Evaluation of Algorithms for Multi-Modality Whole Heart Segmentation: An Open-Access Grand Challenge

Xiahai Zhuang[1,2,*], Lei Li[3,*], Christian Payer[4], Darko Štern[5], Martin Urschler[5], Mattias P. Heinrich[6], Julien Oster[7], Chunliang Wang[8], Örjan Smedby[8], Cheng Bian[9], Xin Yang[10], Pheng-Ann Heng[10], Aliasghar Mortazi[11], Ulas Bagci[11], Guanyu Yang[12], Chenchen Sun[12], Gaetan Galisot[13], Jean-Yves Ramel[13], Thierry Brouard[13], Qianqian Tong[14], Weixin Si[15], Xiangyun Liao[16], Guodong Zeng[17], Zenglin Shi[17], Guoyan Zheng[17], Chengjia Wang[18,19], Tom MacGillivray[19], David Newby[18,19], Kawal Rhode[20], Sebastien Ourselin[20], Raad Mohiaddin[21,22], Jennifer Keegan[21,22], David Firmin[21,22], Guang Yang[21,22,*]

[1] *School of Data Science, Fudan University, 200433, Shanghai, China*
[2] *Fudan-Xinzailing Joint Research Center for Big Data, Fudan University, 200433, Shanghai, China*
[3] *School of Biomedical Engineering, Shanghai Jiao Tong University, 200240, Shanghai, China*
[4] *Institute of Computer Graphics and Vision, Graz University of Technology, 8010, Graz, Austria*
[5] *Ludwig Boltzmann Institute for Clinical Forensic Imaging, 8010, Graz, Austria*
[6] *Institute of Medical Informatics, University of Lubeck, 23562, Lubeck, Germany*
[7] *Inserm, Université de Lorraine, U1254, IADI, Nancy, France*
[8] *School for Technology and Health, KTH Royal Institute of Technology, SE-10044, Stockholm, Sweden*
[9] *School of Biomed. Eng., Health Science Centre, Shenzhen University, 518060, Shenzhen, China*
[10] *Dept. of Comp. Sci. and Eng., The Chinese University of Hong Kong, Hong Kong, China*
[11] *Center for Research in Computer Vision (CRCV), University of Central Florida, 32816, Orlando, U.S.*
[12] *School of Computer Science and Engineering, Southeast University, 210096, Nanjing, China*
[13] *LIFAT (EA6300), Université de Tours, 64 avenue Jean Portalis, 37200, Tours, France*
[14] *School of Computer Science, Wuhan University, 430072, Wuhan, China*
[15] *Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, SIAT, Shenzhen, China*
[16] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 518055, Shenzhen, China*
[17] *Institute for Surgical Technology & Biomechanics, University of Bern, 3014, Bern, Switzerland*
[18] *BHF Centre for Cardiovascular Science, University of Edinburgh, Edinburgh, U.K.*
[19] *Edinburgh Imaging Facility QMRI, University of Edinburgh, Edinburgh, U.K.*
[20] *Department of Imaging Sciences & Biomedical Engineering, Kings College London, London, U.K.*
[21] *Cardiovascular Research Centre, Royal Brompton Hospital, SW3 6NP, London, U.K.*
[22] *National Heart and Lung Institute, Imperial College London, London, SW7 2AZ, London, U.K.*

Abstract

Knowledge of whole heart anatomy is a prerequisite for many clinical applications. Whole heart segmentation (WHS), which delineates substructures of the heart, can be very valuable for modeling and analysis of the anatomy and functions of the heart. However, automating this segmentation can be arduous due to the large variation of the heart shape, and different image qualities of the clinical data. To achieve this goal, a set of training data is generally needed for constructing priors or for training. In addition, it is difficult to perform comparisons between different methods, largely due to differences in the datasets and evaluation metrics used. This manuscript presents the methodologies and evaluation results for the WHS algorithms selected from the submissions to the Multi-Modality Whole Heart Segmentation (MM-WHS) challenge, in conjunction with MICCAI 2017. The challenge provides 120 three-dimensional cardiac images covering the whole heart, including 60 CT and 60 MRI volumes, all acquired in clinical environments with manual delineation. Ten algorithms for CT data and eleven algorithms for MRI data, submitted from twelve groups, have been evaluated. The results show that many of the deep learning (DL) based methods achieved high accuracy, even though the number of training datasets were limited. A number of them also reported poor results in the blinded evaluation, probably due to overfitting in their training. The conventional algorithms, mainly based on multi-atlas segmentation, demonstrated robust and stable performance, even though the accuracy is not as good as the best DL method in CT segmentation. The challenge, including provision of the annotated training data and the blinded evaluation for submitted algorithms on the test data, continues as an ongoing benchmarking resource via its homepage (`www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/`).

1. Introduction

According to the World Health Organization, cardiovascular diseases (CVDs) are the leading cause of death globally (Mendis et al., 2011). Medical imaging has revolutionized the modern medicine and healthcare, and the imaging and computing technologies become increasingly important for the diagnosis and treatments of CVDs. Computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), single photon emission computed tomography (SPECT), and ultrasound (US) have been used extensively for physiologic understanding and diagnostic purposes in cardiology (Kang et al., 2012). Among these, CT and MRI are particularly used to provide clear anatomical information of the heart. Cardiac MRI has the advantages of being free from ionizing radiation, acquiring images with great contrast between soft tissues and relatively high spatial resolutions (Nikolaou et al., 2011). On the other hand, cardiac CT is fast, low cost, and generally of high quality (Roberts et al., 2008).

