Medical Image Analysis 80 (2022) 102469

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Learned iterative segmentation of highly variable anatomy from limited data: Applications to whole heart segmentation for congenital heart disease

Danielle F. Pace[a,b,*], Adrian V. Dalca[a,b], Tom Brosch[c], Tal Geva[d,e], Andrew J. Powell[d,e], Jürgen Weese[c], Mehdi H. Moghari[d,e], Polina Golland[a]

[a] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
[b] A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[c] Philips Research Laboratories, Hamburg, Germany
[d] Department of Cardiology, Boston Children's Hospital, Boston, MA, USA
[e] Department of Pediatrics, Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Article history:
Received 9 July 2021
Revised 26 April 2022
Accepted 29 April 2022
Available online 13 May 2022

Keywords:
Recurrent Neural Network
Whole heart segmentation
Congenital heart disease
Interactive segmentation

ABSTRACT

Training deep learning models that segment an image in one step typically requires a large collection of manually annotated images that captures the anatomical variability in a cohort. This poses challenges when anatomical variability is extreme but training data is limited, as when segmenting cardiac structures in patients with congenital heart disease (CHD). In this paper, we propose an iterative segmentation model and show that it can be accurately learned from a small dataset. Implemented as a recurrent neural network, the model evolves a segmentation over multiple steps, from a single user click until reaching an automatically determined stopping point. We develop a novel loss function that evaluates the entire sequence of output segmentations, and use it to learn model parameters. Segmentations evolve predictably according to growth dynamics encapsulated by training data, which consists of images, partially completed segmentations, and the recommended next step. The user can easily refine the final segmentation by examining those that are earlier or later in the output sequence. Using a dataset of 3D cardiac MR scans from patients with a wide range of CHD types, we show that our iterative model offers better generalization to patients with the most severe heart malformations.

1. Introduction

Congenital heart disease (CHD) includes all heart defects existing at birth, encompassing a wide array of potential cardiac malformations and topological changes (Frescura et al., 2010). The heart of each CHD patient is unique, with different combinations of original heart defects, new atypical connections and implants from prior surgeries, and shape changes from long-term cardiac remodeling (Pandya et al., 2016). Fig. 1 illustrates the wide variability of heart anatomy in CHD. Strong anatomical priors are hard to enforce, and relating information across subjects with dramatically different heart configurations or simulating realistic images is difficult.

Treating severe CHD requires multiple surgeries throughout infancy, childhood and adult life. For surgical planning, clinicians must understand each patient's unique heart anatomy, evaluating the size and location of defects and determining their relationships with other cardiac structures. MR is an attractive preoperative modality (Ntsinjana et al., 2011; Arafati et al., 2019) as it produces high quality images and, unlike CT imaging, does not require ionizing radiation, which is particularly important for children.

However, cardiac MR suffers from (1) low signal-to-noise ratio and spatial resolution (Zhuang et al., 2019), (2) no contrast at many of the valves, thin walls and "holes in the heart" that separate neighboring cardiac chambers and great vessels, and (3) artifacts, especially those surrounding implanted stents. Fig. 2 illustrates these challenges.

There is great interest in patient-specific 3D heart surface models for surgical planning for CHD, whether rendered on a screen or 3D-printed. Using 3D heart surface models promises to yield a greater appreciation of the true locations and sizes of intracardiac structures, aid decision making and consensus, and even lead doctors to alter original surgical plans made based on images

[*] Corresponding author.
E-mail address: dfpace@mgh.harvard.edu (D.F. Pace).

https://doi.org/10.1016/j.media.2022.102469
1361-8415/© 2022 Elsevier B.V. All rights reserved.

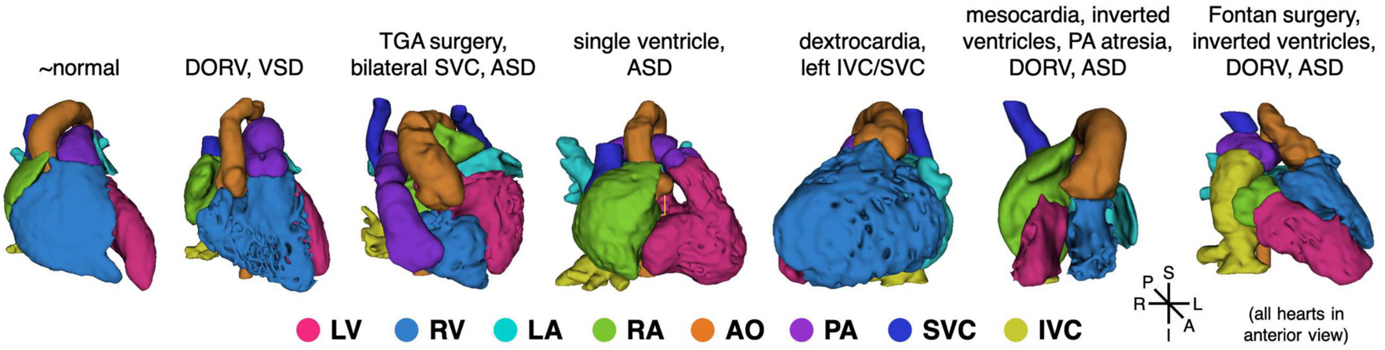
~normal
DORV, VSD
TGA surgery, bilateral SVC, ASD
single ventricle, ASD
dextrocardia, left IVC/SVC
mesocardia, inverted ventricles, PA atresia, DORV, ASD
Fontan surgery, inverted ventricles, DORV, ASD
● LV ● RV ● LA ● RA ● AO ● PA ● SVC ● IVC
S P R L I A
(all hearts in anterior view)

Fig. 1. Example 3D heart surface models of congenital heart disease, which manifests as *size and shape changes* in the chambers or vessels; *abnormal connections*, e.g., DORV (double outlet right ventricle), TGA (transposition of the great arteries), VSD (ventricular septal defect), ASD (atrial septal defect) and Fontan surgery; *duplicated structures*, e.g., bilateral SVC; *missing structures*, e.g., single ventricle; and/or *abnormal structure locations*, e.g., dextrocardia, mesocardia, inverted ventricles and left IVC/SVC.

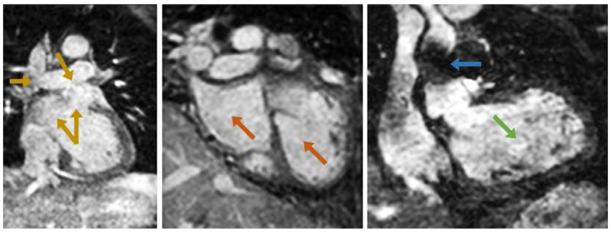

Fig. 2. Example scans illustrating challenges of cardiac MR segmentation for CHD that go beyond anatomical variability. These include noise (green arrow), no contrast at the boundaries of adjacent structures (gold arrows), different structures that locally appear very similar (red arrows), and dark inhomogeneity artifacts from previously implanted stents (blue arrow).

(Lau and Sun, 2018; Valverde et al., 2017a; Bhatla et al., 2017; Garekar et al., 2016; Riesenkampff et al., 2009). However, manual segmentation requires many hours per image. The lack of accurate whole heart segmentation methods for CHD patients currently precludes widespread adoption of 3D heart surface models for surgical planning (Lau and Sun, 2018; Byrne et al., 2016).

To support surgical planning in CHD patients, we focus on whole heart segmentation for CHD patients, which requires outlining the left ventricle (LV), right ventricle (RV), left atrium (LA, including the pulmonary veins), right atrium (RA), aorta (AO), pulmonary artery (PA), superior vena cava (SVC) and inferior vena cava (IVC) (Zhuang, 2013; Zhuang et al., 2019; Peng et al., 2016). We do not focus on segmenting the myocardium. The input to our methods is a 3D cardiac MR scan that captures the heart at a single point in the cardiac cycle (and not a time-series of images), plus a single user click per structure. We propose to support image segmentation via a deep learning model that progressively evolves the segmentation of each structure. Our approach also opens up user opportunities to interact and adjust.

1.1. Prior work

Most whole heart segmentation methods have been developed for patients with relatively normal anatomy, including deformable models (Ecabert et al., 2008; 2011; Peters et al., 2010; Zheng et al., 2008) and atlas-based segmentation (Zhuang et al., 2010; Zhuang and Shen, 2016). More recent convolutional neural network (CNN) investigations include two-step network cascades, multi-planar CNNs, deep supervision, and/or integration of statistical shape priors (Payer et al., 2017; Yang et al., 2018; Wang and Smedby, 2017). The public Multi-Modality Whole Heart Segmentation (MM-WHS) dataset does include CHD images, but these are relatively few (16/120 images), cover a limited number of CHD subtypes, and remain recognized as very difficult to segment (Zhuang et al., 2019). For CHD patients, segmentation of MR images has largely been limited to labeling the blood pool and myocardium (Wolterink et al., 2017; Yu et al., 2017; Dou et al., 2017; Pace et al., 2015). Segmenting each chamber and great vessel as separate labels offers several advantages: it facilitates automatic computation of quantitative metrics of cardiac function that for CHD patients are typically based on manual annotations, such as chamber volumes, ejection fraction and aortic dimensions (Seraphim et al., 2020; Petersen et al., 2019), it yields a more visually intuitive surface model, and the shape variability of each anatomical structure is reduced compared to that of the entire cardiac blood pool. Previously demonstrated methods to segment individual heart structures for CHD either operated on CT images (Xu et al., 2019; Liu et al., 2020b), or only segmented the ventricles in specific CHD subtypes (Zhang et al., 2010; Mansi et al., 2011).

Example interactive segmentation methods include intelligent scissors (Mortensen and Barrett, 1998), graph cuts (Boykov and Jolly, 2001), random walks (Grady, 2006), GrowCut (Vezhnevets and Konouchine, 2005), GraphCut (Rother et al., 2004) and GeoS (Criminisi et al., 2008). These methods are not ideal for whole heart segmentation because the homogeneous blood pool must be partitioned into its component cardiac chambers and great vessels, which are not separated by strong edges or have distinctive intensity distributions. However, interactive segmentation methods based on deep learning can learn more complex features and reduce user interaction. For example, user clicks or scribbles can be transformed into binary, Euclidean distance or geodesic distance maps, and concatenated as additional input channels in a segmentation network (Xu et al., 2016; Sakinis et al., 2019; Wang et al., 2019; Amrehn et al., 2017), or used to update a network's weights to better segment a test image (Wang et al., 2018). Many of these methods focus on back-and-forth interaction with a user to iteratively refine a segmentation. In contrast, we aim to estimate a high-quality segmentation from more limited user interaction, namely one click per structure plus an optional step of choosing amongst a sequence of candidate segmentations.

State-of-the-art segmentation methods train a feedforward CNN to segment an image in one step (Long et al., 2015; Ronneberger et al., 2015; Dalca et al., 2018). An alternative is to iteratively segment an image over multiple steps, at each step conditioning on a previous partial solution to make progress towards the final answer. This is reminiscent of traditional active contours, level sets, hidden Markov models and particle filters (Sonka et al., 2008; Dalca et al., 2011). More recently, deep learning approaches that use iterative segmentation include network cascades (Wachinger et al., 2018; Payer et al., 2017; Valverde et al., 2017b; Havaei et al., 2017), instance segmentation using an internal memory or atten-

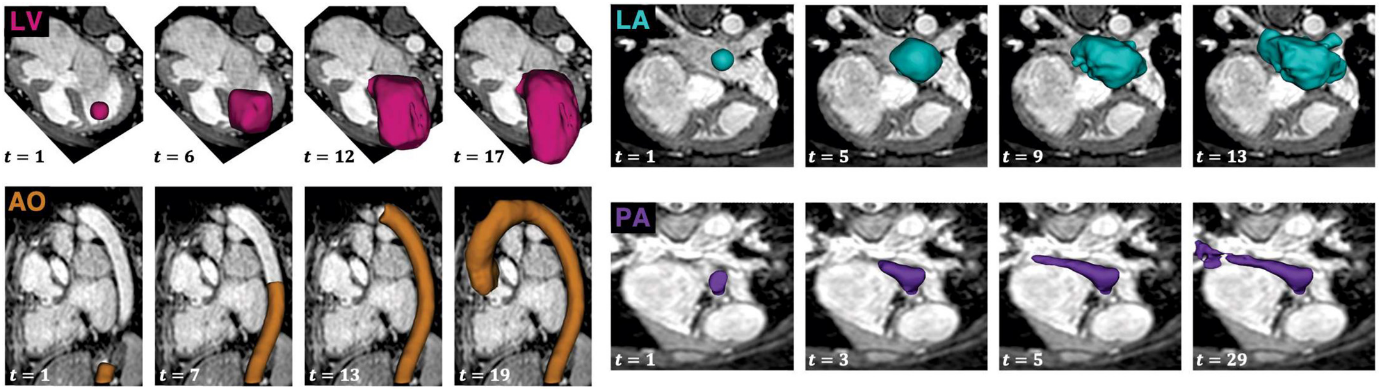

Fig. 3. Results of our iterative segmentation model that evolves segmentations in a predictable way that is defined via training data. Visualizations show 3D surface models in the context of a representative image slice. The variable t refers to the time step of the iterative segmentation model's output segmentations.

tion mechanism (Ren and Zemel, 2017; Romera-Paredes and Torr, 2016; Lessmann et al., 2019), and, most relevant to this paper, recurrent neural networks (RNNs).

