 shows that this number was typically small (we did observe that some finished segmentations can stop evolving without triggering the automatic stopping).

Finally, the benchmarking results in Figure 4-19 shows that inference is fast

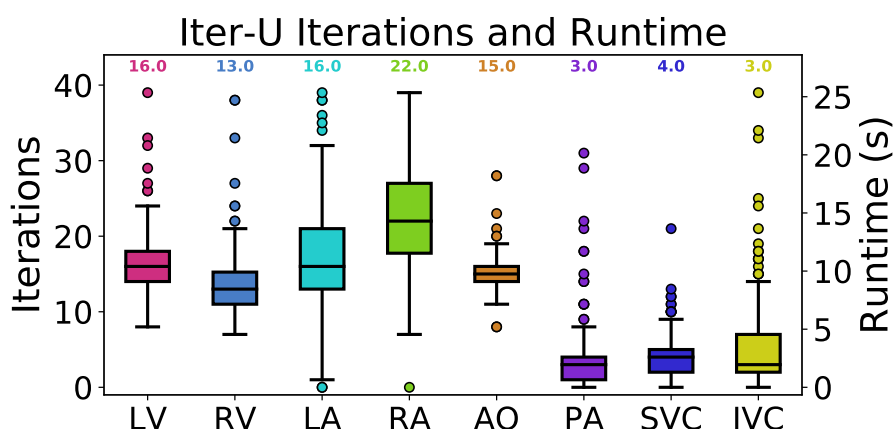


Figure 4-19: Number of iterations and runtime required for iterative segmentation with user-directed stopping. These results are for 40 held-out HVSMR++ subjects after training using 20 HVSMR+ subjects. The median value for each structure is shown at the top of the graph.

enough for our iterative segmentation method to be used interactively. Each iteration required  $0.65 \pm 0.15$  seconds on an NVIDIA TITAN X GPU, and the median time required to segment one structure ranged from 3-22 seconds, leading to a total inference time of less than 2 minutes.

### Failure Cases:

Fig. 4-20 shows some failure cases of iterative segmentation, even after simulating user-based stopping. These are some representative cases in which the **Iter-U** Dice score is relatively low, due to extreme stent artifacts, exceptionally narrow vessels, very large ASDs, and difficulties in segmenting the RV apex. These types of errors are easily observable by a human monitoring the segmentation process, who can resort to manual segmentation in such very difficult cases.

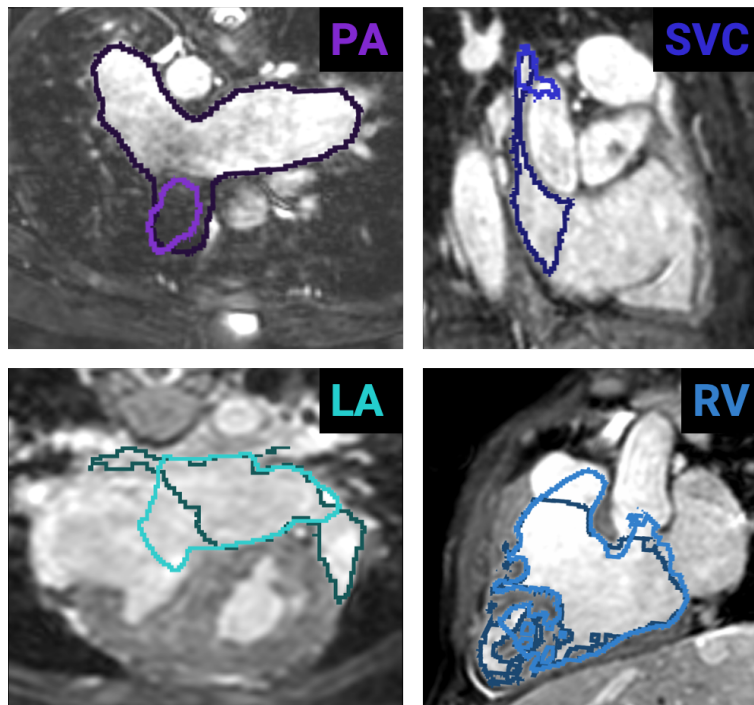


Figure 4-20: Iterative segmentation with user stopping can have low accuracy in especially challenging cases. Ground truth segmentations are dark, **Iter-U** segmentations are lighter.

## 4.5 Discussion

The wide range of potential morphological deformations and topological variations in congenital heart disease makes whole heart segmentation very difficult, especially when training data is limited. Here, we have proposed a novel interactive segmentation method that labels each chamber and great vessel using an iterative model implemented as an RNN. The algorithm is initialized by a single user click per structure, and its stopping point can be automatically determined or user-defined. Notably, our method evolves segmentations in a predictable way, as defined via training data, since our loss function evaluates intermediate segmentations in addition to the final segmentation result.

Our results show that the proposed iterative segmentation model can be learned from a small dataset that does not necessarily include the same pathology present in the test image, generalizing better to patients with the most severe heart malformations compared to conventional methods that directly segment an image in one step. Direct segmentation methods required many more training images to reach the same accuracy as **Iter-A** and **Iter-U** trained with an extremely small and imbalanced dataset (HVS<sub>MR+</sub>). Iterative segmentation with user stopping consistently had the best segmentation accuracy for patients with severe heart defects. The only circumstance in which **Iter-U** did *not* have the highest overall Dice score was in mild and moderate subjects for models trained using the larger HVS<sub>MR++</sub> dataset (Fig. 4-14). In this case, **U-Net+S** was better for mild and moderate subjects (the dominant group) but worse for severe subjects, while **Iter-U** was worse for mild and moderate subjects but better for severe subjects.

After considering how a user would interact with our segmentation algorithm, we chose to encourage the sequence of output segmentations to follow a predetermined pattern. In this Chapter, we showed how this is useful when learned automatic stopping does not choose the best segmentation from the output sequence. We envision our iterative segmentation algorithm being used as follows. After the user clicks once to place a seed point, the **Iter-A** result would be shown. Then, the user would have



the option to either accept the segmentation, or to look forwards or backwards in the sequence for another solution. This minimal amount of additional user interaction yields a large performance improvements for severe subjects, as shown by our results comparing **Iter-U** to **Iter-A** and **U-Net+S**.

There are several other potential benefits of a predictable segmentation evolution for interactive segmentation. Additional user input may be desired for very high accuracy, e.g., via clicks or scribbles [41–44]. Scanning through an entire 3D image to look for errors is tedious and time consuming, whether the user evaluates a completed segmentation or intermediate segmentations such as those in Fig. 4-2, since the user cannot distinguish between areas that would be corrected in subsequent time steps and those which require input. We anticipate that predictable segmentation evolution would make user monitoring and error correction much more straightforward. The user could mentally compare intermediate segmentations to the expected pattern, allowing them to (1) interrogate more limited image regions (especially for vessel segmentation), and (2) give more targeted inputs.

Our results corroborate previous studies in which direct segmentation was outperformed by iterative approaches, including coarse-to-fine network cascades [66,131–134,137] and RNNs [154–158]. Previous authors have provided some intuition behind this. While direct segmentation methods perform inference at each voxel independently, iterative segmentation allows each voxel’s inferred label (or probability distribution over labels) to influence the decisions at its neighbors. This allows the model to learn short- and long-range relationships across voxel labels, and expands the model’s field of view without increasing the number of model parameters [132,154]. In contrast, for compactly shaped structures (e.g., LV, RV, LA, RA, IVC), the distance map input by **U-Net+S** is informative in areas close to the user seed or very far away it, but may be less so at intermediate distances, especially considering variations in object size. The distance map may also be suboptimal when distinguishing between long vessels that are in close proximity (e.g., AO, PA). One unexplored option is to provide a distance map for each input seed, i.e., 8 in total, as additional input channels in **U-Net+S** or for our iterative segmentation method. This may improve the accuracy

of automatic stopping by providing the rough locations of neighboring structures. Finally, we observed that our iterative segmentation typically expanded the connected component attached to the user seed, which could be easily post-processed, while the direct segmentation methods produced multiple islands located both inside and outside the anatomical structure of interest.

Our iterative segmentation method could be applied to any number of growth dynamics, as long as coherent input-output segmentation pairs can be generated for training. For example, our method could be used to generalize spatial propagation RNNs previously proposed to segment the cardiac ventricles slice-by-slice from base to apex [162, 163]. For the four cardiac chambers, another option would be to grow segmentations according to a distance map computed from the ground truth segmentation boundary, which may perform better than spherical growth for elongated structures like the RV. Finally, our iterative method’s flexibility could allow multi-class iterative segmentation to be explored in future (i.e., iterative segmentation in which the number of anatomical labels  $L$  has  $L > 2$ ). Specifically, we found that **U-Net-All** outperformed **U-Net**, which is to be expected because multiclass segmentation uses training data that is labeled more explicitly and enjoys the benefits of multi-task learning. Defining training data in which *multiple* segmentations grow at the same time could allow the model to better learn the spatial relationships between different structures (and not solely within them), and also eliminates the need to resolve conflicts between overlapping binary segmentations. An initial investigation could use the same procedure described above (Section 4.3.2) to create the multiclass segmentations required for training by generating input and output segmentations for each great vessel and cardiac chamber independently. Although this approach will produce input segmentations unlikely to be seen during inference (e.g., having some structures with nearly complete input segmentations and other structures which have barely started to evolve), it may be sufficient to train an accurate model in practice, and could be seen as an extreme form of data augmentation. If needed, a minor extension could limit the randomization to produce more realistic combinations.