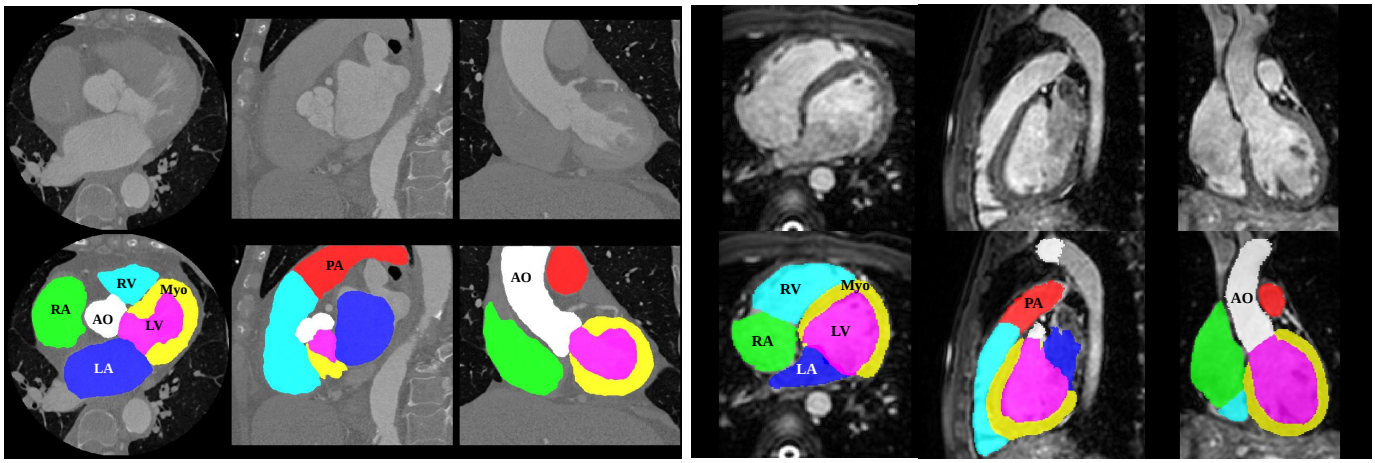
To quantify the morphological and pathological changes, it is commonly a prerequisite to segment the important structures from the cardiac medical images. Whole heart segmentation (WHS) aims to extract each of the individual whole heart substructures, including the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium of LV (Myo), ascending aorta (AO) or the whole aorta, and the pulmonary artery (PA) (Zhuang, 2013), as Fig. 1 shows. The applications of WHS are ample. The results can be used to directly compute the functional indices such as ejection fraction. Additionally, the geometrical information is useful in surgical guidance such as in radio-frequency ablation of the LA. However, the manual delineation of whole heart is labor-intensive and tedious, needing almost 8 hours for a single subject (Zhuang and Shen, 2016). Thus, automating the segmentation from multi-modality images, referred to as MM-WHS, is highly desired but still challenging, mainly due to the following reasons (Zhuang, 2013). First, the shape of the heart varies largely in different subjects or even for the same subject at different cardiac phases, especially for those with pathological and physiological changes. Second, the appearance and image quality can be variable. For example, the enhancement patterns of the CT images can vary significantly for different scanners or acquisition sessions. Also, motion artifacts, poor contrast-to-noise ratio and signal-to-noise ratio, commonly presented in the clinical data, can significantly deteriorate the image quality and consequently challenge the task.

1.1. State-of-the-art for Whole Heart Segmentation

In the last ten years, a variety of WHS techniques have been proposed for cardiac CT and MRI data. The detailed reviews of previously published algorithms can be found in Kang et al. (2012), Zhuang (2013) and Peng et al. (2016). Kang et al. (2012) reviewed several modalities and corresponding segmentation algorithms for the diagnosis and treatments of CVDs. They summarized the roles and characteristics of different modalities of cardiac imaging and the parameter correlation between them. In addition, they categorized the WHS techniques into four kinds, i.e., (1) boundary-driven techniques, (2) region-based techniques, (3) graph-cuts techniques, and (4) model fitting techniques. The advantages and disadvantages of each category were analyzed and summarized. Zhuang (2013) discussed the challenges and methodologies of the fully automatic WHS. Particularly, the work summarized two key techniques, i.e., the construction of prior models and the fitting procedure for segmentation propagation, for achieving this goal. Based on the types of prior models, the segmentation methods can be divided into two groups, namely the deformable model based methods and the atlas-based approaches; and the fitting procedure can be decomposed into three stages, including localizing the whole heart, initializing the substructures, and refining the boundary delineation. Thus, this review paper mainly analyzes the algorithms based on the classification of prior models and fitting algorithms for the WHS from different modality images. Peng et al. (2016) reviewed both the methodologies of WHS and the structural and functional indices of the heart for clinical assessments. In their work, the WHS approaches were classified into three categories, i.e., image-driven techniques, model-driven techniques, and direct estimation.

The three topic review papers mentioned above mainly cover the publications before 2015. A collection of recent works not included by them are summarized in Table 1. Among these works, (Zhuang et al., 2015) proposed an atlas ranking and selection scheme based on conditional entropy for the multi-atlas based WHS of CT. Zhou et al. (2017) developed a set of CT atlases labeled with 15 cardiac substructures. These atlases were then used for automatic WHS of CT via the multi-atlas segmentation (MAS) framework. Cai et al. (2017) developed a method with window width-level adjustment to pre-process CT data, which generates images with clear anatomical structures for WHS. They applied a Gaussian filter-based multi-resolution scheme to eliminate the discontinuity in the down-sampling decomposition for whole heart image registration. Zuluaga et al. (2013) developed a MAS scheme for both CT and MRI WHS. The proposed method ranked and selected optimal atlases based on locally normalised cross correlation. Pace et al. (2015) proposed a patch-based interactive algorithm to extract the heart based on a manual initialization from experts. The method employs active learning to identify the areas that require user interaction. Zhuang and Shen (2016) developed a multi-modality MAS framework for WHS of cardiac MRI, which used a set of atlases built from both CT and MRI. The authors proposed modality invariant metrics for computing the global image similarity and the local

URL: zxh@fudan.edu.cn (Xiahai Zhuang[1,2,*]), lilei.sky@sjtu.edu.cn (Lei Li[3,*]), g.yang@imperial.ac.uk (Guang Yang[21,22,*])



(a) (b)

Figure 1: Examples of cardiac images and WHS results: (a) displays the three orthogonal views of a cardiac CT image and its corresponding WHS result, (b) is from a cardiac MRI image and its WHS. LV: left ventricle; RV: right ventricle; LA: left atrium; RA: right atrium; Myo: myocardium of LV; AO: ascending aorta; PA: pulmonary artery.

similarity. The global image similarity was used to rank and select atlases, from the multi-modality atlas pool, for segmenting a target image; and the local similarity metrics were proposed for the patch-based label fusion, where a multi-scale patch strategy was developed to obtain a promising performance.

In conclusion, WHS based on the MAS framework, referred to as MA-WHS, has been well researched in recent years. MAS segments an unknown target image by propagating and fusing the labels from multiple annotated atlases using registration. The performance relies on the registration algorithms for label propagation and the fusion strategy to combine the segmentation results from the multiple atlases. Both these two key steps are generally computationally expensive.

Recently, a number of deep learning (DL)-based methods have shown great promise in medical image analysis. They have obtained superior performance in various imaging modalities and different clinical applications (Roth et al., 2014; Shen et al., 2017). For cardiac segmentation, Avendi et al. (2016) proposed a DL algorithm for LV segmentation. Ngo et al. (2017) trained multiple layers of deep belief network to localize the LV, and to define the endocardial and epicardial borders, followed by the distance regularised level set. Recently, Tan et al. (2018) designed a fully automated convolutional neural network (CNN) architecture for pixel-wise labeling of both the LV and RV with impressive performance. DL methods have potential of providing faster and more accurate segmentation, compared to the conventional approaches, such as the deformable model based segmentation and MAS method. However, little work has been reported to date using DL for WHS, probability due to the limitation of training data and complexity of the segmentation task.