RNNs are popular for modeling sequential data. They implement the repeated application of a recursive function, using the same learned parameters at each iteration. At each step, the network inputs include information from the previous step via recurrent connections, which can link analogous hidden layers of consecutive iterations or directly connect outputs to hidden units (Goodfellow et al., 2016). Notably, previously proposed RNNs for image segmentation produce unpredictable growth patterns, whether they progressively refine an initially coarse segmentation of the entire object (Pinheiro and Collobert, 2014; McIntosh et al., 2018), model level sets (Le et al., 2018a; 2018b; Chakravarty and Sivaswamy, 2019) or sequentially segment small areas pulled from an internal list of potential regions of interest (Januszewski et al., 2018).

1.2. Approach and contributions

In this paper, we demonstrate a novel segmentation strategy, initialized by a single click per structure, which also enables additional intuitive user interaction that is valuable in our challenging application. Our major contributions are enumerated within the text below.

1. To the best of our knowledge, our work provides the first whole heart segmentation to individually label each cardiac chamber and great vessel in cardiac MR for patients with congenital heart disease.

2. We develop an iterative segmentation model (and RNN implementation) that is trained to evolve a segmentation over multiple steps, until reaching a stopping point that can be automatically determined or defined by a user. See Fig. 3. Vessel segmentations are trained to grow along centerlines and chamber segmentations grow outwards towards the boundary. The model can be trained to follow any desired evolution pattern that is implicitly represented by training data. The algorithm operates directly on the 3D image grid, unlike approaches that learn to progressively trace a contour (Mo et al., 2018; Zhang et al., 2018), perform slice-by-slice analysis (Zheng et al., 2018; Poudel et al., 2017) or propagate information from 2D image patches (Pace et al., 2015).

3. To encourage the output segmentations to grow in a predicable way, we develop a novel loss function that evaluates the entire sequence of output segmentations, adopting a learning framework known as teacher forcing (Williams and Zipser, 1989; Goodfellow et al., 2016). This approach differs from evaluating the final segmentation alone or encouraging every segmentation in the sequence to match the complete ground truth segmentation (Pinheiro and Collobert, 2014; McIntosh et al., 2018). The maximum likelihood loss function factors into a sum over time steps, eliminating the need to back-propagate through time and making training easier than e.g., LSTMs (Shi et al., 2015). We show that this proposed loss can be optimized using a dataset of images alongside input-output pairs of partially completed segmentations. We construct these pairs on-the-fly during training from complete ground truth segmentations.

Due to the challenges of our task, we optionally enable users to interact with the system to intuitively fine-tune the segmentation. After the user clicks once to place a seed, the result from our segmentation with automatic stopping will be shown. The user will have the option to either accept the segmentation, or look for a better segmentation result by looking backwards in the output sequence or asking for more iterations. Our experiments indicate that this minimal amount of additional user interaction may yield large performance improvements. Compared to previous RNN segmentation methods, we anticipate that a user using our model can more easily interact with and edit the segmentation. A user can more easily find a high quality result because the output sequence contains diverse yet automatically-sorted solutions. It is also more efficient to monitor progress, because the region in which growth is expected is spatially limited, which is especially important for 3D images. Finally, if interactive editing of intermediate segmentations via foreground or background clicks or scribbles were to be incorporated in future (Wang et al., 2019; 2018; Xu et al., 2016; Sakinis et al., 2019), our RNN can be restarted at any point (since it does not rely on a memory) and the user can anticipate which areas require input and which would be corrected in subsequent time steps (since it grows segmentations with a predictable pattern).

4. Using a dataset of 3D cardiac MR scans from patients with a wide range of CHD types, we show that existing state-of-the-art segmentation methods fail to handle the extraordinary anatomical variability of CHD, and demonstrate that our iterative model offers better generalization to patients with the most severe heart malformations. In particular, we show advantages when learning from small datasets, in which specific anatomical configurations are often represented by a single sample and hence can be present during inference but not during training. These advantages may go beyond our particular clinical application, as access to limited training data is a very common setting in practice, and efficient interactive segmentation is valuable for direct use or in the process of generating larger training datasets.

This paper expands our preliminary work on segmenting the AO and LV (Pace et al., 2018), arguably the easiest cardiac structures to segment. Here, we extend the method to whole heart segmentation, improve the data augmentation, training and inference strategies, and provide detailed derivations. We present extensive evaluation of the binary segmentation model and demonstrate its

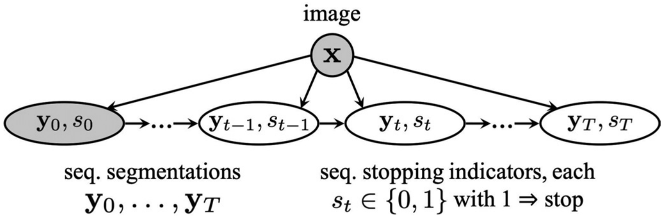
image
$\mathbf{x}$
$\mathbf{y}_0, s_0$
$\mathbf{y}_{t-1}, s_{t-1}$
$\mathbf{y}_t, s_t$
$\mathbf{y}_T, s_T$
seq. segmentations
$\mathbf{y}_0, \dots, \mathbf{y}_T$
seq. stopping indicators, each
$s_t \in \{0,1\}$ with 1 ⇒ stop

Fig. 4. Probabilistic model: given 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.

extension to multiclass iterative segmentation. Finally, we perform a more comprehensive validation in a significantly expanded patient cohort.

2. Iterative segmentation model

Given image $\mathbf{x} : \Omega \rightarrow \mathbb{R}$ and initial segmentation seed $\mathbf{y}_0 : \Omega \rightarrow \{0, \dots, L-1\}$, we seek a segmentation label map $\mathbf{y} : \Omega \rightarrow \{0, \dots, L-1\}$ that parcellates the image into L label maps. In practice, the initial segmentation $\mathbf{y}_0$ is created by centering a small sphere around a seed point placed by the user for each anatomical structure.

2.1. Probabilistic model

We model the segmentation label map $\mathbf{y}$ as the final element in a sequence of segmentations $\mathbf{y}_0, \dots, \mathbf{y}_T$ that captures a growing and evolving portion of the anatomy of interest, where $\mathbf{y}_t : \Omega \rightarrow \{0, \dots, L-1\}$ for time steps $t = 0, \dots, T$. To capture the variable length of the segmentation sequence, 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 $\mathbf{y}_t$. Hence, in practice $s_0 = 0$.

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, as shown in Fig. 4:

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

for $t = 1, \dots, T$, leading to the recursion

$$p(\mathbf{y}_t, s_t|\mathbf{x}, \mathbf{y}_0, s_0) = \sum_{\mathbf{y}_{t-1}} \sum_{s_{t-1}} \underbrace{p(\mathbf{y}_t, s_t|\mathbf{x}, \mathbf{y}_{t-1}, s_{t-1})}_{\text{transition probability}} \cdot \underbrace{p(\mathbf{y}_{t-1}, s_{t-1}|\mathbf{x}, \mathbf{y}_0, s_0)}_{\text{recursive definition}}, \tag{2}$$

for $t = 1, \dots, T$, where $p(\mathbf{y}_0, s_0|\mathbf{x}, \mathbf{y}_0, s_0) = 1$.

2.2. Transition probability model

To complete the recursion in Eq. (2), the transition probability $p(\mathbf{y}_t, s_t|\mathbf{x}, \mathbf{y}_{t-1}, s_{t-1})$ must be defined. We consider $s_{t-1} = 1$ and $s_{t-1} = 0$ separately.

When $s_{t-1} = 1$, the segmentation $\mathbf{y}_{t-1}$ is the final segmentation. The transition model ensures the segmentation remains unchanged:

$$p(\mathbf{y}_t, s_t|\mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 1) = \mathbb{1}[\mathbf{y}_t = \mathbf{y}_{t-1}] \cdot \mathbb{1}[s_t = 1], \tag{3}$$

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

$$\mathbf{h}_t = h(\mathbf{x}, \mathbf{y}_{t-1}) \tag{4}$$

that captures all necessary information from the given image $\mathbf{x}$ and previous segmentation $\mathbf{y}_{t-1}$ to make inferences about $\mathbf{y}_t$ and s_t:

$$p(\mathbf{y}_t, s_t|\mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 0) = p(\mathbf{y}_t, s_t|\mathbf{h}_t). \tag{5}$$

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

$$p(\mathbf{y}_t, s_t|\mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 0) = p(\mathbf{y}_t|\mathbf{h}_t) \cdot p(s_t|\mathbf{h}_t). \tag{6}$$

This conditional independence assumption is justified because 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. Hence, $\mathbf{y}_t$ is not informative for inferences about the stopping indicator s_t given $\mathbf{h}_t$ that captures all necessary 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.

2.3. Learning

We use manually segmented images to learn the parameters $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ of a model for the transition probability $p(\mathbf{y}_t, s_t|\mathbf{x}, \mathbf{y}_{t-1}, s_{t-1} = 0; \boldsymbol{\theta})$. We use $\boldsymbol{\theta} = \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s\}$ to denote the parameters of the learned functions $h(\mathbf{x}, \mathbf{y}_{t-1}; \boldsymbol{\theta}_h)$, $p(\mathbf{y}_t|\mathbf{h}_t; \boldsymbol{\theta}_y)$ and $p(s_t|\mathbf{h}_t; \boldsymbol{\theta}_s)$, respectively.

First we consider 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. The segmentation sequences capture the desired segmentation evolution dynamics.

Adopting the teacher forcing approach (Williams and Zipser, 1989; Goodfellow et al., 2016), we develop a novel loss function that seeks the parameter values which minimize the expected negative log-likelihood over the sequences of segmentations and stopping indicators, conditioned on the image and the initial conditions:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathcal{D}}\Big[-\log p(\mathbf{y}_1, \dots, \mathbf{y}_{T(\mathbf{x})}, s_1, \dots, s_{T(\mathbf{x})}|\mathbf{x}, \mathbf{y}_0, s_0; \boldsymbol{\theta})\Big], \\
&= \mathbb{E}_{\mathcal{D}}\Bigg[\sum_{t=1}^{T(\mathbf{x})} -\log p(\mathbf{y}_t, s_t|\mathbf{x}, \mathbf{y}_{t-1}, s_{t-1}; \boldsymbol{\theta})\Bigg], \qquad (7) \\
&= \mathbb{E}_{\mathcal{D}}\Bigg[\sum_{t=1}^{T(\mathbf{x})} \underbrace{-\log p\big(\mathbf{y}_t|h(\mathbf{x}, \mathbf{y}_{t-1}; \boldsymbol{\theta}_h); \boldsymbol{\theta}_y\big)}_{\text{segmentation loss}} \\
&\qquad \underbrace{-\log p\big(s_t|h(\mathbf{x}, \mathbf{y}_{t-1}; \boldsymbol{\theta}_h); \boldsymbol{\theta}_s\big)}_{\text{stopping indicator loss}}\Bigg]. \qquad (8)
\end{aligned}$$

In Eq. (7), teacher forcing leads to a sum over decoupled time steps, due to the Markov property in Eq. (1). This greatly simplifies training, by eliminating the need for backpropagation through time. Eq. (8) is an expectation over a segmentation loss and a stopping indicator loss. The segmentation $\mathbf{y}_t$ and stopping indicator s_t are predicted jointly, and both of their losses influence the parameters used to compute the latent representation $\mathbf{h}_t = h(\mathbf{x}, \mathbf{y}_{t-1}; \boldsymbol{\theta}_h)$. This multi-task approach often improves learning, and requires fewer parameters compared to training two separate networks (Caruana, 1997; Liu et al., 2020a).