We acknowledge a few limitations of the study presented in this Chapter. A user

study could be informative, both to resolve differences between actual users and the simulated user used by **Iter-U**, and to evaluate different ways in which users can interact with our iterative segmentation model. Second, although we do randomly perturb the seed point when training our iterative model to segment cardiac chambers, we have not yet investigated the impact of varying the seed point location on the accuracy of segmenting the chambers and great vessels. Finally, future work to improve the accuracy of automatic stopping would close the gap between **Iter-A** and **Iter-U**. We did observe that on average, **Iter-A** can have worse performance than direct segmentation for mild and moderate subjects, and was surpassed by **U-Net+S** for severe subjects when the larger HVSMR++ dataset was used for training. However, we argue that the improved performance of our iterative segmentation method for severe subjects was considerable, and much more clinically relevant, since subjects with the most severe heart defects should benefit the most from patient-specific heart models for surgical planning.



# Chapter 5

## Discussion and Conclusions

This thesis work yielded some of the first whole heart segmentation methods for congenital heart disease. However, the interactive segmentation tools that we have developed could be applied to other image segmentation tasks in which anatomical variability is high, regardless of the imaging modality or the organ of interest. Our research therefore has wide potential impact on using images for medical diagnosis, to monitor patients, and to plan interventions. In this chapter, we discuss potential future technical directions and the potential clinical impact of our work.

### 5.1 Technical Directions

Here, we reflect upon our experience and explore possibilities for future research.

#### 5.1.1 User Interaction for Whole Heart Segmentation

During the course of this work, it soon became apparent that carefully considered user interaction can be extremely valuable to solve challenging image analysis problems. Throughout this thesis, we have progressively reduced the amount of user interaction, from dense segmentations on entire short-axis slices or regions within them (Chapter 3) to just eight clicks to localize each major cardiac structure (Chapter 4).

We note two important takeaways for interactive image segmentation. First, sev-

eral options exist for spreading user information to the unannotated portions of the image. One can directly provide the annotations as additional inputs to an algorithm [41–44]; we do this in our patch-based interactive segmentation algorithm and in the **U-Net+S** direct segmentation model. Or, one can spatially propagate user inputs more explicitly, as we do in our learned iterative segmentation model. Our results in Chapter 4 support the spatial propagation approach, which is also taken by several classical interactive segmentation methods such as level sets [171], random walker [172], GrowCut [173] and GeoS [36]. A second takeaway is that allowing the estimated labels at distant voxels to inform inference at other voxels is very advantageous. Traditional feedforward neural networks do not, and we believe that this is a major contributor to the success of our learned iterative segmentation model.

To reduce user interaction for whole heart segmentation in congenital heart disease even further, we could improve our learned iterative segmentation model by predicting likely seed point locations automatically. This may prove challenging when faced with uncertainties in the heart orientation, topology, and presence or absence of each cardiac structure, but the annotation effort for this task is not high and a large training dataset could be created more easily than for image segmentation. As discussed in Chapter 4, the seed point locations were chosen to maximize their potential for automatic detection in future. An iterative procedure could be taken here too, by finding the easiest seed points, growing their structures, and using the results to guide the localization of the next seed(s). This approach has been previously taken for human pose estimation, in which previously predicted joint positions can guide the localization of subsequent joints in the sequence [149]. In fact, segmenting each cardiac structure sequentially mimics how clinicians tend to look at 3D cardiac MRI scans, namely by identifying an easy-to-find structure in one 2D slice and tracking the flow of blood through each connected piece of the anatomy.

### 5.1.2 Dynamic Heart Models

In this thesis, we developed new methods to segment ECG-gated 3D cardiac MRI scans and create static patient-specific heart models. However, the heart’s motion is

an important clinical factor, and “cine” data is routinely required to capture cardiac anatomy over the entire cardiac cycle, either as videos from a slice through the heart or as sequences of 3D images [174].

Segmenting each 3D image in a sequence would yield a series of 3D heart models that can be rendered in quick succession to portray a patient-specific beating heart model. Since the changes in heart shape due to cardiac motion are relatively small compared to the large shape variability in CHD, simply applying our current models to each image separately is likely to be quite successful. However, motion could be a useful cue in discriminating between different cardiac structures, and our models could be upgraded to input multiple 3D images and potentially output multiple different segmentations. In addition, just as prior knowledge of expected organ shapes can be used to improve a segmentation network’s output [175–177], regularization based on prior knowledge of cardiac motion could also be developed in future.

### 5.1.3 Weak Supervision

The accuracy of our iterative segmentation model could potentially be improved in future without additional annotation effort via weak supervision [178, 179]. Before our developments, success in this endeavor would have been unlikely because the accuracy of existing segmentation methods was too low for subjects with severe heart malformations. In particular, we have access to thousands of 3D MRI scans, with which we can associate diagnoses and related prior knowledge for each of the  $\sim 30$  types of heart defects listed in Table 2.1.

Multi-task learning could be used to train models that predict diagnoses for unsegmented images alongside segmentation estimation for training images for which ground truth segmentations are available [180, 181]. Encouraging the model to learn features that discriminate between different diagnoses may be useful for segmentation, and the diagnoses will not need to be known in advance when a new image is to be segmented. (Note that multi-task learning could also be helpful to learn useful features when training classification models for CHD diagnosis tasks because some CHD subtypes, e.g., ASD, common atrium, and inverted ventricles, may induce

relatively small image differences but cause massive changes in the segmentation. Alternatively, an inferred segmentation could be used directly as the sole input to a diagnosis network, provided that it is accurate enough or segmentation errors are discriminative).

In future, additional loss terms tailored for whole heart segmentation could also be formulated based on prior knowledge. For example, this can include soft or hard constraints on which objects should exist in the image, their expected sizes, and the anticipated connectivity between them [182–185]. In our application, each subject’s diagnoses informs the number of structures, their rough anatomical configuration, and their expected size, shape and connectivity. Another useful loss term for semi-supervised learning would be to encourage each structure’s segmentation to have a single connected component [186], since we observed that this is a consistent problem for the direct neural network models explored in Chapter 4.

#### **5.1.4 Evaluation of Segmentation Accuracy**

A final outstanding question is how segmentation accuracy should be assessed [187]. The accuracy required may vary across different areas of the heart; for example more accurate boundaries may be necessary near heart defects. Crucially, evaluation metrics such as cross-entropy and the Dice score that operate on a voxel-by-voxel basis are not sensitive to important shape changes (Figure 5-1). For example, an ASD diagnosis depends on whether there is a gap in a wall that is only a couple of voxels thick. This has consequences both for model training and when different methods are compared. Although generative adversarial networks (GANs) [188, 189] can be applied to image segmentation to assess whether an image-segmentation pair is realistic [190], it remains unclear how to train these models on very small datasets with high variability. Whether the segmentation accurately reflects critical aspects of the true anatomy might be assessed via similarity measures that are more sensitive to global shape changes (e.g., [186, 191, 192]) or that evaluate the area of any shared borders between structures that should or should not be connected. Another option could be to derive accuracy measures from models trained to predict CHD subtypes





Figure 5-1: Comparing image segmentations of the heart’s two ventricles illustrates the deficiencies of the Dice score and average surface distance, and the advantages of a spectral shape similarity measure (nWESD [191, 192]). The segmentations in the top and bottom rows have the same number of incorrect pixels, even though those on the bottom are intuitively much better. For each column, bold scores indicate higher similarity to the original segmentation on the left.

from input segmentations. However, such models would need to be trained using ground truth segmentations, so that reproducible segmentation errors are not used as a cue for classification, and such segmentations are limited in number. Ultimately, any segmentation tool’s required accuracy must be defined by physicians, and its fundamental utility can only be judged after considering its impact on patient care and clinical outcomes.

## 5.2 Clinical Outlook

Whole heart segmentation is chiefly performed to visualize cardiac anatomy or to quantify cardiac function [10, 23]. We discuss both of these aspects below.

### 5.2.1 Visualization and Surgical Planning

The segmentation methods provided in this thesis offer great potential to create patient-specific 3D heart models in the clinic, without extremely tedious manual segmentation. As described in Chapter 1, many case studies indicate the power of patient-specific heart models to support clinical decision making and/or pre-surgery practice. Whether virtual display is sufficient or if 3D-printing offers benefits remains an open question, and may vary from clinician to clinician according to personal preference and the degree to which they “see” with their hands. The extra time required for 3D-printing is not a major concern in this debate, because imaging is typically done a couple of days before surgery (for time-sensitive situations, about three hours are available after imaging while the patient and operating room are prepared, which would require fast 3D-printing). It is also not yet known whether it is preferable to present a model of the cardiac blood pool, a shell model of the heart walls surrounding it, or a model that separately colors each cardiac structure, nor if and how models should be cut to reveal the interior.

### 5.2.2 Cardiac Function and Simulation

Automated whole heart segmentation can also be used to compute important functional indices (e.g., volumes, ejection fraction, myocardial measurements and motion analysis) that are traditionally derived from manual segmentations [21, 22]. Contour representations of valves and defects could be derived from the segmentation boundaries that separate different cardiac structures, which if accurate enough could be used for detailed measuring of the valve annuli [103] or septal defects [193]. Segmentations could also be used to create patient-specific models that incorporate biophysical properties [194, 195] or information from 4D flow MRI and computational fluid dynamics [196]. An even more ambitious idea is virtual surgical simulation [197], which would be aided by the fact that we provide separate segmentations of the different cardiac structures that could then be cut into and manipulated. Finally, it may be possible to create longitudinal heart atlases and derived clinical biomarkers

from CHD patients that are scanned over their lifetime, inspired by analysis that has been applied to the brain [198, 199]. Given a new patient, in future one may even be able to predict the consequences of competing surgical approaches on their cardiac anatomy and future outcome, after evaluating the histories of other subjects with the most similar anatomy or based on outcome prediction models learned from data.