Table 2 summarizes the recent open access datasets for cardiac segmentation, which mainly focus on specific substructures of the heart. Radau et al. (2008); Suinesiaputra et al. (2011); Petitjean et al. (2015); Bernard et al. (2018) organized the challenges for segmenting the left, right or full ventricles. Moghari et al. (2016) organized a challenge for the segmentation of blood pool and myocardium from 3D MRI data. This work aims to offer pre-procedural planning of children with complex congenital heart disease. Tobon-Gomez et al. (2015); Karim et al. (2018) and Zhao and Xiong (2018) provided data for benchmarking algorithms of LA or LA wall segmentation for patients suffering from atrial fibrillation.

1.2. Motivation and Contribution

Due to the above mentioned challenges, we organized the competition of MM-WHS, providing 120 multi-modality whole heart images for developing new WHS algorithms, as well as validating existing ones. We also presented a fair evaluation and comparison framework for participants. In total, twelve groups who submitted their results and methods were selected, and they all agreed to contribute to this work, a benchmark for WHS of two modalities, i.e., CT and MRI. In this work, we introduce the related information, elaborate on the methodologies of these selective submissions, discuss the results and provide insights to the future research.

The rest of this paper is organised as follows. Section 2 provides details of the materials and evaluation framework. Section 3 introduces the evaluated methods for benchmarking. Section 4 presents the results, followed by discussions in Section 5. We conclude this work in Section 6.

Table 1: Summary of previous WHS methods for multi-modality images. Here, the abbreviations are as follows, PIS: patch-based interactive segmentation; FIMH: International Conference on Functional Imaging and Modeling of the Heart; MICCAI: International Conference on Medical Image Computing and Computer-assisted Intervention; MedPhys: Medical Physics; MedIA: Medical Image Analysis; RadiotherOncol: Radiotherapy and Oncology.

Reference	Data	Method	Runtime	Dice
Zuluaga et al. (2013), FIMH	8 CT, 23 MRI	MAS	60 min, 30 min	0.89 ± 0.04, 0.91 ± 0.03
Zhuang et al. (2015), MedPhys	30 CT	MAS	13.2 min	0.92 ± 0.02
Pace et al. (2015), MICCAI	4 MRI	PIS + Active learning	60 min	N/A
Zhuang and Shen (2016), MedIA	20 CT + 20 MRI	Multi-modality MAS	12.58 min	0.90 ± 0.03
Zhou et al. (2017), RadiotherOncol	31 CT	MAS	10 min	0.77 ± 0.07
Cai et al. (2017), Neurocomputing	14 CT	Gaussian filter-based	N/A	0.77 ± 0.07

Table 2: Summary of the previous challenges related to cardiac segmentation from MICCAI society.

Organizers/refernece	Year	Data	Target	Pathology
Radau et al. (2008)	2009	45 cine MRI	LV	hypertrophy, infarction
Suinesiaputra et al. (2011)	2011	200 cine MRI	LV	myocardial infarction
Petitjean et al. (2015)	2012	48 cine MRI	RV	congenital heart disease
Tobon-Gomez et al. (2015)	2013	30 CT + 30 MRI	LA	atrial fibrillation
Karim et al. (2018)	2016	10 CT + 10 MRI	LA wall	atrial fibrillation
Moghari et al. (2016)	2016	20 MRI	Blood pool, Myo	congenital heart disease
Bernard et al. (2018)	2017	150 cine MRI	Ventricles	infarction, dilated/ hypertrophic cardiomyopathy, abnormal RV
Zhao and Xiong (2018)	2018	150 LGE-MRI	LA	atrial fibrillation

2. Materials and setup

2.1. Data Acquisition

The cardiac CT/CTA data were acquired from two state-of-the-art 64-slice CT scanners (Philips Medical Systems, Netherlands) using a standard coronary CT angiography protocol at two sites in Shanghai, China. All the data cover the whole heart from the upper abdomen to the aortic arch. The in-plane resolution of the axial slices is 0.78×0.78 mm, and the average slice thickness is 1.60 mm.

The cardiac MRI data were obtained from two hospitals in London, UK. One set of data were acquired from St. Thomas Hospital on a 1.5T Philips scanner (Philips Healthcare, Best, The Netherlands), and the other were from Royal Brompton Hospital on a Siemens Magnetom Avanto 1.5T scanner (Siemens Medical Systems, Erlangen, Germany). In both sites we used the 3D balanced steady state free precession (b-SSFP) sequence for whole heart imaging, and realized free breathing scans by enabling a navigator beam before data acquisition for each cardiac phase. The data were acquired at a resolution of around $(1.6 \sim 2) \times (1.6 \sim 2) \times (2 \sim 3.2)$ mm, and reconstructed to half of its acquisition resolution, i.e., about $(0.8 \sim 1) \times (0.8 \sim 1) \times (1 \sim 1.6)$ mm.

Both cardiac CT and cardiac MRI data were acquired in real clinical environment. The pathologies of patients cover a wide range of cardiac diseases, including myocardium infarction, atrial fibrillation, tricuspid regurgitation, aortic valve stenosis, Alagille syndrome, Williams syndrome, dilated cardiomyopathy, aortic coarctation, Tetralogy of Fallot. The subjects for MRI scans also include a small number of normal controls.

All the CT and MRI data have been anonymized in agreement with the local regional ethics committee before being released to the MM-WHS challenge. In total, we provided 120 multi-modality whole heart images from multiple sites, including 60 cardiac CT and 60 cardiac MRI. Note that the data were collected from clinical environments, so the image quality was variable. This enables to assess the validation and robustness of the developed algorithms with representative clinical data, rather than with selected best quality images.

2.2. Definition and Gold Standard

The WHS studied in this work aims to delineate and extract the seven substructures of the heart, into separate individuals (Zhuang, 2013). These seven structures include the following,

(1) the LV blood cavity, also referred to as LV;
(2) the RV blood cavity, also referred to as RV;
(3) the LA blood cavity, also referred to as LA;
(4) the RA blood cavity, also referred to as RA;
(5) the myocardium of the LV (Myo) and the epicardium (Epi), defined as the epicardial surface of the LV;
(6) the AO trunk from the aortic valve to the superior level of the atria, also referred to as AO;
(7) the PA trunk from the pulmonary valve to the bifurcation point, also referred to as PA.

The four blood pool cavities, i.e., LV, RV, LA and RA, are also referred to as the four chambers.

Manual labeling was adopted for generating the gold standard segmentation. They were done by clinicians or by students majoring in biomedical engineering or medical

Table 3: Summary of submitted methods. Asterisk (*) indicates the results that were submitted after the challenge deadline.