Since the loss is a sum over decoupled time steps, training data of entire predefined output sequences is 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 segmentations $\mathbf{y}_{in}$ and $\mathbf{y}_{out}$ correspond to consecutive time steps and s denotes whether $\mathbf{y}_{out}$ is a complete segmentation:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}'}\left[\underbrace{-\log p\big(\mathbf{y}_{out}|h(\mathbf{x},\mathbf{y}_{in};\boldsymbol{\theta}_h);\boldsymbol{\theta}_y\big)}_{\text{segmentation loss}}\underbrace{-\log p\big(s|h(\mathbf{x},\mathbf{y}_{in};\boldsymbol{\theta}_h);\boldsymbol{\theta}_s\big)}_{\text{stopping indicator loss}}\right]. \tag{9}$$

These input-output pairs can be generated on-the-fly during training (more details are provided in Section 3.2).

In the rest of this section, we make typical modeling choices to define the segmentation and stopping indicator losses.

2.3.1. 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. Let $\mathbf{y}_{out}$ be a one-hot ground truth segmentation and $\hat{\mathbf{y}}_{out}$ be a predicted segmentation probability map, i.e., we use $\hat{\mathbf{y}}_{out}$ as a shorthand for $p(\mathbf{y}_{out}|h(\mathbf{x},\mathbf{y}_{in};\boldsymbol{\theta}_h);\boldsymbol{\theta}_y)$. The segmentation loss in Eq. (9) is a voxel-wise categorical cross-entropy loss with spatially varying weights $\omega_{\mathbf{y}_{out},l}(\mathbf{v})$:

$$\mathcal{L}_{seg}\big(\mathbf{y}_{out},\hat{\mathbf{y}}_{out}\big) = \sum_{\mathbf{v}\in\Omega}\sum_{l=0}^{L-1} -\omega_{\mathbf{y}_{out},l}(\mathbf{v})\cdot\mathbf{y}_{out,l}(\mathbf{v})\cdot\log\hat{\mathbf{y}}_{out,l}(\mathbf{v}),$$
$$\omega_{\mathbf{y}_{out},l}(\mathbf{v}) = \omega_l + \omega_{\mathbf{y}_{out}}(\mathbf{v}). \tag{10}$$

We use spatially varying weights with two goals in mind. The first term addresses class rebalancing, where each weight $\omega_l = (1/f_l)/\sum_{l'}(1/f_{l'})$ is a normalized inverse label frequency in the training data's target segmentations. The second goal is to encourage segmentations to "snap" to image boundaries, by more strongly penalizing errors near ground truth segmentation borders, hence the dependence on the ground truth segmentation $\mathbf{y}_{out}$ (Ronneberger et al., 2015; Roy et al., 2017). We introduce a weight map $\omega_{\mathbf{y}_{out}}:\Omega\to\{0,\omega_0\}$ that contains a constant boundary weight $\omega_0 > 0$ for voxels located within d_0 voxels of any boundary in the ground truth segmentation $\mathbf{y}_{out}$, and zero otherwise.

2.3.2. Stopping indicator loss

The distribution of the stopping indicator s is Bernoulli. The stopping indicator loss in Eq. (9) is a binary cross-entropy loss, which we again weight for class rebalancing. Let s be a ground truth binary stopping indicator and $\hat{s}$ be a predicted stopping probability, i.e., we use $\hat{s}$ as a shorthand for $p(s|h(\mathbf{x},\mathbf{y}_{in};\boldsymbol{\theta}_h);\boldsymbol{\theta}_s)$. We have

$$\mathcal{L}_{stop}(s,\hat{s}) = -(1-\omega_s)\cdot s\log\hat{s} - \omega_s\cdot(1-s)\log(1-\hat{s}), \tag{11}$$

where the class rebalancing weight ω_s is the estimated proportion of training instances in which the stopping indicator equals 1.

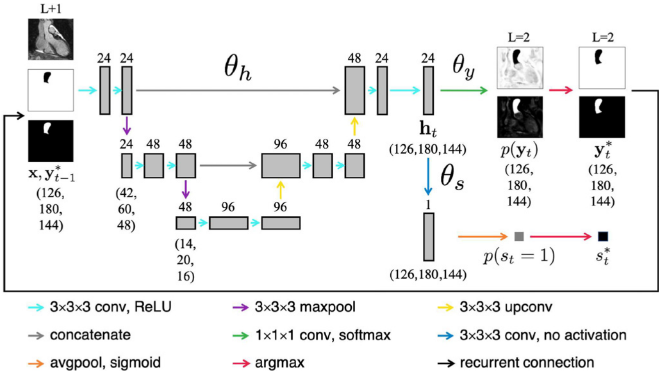

Fig. 5. Our RNN jointly evolves the segmentation and predicts the stopping indicator. The main block is a trained U-Net modified to model $p(\mathbf{y}_t,s_t|\mathbf{x},\mathbf{y}_{t-1},s_{t-1}=0;\boldsymbol{\theta})$, where the final bank of $C=24$ features forms the latent representation $\mathbf{h}_t = h(\mathbf{x},\mathbf{y}_{t-1};\boldsymbol{\theta}_h)$. This particular example visualizes the model for binary segmentation.

2.4. Inference

Since the recursion in Eq. (2) is computationally intractable due to the summation over all possible segmentations $\mathbf{y}_{t-1}$, we follow the widely accepted practice of using point estimates (Iglesias et al., 2013) to infer $\mathbf{y}_t$ and s_t directly from the most likely previous binary segmentation $\mathbf{y}^*_{t-1}$ and binary stopping indicator s^*_{t-1}:

$$\begin{aligned}
&p(\mathbf{y}_t,s_t|\mathbf{x},\mathbf{y}_0,s_0;\boldsymbol{\theta})\\
&=\sum_{\mathbf{y}_{t-1}}\sum_{s_{t-1}} p(\mathbf{y}_t,s_t|\mathbf{x},\mathbf{y}_{t-1},s_{t-1};\boldsymbol{\theta})\cdot p(\mathbf{y}_{t-1},s_{t-1}|\mathbf{x},\mathbf{y}_0,s_0;\boldsymbol{\theta})\\
&\approx p(\mathbf{y}_t,s_t|\mathbf{x},\mathbf{y}^*_{t-1},s^*_{t-1};\boldsymbol{\theta}),\\
&\text{where } \mathbf{y}^*_{t-1},s^*_{t-1} = \operatorname*{argmax}_{\mathbf{y}_{t-1},s_{t-1}} p(\mathbf{y}_{t-1},s_{t-1}|\mathbf{x},\mathbf{y}_0,s_0;\boldsymbol{\theta}).
\end{aligned} \tag{12}$$

Eq. (12) is a mode approximation of Eq. (2). It is accurate whenever $p(\mathbf{y}^*_{t-1},s^*_{t-1}|\mathbf{x},\mathbf{y}_0,s_0;\boldsymbol{\theta})\approx 1$, i.e., when the distribution $p(\mathbf{y}_{t-1},s_{t-1}|\mathbf{x},\mathbf{y}_0,s_0;\boldsymbol{\theta})\approx\delta(\mathbf{y}_{t-1}-\mathbf{y}^*_{t-1})\cdot\delta(s_{t-1}-s^*_{t-1})$, where δ is the Dirac delta function, as the non-maximal members of the sum in Eq. (2) are negligible compared with the maximal one.

When $s^*_{t-1}=1$, Eqs. (3) and (12) yield the approximation

$$p(\mathbf{y}_t,s_t|\mathbf{x},\mathbf{y}_0,s_0;\boldsymbol{\theta}) \approx \mathbb{1}\big[\mathbf{y}_t=\mathbf{y}^*_{t-1}\big]\cdot\mathbb{1}(s_t=1). \tag{13}$$

We adopt the maximum *a posteriori* (MAP) approach and continue the recursion 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. While here we use the MAP criterion for stopping, one could also choose a different threshold of the posterior probability of the stopping indicator based on domain knowledge or empirical results. A user can override this automatic stopping prediction by choosing an earlier segmentation or asking for more iterations.

3. Recurrent neural network

Our RNN implements the recursion

$$\begin{aligned}
\mathbf{h}_t &= h(\mathbf{x},\mathbf{y}^*_{t-1};\boldsymbol{\theta}_h),\\
\mathbf{y}^*_t &= \operatorname*{argmax}_{\mathbf{y}_t} p(\mathbf{y}_t|\mathbf{h}_t;\boldsymbol{\theta}_y),\\
s^*_t &= \operatorname*{argmax}_{s_t} p(s_t|\mathbf{h}_t;\boldsymbol{\theta}_s),
\end{aligned} \tag{14}$$

until $s^*_t=1$, at which point $\mathbf{y}^*_t$ is the final solution.

3.1. RNN architecture

Our RNN is depicted in Fig. 5. It is constructed by joining copies of a 3D U-Net architecture (Ronneberger et al., 2015) that we modify to model $p(\mathbf{y}_t,s_t|\mathbf{x},\mathbf{y}_{t-1},s_{t-1}=0;\boldsymbol{\theta})$. The U-Net has $L+1$ input channels for 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 via a recurrent connection, and the stopping indicator s^*_t.

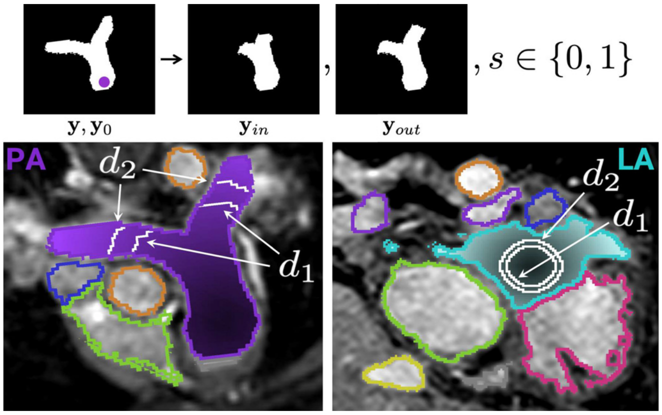

Fig. 6. Input and output partial segmentations and binary stopping indicators ($\mathbf{y}_{in}, \mathbf{y}_{out}, s$) are generated on-the-fly during training from ground truth complete segmentations $\mathbf{y}$ and seeds $\mathbf{y}_0$.

Recall that in the U-Net architecture, a final bank of learned feature maps is used to produce the output segmentation probability map. In our RNN, these learned feature maps form the latent representation $\mathbf{h}_t = h(\mathbf{x}, \mathbf{y}^*_{t-1}; \boldsymbol{\theta}_h)$. Note that the size of $\mathbf{h}_t$ equals the dimensions of image $\mathbf{x}$ multiplied by the number of channels C (i.e., this is not a bottleneck layer).

3.2. Training data generation

The training data $\mathcal{D}' = \{\mathbf{x}, \mathbf{y}_{in}, \mathbf{y}_{out}, s\}$ should capture the application-dependent segmentation evolution pattern that the RNN should learn to produce. There are many ways in which this training data can be generated. In our case, every training image has a ground truth complete segmentation $\mathbf{y}$ and an example seed $\mathbf{y}_0$ for each anatomical label. We consider binary segmentation in this section, and discuss extensions to multiclass segmentation in Section 3.4. During each epoch, we automatically generate one sample from $\mathcal{D}'$ for each training image (Fig. 6) using a different mechanism for great vessels and cardiac chambers.

We train models for great vessel segmentation that grow along their centerline at a constant rate. Before training, we precompute a distance map that can be randomly thresholded to sample the partial segmentations used for training. We use fast marching (Sethian, 1996) to create a geodesic distance map that is zero in the background (and hence embeds the ground truth complete segmentation), 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, we threshold at a distance d_1 chosen uniformly at random to form $\mathbf{y}_{in}$, and then at $d_2 = d_1 + d_s$ to form $\mathbf{y}_{out}$, where d_s is the desired step size. Note that this precomputed distance map relies on the ground truth segmentation and therefore is unavailable during inference.