### 5.3 Conclusions

In this thesis, we have developed new interactive image segmentation methods for clinical problems that involve extreme anatomical variability and little training data. We have developed the first datasets and methods for whole heart segmentation for patients with congenital heart disease. This includes a patch-based interactive segmentation method that can incorporate active learning, and a learned iterative segmentation model that can generalize from very small datasets to severe cardiac pathologies. Our contributions have potential clinical impact on improving the utility of preoperative imaging for surgical planning in congenital heart disease, and on medical image analysis for highly variable diseases more generally.



# Appendix A

## Learning Iterative Segmentation: Supplemental Material

### A.0.1 Network Architectures

Architectural parameters describing each network are provided in Table A.1. These parameters were tuned independently, for each of the networks (**U-Net-All**, **U-Net**, **U-Net+S**, **Iter-A**) and cardiac structures (LV, RV, LA, RA, AO, PA, SVC, IVC). However, the best-performing network architecture was the same for all of the binary U-nets and for all cardiac structures.

Table A.1: Network architectures

Method	Input Size	Levels	$C$	Maxpool
<b>U-Net-All</b>	120×168×132	4	24	$3^3, 2^3, 2^3$
<b>U-Net</b>	126×180×144	3	24	$3^3, 3^3$
<b>U-Net+S</b>	126×180×144	3	24	$3^3, 3^3$
<b>Iter-A</b>	126×180×144	3	24	$3^3, 3^3$
	Receptive field			Parameters
<b>U-Net-All</b>	$128 \times 128 \times 128$			3,610,201
<b>U-Net</b>	$68 \times 68 \times 68$			872,162
<b>U-Net+S</b>	$68 \times 68 \times 68$			873,458
<b>Iter-A</b>	$68 \times 68 \times 68$			874,107

## A.0.2 Training Data Generation

The step sizes ( $d_s$ ) for each structure are given in Table A.2.

Table A.2: Step sizes  $d_s$

	LV	RV	LA	RA	AO	PA	SVC	IVC
$d_s$	3	3.5	3	2.5	10	5	3	3

### Great Vessel Segmentations:

The seed points are not randomized during training, since the vessels are quite narrow in cross-section, and slightly changing the seed point’s location along the centerline does not impact accuracy. The complete segmentation  $\mathbf{y}$ ’s distance map  $d_{\mathbf{y}}(\cdot)$  (from fast marching) is randomly thresholded to form the input partial segmentation  $\mathbf{y}_{in}$ , which is then corrupted as described in Section A.0.3. The maximum value  $d_1$  is found after intersecting the corrupted  $\mathbf{y}_{in}$  with  $d_{\mathbf{y}}(\cdot)$ , ignoring any new free-floating blobs, and  $d(\cdot)$  is thresholded at  $d_2 = d_1 + d_s$  to form  $\mathbf{y}_{out}$ . The stopping indicator  $s = 1$  if  $d_2$  is at least  $(100 - 1/2 \cdot d_s)\%$  finished compared to the maximum distance in  $d_{\mathbf{y}}(\cdot)$ .

### Chamber Segmentations:

For both **U-Net+S** and iterative segmentation, during training a seed point  $\mathbf{y}_0'$  is randomized within 25% of the maximum possible distance (in the complete segmentation  $\mathbf{y}$ ) from the example seed point  $\mathbf{y}_0$ . The input partial segmentation  $\mathbf{y}_{in}$  is generated by intersecting a sphere that is centered on  $\mathbf{y}_0'$  and has random radius  $d_1$  with  $\mathbf{y}$ , and then corrupting it as described in Section A.0.3. The partial output segmentation  $\mathbf{y}_{out}$  is created by keeping only the island connected to the initial seed  $\mathbf{y}_0'$ , applying  $d_s$  dilations, and masking with  $\mathbf{y}$ . The stopping indicator  $s = 1$  if at least  $(100 - 1/2 \cdot d_s)\%$  of the voxels in  $\mathbf{y}_{out}$  are finished as compared to  $\mathbf{y}$ .

### A.0.3 Data Augmentation

#### Image Transformations:

The random affine transformations uses rotation uniformly distributed in  $[-7^\circ, 7^\circ]$ , translation uniformly distributed in  $[-5, 5]$  voxels, shear with shear factor uniformly distributed in  $[0.9, 1.1]$ , and scaling with scale factor uniformly distributed in  $[1 - s, 1 + s]$ , where  $s$  was tuned for each experiment (Table A.3) since different cardiac structures are expected to vary in size more than others.

Table A.3: Scale factor  $s$  bounds for affine transformations

Method	LV	RV	LA	RA	AO	PA	SVC	IVC
<b>U-Net-All</b>	0.2							
<b>U-Net</b>	0.2	0.1	0.2	0.1	0.1	0.2	0.2	0.2
<b>U-Net+S</b>	0.1	0.1	0.2	0.2	0.1	0.2	0.2	0.2
<b>Iter-A</b>	0.2	0.1	0.1	0.1	0.1	0.2	0.1	0.2

The nonlinear transformations are created by randomizing displacement vectors on a  $4 \times 6 \times 4$  grid (approximately every 30 voxels). The maximum displacement allowed in each direction,  $m$ , is chosen uniformly at random in  $[3, 6]$  voxels, so that some images will be highly deformed and some will not. The coarse displacements are sampled uniformly in  $[-m, m]$ , the displacement field is resampled onto the original image grid via bicubic interpolation, and the images are warped using linear interpolation for images and nearest neighbor interpolation for segmentations.

The constant intensity shifts are uniformly distributed in  $[-1.0005, 0.7395]$ . These bounds were chosen to ensure that, for each cardiac structure, an estimate of the overall intensity histogram after data augmentation covered the intensity distribution for each individual image well. The additive Gaussian noise has mean 0 and standard deviation uniformly distributed in  $[0, 0.05]$ , so that the resulting images have varying amounts of noise.

### Random dark and bright blobs for AO and PA:

To mimic dark inhomogeneity artifacts caused by stents and the heterogenous background appearance, random dark blobs are created inside of the AO and PA, and random dark or bright blobs are created just outside of the AO and PA. The probability of creating a blob inside a structure equals the probability of creating a blob outside a structure (Table A.4). At most one blob can be created per structure. For our iterative segmentation method, blobs are not created if the output segmentation is 90% finished or more.

Table A.4: Probability of creating a random intensity-shifted blob

Method	Train on HVSMR+, Test on HVSMR++		Train and Test on Subsets of HVSMR++	
	AO	PA	AO	PA
<b>U-Net-All</b>	0		0	
<b>U-Net</b>	0	0	0.2	0.2
<b>U-Net+S</b>	0.1	0.2	0.2	0.1
<b>Iter-A</b>	0.2	0.1	0.2	0.2

All blobs are initialized as a cube. The cube’s radius is chosen uniformly at random (in  $[3, 45]$  voxels for inside blobs and in  $[17, 45]$  voxels for outside blobs). For direct segmentation methods (**U-Net-All**, **U-Net** and **U-Net+S**), the cube’s center is chosen uniformly at random anywhere within the ground truth segmentation for inside blobs, and anywhere in the border region outside of the ground truth segmentation for outside blobs (at most 5 voxels from the boundary). For our iterative segmentation method, coordinates close to the area in which the segmentation evolved (i.e., in the difference region between the output and input segmentations) have a higher probability of being chosen as the cube’s center. Finally, the cubes are deformed into blobs, by randomizing displacement vectors on a  $15 \times 22 \times 18$  grid (approximately every 8 voxels), with the maximum displacement allowed in each direction,  $m$ , chosen uniformly at random in  $[5, 15]$  voxels.

All blobs created inside a structure are dark. If a blob is created outside a struc-



ture, it is dark with probability 0.5 and bright with probability 0.5. The intensity of a dark blob is initially chosen uniformly at random between the minimum image intensity and the minimum image intensity plus 40% of the difference between the average and minimum intensities. The intensity of a bright blob is initially chosen uniformly at random between the maximum image intensity minus 80% of the difference between the maximum and average intensities, and the maximum image intensity minus 40% of the difference between the maximum and average intensities. Additive Gaussian noise with mean 0 and standard deviation 0.02 is applied, and average pooling with a pool size of 11 is used to blur the blob’s initially binary representation. The ground truth segmentation is subtracted from outside blobs such that no voxel in an outside blob is within 2 voxels of the ground truth segmentation boundary. Finally, the blob is averaged with the original image.

### **Corrupting Partial Input Segmentations:**

When training our interactive segmentation model, partial input segmentations  $\mathbf{y}_{in}$  are corrupted in two ways:

First, the segmentations undergo random nonlinear transformations, without similarly deforming the underlying image. Deformations are applied by randomizing displacement vectors on a  $9 \times 12 \times 10$  grid (approximately every 15 voxels), with the maximum displacement allowed in each direction,  $m$ , chosen uniformly at random in  $[0.6, 6]$  voxels.