Teams	Tasks	Key elements in methods	Teams	Tasks	Key elements in methods
GUT	CT, MRI	Two-step CNN, combined with anatomical label configurations.	UOL	MRI	MAS and discrete registration, to adapt the large shape variations.
KTH	CT, MRI	Multi-view U-Nets combining hierarchical shape prior.	CUHK1	CT, MRI	3D FCN with the gradient flow optimization and Dice loss function.
SEU	CT	Conventional MAS-based method.	CUHK2	CT, MRI	Hybrid loss guided FCN.
UCF	CT, MRI	Multi-object multi-planar CNN with an adaptive fusion method.	UT	CT, MRI	Local probabilistic atlases coupled with a topological graph.
SIAT	CT, MRI	3D U-Net learning learn multi-modality features.	UB2*	MRI	Multi-scale fully convolutional DenseNets.
UB1*	CT, MRI	Dilated residual networks.	UOE*	CT, MRI	Two-stage concatenated U-Net.

physicists who were familiar with the whole heart anatomy, slice-by-slice using the ITK-SNAP software (Yushkevich et al., 2006). Each manual segmentation result was examined by a senior researchers specialized in cardiac imaging with experience of more than five years, and modifications have been take if revision was necessary. Also, the sagittal and coronal views were visualised simultaneously to check the consistency and smoothness of the segmentation, although the manual delineation was mainly performed in the axial views. For each image, it takes approximately 6 to 10 hours for the observer to complete the manual segmentation of the whole heart.

2.3. Evaluation Metrics

We employed four widely used metrics to evaluate the accuracy of a segmentation result, including the Dice score (Kittler et al., 1998), Jaccard index (Jaccard, 1901), surface-to-surface distance (SD), and Hausdorff Distance (HD). For WHS evaluation, we adopted the generalized version of them, the normalized metrics with respect to the size of substructures. They are expected to provide more objective measurements (Crum et al., 2006; Zhuang, 2013).

For each modality, the data were split into two sets, i.e., the training set (20 CT and 20 MRI) and the test set (40 CT and 40 MRI). For the training data, both the images and the corresponding gold standard were released to the participants for building, training and cross-validating their models. For the test data, only the CT and MRI images were released. Once the participants developed their algorithms, they could submit their segmentation results on the test data to the challenge moderators for a final independent evaluation. To make a fair comparison, the challenge organizers only allowed maximum of two evaluations for one algorithm.

2.4. Participants

Twelve algorithms (teams) were selected for this benchmark work. Nine of them provided results for both CT and MRI data, one experimented only on the CT data and two worked solely on the MRI data.

All of the 12 teams agreed to include their results in this paper. To simplify the description below, we used the team abbreviations referring to both the teams and their corresponding methods and results. The evaluated methods are elaborated on in Section 3, and the key contributions of the teams are summarized in Table 3. Note that the three methods, indicated with Asterisk (*), were submitted after the challenge deadline for performance ranking.

3. Evaluated Methods

In this section, we elaborate on the twelve benchmarked algorithms. Table 3 provides the summary for reference.

3.1. Graz University of Technology (GUT)

Payer et al. (2017) proposed a fully automatic whole heart segmentation, based on multi-label CNN and using volumetric kernels, which consists of two separate CNNs: one to localize the heart, referred to as localization CNN, and the other to segment the fine detail of the whole heart structure within a small region of interest (ROI), referred to as segmentation CNN. The localization CNN is designed to predict the approximate centre of the bounding box around all heart substructures, based on the U-Net (Ronneberger et al., 2015) and heatmap regression (Payer et al., 2016). A fixed physical size ROI is then cropped around the predicted center, ensuring that it can enclose all interested substructures of the heart. Within the cropped ROI, the multi-label segmentation CNN predicts the label of each pixel. In this method, the segmentation CNN works on high-resolution ROI, while the localization CNN works on the low resolution images. This two-step CNN pipeline helps to mitigate the intensive memory and runtime generally required by the volumetric kernels equipped 3D CNNs.

3.2. University of Lubeck (UOL)

Heinrich and Oster (2017) proposed a multi-atlas registration approach for WHS of MRI, as Fig. 2 shows. This method adopts a discrete registration, which can capture large shape variations across different scans (Heinrich et al., 2013a). Moreover, it can ensure the alignment of anatomical structures by using dense displacement sampling and graphical model-based optimization (Heinrich et al., 2013b). Due to the use of contrast-invariant features (Xu et al., 2016), the multi-atlas registration can implicitly deal with the challenging varying intensity distributions due to different acquisition protocols. Within this method, one can

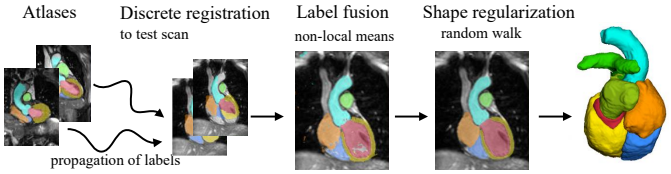


Figure 2: Multi-atlas registration and label fusion with regularization proposed by Heinrich and Oster (2017).

register all the training atlases to an unseen test image. The warped atlas label images are then combined by means of weighted label fusion. Finally, an edge-preserving smoothing of the generated probability maps is performed using the multi-label random walk algorithm, as implemented and parameterized in Heinrich and Blendowski (2016).

3.3. KTH Royal Institute of Technology (KTH)

Wang and Smedby (2017) propose an automatic WHS framework combined CNN with statistical shape priors. The additional shape information, also called shape context (Mahbod et al., 2018), is used to provide explicit 3D shape knowledge to the CNN. The method uses a random forest based landmark detection to detect the ROI. The statistical shape models are created using the segmentation masks of the 20 training CT images. The probability map is generated from three 2D U-Nets learned from the multi-view slices of the 3D training images. To estimate the shape of each subregion of heart, a hierarchical shape prior guided segmentation algorithm (Wang and Smedby, 2014) is then performed on the probability map. This shape information is represented using volumetric shape models, i.e., signed distance maps of the corresponding shapes. Finally, the estimated shape information is used as an extra channel, to train a new set of multi-view U-Nets for the final segmentation of whole heart.

3.4. The Chinese University of Hong Kong, Method No. 1 (CUHK1)

Yang et al. (2017b) apply a general and fully automatic framework based on a 3D fully convolutional network (FCN). The framework is reinforced in the following aspects: First, an initialization is achieved by inheriting the knowledge from a 3D convolutional networks trained on the large-scale Sports-1M video dataset (Tran et al., 2015). Then, the gradient flow is applied by shortening the back-propagation path and employing several auxiliary loss functions on the shallow layers of the network. This is to tackle the low efficiency and over-fitting issues when directly train the deep 3D FCNs, due to the gradient vanishing problem in shallow layers. Finally, the Dice similarity coefficient based loss function (Milletari et al., 2016) is included into a multi-class variant to balance the training for all classes.