Chamber segmentations are trained to dilate outward at a constant rate. During training, we first randomly perturb the seed point by moving $\mathbf{y}_0$ within the chamber's center region, and then generate two concentric spheres centered on it: the radius d_1 of the smaller sphere is chosen uniformly at random, and the larger radius is $d_2 = d_1 + d_s$. Both spheres are intersected with $\mathbf{y}$ to form ($\mathbf{y}_{in}, \mathbf{y}_{out}$).

Finally, the ground truth binary stopping indicator s is computed by comparing $\mathbf{y}_{out}$ with $\mathbf{y}$.

The seed points to be clicked by the user[1] were chosen to maximize the potential for automatic detection in future. For example, the aortic seed could have been placed at the aortic valve, and segmentations grown away from the heart. 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 towards one of the inter-structure boundaries that separate the global blood pool. The lack of contrast at these borders provides a challenging test case for automatic stopping.

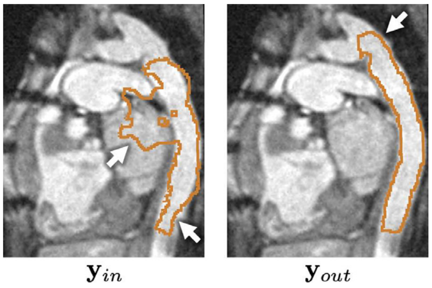

Fig. 7. Data augmentation creates corrupted inputs $\mathbf{y}_{in}$ and uncorrupted outputs $\mathbf{y}_{out}$ so the trained RNN is robust to errors in its intermediate results.

3.3. Data augmentation

We apply random affine and nonlinear transformations, left-right and anterior-posterior flips (relevant due to dextrocardia and other cardiac malpositions in CHD), constant intensity shifts and additive Gaussian noise.

Cardiac MR has inhomogeneity artifacts around implanted stents and a heterogeneous background due to inconsistent surrounding vasculature. We perform additional data augmentation for the AO and PA by adding random dark regions inside the vessels and random dark or bright regions next to them.

Finally, if our modified U-Net is trained using error-free input segmentations $\mathbf{y}_{in}$, then it may not operate well when performing inference on its own imperfect outputs at test time. We address this by corrupting each label of $\mathbf{y}_{in}$ using random nonrigid deformations and also add random foreground blobs that vary in number, location and size (Fig. 7). The output segmentation $\mathbf{y}_{out}$ remains unchanged. During training, the network must learn to correct errors in its input while simultaneously growing the segmentation appropriately. Hence, the trained model will be more robust when it operates recursively.

3.4. Multiclass learned iterative segmentation

Although we primarily focus on binary segmentation, we also demonstrate the application of our framework to multiclass iterative segmentation (i.e., $L > 2$). Multiclass segmentation enjoys the benefits of multi-task learning, requires training only one model that segments multiple anatomical structures, and eliminates the need to subsequently resolve conflicts between overlapping binary segmentations. Moreover, an iterative model that learns to simultaneously grow multiple segmentations might better learn the spatial relationships between them. For the most part, the training data generation for multiclass iterative segmentation can be achieved by separately processing the data for each anatomical label as described above for binary segmentation, with a few additional considerations. In our implementation, each multiclass $\mathbf{y}_{in}$ contains input partial segmentations of each structure that are approximately the same (random) percentage complete. We continue to use a single binary stopping indicator s. During data augmentation, corrupting the segmentation of a structure in $\mathbf{y}_{in}$ often changes the segmentations of neighboring structures, so we avoid biasing the degree to which each anatomical label is distorted by corrupting the foreground channels of $\mathbf{y}_{in}$ in a random order and

[1] LV, RV, LA, RA: center region; AO: bottom of descending aorta; PA: bottom of main PA trunk; SVC: superior end; IVC: center of hepatic segment.

making a structure's label map immutable after it has been corrupted.

4. Evaluation

We evaluated our iterative segmentation model for the task of whole heart segmentation in patients with CHD, comparing to several automatic and interactive learning-based methods that directly segment an image in one step.

4.1. Data

The first 20 cardiac MR scans come from our group's public HVSMR challenge dataset (http://segchd.csail.mit.edu, (Pace et al., 2015)). We retrospectively retrieved 40 additional images from the clinical archive for this study, thus creating a larger dataset of 60 images with manual segmentations of the LV, RV, LA, RA, AO, PA, SVC and IVC. This is a unique dataset acquired during clinical practice, and reflects the substantial manual labeling effort required for whole heart segmentation in CHD.

All images were acquired at Boston Children's Hospital on a 1.5T scanner (Philips Achieva) using ECG and respiratory navigator gating, the vast majority using a free-breathing SSFP pulse sequence (TR$\sim$3.4-4.7 ms, TE$\sim$1.7-2.4 ms, α $\sim$60-90°) (Moghari et al., 2017). A few images were acquired using alternate MR protocols; image appearance differences are relatively minor. Gadolinium-based contrast agent (Ablavar or Gadovist) was used in some patients at the discretion of clinicians. The image resolution averaged $0.9 \times 0.9 \times 0.85$mm. All images were cropped around the heart and resized to $126 \times 180 \times 144$. Intensity normalization was performed using estimates of the blood pool and lung intensity that were automatically derived from intensity histograms within predetermined image bounding boxes (see Supplementary Material). Subject age ranged from <1 to 55 years old.

Ground truth segmentation required many hours per image. The original HVSMR dataset included segmentations of the blood pool and myocardium only. Trained raters manually divided each blood pool surface model using 3D Slicer (Fedorov et al., 2012) by fitting local separating planes at the interfaces between structures. The 40 new images were segmented by combining manual contours of the interfaces between different heart structures made using a valve annotation module (Scanlan et al., 2018; Nguyen et al., 2019) with segmentations from a 3D U-Net trained using the HVSMR dataset, and performing extensive manual cleanup to fix errors and avoid bias towards the network's output. All segmentations were validated by hospital experts when the correct segmentation was ambiguous. For each cardiac structure, a simulated user click was created using morphological and center-of-mass calculations.

Under the advice of a cardiologist, all images were categorized as mild, moderate or severe, according to each heart's anatomical malformations (not clinical prognosis)[2] Note that even moderate cases have significant defects. Most subjects have a unique combination of heart defects (45% of moderate subjects and 89% of severe subjects).

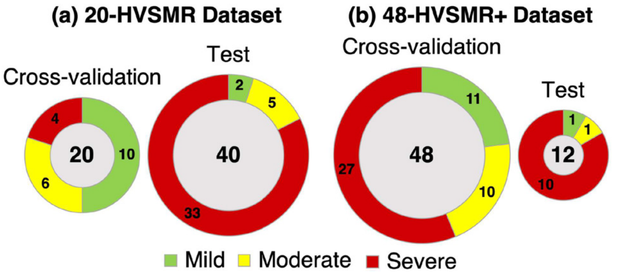

Fig. 8. Size and composition of the **20–HVSMR** and **48–HVSMR+** training datasets.

4.2. Experimental setup

We compared seven segmentation approaches, including five interactive and two fully-automatic methods. All of the interactive methods generate a full segmentation from a single user click per structure (i.e., there is no iterative back-and-forth with a user). (1) **Iter–A** is our iterative binary segmentation with automatic stopping. (2) **Iter–U** uses the same trained model as **Iter–A** but simulates a user who chooses the stopping point (by keeping the best segmentation from the first 40 iterations), included to evaluate how a slight increase in user interaction may improve segmentation accuracy. (3) **Iter–A–All** is our multiclass iterative segmentation with automatic stopping, and (4) **Iter–U–All** is our multiclass iterative segmentation with simulated user stopping (by keeping the segmentation from the first 40 iterations that has the best mean accuracy over all eight cardiac structures). (5) **U-Net–All** is a U-Net for fully-automatic multiclass segmentation of all 8 structures, included as a state-of-the-art reference. (6) **U–Net** is a U-Net for fully-automatic binary segmentation of each anatomical structure, included as a benchmark for binary segmentation. (7) **U–Net+S** is a U-Net for interactive binary segmentation which also inputs a Euclidean distance map to the user seed, included to evaluate the value of input seed points. We emphasize that **U-Net+S** is an interactive method as it relies on a distance map to a user-specified seed, and is nearly identical to prior state-of-the-art methods when generating an entire segmentation from a single user click (Xu et al., 2016; Sakinis et al., 2019). We also note that although the training data for the multiclass iterative segmentation models **Iter–A/U–All** is generated on-the-fly for each anatomical label separately, these are true multiclass models that input $L+1$ input channels and produce L output channels in a single forward pass for each iteration.

In total, we trained 26 models (1 **U–Net–All** model, 1 **Iter–A–All** model, plus 8 structures × 3 models for **Iter–A, U–Net** and **U-Net+S**. Note **Iter–U** uses the trained model from **Iter–A** and **Iter–U–All** uses the trained model from **Iter–A–All**). All models used the same data augmentation procedures for image transformations, intensity adjustments and MR artifact simulation. User interaction was simulated in our experiments, including the clicks to place the seed points required by **U–Net+S, Iter–A/U** and **Iter–A/U–All** and the user stopping required by **Iter–U** and **Iter–U–All**. In particular, **U–Net+S** used the same procedures as **Iter–A/U** to simulate input seeds, after which the Euclidean distance map is computed. For more details on network architectures, training data generation, data augmentation and learning procedures, including chosen parameters, please see the Supplementary Material.

We evaluated the impact of training dataset size as follows (see Fig. 8). First, we trained **20–HVSMR** models on the original HVSMR scans, and tested on the 40 new images. These experiments evaluate generalization from a very small dataset, which is biased towards more normal anatomy, to more severe cases. Second, we trained **48–HVSMR+** models using the 20 HVSMR images and 28 new images, and tested on 12 new images. These 12 images repre-

[2] **Mild**: Roughly normal anatomy, prior CHD surgery with restoration of normal anatomy, and/or a mildly or moderately dilated chamber or vessel, **Moderate**: Septal defect (VSD, ASD), abnormal connectivity (DORV, D-Loop TGA), bilateral SVC, severely dilated chamber/vessel, and/or connective tissue disorder causing tortuous vessels, **Severe**: Heart malpositions or situs inversus (dextrocardia, mesocardia, inverted ventricles or atria, left/central IVC or SVC), L-loop TGA, common atrium, single ventricle, and/or major prior reconstructive surgery resulting in highly abnormal anatomy (atrial switch, Rastelli, Glenn, Fontan). **Note:** A dilated chamber or vessel was counted only if it was the sole diagnosis.

Table 1
Overall Dice scores for held-out test images. For mild / moderate subjects, all methods have comparable performance. For severe subjects, our binary iterative segmentation is superior, and our multiclass iterative segmentation is superior to the baselines when the training dataset is small and imbalanced (**20–HVSMR**). Dice scores are averaged over all 8 cardiac structures and shown as mean ± standard deviation. Where shown, p-values indicate statistically significant differences between the given iterative method and **U–Net+S** ($^{*}p$) or **U–Net–All** ($^{**}p$) in a paired t-test with a threshold of 0.05. Table S1 of the Supplementary Material gives results for each cardiac structure in detail.

	Mild / Moderate Subjects				Severe Subjects			
	20-HVSMR		**48-HVSMR+**		**20-HVSMR**		**48-HVSMR+**	
U-Net-All	87.7±14.6		90.7±7.5		64.6±31.9		84.9±18.0	
U-Net	87.3±9.5		90.5±5.3		55.8±36.1		75.6±28.0	
U-Net+S	85.4±17.3		89.7±7.7		68.2±27.8		82.0±19.7	
Iter-A	88.1±7.7		90.4±4.3		79.0±17.2	$^{*}p < 10^{-11}$, $^{**}p < 10^{-12}$	85.1±15.6	
Iter-U	**91.1±4.1**	$\boldsymbol{^{*}p < 10^{-2}}$	**92.6±3.0**		**85.3±12.8**	$\boldsymbol{^{*}p < 10^{-22}}$, $\boldsymbol{^{**}p < 10^{-22}}$	**88.2±14.0**	$\boldsymbol{^{*}p < 10^{-2}}$
Iter-A-All	85.6±14.3		88.2±9.0		77.3±22.6	$^{*}p < 10^{-7}$, $^{**}p < 10^{-9}$	85.2±17.8	
Iter-U-All	87.4±13.6		90.3±7.3		78.9±22.0	$^{*}p < 10^{-10}$, $^{**}p < 10^{-11}$	86.8±18.1	

sent 20% of the total dataset size and also focus on generalization to severe subjects, but include one mild and one moderate subject as sanity checks. These experiments assess an algorithm's accuracy when a larger and more balanced dataset is available for training, although we note that the dataset of 48 images is still small. All training was done using 4-fold cross validation, where each fold had an approximately equal number of mild, moderate and severe cases. An ensemble of the resulting four networks yielded the final prediction on the test images.