Second, we insert random foreground blobs that vary in number, location and size. These blobs can be attached to the segmentation, or free-floating. They are created by first randomizing the number of each type of blob uniformly in  $[0, 4]$ . Then for each blob, a cube is created with a random center location, and radius chosen uniformly at random from the bounds listed in Table A.5. Finally, the cubes are deformed into blobs, by randomizing displacement vectors on a  $15 \times 22 \times 18$  grid (approximately every 8 voxels), with the maximum displacement allowed in each direction,  $m$ , chosen uniformly at random in  $[2.5, 10]$  voxels.

Table A.5: Cube size bounds for foreground blobs (0 indicates no foreground blobs)

	LV	RV	LA	RA	AO	PA	SVC	IVC
Min	5	5	5	5	5	5	5	0
Max	60	60	60	60	35	35	35	0

#### A.0.4 Learning

The model’s weight and bias parameters were initialized using the default Keras initializers.

Our loss function more strongly penalizes errors near segmentation boundaries, as described in eq. (4.12). We used the segmentation boundary distance  $d_0 = 5$  and the segmentation boundary weights  $\omega_0$  in Table A.6.

Table A.6: Segmentation boundary weights  $\omega_0$

Method	LV	RV	LA	RA	AO	PA	SVC	IVC
<b>U-Net-All</b>	10							
<b>U-Net</b>	50	50	50	50	50	50	50	50
<b>U-Net+S</b>	10	10	10	10	20	20	20	20
<b>Iter-A</b>	30	30	30	30	50	50	50	50

## A.0.5 Additional Figures

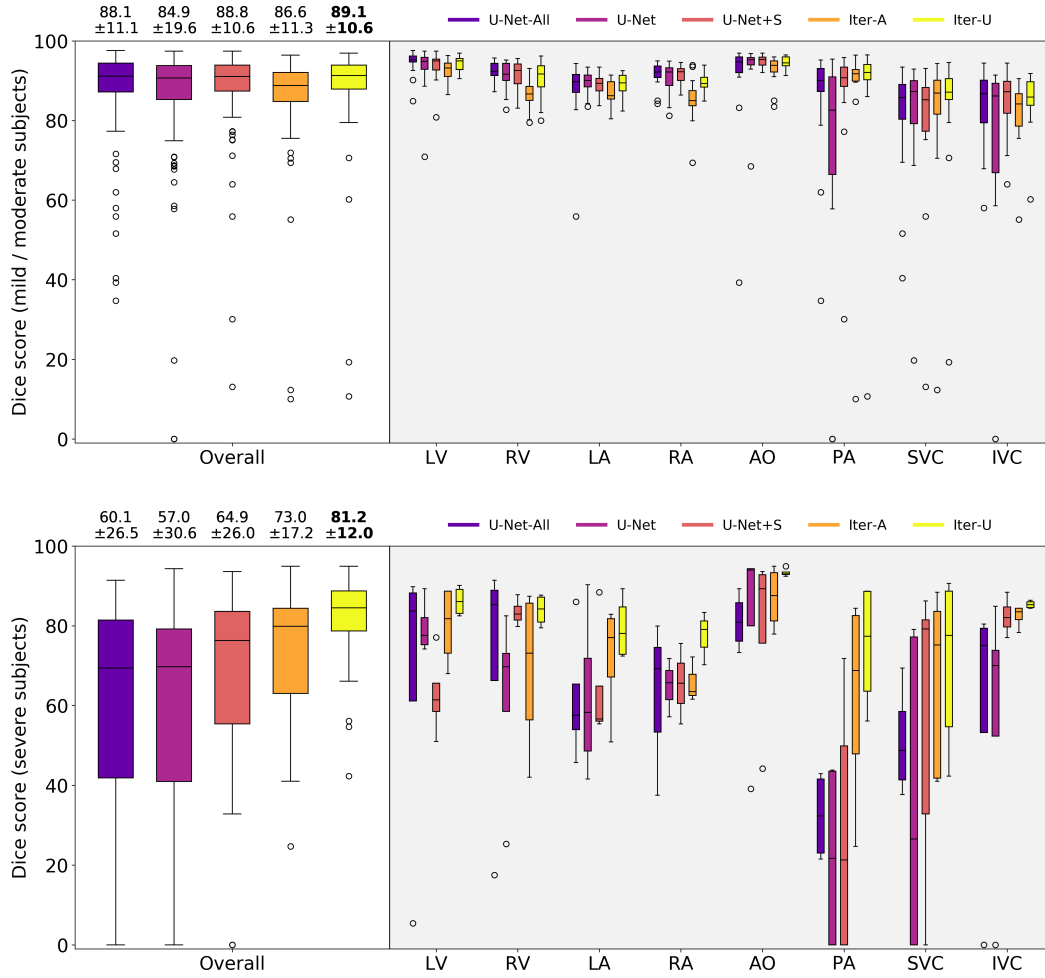


Figure A-1: HVS MR+ cross-validation summary statistics (Dice score). (Top) For mild and moderate subjects, all five methods had comparable performance. (Bottom) For severe subjects, the iterative segmentation methods (**Iter-A** and **Iter-U**) were superior, especially iterative segmentation with user stopping. For each method, the mean and standard deviation of the overall Dice score is shown at the top of the graph, with the best mean score in bold.



# Bibliography

- [1] C. Frescura, E. V. Büchel, S. Y. Ho, and G. Thiene. Anatomical and pathophysiological classification of congenital heart disease. In F. Saremi, S. Achenbach, E. Arbustini, and J. Narula, editors, *Revisiting Cardiac Anatomy: A Computed-Tomography-Based Atlas and Reference*, chapter 2, pages 40–75. Blackwell Publishing, Chichester, UK, 2010.
- [2] Centers for Disease Control and Prevention (CDC). Trends in infant mortality attributable to birth defects – United States, 1980-1995. *Morbidity and Mortality Weekly Report (MMWR)*, 47(37):773–778, 1998.
- [3] M. E. Oster, K. A. Lee, M. A. Honein, T. Riehle-Colarusso, M. Shin, and A. Correa. Temporal trends in survival among infants with critical congenital heart defects. *Pediatrics*, 131(5):e1502–e1508, 2013.
- [4] A. J. Marelli, A. S. Mackie, R. Ionescu-Ittu, E. Rahme, and L. Pilote. Congenital heart disease in the general population: Changing prevalence and age distribution. *Circulation*, 115(2):163–172, 2007.
- [5] K. A. Young, J. A. Wise, P. DeSaix, D. H. Kruse, B. Poe, E. Johnson, J. E. Johnson, O. Korol, J. G. Betts, and M. Womble. *Anatomy and Physiology by OpenStax*. XanEdu Publishing Inc, Houston, TX, USA, 2013.
- [6] A. M. Gaca, J. J. Jaggars, L. T. Dudley, and G. S. Bisset. Repair of congenital heart disease: A primer - Part 1. *Radiology*, 247(3):617–631, 2008.
- [7] K. M. Farooqi, J. C. Nielsen, S. C. Uppu, S. Srivastava, I. A. Parness, J. Sanz, and K. Nguyen. Use of 3-Dimensional printing to demonstrate complex intracardiac relationships in double-outlet right ventricle for surgical planning. *Circulation: Cardiovascular Imaging*, 8(5):e003043, 2015.
- [8] B. Pandya, S. Cullen, and F. Walker. Congenital heart disease in adults. *BMJ*, 354:i3905, 2016.
- [9] H. N. Ntsinjana, M. L. Hughes, and A. M. Taylor. The role of cardiovascular magnetic resonance in pediatric congenital heart disease. *Journal of Cardiovascular Magnetic Resonance*, 13:51, 2011.

- [10] A. Arafati, P. Hu, J. P. Finn, C. Rickers, A. L. Cheng, H. Jafarkhani, and A. Kheradvar. Artificial intelligence in pediatric and adult congenital cardiac MRI: An unmet clinical need. *Cardiovascular Diagnosis and Therapy*, 9(Suppl 2):S310–S325, 2019.
- [11] P. Bhatla, J. T. Tretter, A. Ludomirsky, M. Argilla, L. A. Latson, S. Chakravarti, P. C. Barker, S.-J. Yoo, D. B. McElhinney, N. Wake, and R. S. Mosca. Utility and scope of rapid prototyping in patients with complex muscular ventricular septal defects or double-outlet right ventricle: Does it alter management decisions? *Pediatric Cardiology*, 38(1):103–114, 2017.
- [12] I. Lau and Z. Sun. Three-dimensional printing in congenital heart disease: A systematic review. *Journal of Medical Radiation Sciences*, 65(3):226–236, 2018.
- [13] I. Valverde, G. Gomez-Ciriza, T. Hussain, C. Suarez-Mejias, M. N. Velasco-Forte, N. Byrne, A. Ordoñez, A. Gonzalez-Calle, D. Anderson, M. G. Hazekamp, A. A. W. Roest, J. Rivas-Gonzalez, S. Uribe, I. El-Rassi, J. Simpson, O. Miller, E. Ruiz, I. Zabala, A. Mendez, B. Manso, P. Gallego, F. Prada, M. Cantinotti, L. Ait-Ali, C. Merino, A. Parry, N. Poirier, G. Greil, R. Razavi, T. Gomez-Cia, and A.-R. Hosseinpour. Three-dimensional printed models for surgical planning of complex congenital heart defects: An international multicentre study. *European Journal of Cardio-Thoracic Surgery*, 52(6):1139–1148, 2017.
- [14] S. Garekar, A. Bharati, M. Chokhandre, S. Mali, B. Trivedi, V. P. Changela, N. Solanki, S. Gaikwad, and V. Agarwal. Clinical application and multidisciplinary assessment of three dimensional printing in double outlet right ventricle with remote ventricular septal defect. *World Journal for Pediatric & Congenital Heart Surgery*, 7(3):344–350, 2016.
- [15] E. Riesenkampff, U. Rietdorf, I. Wolf, B. Schnackenburg, P. Ewert, M. Huebler, V. Alexi-Meskishvili, R. H. Anderson, N. Engel, H.-P. Meinzer, R. Hetzer, F. Berger, and T. Kuehne. The practical clinical value of three-dimensional models of complex congenitally malformed hearts. *The Journal of Thoracic and Cardiovascular Surgery*, 138(3):571–580, 2009.
- [16] D. Schmauss, S. Haeberle, C. Hagl, and R. Sodian. Three-dimensional printing in cardiac surgery and interventional cardiology: A single-centre experience. *European Journal of Cardio-Thoracic Surgery*, 47(6):1044–1052, 2015.
- [17] I. Shiraishi, M. Yamagishi, K. Hamaoka, M. Fukuzawa, and T. Yagihara. Simulative operation on congenital heart disease using rubber-like urethane stereolithographic biomodels based on 3D datasets of multislice computed tomography. *European Journal of Cardio-Thoracic Surgery*, 37(2):302–306, 2010.
- [18] J. R. Ryan, T. G. Moe, R. Richardson, D. H. Frakes, J. J. Nigro, and S. Pophal. A novel approach to neonatal management of tetralogy of fallot, with pulmonary atresia and multiple aortopulmonary collaterals. *JACC: Cardiovascular Imaging*, 8(1):103–104, 2015.