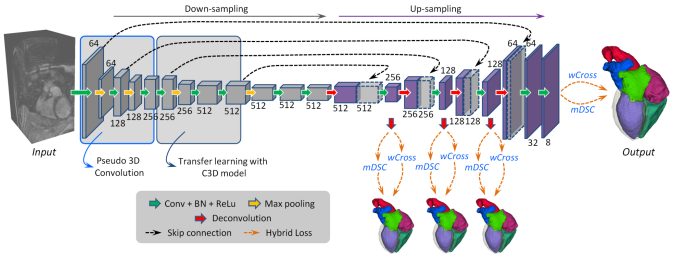


Figure 3: A schematic illustration of the method developed by Yang et al. (2017c). Digits represent the number of feature volumes in each layer. Volume with dotted line is for concatenation.

3.5. University of Central Florida (UCF)

Mortazi et al. (2017a) propose a multi-object multi-planar CNN (MO-MP-CNN) method based on an encoder-decoder CNN. The multiple CNNs (Mortazi et al., 2017b) are trained from three different views, i.e., axial, sagittal, and coronal views, in 2D manners. An adaptive fusion method is then employed to combine the multiple outputs to refine the delineation. Furthermore, they apply the connected component analysis (CCA) on the final segmentation, to estimate the reliable (true positive) and unreliable (false positives) regions. Let n denotes the number of classes in the images and m denotes the number of components in each class, then the CCA could be performed as follows,

$$\begin{aligned} CCA(S) = \{S_{11}, \cdots, S_{nm} | \cup S_{ij} = \mathbf{o}\} \& \\ \{S_{11}, \cdots, S_{nm} | \cap S_{ij} = \phi\}, \end{aligned} \tag{1}$$

where S indicates the segmentation result. The differences between the reliable and unreliable regions are used to guide the reliability of the segmentation process, namely the higher the difference, the more reliable the segmentation.

3.6. The Chinese University of Hong Kong, Method No. 2 (CUHK2)

Yang et al. (2017c) employ a 3D FCN for an end-to-end dense labeling, as Fig. 3 shows. The proposed network is coupled with several auxiliary loss functions in a deep supervision mechanism, to tackle the potential gradient vanishing problem and class imbalance in training. The network learns a spatial-temporal knowledge from a large-scale video dataset, and then transfer to initialize the shallow convolutional layers in the down-sampling path (Tran et al., 2015). For the class imbalance issue, a hybrid loss is proposed (Milletari et al., 2016), combining two complementary components: (1) volume-size weighted cross entropy loss (*wCross*) to preserve branchy details such as PA trunks. (2) multi-class Dice similarity coefficient loss (*mDSC*) to compact anatomy segmentation. Then, the proposed network can be well trained to simultaneously segment different classes of heart substructures, and generate a segmentation in a dense but detail-preserved format.

Table 4: Results of the ten evaluated algorithms on CT dataset. SD: surface-to-surface distance; HD: Hausdorff Distance; DL: deep learning-based method; MAS: conventional method based on multi-atlas segmentation. Asterisk (*) indicates the results were submitted after the challenge deadline.

Teams	Dice	Jaccard	SD (mm)	HD (mm)	DL/MAS
GUT	$\mathbf{0.908 \pm 0.086}$	$\mathbf{0.832 \pm 0.037}$	$\mathbf{1.117 \pm 0.250}$	$\mathbf{25.242 \pm 10.813}$	DL
KTH	0.894 ± 0.030	0.810 ± 0.048	1.387 ± 0.516	31.146 ± 13.203	DL
CUHK1	0.890 ± 0.049	0.805 ± 0.074	1.432 ± 0.590	29.006 ± 15.804	DL
CUHK2	0.886 ± 0.047	0.798 ± 0.072	1.681 ± 0.593	41.974 ± 16.287	DL
UCF	0.879 ± 0.079	0.792 ± 0.106	1.538 ± 1.006	28.481 ± 11.434	DL
SEU	0.879 ± 0.023	0.784 ± 0.036	1.705 ± 0.399	34.129 ± 12.528	MAS
SIAT	0.849 ± 0.061	0.742 ± 0.086	1.925 ± 0.924	44.880 ± 16.084	DL
UT	0.838 ± 0.152	0.742 ± 0.161	4.812 ± 13.604	34.634 ± 12.351	MAS
UB1*	0.887 ± 0.030	0.798 ± 0.048	1.443 ± 0.302	55.426 ± 10.924	DL
UOE*	0.806 ± 0.159	0.697 ± 0.166	4.197 ± 7.780	51.922 ± 17.482	DL
Average	0.859 ± 0.108	0.763 ± 0.118	3.259 ± 9.748	34.382 ± 12.468	MAS
	0.875 ± 0.083	0.784 ± 0.010	1.840 ± 2.963	38.510 ± 17.890	DL
	0.872 ± 0.087	0.780 ± 0.102	2.124 ± 5.133	37.684 ± 17.026	ALL

Table 5: Results of the eleven evaluated algorithms on MRI dataset. SD: surface-to-surface distance; HD: Hausdorff Distance; DL: deep learning-based method; MAS: conventional method based on multi-atlas segmentation. Asterisk (*) indicates the results were submitted after the challenge deadline.

Teams	Dice	Jaccard	SD (mm)	HD (mm)	DL/MAS
UOL	0.870 ± 0.035	0.772 ± 0.054	1.700 ± 0.649	$\mathbf{28.535 \pm 13.220}$	MAS
GUT	0.863 ± 0.043	0.762 ± 0.064	1.890 ± 0.781	30.227 ± 14.046	DL
KTH	0.855 ± 0.069	0.753 ± 0.094	1.963 ± 1.012	30.201 ± 13.216	DL
UCF	0.818 ± 0.096	0.701 ± 0.118	3.040 ± 3.097	40.092 ± 21.119	DL
UT	0.817 ± 0.059	0.695 ± 0.081	2.420 ± 0.925	30.938 ± 12.190	MAS
CUHK2	0.810 ± 0.071	0.687 ± 0.091	2.385 ± 0.944	33.101 ± 13.804	DL
CUHK1	0.783 ± 0.097	0.653 ± 0.117	3.233 ± 1.783	44.837 ± 15.658	DL
SIAT	0.674 ± 0.182	0.532 ± 0.178	9.776 ± 6.366	92.889 ± 18.001	DL
UB2*	$\mathbf{0.874 \pm 0.039}$	$\mathbf{0.778 \pm 0.060}$	$\mathbf{1.631 \pm 0.580}$	28.995 ± 13.030	DL
UB1*	0.869 ± 0.058	0.773 ± 0.079	1.757 ± 0.814	30.018 ± 14.156	DL
UOE*	0.832 ± 0.081	0.720 ± 0.105	2.472 ± 1.892	41.465 ± 16.758	DL
Average	0.844 ± 0.047	0.734 ± 0.072	2.060 ± 0.876	29.737 ± 12.771	MAS
	0.820 ± 0.107	0.707 ± 0.127	3.127 ± 3.640	41.314 ± 24.711	DL
	0.824 ± 0.102	0.711 ± 0.125	2.933 ± 3.339	39.209 ± 23.435	ALL