We implemented our method using Keras (Chollet et al., 2015) with a Tensorflow backend (Abadi et al., 2015). Optimization was done using Adam (Kingma and Ba, 2015), with a learning rate of 10^{-4} and a batch size of 1, for 2000 epochs. Inference was fast enough for our iterative segmentation method to be used interactively. For binary iterative segmentation, each iteration required 0.65±0.15 seconds on an NVIDIA TITAN X GPU and the mean time required to segment one structure ranged from 2 to 18 seconds, leading to a total inference time of approximately 1 minute for whole heart segmentation. For binary iterative segmentation, we expect users to interact with the algorithm as each structure is segmented sequentially.

4.3. Evaluation metrics

We employ the Dice score to quantify the volume overlap between the ground truth and predicted segmentations.

All segmentations were post-processed to keep the island containing the user seed if the given method inputs a seed and the segmentation contained it, or the largest connected component otherwise. For the iterative segmentation methods, this is done after every step of inference. Vessel segmentations that are slightly too long or too short are no less clinically useful (Zhuang et al., 2019). Hence, we included "optional zones" for the AO, PA, SVC, IVC and pulmonary veins, which specified both a minimum required vessel length and a permitted continuation. Optional areas were subtracted from both the ground truth and algorithm segmentations before computing the Dice score, so that only the required regions were compared.

4.4. Segmentation accuracy

Table 1 and Fig. 9 report segmentation accuracy on held-out test images. Our proposed binary iterative segmentation with user stopping (**Iter–U**) consistently had the highest accuracy.

Mild and Moderate Subjects: We combine results from mild and moderate subjects because they were very similar: Welch's t-tests for independent samples with a threshold of 0.05 showed no significant differences between the overall Dice scores on mild versus moderate test images for any combination of segmentation algorithm and training dataset. All seven models performed well in mild/moderate subjects. Despite the small dataset sizes, good generalization is possible with or without user interaction since these hearts exhibit low variability and are well represented in both datasets.

Severe Subjects: Extreme cardiac malformations make segmenting severe subjects a much more challenging task. First, the user seed provides a signal for object localization that proved very useful since CHD often involves abnormal positioning of cardiac structures in the body, and **U–Net+S** outperformed **U–Net** for both training datasets.

Binary iterative segmentation had the highest accuracy. Note that **Iter–A** requires the same level of user input as **U–Net+S**, but had a better mean segmentation accuracy. The multiclass segmentation baseline (**U–Net–All**) did show some advantages compared to binary segmentation, but nevertheless had lower mean accuracy than **Iter–A** and **Iter–U**. In addition, **Iter–A** and **Iter–U** had smaller Dice score variances and suffered from fewer catastrophic failures compared to the baselines. In cases where **Iter–A** does not identify the ideal stopping point, we have evidence that iterative segmentation may be improved by a small additional amount of user interaction to choose the final segmentation (**Iter–U**). Including this interaction brought the mean Dice score in severe subjects to greater than 85, within ∼6 Dice points of the mean performance in mild and moderate subjects. As expected, accuracy improved with more training data, but the accuracy of binary iterative segmentation was less sensitive to the training dataset size than the three direct segmentation methods (**U–Net–All, U–Net** and **U–Net+S**). The benefits of iterative segmentation were especially pronounced in the smaller **20–HVSMR** dataset, in which paired t-tests with a threshold of 0.05 showed statistically significant differences between **Iter–U** and **U–Net+S** for all 8 cardiac structures, and between **Iter–A** and **U–Net+S** for all structures except the AO and SVC. Similarly, there were statistically significant differences between **Iter–U** and **U–Net–All** for all structures except the AO, and between **Iter–A** and **U–Net–All** for the LV, LA, RA, SVC and IVC. For both training datasets, the PA was the most difficult structure to segment, as corroborated by the results from a recent whole heart segmentation challenge (Zhuang et al., 2019). For **20–HVSMR**, the RV, RA and SVC were also relatively difficult to segment.

Multiclass iterative segmentation shows similar trends. Most importantly, for the **20–HVSMR** dataset both **Iter–A–All** and **Iter–U–All** showed statistically significant improvements over all three baseline algorithms, including the multiclass **U–Net–All**. However, for multiclass iterative segmentation simulated user stopping did not improve the results as much as it did for binary iterative segmentation. Regarding individual cardiac structures, paired t-tests with a threshold of 0.05 showed significantly significant differences between **Iter–U–All** and **U–Net+S** for all structures except the AO and IVC, and between **Iter–A–All** and **U–Net+S** for the LV, RV, LA, PA and SVC. Similarly, there were statistically significant

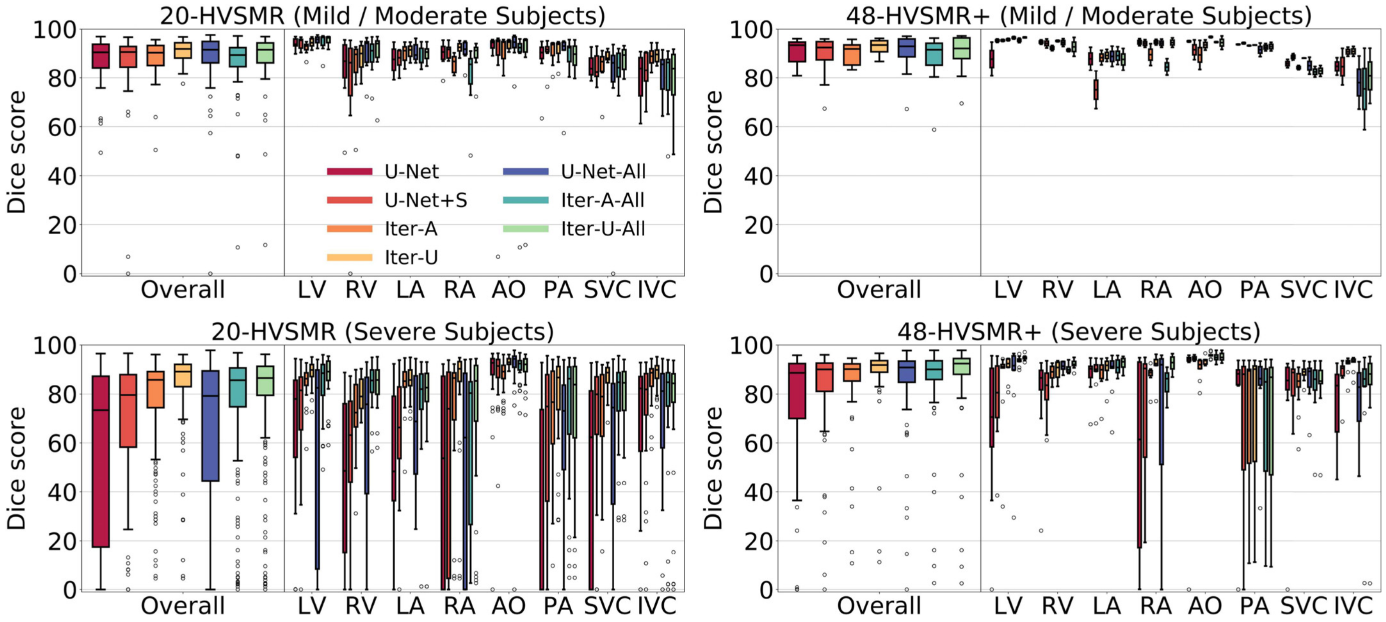

Fig. 9. Overall and structure-specific Dice scores for held-out test images. For mild / moderate subjects, all methods have comparable performance. For severe subjects, our binary iterative segmentation is superior, and our multiclass iterative segmentation is superior to the baselines when the training dataset is small and imbalanced (**20–HVSMR**).

differences between **Iter–U–All** and **U–Net–All** for the LV, RV, LA, RA and SVC, and between **Iter–A–All** and **U–Net–All** for the LV, RV, LA and SVC. We again found that accuracy improved when using the larger **48–HVSMR+** dataset. Here, the overall Dice scores of **Iter–A–All** and **Iter–U–All** were similar to or better than those of the three baselines, although no statistically significant improvements were found. For multiclass iterative segmentation, the PA was again difficult to segment for both training datasets. For **20–HVSMR**, the RA and IVC had the lowest accuracy.

Finally, we compare multiclass to binary iterative segmentation. Although the binary iterative segmentation methods most often have a higher mean Dice score than their corresponding multiclass iterative segmentation methods, the difference is typically not statistically significant in paired t-tests with a threshold of 0.05. The exception is between **Iter–U** and **Iter–U–All** for models trained with the smaller **20–HVSMR** dataset in both mild / moderate and severe subjects. In severe subjects, for the **20–HVSMR** dataset, **Iter–U–All** was significantly better than **Iter–U** for the RV but significantly worse for the LA, AO, SVC and IVC, and **Iter–A–All** was significantly better than **Iter–A** for the RV and AO but significantly worse for the LA and IVC. For the **48–HVSMR+** dataset, **Iter–U–All** showed statistically significant advantages for the LV and AO, and **Iter–A–All** showed statistically significant advantages for the AO. No significant disadvantages were found for severe subjects in **48–HVSMR+** between corresponding binary and multiclass algorithms.

In the remaining evaluations, we focus on the performance of **Iter–U** using the **48–HVSMR+** training dataset, since it had the best accuracy.

4.5. Example segmentations and failure cases

Example **Iter–U** segmentations are shown in Fig. 10. The bottom rows illustrate some failure cases.In the LV, RV, LA and RA, the **Iter–U** segmentation sometimes grew through a septal defect or an adjoining valve, with a coincident under-segmentation of the basal and apical ventricles, the main LA chamber or the pulmonary veins. Chamber segmentations could also be under-segmented near septal defects. LA segmentations sometimes suffered from pulmonary veins that were missing or excessively long, while RA segmentations were sometimes mis-segmented at the IVC boundary. The AO surface models were highly accurate, but could be too bumpy near dark inhomogeneity artifacts. The main failure case for the PA was a segmentation that could not grow through a narrow main PA (MPA) trunk for patients with a PA band, PA stenosis, or whose images had artifacts in the MPA region. The left or right PA branch sometimes grew past the maximum length specified by the "optional zones". Finally, slight differences compared to the ground truth could be found at the interface between the SVC/IVC and the RA, or between the SVC/IVC and the attached PA for patients with prior Glenn or Fontan surgeries.

4.6. User interaction: automatic versus user stopping

Some examples in which **Iter–U** outperformed **Iter–A** are shown in Fig. 11. Recall that in our proposed setup, the **Iter–A** result is initially shown, and the user can accept it or look forwards or backwards in the output sequence for a better result. The error in the number of iterations predicted via automatic stopping directly causes the decrease in accuracy for **Iter–A** compared to **Iter–U**, as well as the number of additional clicks required of the user. However, Fig. 12 shows that this number was typically small and often zero, i.e., a user would not have to search far, if at all, from the automatically proposed solution.

4.7. Do the iterative models learn the desired trajectories?

Our iterative segmentation algorithm is trained to produce sequences of output segmentations that follow a prescribed growth pattern. To evaluate this evolution, for each cardiac structure in the test images we compared each intermediate segmentation in the predicted output sequence $\mathbf{y}_1^*, \ldots, \mathbf{y}_t^*, \ldots$ (for all t up to and including the stopping point chosen by the simulated user in **Iter–U**) to those in the ground truth sequence that begins with the same initial segmentation $\mathbf{y}_0$ and grows according to the desired step size (as described in Section 3.2). Results are shown in Fig. 13. High Dice scores (approximately 80) over all subjects indicates good correspondences between the generated and ideal output sequences.