- [19] T. Loke, A. Krieger, C. Sable, and L. Olivieri. Novel uses for three-dimensional printing in congenital heart disease. *Current Pediatrics Reports*, 4(2):28–34, 2016.
- [20] J. P. Costello, L. J. Olivieri, L. Su, A. Krieger, F. Alfares, O. Thabit, M. B. Marshall, S.-J. Yoo, P. C. Kim, R. A. Jonas, and D. S. Nath. Incorporating three-dimensional printing into a simulation-based congenital heart disease and critical care training curriculum for resident physicians. *Congenital Heart Disease*, 10(2):185–190, 2015.
- [21] A. Seraphim, K. D. Knott, J. Augusto, A. N. Bhuvu, C. Manisty, and J. C. Moon. Quantitative cardiac MRI. *Journal of Magnetic Resonance Imaging*, 51(3):693–711, 2020.
- [22] S. E. Petersen, M. Y. Khanji, S. Plein, P. Lancellotti, and C. Bucciarelli-Ducci. European Association of Cardiovascular Imaging expert consensus paper: a comprehensive review of cardiovascular magnetic resonance normal values of cardiac chamber size and aortic root in adults and recommendations for grading severity. *European Heart Journal - Cardiovascular Imaging*, 20(12):1321–1331, 2019.
- [23] X. Zhuang. Challenges and methodologies of fully automatic whole heart segmentation: A review. *Journal of Healthcare Engineering*, 4(3):371–408, 2013.
- [24] X. Zhuang, L. Li, C. Payer, D. Ååtern, M. Urschler, M. P. Heinrich, J. Oster, C. Wang, Ö. Smedby, C. Bian, X. Yang, P.-A. Heng, A. Mortazi, U. Bagci, G. Yang, C. Sun, G. Galisot, J.-Y. Ramel, T. Brouard, Q. Tong, W. Si, X. Liao, G. Zeng, Z. Shi, G. Zheng, C. Wang, T. MacGillivray, D. Newby, K. Rhode, S. Ourselin, R. Mohiaddin, J. Keegan, D. Firmin, and G. Yang. Evaluation of algorithms for multi-Modality whole heart segmentation: An open-access grand challenge. *Medical Image Analysis*, 58:101537, 2019.
- [25] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 29(2):155–195, 2016.
- [26] N. Byrne, M. Velasco Forte, A. Tandon, I. Valverde, and T. Hussain. A systematic review of image segmentation methodology, used in the additive manufacture of patient-specific 3D printed models of the cardiovascular system. *JRSM Cardiovascular Disease*, 5:2048004016645467, 2016.
- [27] S. Jacobs, R. Grunert, F. W. Mohr, and V. Falk. 3D-imaging of cardiac structures using 3D heart models for planning in heart surgery: A preliminary study. *Interactive Cardiovascular and Thoracic Surgery*, 7(1):6–9, 2008.

- [28] I. Valverde, G. Gomez, A. Gonzalez, C. Suarez-Mejias, A. Adsuar, J. F. Coserria, S. Uribe, T. Gomez-Cia, and A. R. Hosseinpour. Three-dimensional patient-specific cardiac model for surgical planning in Nikaidoh procedure. *Cardiology in the Young*, 25(4):698–704, 2014.
- [29] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins. Next-generation machine learning for biological networks. *Cell*, 173(7):1581–1592, 2018.
- [30] Y. Zhang and C. Ling. A strategy to apply machine learning to small datasets in materials science. *npj Computational Materials*, 4(25):1–8, 2018.
- [31] Y. Gao and K. M. Mosalam. Deep transfer learning for image-based structural damage recognition. *Computer-Aided Civil and Infrastructure Engineering*, 33(9):748–768, 2018.
- [32] R. Hanna, H. Barschdorf, T. Klinder, F. M. Weber, M. W. Krueger, O. Dössel, and C. Lorenz. A hybrid method for automatic anatomical variant detection and segmentation. In *Functional Imaging and Modeling of the Heart (FIMH)*, volume 6666 of *Lecture Notes in Computer Science*, pages 333–340. 2011.
- [33] D. Kutra, A. Saalbach, H. Lehmann, A. Groth, S. P. M. Dries, M. W. Krueger, O. Dössel, and J. Weese. Automatic multi-model-based segmentation of the left atrium in cardiac MRI scans. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 7511 of *Lecture Notes in Computer Science*, pages 1–8. 2012.
- [34] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 105–112, 2001.
- [35] S. Andrews, G. Hamarneh, and A. Saad. Fast random walker with priors using precomputation for interactive medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 6363 of *Lecture Notes in Computer Science*, pages 9–16. 2010.
- [36] A. Criminisi, T. Sharp, and A. Blake. GeoS: Geodesic image segmentation. In *European Conference on Computer Vision (ECCV)*, volume 5302 of *Lecture Notes in Computer Science*, pages 99–112, 2008.
- [37] L. Zhu, I. Kolesov, Y. Gao, R. Kikinis, and A. Tannenbaum. An effective interactive medical image segmentation method using fast GrowCut. In *MICCAI Workshop on Interactive Medical Image Computing (IMIC)*, 2014.
- [38] P. Karasev, I. Kolesov, K. Fritscher, P. Vela, P. Mitchell, and A. Tannenbaum. Interactive medical image segmentation using PDE control of active contours. *IEEE Transactions on Medical Imaging*, 32(11):2127–2139, 2013.



- [39] Y. Gao, R. Kikinis, S. Bouix, M. Shenton, and A. Tannenbaum. A 3D interactive multi-object segmentation tool using local robust statistics driven active contours. *Medical Image Analysis*, 16(6):1216–1227, 2012.
- [40] O. Shitrit, T. Hershkovich, T. Shalmon, I. Shelef, and T. Riklin-Raviv. Probabilistic model for 3D interactive segmentation. In *MICCAI Workshop on Interactive Medical Image Computing (IMIC)*, 2014.
- [41] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren. DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1559–1572, 2019.
- [42] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging*, 37(7):1562–1573, 2018.
- [43] M. Amrehn, S. Gaube, M. Unberath, F. Schebesch, T. Horz, M. Strumia, S. Steidl, M. Kowarschik, and A. Maier. UI-Net: Interactive artificial neural networks for iterative image segmentation based on a user model. In *Eurographics Workshop on Visual Computing for Biology and Medicine (EG VCBM)*, pages 143–147, 2017.
- [44] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang. Deep interactive object selection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 373–381, 2016.
- [45] T. McInerney and D. Terzopoulos. Deformable models in medical image analysis: A survey. *Medical Image Analysis*, 1(2):91–108, 1996.
- [46] T. Heimann and H. P. Meinzer. Statistical shape models for 3D medical image segmentation: A review. *Medical Image Analysis*, 13(4):543–563, 2009.
- [47] J. E. Iglesias and M. R. Sabuncu. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219, 2015.
- [48] J. Peters, O. Ecabert, C. Meyer, R. Kneser, and J. Weese. Optimizing boundary detection via simulated search with applications to multi-modal heart segmentation. *Medical Image Analysis*, 14(1):70–84, 2010.
- [49] O. Ecabert, J. Peters, H. Schramm, C. Lorenz, J. von Berg, M. J. Walker, M. Vembar, M. E. Olszewski, K. Subramanyan, G. Lavi, and J. Weese. Automatic model-based segmentation of the heart in CT images. *IEEE Transactions on Medical Imaging*, 27(9):1189–1201, 2008.
- [50] O. Ecabert, J. Peters, M. J. Walker, T. Ivanc, C. Lorenz, J. v. Berg, J. Lessick, M. Vembar, and J. Weese. Segmentation of the heart and great vessels in CT