3.7. Southeast University (SEU)

Yang et al. (2017a) develop a MAS-based method for WHS of CT images. The proposed method consists of the following major steps. Firstly, a ROI detection is performed on atlas images and label images, which are downsampled and resized to crop and generate a heart mask. Then, an affine registration is used to globally align the target image with the atlas images, followed by a non-rigid registration to refine alignment of local details. In addition, an atlas ranking step is applied by using mutual information as the similarity criterion, and those atlases with low similarity are discarded. A non-rigid registration is further performed by minimizing the dissimilarity within the heart substructures using the adaptive stochastic gradient descent method. Finally, the propagated labels are fused with different weights according to the similarities between the deformed atlases and the target image.

3.8. University of Tours (UT)

Galisot et al. (2017) propose an incremental and interactive WHS method, combining several local probabilistic atlases based on a topological graph. The training images are used to construct the probabilistic atlases, for each of the substructures of the heart. The graph is used to encode the priori knowledge to incrementally extract different ROIs. The priori knowledge about the shape and intensity distributions of substructures is stored as features to the nodes of the graph. The spatial relationships between these anatomical structures are also learned and stored as the profile of edges of the graph. In the case of multi-modality data, multiple graphs are constructed, for example two graphs are built for the CT and MRI images, respectively. A pixelwise classification method combining hidden Markov random field is developed to integrate the probability map information. To correct the misclassifications, a post-correction is performed based on the Adaboost scheme.

3.9. Shenzhen Institutes of Advanced Technology (SIAT)

Tong et al. (2017) develop a deeply-supervised end-to-end 3D U-Net for fully automatic WHS. The training dataset are artificially augmented by considering each ROI of the heart substructure independently. To reduce false positives from the surrounding tissues, a 3D U-Net is firstly trained to coarsely detect and segment the whole heart structure. To take full advantage of multi-modality information so that features of different substructures could be better extracted, the cardiac CT and MRI data are

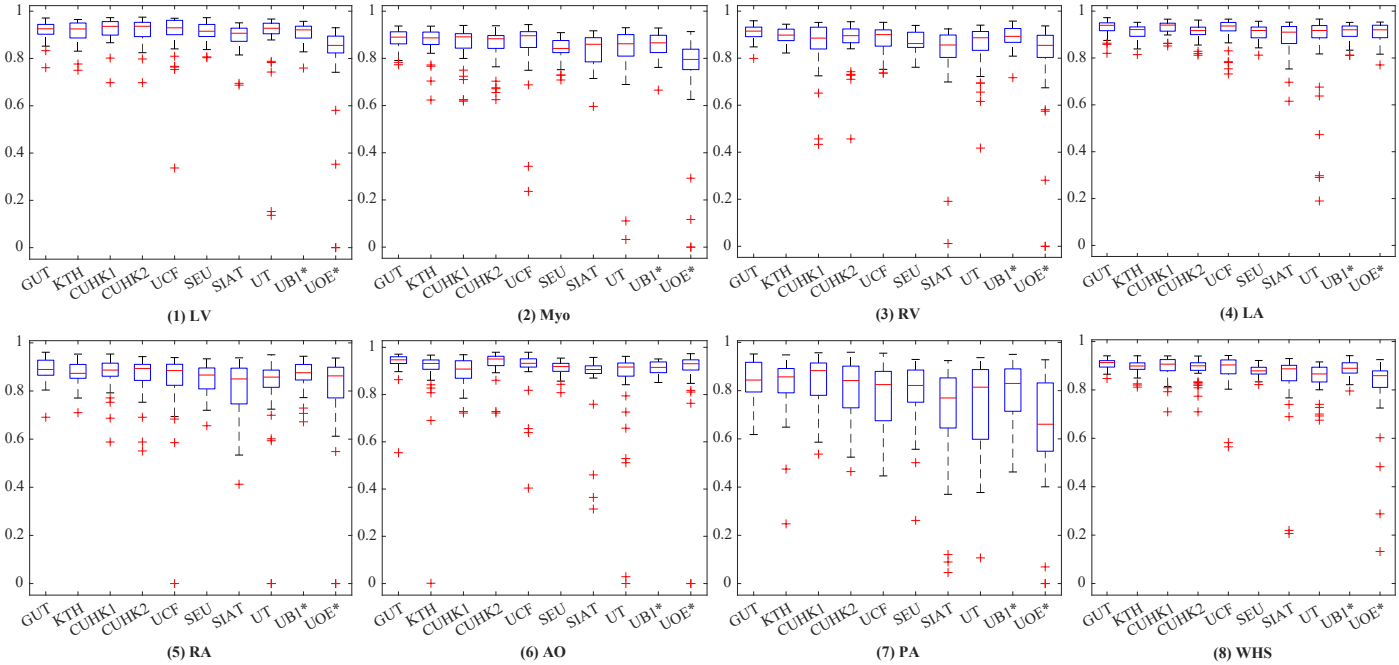


Figure 4: Boxplot of Dice scores of the whole heart segmentation on CT dataset by the ten methods.

fused. Both the size and the intensity range of the different modality images are normalized before training the 3D U-Net model. Finally, the detected ROI is refined to achieve the final WHS, which is performed by a pixel-wise classification fashion using the 3D U-Net.

3.10. University of Bern, Method No. 1 (UB1)*

Shi et al. (2018) design a pixel-wise dilated residual networks, referred to as Bayesian VoxDRN, to segment the whole heart structures from 3D MRI images. It can be used to generate a semantic segmentation of an arbitrary-sized volume of data after training. Conventional FCN methods integrate multi-scale contextual information by reducing the spatial resolution via successive pooling and sub-sampling layers, for semantic segmentation. By contrast, the proposed method achieves the same goal using dilated convolution kernels, without decreasing the spatial resolution of the network output. Additionally, residual learning is incorporated as pixel-wise dilated residual modules to alleviate the degrading problem, and the WHS accuracy can be further improved by avoiding gridding artifacts introduced by the dilation (Yu et al., 2017).