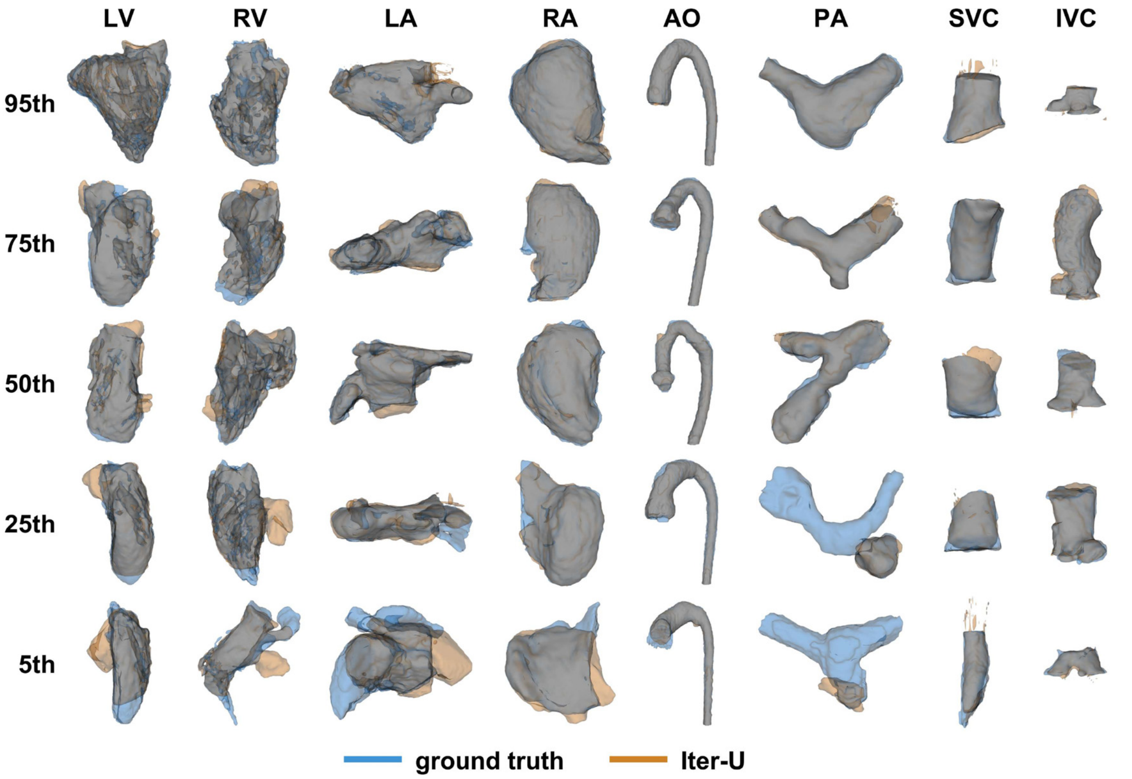

Fig. 10. Inspection of representative **Iter–U** segmentations of severe subjects predominantly shows good overlap (gray), with some examples of under-segmentation (blue) and over-segmentation (orange). Results are from test subjects after training using **48–HVSMR+**. Subjects were chosen for visualization according to the Dice score at the 95th, 75th, 50th, 25th and 5th percentiles.

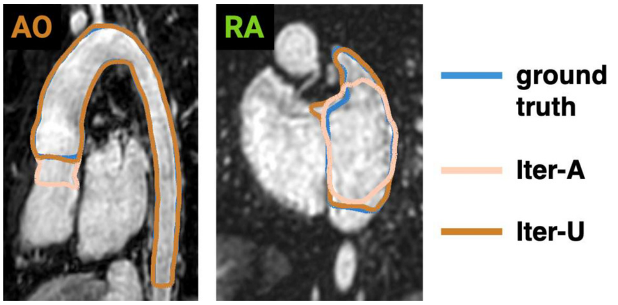

Fig. 11. Examples from severe subjects for which our iterative segmentation allowed the simulated user to choose a better segmentation than that predicted by automatic stopping. The user can select an earlier segmentation if the automatic segmentation is too large, or ask for more iterations if is under-segmented.

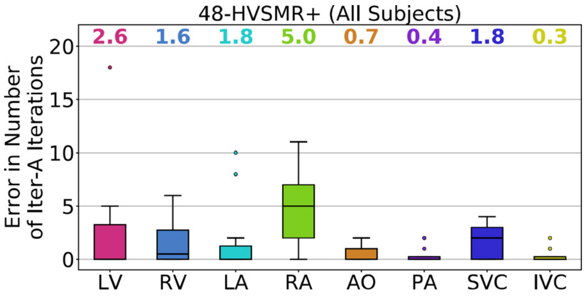

Fig. 12. The automatically detected number of iterations (used by **Iter–A**) was typically close to the ideal number of iterations (used by **Iter–U**). Results are from all 12 test subjects after training using **48–HVSMR+**. The mean error for each structure is shown at the top of the graph.

This adds further support for the accuracy of the learned iterative process, in addition to the improved accuracy of the final output segmentations.

4.8. Impact of input seed location

We used cardiac chamber segmentation to evaluate the robustness of **Iter–U** and **Iter–A** to the location of the input seed. For each test image, we varied the ideal "center" seed point location (derived from morphological and center-of-mass calculations as described above), sampling 20 "moving" seed points uniformly at random within a spherical region whose radius was set to 25%, 30%, 40% and 50% of the maximum possible distance to the ground truth boundary. Note that 25% is the same amount of perturbation as was used when randomly varying the seed point location during training. Results are shown in Fig. 14. Varying the input seed location within the 25%, 30%, 40% and 50% regions decreased the mean Dice score for the cardiac chambers by 1.4, 1.6, 2.2 and 2.7 points for **Iter–U**, respectively, and by 1.5, 2.2, 3.2 and 4.2 points for **Iter–A**. RA segmentation was the most sensitive. For **Iter–U**, Welch's *t*-tests for independent samples with a threshold of 0.05 showed no significant differences between the results for moving vs. center seed point locations for any cardiac chamber or moving seed region size. For **Iter–A**, the only significant differences were for the RV's 50% region and the RA's 25%, 30%, 40% and 50% regions.

4.9. Impact of congenital heart defects

A detailed examination of the impact of different heart defects, prior surgeries and MR artifacts on segmentation accuracy is given by Fig. 15. The PA was the most difficult structure to segment in the presence of cardiac malformations. Counterintuitively, segmentation accuracy increased in the LA for severe subjects compared to

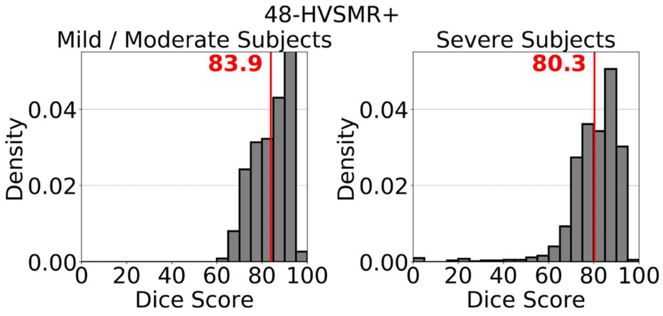

Fig. 13. The distribution of Dice scores between predicted and ground truth intermediate segmentations in **Iter–U**'s output sequences show high overlap. Results are from all 12 test subjects after training using **48–HVSMR+**. The mean Dice score for each group of subjects is shown at the top of the graph.

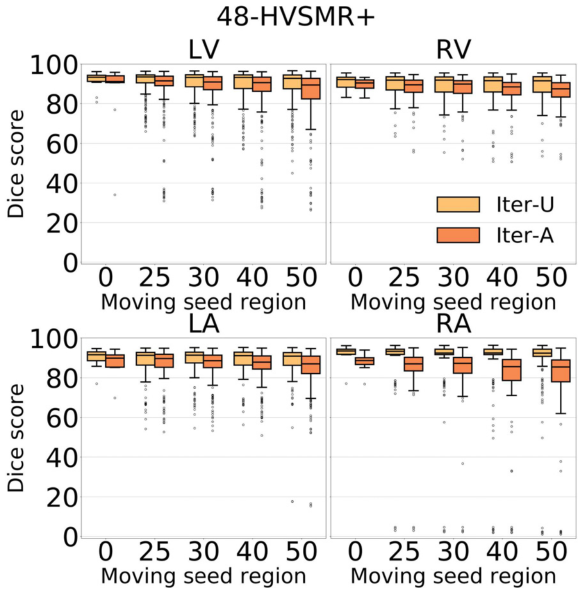

Fig. 14. Perturbing the input seed had a minor impact on segmentation accuracy for **Iter–U** for all cardiac chambers and for **Iter–A** for all cardiac chambers except the RA. Results are from all 12 test subjects after training using **48–HVSMR+**.

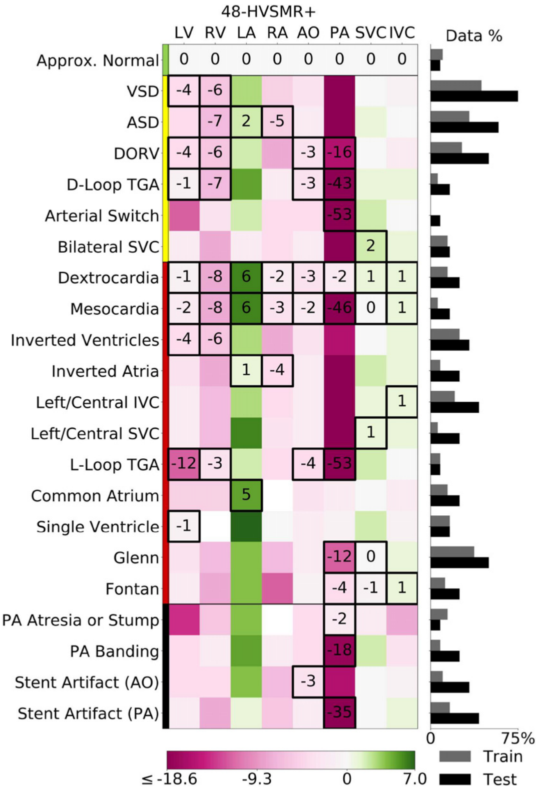

Fig. 15. Comparing the mean Dice score for patients with heart defects, prior surgeries and MR artifacts to that of normal subjects shows that these conditions variously impact the accuracy of **Iter–U**. Green squares indicate higher accuracy than in normal subjects, pink squares indicate lower accuracy. The darkest pink intensity is set to -20% of the overall Dice score in normal subjects. The left vertical line indicates mild (green), moderate (yellow), severe (red) and coincident (black) abnormalities. Boxes indicate which structures are expected to be affected by each cardiac abnormality, but additional structures may be impacted if other conditions co-occur with the given abnormality. On the right, gray and black horizontal bars indicate prevalence in the **48–HVSMR+** cross validation and test splits, respectively. Results are from all 12 test subjects after training using **48–HVSMR+**.

normal subjects. Segmentation accuracy in the SVC and IVC were the least impacted by heart defects, while the LV, RV, RA and AO were moderately impacted. These results indicate where more data or specialized data augmentation schemes might be helpful in future, e.g., to address difficulties in segmenting the PA (under various conditions) and subjects with L-Loop TGA or prior arterial switch.

4.10. Additional ablation studies

The results of additional ablation studies examining the impact of (1) the step size d_s, (2) the new data augmentation step compared to (Pace et al., 2018), and (3) segmentation post-processing can be found in the Supplementary Material.

5. Discussion

We show that it is possible to learn a whole heart segmentation model for cardiac MR images from patients with congenital heart disease, despite limited labeled data and high anatomical variability. The proposed iterative segmentation method provides a new approach for very difficult segmentation problems in which some user interaction remains necessary, for example when creating new labeled datasets.