- images using a model-based adaptation framework. *Medical Image Analysis*, 15(6):863–876, 2011.
- [51] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu. Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features. *IEEE Transactions on Medical Imaging*, 27(11):1668–1681, 2008.
- [52] H. A. Kirisli, M. Schaap, S. Klein, S. L. Papadopoulou, M. Bonardi, C. H. Chen, A. C. Weustink, N. R. Mollet, E. J. Vonken, R. J. v. d. Geest, T. v. Walsum, and W. J. Niessen. Evaluation of a multi-atlas based method for segmentation of cardiac CTA data: A large-scale, multicenter, and multivendor study. *Medical Physics*, 37(12):6279–6291, 2010.
- [53] X. Zhuang, K. Rhode, R. Razavi, D. Hawkes, and S. Ourselin. A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. *IEEE Transactions on Medical Imaging*, 29(9):1612–1625, 2010.
- [54] M. A. Zuluaga, M. J. Cardoso, M. Modat, and S. Ourselin. Multi-atlas propagation whole heart segmentation from MRI and CTA using a local normalised correlation coefficient criterion. In *Functional Imaging and Modeling of the Heart (FIMH)*, volume 7945 of *Lecture Notes in Computer Science*, pages 174–181. 2013.
- [55] X. Zhuang, W. Bai, J. Song, S. Zhan, X. Qian, W. Shi, Y. Lian, and D. Rueckert. Multiatlas whole heart segmentation of CT data using conditional entropy for atlas ranking and selection. *Medical Physics*, 42(7):3822–3833, 2015.
- [56] X. Zhuang and J. Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical Image Analysis*, 31:77–87, 2016.
- [57] M. A. Zuluaga, N. Burgos, A. F. Mendelson, A. M. Taylor, and S. Ourselin. Voxelwise atlas rating for computer assisted diagnosis: Application to congenital heart diseases of the great arteries. *Medical Image Analysis*, 26(1):185–194, 2015.
- [58] X. Albà, M. Pereañez, C. Hoogendoorn, A. J. Swift, J. M. Wild, A. F. Frangi, and K. Lekadir. An algorithm for the segmentation of highly abnormal hearts using a generic statistical shape model. *IEEE Transactions on Medical Imaging*, 35(3):845–859, 2016.
- [59] W. Shi, X. Zhuang, H. Wang, S. Duckett, D. Oregan, P. Edwards, S. Ourselin, and D. Rueckert. Automatic segmentation of different pathologies from cardiac cine MRI using registration and multiple component EM estimation. In *Functional Imaging and Modeling of the Heart (FIMH)*, volume 6666 of *Lecture Notes in Computer Science*, pages 163–170. 2011.

- [60] A. Eslami, A. Karamalis, A. Katouzian, and N. Navab. Segmentation by retrieval with guided random walks: Application to left ventricle segmentation in MRI. *Medical Image Analysis*, 17(2):236–253, 2013.
- [61] M. R. Avendi, A. Kheradvar, and H. Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Medical Image Analysis*, 30:108–119, 2016.
- [62] M. Depa, M. R. Sabuncu, G. Holmvang, R. Nezafat, E. J. Schmidt, and P. Golland. Robust atlas-based segmentation of highly variable anatomy: Left atrium segmentation. In *Statistical Atlases and Computational Models of the Heart (STACOM)*, volume 6364 of *Lecture Notes in Computer Science*, pages 85–94. 2010.
- [63] Y. Zheng, D. Yang, M. John, and D. Comaniciu. Multi-part modeling and segmentation of left atrium in C-arm CT for image-guided ablation of atrial fibrillation. *IEEE Transactions on Medical Imaging*, 33(2):318–331, 2014.
- [64] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.
- [65] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. 2015.
- [66] C. Payer, D. Stern, H. Bischof, and M. Urschler. Multi-label whole heart segmentation using CNNs and anatomical label configurations. In *Statistical Atlases and Computational Models of the Heart (STACOM)*, volume 10663 of *Lecture Notes in Computer Science*, pages 190–198. 2017.
- [67] Q. Tong, M. Ning, W. Si, X. Liao, and J. Qin. 3D deeply-supervised U-Net based whole heart segmentation. In *Statistical Atlases and Computational Models of the Heart (STACOM)*, volume 10663 of *Lecture Notes in Computer Science*, pages 224–232, 2018.
- [68] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng. Hybrid loss guided convolutional networks for whole heart parsing. In *Statistical Atlases and Computational Models of the Heart (STACOM)*, volume 10663 of *Lecture Notes in Computer Science*, pages 215–223, 2018.
- [69] C. Wang and O. Smedby. Automatic whole heart segmentation using deep learning and shape context. In *Statistical Atlases and Computational Models of the Heart (STACOM)*, volume 10663 of *Lecture Notes in Computer Science*, pages 242–249. 2017.

- [70] A. Mortazi, J. Burt, and U. Bagci. Multi-planar deep segmentation networks for cardiac substructures from MRI and CT. In *Statistical Atlases and Computational Models of the Heart (STACOM)*, volume 10663 of *Lecture Notes in Computer Science*, pages 199–206, 2018.
- [71] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng. 3D convolutional networks for fully automatic fine-grained whole heart partition. In *Statistical Atlases and Computational Models of the Heart (STACOM)*, volume 10663 of *Lecture Notes in Computer Science*, pages 181–189, 2018.
- [72] H. Zhang, A. Wahle, R. Johnson, T. Scholz, and M. Sonka. 4-D cardiac MR image analysis: Left and right ventricular morphology and function. *IEEE Transactions on Medical Imaging*, 29(2):350–364, 2010.
- [73] D. H. Ye, B. Desjardins, J. Hamm, H. Litt, and K. M. Pohl. Regional manifold learning for disease classification. *IEEE Transactions on Medical Imaging*, 33(6):1236–1247, 2014.
- [74] J. L. Bruse, K. McLeod, G. Biglino, H. N. Ntsinjana, C. Capelli, T.-Y. Hsia, M. Sermesant, X. Pennec, A. M. Taylor, and S. Schievano. A non-parametric statistical shape model for assessment of the surgically repaired aortic arch in coarctation of the aorta: How normal is abnormal? In *Statistical Atlases and Computational Models of the Heart (STACOM)*, volume 9534 of *Lecture Notes in Computer Science*, pages 21–29. 2015.
- [75] T. Mansi, I. Voigt, B. Leonardi, X. Pennec, S. Durrleman, M. Sermesant, H. Delingette, A. Taylor, Y. Boudjemline, G. Pongiglione, and N. Ayache. A statistical model for quantification and prediction of cardiac remodelling: Application to Tetralogy of Fallot. *IEEE Transactions on Medical Imaging*, 30(9):1605–1616, 2011.
- [76] K. Punithakumar, M. Noga, I. Ben Ayed, and P. Boulanger. Right ventricular segmentation in cardiac MRI with moving mesh correspondences. *Computerized Medical Imaging and Graphics*, 43:15–25, 2015.
- [77] K. Gilbert, B. R. Cowan, A. Suinesiaputra, C. Occleshaw, and A. A. Young. Rapid D-affine biventricular cardiac function with polar prediction. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 8674 of *Lecture Notes in Computer Science*, pages 546–553. 2014.
- [78] K. Ralovich, L. Itu, D. Vitanovski, P. Sharma, R. Ionasec, V. Mihalef, W. Krawtschuk, Y. Zheng, A. Everett, G. Pongiglione, B. Leonardi, R. Ringel, N. Navab, T. Heimann, and D. Comaniciu. Noninvasive hemodynamic assessment, treatment outcome prediction and follow-up of aortic coarctation from MR imaging. *Medical Physics*, 42(5):2143–2156, 2015.

- [79] Y. Zhang, D. Kwon, and K. M. Pohl. Computing group cardinality constraint solutions for logistic regression problems. *Medical Image Analysis*, 35:58–69, 2017.
- [80] M. A. Zuluaga, K. Bhatia, B. Kainz, M. H. Moghari, and D. F. Pace. *Reconstruction, Segmentation, and Analysis of Medical Images*, volume 10129 of *Lecture Notes in Computer Science*. 2016.
- [81] D. Pace, A. Dalca, T. Geva, A. Powell, M. Moghari, and P. Golland. Interactive whole-heart segmentation in congenital heart disease. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pages 80–88. 2015.
- [82] D. F. Pace, A. V. Dalca, T. Brosch, T. Geva, A. J. Powell, J. Weese, M. H. Moghari, and P. Golland. Iterative segmentation from limited training data: Applications to congenital heart disease. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA)*, volume 11045 of *Lecture Notes in Computer Science*, pages 334–342. 2018.
- [83] M. McCormick, X. Liu, J. Jomier, C. Marion, and L. Ibanez. ITK: Enabling reproducible research and open science. *Frontiers in Neuroinformatics*, 8:13, 2014.
- [84] B. Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012.
- [85] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, and R. Kikinis. 3D Slicer as an image computing platform for the Quantitative imaging network. *Magnetic Resonance Imaging*, 30(9):1323–1341, 2012.
- [86] A. Rosset, L. Spadola, and O. Ratib. OsiriX: An open-source software for navigating in multidimensional DICOM images. *Journal of Digital Imaging*, 17(3):205–216, 2004.
- [87] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011.
- [88] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.
- [89] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren. NiftyNet: A deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 158:113–122, 2018.