3.11. University of Bern, Method No. 2 (UB2)*

This method includes a multi-scale pixel-wise fully convolutional Dense-Nets (MSVoxFCDN) for 3D WHS of MRI images, which could directly map a whole volume of data to its volume-wise labels after training. The multi-scale context and multi-scale deep supervision strategies are adopted to enhance feature learning. The deep neural network is an encoder (contracting path)-decoder (expansive path) architecture. The encoder is focused on feature learning, while the decoder is used to generate the segmentation results. Skip connection is employed to recover spatial context loss in the down-sampling path. To further boost feature learning in the contracting path, multi-scale contextual information is incorporated. Two down-scaled branch classifiers are inserted into the network to alleviate the potential gradient vanishing problem. Thus, more efficient gradients can be back-propagated from loss function to the shallow layers.

3.12. University of Edinburgh (UOE)*

Wang and Smedby (2017) develop a two-stage concatenated U-Net framework that simultaneously learns to detect a ROI of the heart and classifies pixels into different substructures without losing the original resolution. The first U-Net uses a down-sampled 3D volume to produce a coarse prediction of the pixel labels, which is then resampled to the original resolution. The architecture of the second U-Net is inspired by the SRCNN (Dong et al., 2016) with skipping connections and recursive units (Kim et al., 2016). It inputs a two-channel 4D volume, consisting of the output of the first U-Net and the original data. In the test phase, a dynamic-tile layer is introduced between the two U-Nets to crop a ROI from both the input and output volume of the first U-Net. This layer is removed when performing end-to-end training to simplify the implementation. Unlike the other U-Net based architecture, the proposed method can directly perform prediction on the images with their original resolution, thanks to the SRCNN-like network architecture.

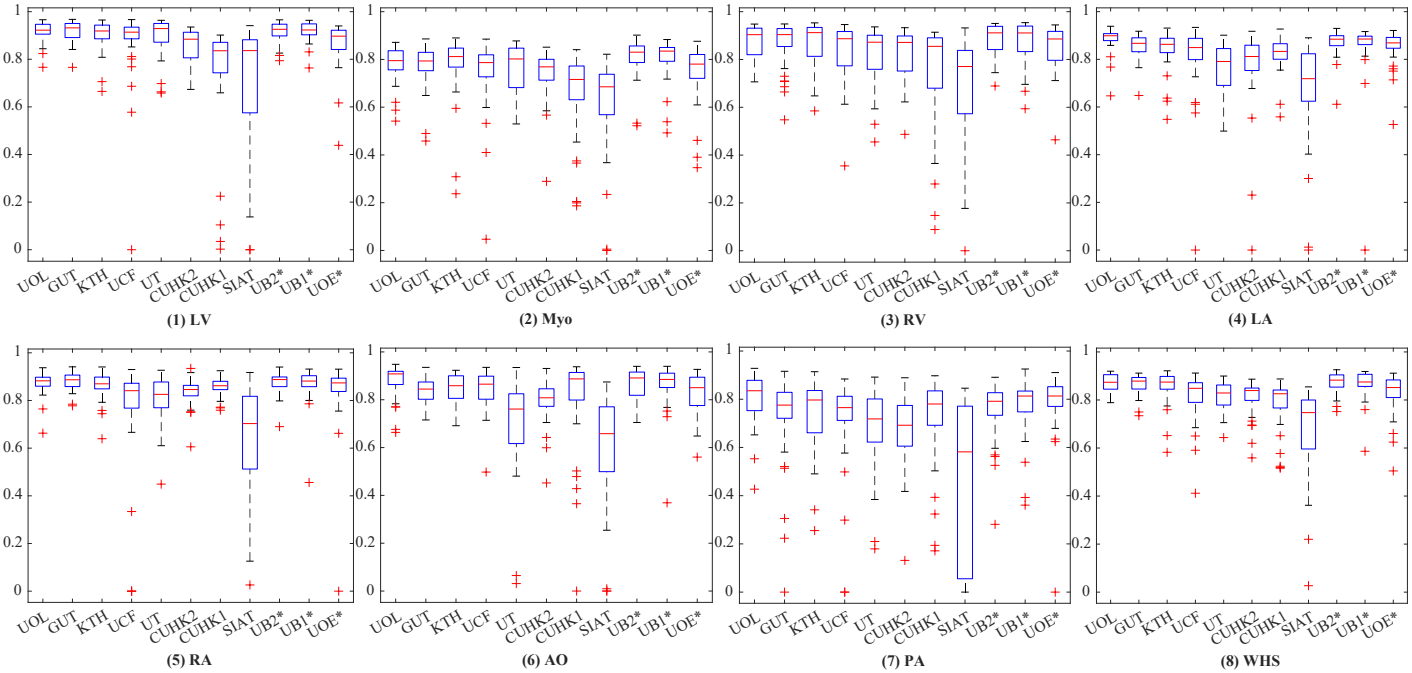


Figure 5: Boxplot of Dice scores of the whole heart segmentation on MRI dataset by the eleven methods.

4. Results

Table 4 and Table 5 present the quantitative results of the evaluated algorithms on CT and MRI dataset, respectively.

For the CT data, the results are generally promising, and the best Dice score (0.91 ± 0.09) was achieved by GUT, which is a DL-based algorithm with anatomical label configurations. The DL-based methods generally obtained better accuracies than the MAS-based approaches in terms of Jaccard, Dice, and SD metrics, though this conclusion was not applied when the HD metric is used. Particularly, one can find that the mean of HD from the two MAS methods was not worse than that of the other eight DL-based approaches.

For MRI data, the best Dice score of the WHS (0.87 ± 0.04) was obtained by UB2*, which is a DL-based method and a delayed submission; and the best HD (28.535 ± 13.220 mm) was achieved by UOL, a MAS-based algorithm. Here, the average accuracy of MAS (two teams) was better than that of the DL-based segmentation (nine teams) in all evaluation metrics. However, the performance across different DL methods could vary a lot, similar to the results from the CT experiment. For example, the top four DL methods, i.e., GUT, KTH, UB1* and UB2*, obtained comparable accuracy to that of UOL, but the other DL approaches could generate much poorer results.

Fig. 4 shows the boxplots of the evaluated algorithms on CT data. One can see that they achieved relatively accurate segmentation for all substructures of the heart, except for the PA whose variability in terms of shape and appearance is notably greater. For GUT, KTH, CUHK1, UB1*, and CUHK2, the delineation of PA is reasonably good with the mean Dice score larger than 0.8. Fig. 5 presents the boxplots on the MRI data. The five methods, i.e., UB2*, UOL, UB1*, GUT, and KTH, all demonstrate good Dice scores on the segmentation of four chambers and LV myocardium. Similar to the conclusion drawn from Table 4 and Table 5, the segmentation on the CT images is generally better than that on the MRI data as indicated by the quantitative evaluation metrics.