Iterative segmentation with simulated user stopping consistently had the best overall segmentation accuracy in our experiments. Moreover, iterative segmentation with automatic stopping outperformed all of the direct segmentation methods for subjects with severe cardiac malformations. All algorithms were highly accurate for patients with mild or moderate heart defects. This is because they make up a large proportion of both training datasets while exhibiting lower anatomical variability. However, the direct segmentation methods broke down in severe subjects, especially when the training dataset was very small and imbalanced (**20–HVSMR**). The much improved performance of **Iter–A/U** and **Iter–A/U–All** in this situation is impressive since only 4 severe subjects were available for training. Iterative segmentation was less sensitive to dataset size, produced output segmentations whose predictability enables simpler user interaction, coped well with variability in the input seed location, and binary iterative segmentation computed a whole heart segmentation in less than one minute with just a few seconds per structure.

We envision that our iterative segmentation RNN can be used within a complete solution to interactive segmentation as follows:

1. The user places one seed click per structure.
2. Run **Iter–A**, which produces a segmentation with automatic stopping.
3. In rare cases of significant **Iter–A** failures (e.g., a PA doesn't grow through a stent or MR artifact region): the user can go back in the output segmentation sequence to identify the first

point of failure and fix it (e.g., draw the PA through the difficult region) using existing interactive segmentation methods (Wang et al., 2018; 2019; Xu et al., 2016; Sakinis et al., 2019) or manual editing. Our iterative segmentation RNN can then be easily restarted to continue growing the segmentation, since at each step it requires only an image and a partially completed segmentation (without relying on a memory).

4. If the **Iter–A** result is not ideal, the user can look a few steps forward or backward in the segmentation, yielding the **Iter–U** result.
5. Finally, as for any algorithmic segmentation, the user can make any final required edits using existing interactive segmentation methods (Wang et al., 2018; 2019; Xu et al., 2016; Sakinis et al., 2019).

In this way, user effort is minimized because s/he knows what the segmentation should look like as it progresses, and can make minimal corrections at intermediate phases if needed. In addition, the user does not have to look at the entire image to monitor progress, since updates are limited in spatial extent, which is especially important for 3D images.

Our results corroborate previous studies in which direct segmentation was outperformed by iterative approaches (Pinheiro and Collobert, 2014; McIntosh et al., 2018; Le et al., 2018a; 2018b; Chakravarty and Sivaswamy, 2019; Januszewski et al., 2018). Conventional feedforward neural networks perform inference at each voxel independently. In contrast, iterative segmentation models can learn both local structure and long-range dependencies in the output domain (Havaei et al., 2017; Pinheiro and Collobert, 2014), since inference at each pixel is informed by label estimates from its surroundings, perhaps more effectively propagating information from distant landmarks. Binary iterative segmentation can learn spatial correlations between subparts of a single anatomical structure, while multiclass iterative segmentation can learn correlations between them. Iterative segmentation also implicitly expands the model's field of view without increasing its complexity. Finally, we observed that our iterative segmentation model successfully learned to slowly expand a single connected component connected to the user seed, leading to simple post-processing, while the direct segmentation methods produced multiple islands inside and outside the anatomical structure of interest. In particular, the distance map used by **U–Net+S** encodes the same user click as **Iter–A/U**, but may be less informative at intermediate distances from the seed point for compact structures that vary in size (e.g., LV, RV, LA, RA, IVC), or when distinguishing between long vessels that lie in close proximity (e.g., AO vs. PA).

Our iterative segmentation method could be applied to any number of growth dynamics, as long as appropriate 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 (Zheng et al., 2018; Poudel et al., 2017). For the four cardiac chambers, a possible alternative is to train the network 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.

We acknowledge a few limitations of this study and ideas for future research. Additional investigations into automatic stopping could close the gap between **Iter–A**/**Iter–A–All** and **Iter–U**/**Iter–U–All**. Our experiments show that some structures remain difficult to segment, including the PA in general and the LV, RV, RA and IVC for certain pathologies. In our experiments, the multiclass iterative segmentation models generally performed worse than the binary models, and did not realize as large an advantage from simulated user stopping versus automatic stopping. They were also sometimes outperformed by **U–Net–All** in mild / moderate subjects. This may indicate that there exists a good stopping point for each structure in the sequences of binary segmentations for the simulated user to select, but that the multiclass output sequences may not necessarily have a segmentation that simultaneously contains the best segmentation for each structure. Developing variants that further extend the multiclass iterative segmentation method to predict a separate stopping point for each cardiac structure is hence a promising area for future research. All procedures and parameters for training data generation, data augmentation, and learning were tuned for binary iterative segmentation and then subsequently applied to multiclass iterative segmentation, and might be further optimized for the multiclass setting. That said, from a user interaction perspective is possible that segmenting each structure sequentially may be simpler than interacting with a multiclass segmentation. Incorporating additional input channels, for example containing geodesic distance maps to foreground and background user annotations (Wang et al., 2019), may further improve accuracy for **U–Net+S, Iter–A/U** and **Iter–A/U–All**. A variable step size for iterative segmentation could be tuned by the user as an additional input parameter or automatically adapted by the network itself, which could reduce the required number of steps for easy cases. An empirical study that evaluates different RNN architectures is an interesting direction for future research. Finally, a user study would be informative to quantify any differences in segmentation accuracy between actual users versus the user simulation used in our experiments. A user study could also investigate our hypothesis that interaction with the sequence of predictably evolving segmentations output by our iterative model is faster and more intuitive than interactive segmentation methods that require substantial back-and-forth with the user over the entire 3D volume.

In addition to enabling 3D visualization of cardiac anatomy for surgical planning, automated whole heart segmentation could also be used to compute important functional indices that are currently derived from manual segmentations in CHD patients. Applied to 3D+time imaging in future, fast and accurate segmentation would also enable research into dynamic heart surface models for surgical planning and into simulating post-surgical hemodynamics, assessing joint atrio-ventricular function, and quantifying vessel wall stiffness, perhaps via patient-specific models that incorporate biophysical properties (Weese et al., 2013; Suinesiaputra et al., 2016) or information from 4D flow MRI and computational fluid dynamics (Lawley et al., 2018). Finally, once patient-specific anatomy and function can be captured and summarized, it may be possible to learn outcome prediction models from data to forecast the consequences of competing surgical approaches in CHD.

6. Conclusions

We propose a learned iterative segmentation model, implemented as a recurrent neural network, that inputs a single user click and evolves a segmentation over multiple steps. We develop a novel loss function and use it to learn the model's parameters so that the growth pattern of its output segmentations corresponds to that defined by training data. The final output segmentation can be inferred automatically, or can be easily adjusted by the user. We use our iterative model to demonstrate the first whole heart segmentation in cardiac MR for patients with congenital heart disease, and use this challenging application to show our model's superiority over conventional feedforward neural networks when anatomical variability is wide and training datasets are small. The resulting method enables patient-specific 3D heart surface models for surgical planning for CHD.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project was supported by NIH NIBIB NAC P41EB015902, NIH NHLBI 1R01HL149807-01, NIH 1R56AG064027, NIH 1R01AG064027, Philips Inc., NSERC CGS-D, Wistron Corporation, the Boston Children's Hospital Translational Research Program and Office of Faculty Development, Harvard Catalyst, the Charles H. Hood Foundation and the American Heart Association. We would also like to thank Alireza Akhondi-Asl, Hannah Contreras, Bruce Fischl, Patricia Gao, Shruti Ghelani, Sunil Ghelani, Christian Herz, Matthew Jolley, Andras Lasso, Imon Rahaman and Yue Zhang.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2022.102469.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Amrehn, M., Gaube, S., Unberath, M., Schebesch, F., Horz, T., Strumia, M., Steidl, S., Kowarschik, M., Maier, A., 2017. 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), pp. 143–147.

Arafati, A., Hu, P., Finn, J.P., Rickers, C., Cheng, A.L., Jafarkhani, H., Kheradvar, A., 2019. Artificial intelligence in pediatric and adult congenital cardiac MRI: an unmet clinical need. Cardiovasc. Diagn. Ther. 9 (Suppl 2), S310–S325.

Bhatla, P., Tretter, J.T., Ludomirsky, A., Argilla, M., Latson, L.A., Chakravarti, S., Barker, P.C., Yoo, S.-J., McElhinney, D.B., Wake, N., Mosca, R.S., 2017. Utility and scope of rapid prototyping in patients with complex muscular ventricular septal defects or double-outlet right ventricle: does it alter management decisions? Pediatr. Cardiol. 38 (1), 103–114.

Boykov, Y., Jolly, M., 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: IEEE International Conference on Computer Vision (ICCV), Vol. 1, pp. 105–112.

Byrne, N., Velasco Forte, M., Tandon, A., Valverde, I., Hussain, T., 2016. A systematic review of image segmentation methodology, used in the additive manufacture of patient-specific 3D printed models of the cardiovascular system. JRSM Cardiovasc. Dis. 5. 2048004016645467.

Caruana, R., 1997. Multitask learning. Mach. Learn. 28, 41–75.

Chakravarty, A., Sivaswamy, J., 2019. RACE-Net: a recurrent neural network for biomedical image segmentation. IEEE J. Biomed. Health Inform. 23 (3), 1151–1162.

Chollet, F., et al., 2015. Keras. https://keras.io.

Criminisi, A., Sharp, T., Blake, A., 2008. GeoS: Geodesic image segmentation. In: European Conference on Computer Vision (ECCV). In: Lecture Notes in Computer Science, Vol. 5302, pp. 99–112.

Dalca, A., Danagoulian, G., Kikinis, R., Schmidt, E., Golland, P., 2011. Segmentation of nerve bundles and ganglia in spine MRI using particle filters. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 6893, pp. 537–545.

Dalca, A.V., Guttag, J., Sabuncu, M.R., 2018. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9290–9299.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017. 3D Deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. 41, 40–54.

Ecabert, O., Peters, J., Schramm, H., Lorenz, C., von Berg, J., Walker, M.J., Vembar, M., Olszewski, M.E., Subramanyan, K., Lavi, G., Weese, J., 2008. Automatic model-based segmentation of the heart in CT images. IEEE Trans. Med. Imaging 27 (9), 1189–1201.

Ecabert, O., Peters, J., Walker, M.J., Ivanc, T., Lorenz, C., Berg, J.v., Lessick, J., Vembar, M., Weese, J., 2011. Segmentation of the heart and great vessels in CT images using a model-based adaptation framework. Med. Image Anal. 15 (6), 863–876.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J.V., Pieper, S., Kikinis, R., 2012. 3D Slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging 30 (9), 1323–1341.

Frescura, C., Büchel, E.V., Ho, S.Y., Thiene, G., 2010. Anatomical and Pathophysiological Classification of Congenital Heart Disease. In: Saremi, F., Achenbach, S., Arbustini, E., Narula, J. (Eds.), Revisiting Cardiac Anatomy: A Computed-Tomography-Based Atlas and Reference. Blackwell Publishing, Chichester, UK, pp. 40–75.

Garekar, S., Bharati, A., Chokhandre, M., Mali, S., Trivedi, B., Changela, V.P., Solanki, N., Gaikwad, S., Agarwal, V., 2016. Clinical application and multidisciplinary assessment of three dimensional printing in double outlet right ventricle with remote ventricular septal defect. World J. Pediatr. Congenital Heart Surg. 7 (3), 344–350.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT Press, Cambridge, MA, USA.

Grady, L., 2006. Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 28 (11), 1768–1783.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.

Iglesias, J.E., Sabuncu, M.R., Leemput, K.V., 2013. Improved inference in bayesian segmentation using monte carlo sampling: application to hippocampal subfield volumetry. Med. Image Anal. 17 (7), 766–778.

Januszewski, M., Kornfeld, J., Li, P., Pope, A., Blakely, T., Lindsey, L., Maitin-Shepard, J., Tyka, M., Denk, W., Jain, V., 2018. High-precision automated reconstruction of neurons with flood-filling networks. Nat. Methods 15 (8), 605–610.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: International Conference for Learning Representations (ICLR).

Lau, I., Sun, Z., 2018. Three-dimensional printing in congenital heart disease: a systematic review. J. Med. Radiat. Sci. 65 (3), 226–236.

Lawley, C.M., Broadhouse, K.M., Callaghan, F.M., Winlaw, D.S., Figtree, G.A., Grieve, S.M., 2018. 4D Flow magnetic resonance imaging: role in pediatric congenital heart disease. Asian Cardiovasc. Thoracic Annals 26 (1), 28–37.

Le, T.H.N., Gummadi, R., Savvides, M., 2018. Deep recurrent level set for segmenting brain tumors. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 11072, pp. 646–653.