- [90] H. Harvey and B. Glocker. A standardised approach for preparing imaging data for machine learning tasks in radiology. In E. R. Ranschaert, S. Morozov, and P. R. Algra, editors, *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*, chapter 6, pages 61–72. Springer Nature Switzerland AG, Cham, Switzerland, 2019.
- [91] A. Kesner, R. Laforest, R. Otazo, K. Jennifer, and T. Pan. Medical imaging data in the digital innovation age. *Medical Physics*, 45(4):e40–e52, 2018.
- [92] M. J. Willeminck, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.
- [93] P. M. A. van Ooijen. Quality and curation of medical images and data. In E. R. Ranschaert, S. Morozov, and P. R. Algra, editors, *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*, chapter 17, pages 247–255. Springer Nature Switzerland AG, Cham, Switzerland, 2019.
- [94] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. The Alzheimer’s Disease Neuroimaging Initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.
- [95] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Molyer, D. Vukcevic, O. Delaneau, J. O’Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.
- [96] M. A. Ikram, G. G. O. Brusselle, S. D. Murad, C. M. van Duijn, O. H. Franco, A. Goedegebure, C. C. W. Klaver, T. E. C. Nijsten, R. P. Peeters, B. H. Stricker, H. Tiemeier, A. G. Uitterlinden, M. W. Vernooij, and A. Hofman. The Rotterdam Study: 2018 update on objectives, design and main results. *European Journal of Epidemiology*, 32(9):807–850, 2017.
- [97] M. D. Kohli, R. M. Summers, and J. R. Geis. Medical image data and datasets in the era of machine learning - Whitepaper from the 2016 C-MIMI meeting dataset session. *Journal of Digital Imaging*, 30(4):392–399, 2017.
- [98] Y. Landau and N. Kiryati. Dataset growth in medical image analysis research. *arXiv:1908.07765 [cs, eess]*, August 2019.
- [99] N. Heller, J. Rickman, C. Weight, and N. Papanikolopoulos. The role of publicly available data in MICCAI papers from 2014 to 2018. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention (LABELS)*, volume 11851 of *Lecture Notes in Computer Science*, pages 70–77, 2019.

- [100] M. H. Moghari, T. Geva, and A. J. Powell. Prospective heart tracking for whole-heart magnetic resonance angiography. *Magnetic Resonance in Medicine*, 77(2):759–765, 2017.
- [101] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.
- [102] A. B. Scanlan, A. V. Nguyen, A. Ilina, A. Lasso, L. Cripe, A. Jegatheeswaran, E. Silvestro, F. X. McGowan, C. E. Mascio, S. Fuller, T. L. Spray, M. S. Cohen, G. Fichtinger, and M. A. Jolley. Comparison of 3D echocardiogram-derived 3D printed valve models to molded models for simulated repair of pediatric atrioventricular valves. *Pediatric Cardiology*, 39(3):538–547, 2018.
- [103] A. V. Nguyen, A. Lasso, H. H. Nam, J. Faerber, A. H. Aly, A. M. Pouch, A. B. Scanlan, F. X. McGowan, L. Mercer-Rosa, M. S. Cohen, J. Simpson, G. Fichtinger, and M. A. Jolley. Dynamic three-dimensional geometry of the tricuspid valve annulus in hypoplastic left heart syndrome with a Fontan circulation. *Journal of the American Society of Echocardiography*, 32(5):655–666.e13, 2019.
- [104] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 327–340, 2001.
- [105] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, 28(3):24:1–24:11, 2009.
- [106] T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, and D. Rueckert. Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling. *NeuroImage*, 76:11–23, 2013.
- [107] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954, 2011.
- [108] F. Rousseau, P. Habas, and C. Studholme. A supervised patch-based approach for human brain labeling. *IEEE Transactions on Medical Imaging*, 30(10):1852–1862, 2011.
- [109] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solorzano. Combination strategies in multi-atlas image segmentation: application to brain MR data. *IEEE Transactions on Medical Imaging*, 28(8):1266–1277, 2009.
- [110] W. Shi, H. Lombaert, W. Bai, C. Ledig, X. Zhuang, A. Marvao, T. Dawes, D. O’Regan, and D. O’Regan. Multi-atlas spectral PatchMatch: Application

- to cardiac image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 8673 of *Lecture Notes in Computer Science*, pages 348–355. 2014.
- [111] W. Bai, W. Shi, C. Ledig, and D. Rueckert. Multi-atlas segmentation with augmented features for cardiac MR images. *Medical Image Analysis*, 19(1):98–109, 2015.
- [112] Z. Wang, K. K. Bhatia, B. Glocker, A. Marvao, T. Dawes, K. Misawa, K. Mori, and D. Rueckert. Geodesic patch-based segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 8673 of *Lecture Notes in Computer Science*, pages 666–673. 2014.
- [113] W. Bai, W. Shi, D. O’Regan, T. Tong, H. Wang, S. Jamil-Copley, N. Peters, and D. Rueckert. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images. *IEEE Transactions on Medical Imaging*, 32(7):1302–1315, 2013.
- [114] R. Giraud, V.-T. Ta, N. Papadakis, J. V. Manjón, D. L. Collins, and P. Coupé. An optimized PatchMatch for multi-scale and multi-feature label fusion. *NeuroImage*, 124, Part A:770–782, 2016.
- [115] E. Alcain, A. Torrado-Carvajal, A. S. Montemayor, and N. Malpica. Real-time patch-based medical image gation by GPU computing. *Journal of Real-Time Image Processing*, 13(1):193–204, 2017.
- [116] P. Beerbaum, P. Barth, S. Kropf, S. Sarikouch, A. Kelter-Kloepping, D. Franke, M. Gutberlet, and T. Kuehne. Cardiac function by MRI in congenital heart disease: Impact of consensus training on interinstitutional variance. *Journal of Magnetic Resonance Imaging*, 30(5):956–966, 2009.
- [117] B. Settles. *Active Learning*. Morgan & Claypool Publishers, San Rafael, CA, USA, 2012.
- [118] D. Chyzhyk, R. Dacosta-Aguayo, M. Mataró, and M. Graña. An active learning approach for stroke lesion segmentation on multimodal MRI data. *Neurocomputing*, 150, Part A:26–36, 2015.
- [119] D. Mahapatra, P. J. Schüffler, J. A. W. Tielbeek, F. M. Vos, and J. M. Buhmann. Semi-supervised and active learning for automatic segmentation of Crohn’s disease. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 8150 of *Lecture Notes in Computer Science*, pages 214–221. 2013.
- [120] A. Top, G. Hamarneh, and R. Abugharbieh. Active learning for interactive 3D image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 6893 of *Lecture Notes in Computer Science*, pages 603–610. 2011.



- [121] H. Veeraraghavan and J. Miller. Active learning guided interactions for consistent image segmentation with reduced user interactions. In *IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro*, pages 1645–1648, 2011.
- [122] B. Wang, K. W. Liu, K. M. Prastawa, A. Irima, P. M. Vespa, J. D. van Horn, P. T. Fletcher, and G. Gerig. 4D active cut: An interactive tool for pathological anatomy modeling. In *IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro*, pages 529–532, 2014.
- [123] S. Yifrah, E. Zadicario, T. Ju, and D. Cohen-Or. An algorithm for suggesting delineation planes for interactive segmentation. In *IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro*, pages 361–364, 2014.
- [124] A. V. Dalca, A. Bobu, N. S. Rost, and P. Golland. Patch-based discrete registration of clinical brain images. In *Patch-Based Techniques in Medical Imaging (Patch-MI)*, volume 9993 of *Lecture Notes in Computer Science*, pages 60–67, 2016.
- [125] H. Zhu, H. Cheng, X. Yang, and Y. Fan. Metric learning for multi-atlas based segmentation of hippocampus. *Neuroinformatics*, 15(1):41–50, 2017.
- [126] H. Yang, J. Sun, H. Li, L. Wang, and Z. Xu. Deep fusion net for multi-atlas segmentation: Application to cardiac MR images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9901 of *Lecture Notes in Computer Science*, pages 521–528, 2016.
- [127] L. Yu, X. Yang, J. Qin, and P.-A. Heng. 3D FractalNet: Dense volumetric segmentation for cardiovascular MRI volumes. In *Reconstruction, Segmentation, and Analysis of Medical Images (HVS MR)*, volume 10129 of *Lecture Notes in Computer Science*, pages 103–110, 2017.
- [128] J. Wolterink, T. Leiner, M. Viergever, and I. Išgum. Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease. In *Reconstruction, Segmentation, and Analysis of Medical Images (HVS MR)*, volume 10129 of *Lecture Notes in Computer Science*, pages 95–102, 2017.
- [129] A. V. Dalca, J. Guttag, and M. R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9290–9299, 2018.
- [130] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.
- [131] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver, and X. Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

- [132] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.
- [133] C. Wachinger, M. Reuter, and T. Klein. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170:434–445, 2018.
- [134] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble.  $\Omega$ -Net (Omega-Net): Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks. *Medical Image Analysis*, 48:95–106, 2018.
- [135] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Medical Image Analysis*, 45:94–107, 2018.
- [136] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille. A fixed-point model for pancreas segmentation in abdominal CT scans. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 10433 of *Lecture Notes in Computer Science*, pages 693–701, 2017.
- [137] H. Chen, Y. Zheng, J.-H. Park, P.-A. Heng, and S. K. Zhou. Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9901 of *Lecture Notes in Computer Science*, pages 487–495, 2016.
- [138] M. Ren and R. Zemel. End-to-end instance segmentation with recurrent attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6656–6664, 2017.
- [139] B. Romera-Paredes and P. H. S. Torr. Recurrent instance segmentation. In *European Conference on Computer Vision (ECCV)*, volume 9910 of *Lecture Notes in Computer Science*, pages 312–329, 2016.
- [140] N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Medical Image Analysis*, 53:142–155, 2019.
- [141] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI Conference on Artificial Intelligence*, pages 4225–4232, 2017.
- [142] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with LSTM recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3547–3555, 2015.