Le, T.H.N., Quach, K.G., Luu, K., Duong, C.N., Savvides, M., 2018. Reformulating level sets as deep recurrent neural network approach to semantic segmentation. IEEE Trans. Image Process. 27 (5), 2393–2407.

Lessmann, N., van Ginneken, B., de Jong, P.A., Išgum, I., 2019. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. Med. Image Anal. 53, 142–155.

Liu, M., Li, F., Yan, H., Wang, K., Ma, Y., Shen, L., Xu, M., 2020. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease. NeuroImage 208, 116459.

Liu, T., Tian, Y., Zhao, S., Huang, X., 2020. Graph Reasoning and Shape Constrains for Cardiac Segmentation in Congenital Heart Defect. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 12264, pp. 607–616.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.

Mansi, T., Voigt, I., Leonardi, B., Pennec, X., Durrleman, S., Sermesant, M., Delingette, H., Taylor, A., Boudjemline, Y., Pongiglione, G., Ayache, N., 2011. A statistical model for quantification and prediction of cardiac remodelling: application to tetralogy of Fallot. IEEE Trans. Med. Imaging 30 (9), 1605–1616.

McIntosh, L., Maheswaranathan, N., Sussillo, D., Shlens, J., 2018. Recurrent segmentation for variable computational budgets. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1729–172909.

Mo, Y., Liu, F., McIlwraith, D., Yang, G., Zhang, J., He, T., Guo, Y., 2018. The Deep Poincaré map: a Novel Approach for Left Ventricle Segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 11073, pp. 561–568.

Moghari, M.H., Geva, T., Powell, A.J., 2017. Prospective heart tracking for whole-heart magnetic resonance angiography. Magn. Reson. Med. 77 (2), 759–765.

Mortensen, E.N., Barrett, W.A., 1998. Interactive segmentation with intelligent scissors. Graph. Models Image Process. 60 (5), 349–384.

Nguyen, A.V., Lasso, A., Nam, H.H., Faerber, J., Aly, A.H., Pouch, A.M., Scanlan, A.B., McGowan, F.X., Mercer-Rosa, L., Cohen, M.S., Simpson, J., Fichtinger, G., Jolley, M.A., 2019. Dynamic three-dimensional geometry of the tricuspid valve annulus in hypoplastic left heart syndrome with a Fontan circulation. J. Am. Soc. Echocardiograph. 32 (5), 655–666.

Ntsinjana, H.N., Hughes, M.L., Taylor, A.M., 2011. The role of cardiovascular magnetic resonance in pediatric congenital heart disease. J. Cardiovasc. Magn. Resonance 13, 51.

Pace, D., Dalca, A., Geva, T., Powell, A., Moghari, M., Golland, P., 2015. Interactive Whole-Heart Segmentation in Congenital Heart Disease. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 9351, pp. 80–88.

Pace, D.F., Dalca, A.V., Brosch, T., Geva, T., Powell, A.J., Weese, J., Moghari, M.H., Golland, P., 2018. 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). In: Lecture Notes in Computer Science, Vol. 11045, pp. 334–342.

Pandya, B., Cullen, S., Walker, F., 2016. Congenital heart disease in adults. BMJ 354, i3905.

Payer, C., Stern, D., Bischof, H., Urschler, M., 2017. Multi-label Whole Heart Segmentation Using CNNs and Anatomical Label Configurations. In: Statistical Atlases and Computational Models of the Heart (STACOM). In: Lecture Notes in Computer Science, Vol. 10663, pp. 190–198.

Peng, P., Lekadir, K., Gooya, A., Shao, L., Petersen, S.E., Frangi, A.F., 2016. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. Magn. Reson. Mater. Phys., Biol. Med. 29 (2), 155–195.

Peters, J., Ecabert, O., Meyer, C., Kneser, R., Weese, J., 2010. Optimizing boundary detection via simulated search with applications to multi-modal heart segmentation. Med. Image Anal. 14 (1), 70–84.

Petersen, S.E., Khanji, M.Y., Plein, S., Lancellotti, P., Bucciarelli-Ducci, C., 2019. 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. Eur. Heart J. 20 (12), 1321–1331.

Pinheiro, P., Collobert, R., 2014. Recurrent Convolutional Neural Networks for Scene Labeling. In: International Conference on Machine Learning. In: Proceedings of Machine Learning Research, Vol. 32, pp. 82–90.

Poudel, R.P.K., Lamata, P., Montana, G., 2017. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In: Reconstruction, Segmentation, and Analysis of Medical Images (RAMBO). In: Lecture Notes in Computer Science, Vol. 10129, pp. 83–94.

Ren, M., Zemel, R., 2017. End-to-end instance segmentation with recurrent attention. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6656–6664.

Riesenkampff, E., Rietdorf, U., Wolf, I., Schnackenburg, B., Ewert, P., Huebler, M., Alexi-Meskishvili, V., Anderson, R.H., Engel, N., Meinzer, H.-P., Hetzer, R., Berger, F., Kuehne, T., 2009. The practical clinical value of three-dimensional models of complex congenitally malformed hearts. J. Thorac. Cardiovasc. Surg. 138 (3), 571–580.

Romera-Paredes, B., Torr, P.H.S., 2016. Recurrent instance segmentation. In: European Conference on Computer Vision (ECCV). In: Lecture Notes in Computer Science, Vol. 9910, pp. 312–329.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 9351, pp. 234–241.

Rother, C., Kolmogorov, V., Blake, A., 2004. “GrabCut”: interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23 (3), 309–314.

Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C., 2017. Error Corrective Boosting for Learning Fully Convolutional Networks with Limited Data. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 10435, pp. 231–239.

Sakinis, T., Milletari, F., Roth, H., Korfiatis, P., Kostandy, P.M., Philbrick, K., Akkus, Z., Xu, Z., Xu, D., Erickson, B.J., 2019. Interactive segmentation of medical images through fully convolutional neural networks. arXiv preprint arXiv: 1903.08205.

Scanlan, A.B., Nguyen, A.V., Ilina, A., Lasso, A., Cripe, L., Jegatheeswaran, A., Silvestro, E., McGowan, F.X., Mascio, C.E., Fuller, S., Spray, T.L., Cohen, M.S., Fichtinger, G., Jolley, M.A., 2018. Comparison of 3D echocardiogram-derived 3D printed valve models to molded models for simulated repair of pediatric atrioventricular valves. Pediatr. Cardiol. 39 (3), 538–547.

Seraphim, A., Knott, K.D., Augusto, J., Bhuva, A.N., Manisty, C., Moon, J.C., 2020. Quantitative cardiac MRI. J. Magn. Reson. Imaging 51 (3), 693–711.

Sethian, J.A., 1996. A fast marching level set method for monotonically advancing fronts. Proc. Natl. Acad. Sci. 93 (4), 1591–1595.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., Woo, W.-c., 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Neural Information Processing Systems (NIPS), pp. 802–810.

Sonka, M., Hlavac, V., Boyle, R., 2008. Image processing, analysis and machine vision. Cengage Learning, Stamford, CT, USA.

Suinesiaputra, A., McCulloch, A.D., Nash, M.P., Pontre, B., Young, A.A., 2016. Cardiac image modelling: breadth and depth in heart disease. Med. Image Anal. 33, 38–43.

Valverde, I., Gomez-Ciriza, G., Hussain, T., Suarez-Mejias, C., Velasco-Forte, M.N., Byrne, N., Ordoñez, A., Gonzalez-Calle, A., Anderson, D., Hazekamp, M.G., Roest, A.A.W., Rivas-Gonzalez, J., Uribe, S., El-Rassi, I., Simpson, J., Miller, O., Ruiz, E., Zabala, I., Mendez, A., Manso, B., Gallego, P., Prada, F., Cantinotti, M., Ait-Ali, L., Merino, C., Parry, A., Poirier, N., Greil, G., Razavi, R., Gomez-Cia, T., Hosseinpour, A.-R., 2017. Three-dimensional printed models for surgical planning of complex congenital heart defects: an international multicentre study. Eur. J. Cardio-Thoracic Surg. 52 (6), 1139–1148.

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Oliver, A., Lladó, X., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage 155, 159–168.

Vezhnevets, V., Konouchine, V., 2005. ”GrowCut” - Interactive Multi-label N-D Image Segmentation by Cellular Automata. In: International Conference on Computer Graphics and Vision (GraphiCon), pp. 150–156.

Wachinger, C., Reuter, M., Klein, T., 2018. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. NeuroImage 170, 434–445.

Wang, C., Smedby, O., 2017. Automatic Whole Heart Segmentation Using Deep Learning and Shape Context. In: Statistical Atlases and Computational Models of the Heart (STACOM). In: Lecture Notes in Computer Science, Vol. 10663, pp. 242–249.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T., 2018. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans. Med. Imaging 37 (7), 1562–1573.

Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Deepigeos: a deep interactive geodesic framework for medical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 41 (7), 1559–1572.

Weese, J., Groth, A., Nickisch, H., Barschdorf, H., Weber, F.M., Velut, J., Castro, M., Toumoulin, C., Coatrieux, J.L., Craene, M.D., Piella, G., Tobón-Gomez, C., Frangi, A.F., Barber, D.C., Valverde, I., Shi, Y., Staicu, C., Brown, A., Beerbaum, P., Hose, D.R., 2013. Generating anatomical models of the heart and the aorta from medical images for personalized physiological simulations. Med. Biol. Eng. Comput. 51 (11), 1209–1219.

Williams, R., Zipser, D., 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. 1 (2), 270–280.

Wolterink, J., Leiner, T., Viergever, M., Išgum, I., 2017. Dilated Convolutional Neural Networks for Cardiovascular MR Segmentation in Congenital Heart Disease. In: Reconstruction, Segmentation, and Analysis of Medical Images (HVSMR). In: Lecture Notes in Computer Science, Vol. 10129, pp. 95–102.

Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.S., 2016. Deep interactive object selection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 373–381.

Xu, X., Wang, T., Shi, Y., Yuan, H., Jia, Q., Huang, M., Zhuang, J., 2019. Whole Heart and Great Vessel Segmentation in Congenital Heart Disease Using Deep Neural Networks and Graph Matching. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 11765, pp. 477–485.

Yang, X., Bian, C., Yu, L., Ni, D., Heng, P.-A., 2018. Hybrid loss guided convolutional networks for whole heart parsing. In: Statistical Atlases and Computational Models of the Heart (STACOM). In: Lecture Notes in Computer Science, Vol. 10663, pp. 215–223.

Yu, L., Cheng, J.-Z., Dou, Q., Yang, X., Chen, H., Qin, J., Heng, P.-A., 2017. Automatic 3D Cardiovascular MR Segmentation with Densely-Connected Volumetric ConvNets. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 10434, pp. 287–295.

Zhang, H., Wahle, A., Johnson, R., Scholz, T., Sonka, M., 2010. 4-D Cardiac MR image analysis: left and right ventricular morphology and function. IEEE Trans. Med. Imaging 29 (2), 350–364.

Zhang, P., Wang, F., Zheng, Y., 2018. Deep Reinforcement Learning for Vessel Centerline Tracing in Multi-Modality 3D Volumes. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). In: Lecture Notes in Computer Science, Vol. 11073, pp. 755–763.

Zheng, Q., Delingette, H., Duchateau, N., Ayache, N., 2018. 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. IEEE Trans. Med. Imaging 37 (9), 2137–2148.

Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D., 2008. Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features. IEEE Trans. Med. Imaging 27 (11), 1668–1681.

Zhuang, X., 2013. Challenges and methodologies of fully automatic whole heart segmentation: a review. J. Healthc. Eng. 4 (3), 371–408.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., Yang, X., Heng, P.-A., Mortazi, A., Bagci, U., Yang, G., Sun, C., Galisot, G., Ramel, J.-Y., Brouard, T., Tong, Q., Si, W., Liao, X., Zeng, G., Shi, Z., Zheng, G., Wang, C., MacGillivray, T., Newby, D., Rhode, K., Ourselin, S., Mohiaddin, R., Keegan, J., Firmin, D., Yang, G., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med. Image Anal. 58, 101537.

Zhuang, X., Rhode, K., Razavi, R., Hawkes, D., Ourselin, S., 2010. A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. IEEE Trans. Med. Imaging 29 (9), 1612–1625.

Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med. Image Anal. 31, 77–87.