- [143] F. Visin, A. Romero, K. Cho, M. Matteucci, M. Ciccone, K. Kastner, Y. Bengio, and A. Courville. ReSeg: A recurrent neural network-based model for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) DeepVision: Deep Learning in Computer Vision Workshop*, pages 426–433, 2016.
- [144] W. Xue, I. B. Nachum, S. Pandey, J. Warrington, S. Leung, and S. Li. Direct estimation of regional wall thicknesses via residual recurrent neural network. In *Information Processing in Medical Imaging (IPMI)*, volume 10265 of *Lecture Notes in Computer Science*, pages 505–516, 2017.
- [145] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1747–1756, 2016.
- [146] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1462–1471, 2015.
- [147] M. Liang and X. Hu. Recurrent convolutional neural network for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3367–3375, 2015.
- [148] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4742, 2016.
- [149] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, volume 9908 of *Lecture Notes in Computer Science*, pages 728–743, 2016.
- [150] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, 2015.
- [151] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision*. Cengage Learning, Stamford, CT, USA, 2008.
- [152] A. Dalca, G. Danagoulian, R. Kikinis, E. Schmidt, and P. Golland. Segmentation of nerve bundles and ganglia in spine MRI using particle filters. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 6893 of *Lecture Notes in Computer Science*, pages 537–545, 2011.
- [153] L. McIntosh, N. Maheswaranathan, D. Sussillo, and J. Shlens. Recurrent segmentation for variable computational budgets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1729–172909, 2018.

- [154] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 82–90. 2014.
- [155] T. H. N. Le, K. G. Quach, K. Luu, C. N. Duong, and M. Savvides. Reformulating level sets as deep recurrent neural network approach to semantic segmentation. *IEEE Transactions on Image Processing*, 27(5):2393–2407, 2018.
- [156] T. H. N. Le, R. Gummadi, and M. Savvides. Deep recurrent level set for segmenting brain tumors. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11072 of *Lecture Notes in Computer Science*, pages 646–653, 2018.
- [157] A. Chakravarty and J. Sivaswamy. RACE-net: A recurrent neural network for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 23(3):1151–1162, 2019.
- [158] M. Januszewski, J. Kornfeld, P. Li, A. Pope, T. Blakely, L. Lindsey, J. Maitin-Shepard, M. Tyka, W. Denk, and V. Jain. High-precision automated reconstruction of neurons with flood-filling networks. *Nature Methods*, 15(8):605–610, 2018.
- [159] R. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989.
- [160] Y. Mo, F. Liu, D. McIlwraith, G. Yang, J. Zhang, T. He, and Y. Guo. The deep Poincaré map: A novel approach for left ventricle segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11073 of *Lecture Notes in Computer Science*, pages 561–568. 2018.
- [161] P. Zhang, F. Wang, and Y. Zheng. Deep reinforcement learning for vessel centerline tracing in multi-modality 3D volumes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11073 of *Lecture Notes in Computer Science*, pages 755–763. 2018.
- [162] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache. 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE Transactions on Medical Imaging*, 37(9):2137–2148, 2018.
- [163] R. P. K. Poudel, P. Lamata, and G. Montana. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In *Reconstruction, Segmentation, and Analysis of Medical Images (RAMBO)*, volume 10129 of *Lecture Notes in Computer Science*, pages 83–94, 2017.
- [164] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.
- [165] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, and M. Xu. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer’s disease. *NeuroImage*, 208:116459, 2020.

- [166] A. G. Roy, S. Conjeti, D. Sheet, A. Katouzian, N. Navab, and C. Wachinger. Error corrective boosting for learning fully convolutional networks with limited data. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 10435 of *Lecture Notes in Computer Science*, pages 231–239. 2017.
- [167] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.
- [168] F. Chollet et al. Keras. <https://keras.io>, 2015.
- [169] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [170] M. D. Zeiler. ADADELTA: An adaptive learning rate method. *arXiv:1212.5701 [cs]*, December 2012.
- [171] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on Image Processing.*, 10(2):266–277, 2001.
- [172] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.
- [173] V. Vezhnevets and V. Konouchine. "GrowCut" - interactive multi-label N-D image segmentation by cellular automata. In *International Conference on Computer Graphics and Vision (GraphiCon)*, pages 150–156. 2005.
- [174] M. H. Moghari, A. Barthur, M. E. Amaral, T. Geva, and A. J. Powell. Free-breathing whole-heart 3D cine magnetic resonance imaging with prospective respiratory motion compensation. *Magnetic Resonance in Medicine*, 80(1):181–189, 2018.
- [175] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. d. Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker, and D. Rueckert. Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation. *IEEE Transactions on Medical Imaging*, 37(2):384–395, 2018.
- [176] H. Ravishankar, R. Venkataramani, S. Thiruvankadam, P. Sudhakar, and V. Vaidya. Learning and incorporating shape models for semantic segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 10433 of *Lecture Notes in Computer Science*, pages 203–211, 2017.

- [177] T. Brosch, J. Peters, A. Groth, T. Stehle, and J. Weese. Deep learning-based boundary detection for model-based segmentation with application to MR prostate segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11073 of *Lecture Notes in Computer Science*, pages 515–522, 2018.
- [178] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019.
- [179] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.
- [180] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro. Y-Net: Joint segmentation and classification for diagnosis of breast biopsy images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11071 of *Lecture Notes in Computer Science*, pages 893–901, 2018.
- [181] M. Li, W. Zhang, G. Yang, C. Wang, H. Zhang, H. Liu, W. Zheng, and S. Li. Recurrent aggregation learning for multi-view echocardiographic sequences segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11765 of *Lecture Notes in Computer Science*, pages 678–686, 2019.
- [182] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1796–1804, 2015.
- [183] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. Ben Ayed. Constrained-CNN losses for weakly supervised segmentation. *Medical Image Analysis*, 54:88–99, 2019.
- [184] Y. Zhou, Z. Li, S. Bai, C. Wang, X. Chen, M. Han, E. Fishman, and A. L. Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10671–10680, 2019.
- [185] P.-A. Ganaye, M. Sdika, and H. Benoit-Cattin. Semi-supervised learning for segmentation under semantic constraint. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11072 of *Lecture Notes in Computer Science*, pages 595–602, 2018.
- [186] J. R. Clough, I. Oksuz, N. Byrne, J. A. Schnabel, and A. P. King. Explicit topological priors for deep-learning based image segmentation using persistent homology. In *Information Processing in Medical Imaging (IPMI)*, volume 11492 of *Lecture Notes in Computer Science*, pages 16–28, 2019.

- [187] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.
- [188] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS) 27*, pages 2672–2680, 2014.
- [189] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.
- [190] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. In *Advances in Neural Information Processing Systems (NIPS) Workshop on Adversarial Training*, 2016.
- [191] E. Konukoglu, B. Glocker, D. H. Ye, A. Criminisi, and K. M. Pohl. Discriminative segmentation-based evaluation through shape dissimilarity. *IEEE Transactions on Medical Imaging*, 31(12):2278–2289, 2012.
- [192] E. Konukoglu, B. Glocker, A. Criminisi, and K. M. Pohl. WESD-Weighted Spectral Distance for measuring shape dissimilarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2284–2297, 2013.
- [193] Holmvang Godtfred, Palacios Igor F., Vlahakes Gus J., Dinsmore Robert E., Miller Stephen W., Liberthson Richard R., Block Peter C., Ballen Barbara, Brady Thomas J., and Kantor Howard L. Imaging and sizing of atrial septal defects by magnetic resonance. *Circulation*, 92(12):3473–3480, 1995.
- [194] J. Weese, A. Groth, H. Nickisch, H. Barschdorf, F. M. Weber, J. Velut, M. Castro, C. Toumoulin, J. L. Coatrieux, M. D. Craene, G. Piella, C. Tobón-Gomez, A. F. Frangi, D. C. Barber, I. Valverde, Y. Shi, C. Staicu, A. Brown, P. Beerbaum, and D. R. Hose. Generating anatomical models of the heart and the aorta from medical images for personalized physiological simulations. *Medical & Biological Engineering & Computing*, 51(11):1209–1219, 2013.
- [195] A. Suinesiaputra, A. D. McCulloch, M. P. Nash, B. Pontre, and A. A. Young. Cardiac image modelling: Breadth and depth in heart disease. *Medical Image Analysis*, 33:38–43, 2016.
- [196] C. M. Lawley, K. M. Broadhouse, F. M. Callaghan, D. S. Winlaw, G. A. Figtree, and S. M. Grieve. 4D flow magnetic resonance imaging: Role in pediatric congenital heart disease. *Asian Cardiovascular & Thoracic Annals*, 26(1):28–37, 2018.
- [197] T. S. Sørensen, J. Mosegaard, S. Kislinskiy, and G. F. Greil. Virtual surgery in congenital heart disease. In F. Saremi, editor, *Cardiac CT and MR for Adult Congenital Heart Disease*, chapter 23, pages 515–523. Springer, New York Heidelberg Dordrecht London, 2014.

- [198] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, 2012.
- [199] G. Li, L. Wang, P.-T. Yap, F. Wang, Z. Wu, Y. Meng, P. Dong, J. Kim, F. Shi, I. Rekik, W. Lin, and D. Shen. Computational neuroanatomy of baby brains: A review. *NeuroImage*, 185:906–925, 2019.