Fig. 6 shows the 3D visualization of the cases with the median and worst WHS Dice scores by the evaluated methods on the CT data. Most of the median cases look reasonablely good, though some contain patchy noise; and the worst cases require significant improvements. Specifically, UOE* median case contains significant amount of misclassification in AO, and parts of the LV are labeled as LA in the UOE* and SIAT median cases. In the worst cases, the CUHK1 and CUHK2 results do not have a complete shape of the RV; KTH and SIAT contain a large amount of misclassification, particularly in myocardium; UCF mistakes the RA as LV; UOE* only segments the LA, and UT generates a result with wrong orientation.

Fig. 7 visualizes the median and worst results on MRI WHS. Compared with the CT results, even the median cases of MRI cases are poor. For example, the SIAT method could perform well on most of the CT cases, but failed to generate acceptable results for most of the MRI images, including the median case presented in the figure. The worst cases of UOE*, CUHK2 and UB1 miss at least one substructure, and UCF and SIAT results do not contain any complete substructure of the whole heart. In conclusion, the CT segmentation results look better than the

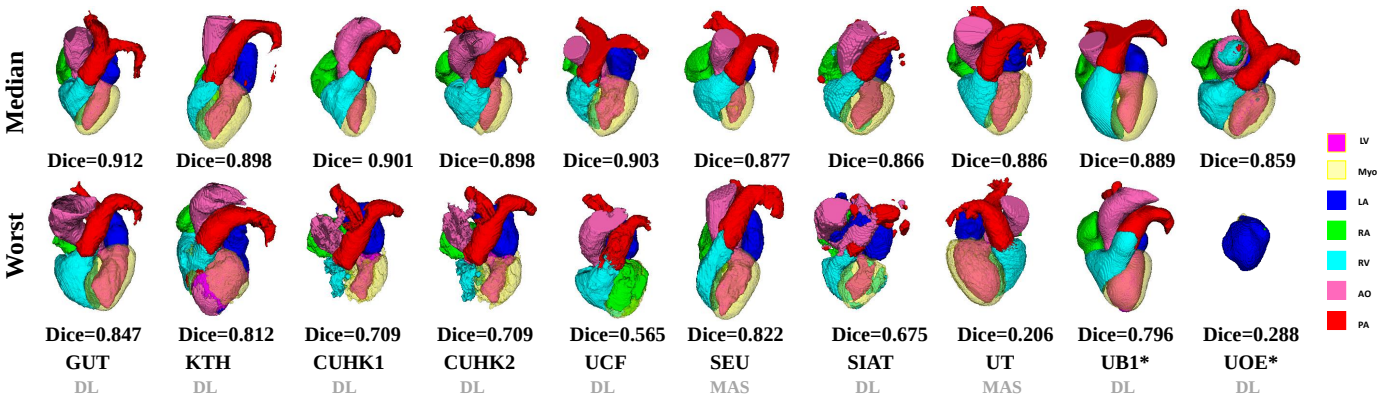


Figure 6: 3D visualization of the WHS results of the median and worse cases in the CT test dataset by the ten evaluated methods. The color bar indicates the correspondence of substructures. Note that the colors of Myo and LV in 3D visualization do not look exactly the same as the keys in the color bar, due to the 50% transparency setting for Myo rendering and the addition effect from two colors (LV and 50% Myo) for LV rendering, respectively.

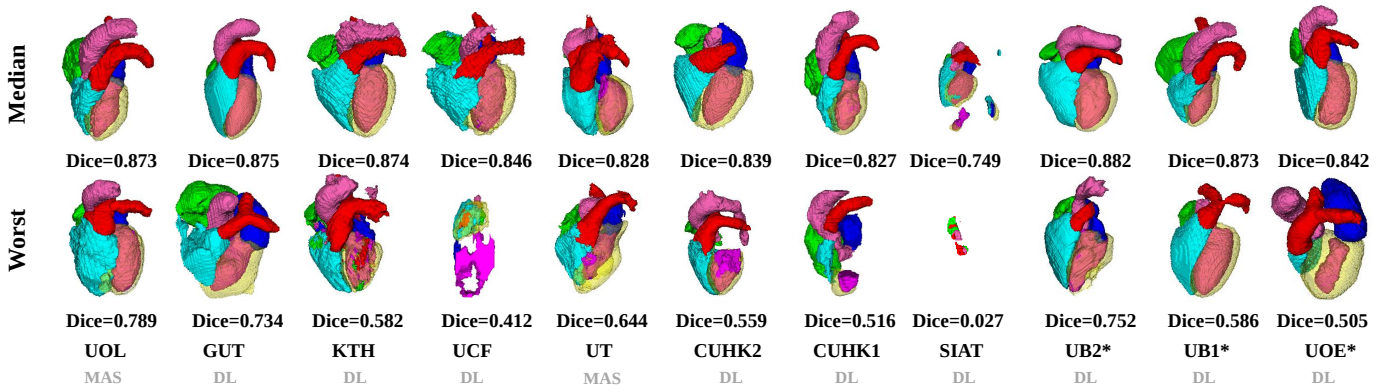


Figure 7: 3D visualization of the WHS results of the median and worse cases in the MRI test dataset by the eleven evaluated methods.

MRI results, which is consistent with the quantitative results. Also, one can conclude from Fig. 6 and Fig. 7 that the resulting shape from the MAS-based methods looks more realistic, compared to the DL-based algorithms, even though the segmentation could sometimes be very poor or even a failure, such as the worst MRI case by UOL and the worst CT case by UT.

5. Discussion

5.1. Overall performance of the evaluated algorithms

The mean Dice scores of the evaluated methods for MM-WHS are respectively 0.872 ± 0.087 (CT) and 0.824 ± 0.102 (MRI), and the best average Dices from one team are respectively 0.908 ± 0.086 (CT by GUT) and 0.874 ± 0.039 (MRI by UB2*). Table 4 and Table 5 provide the average numbers of the other evaluation metrics, for the different methodological categories and different imaging modalities. In general, the benchmarked algorithms obtain better WHS accuracies for CT than for MRI, using the four metrics. In addition, the mean Dice scores of MAS-based methods are 0.859 ± 0.108 (CT) and 0.844 ± 0.047 (MRI), and those of DL-based methods are 0.875 ± 0.083 (CT) and 0.820 ± 0.107 (MRI). DL-based WHS methods obtain better mean accuracies, but the MAS-based approaches tend to generate results with more realistic heart shapes.

Furthermore, the segmentation accuracies reported for the four chambers are generally good, but the segmentation of the other substructures demonstrates more challenges. For example, one can see from Fig. 4 and Fig. 5 that in CT WHS the PA segmentation is much poorer compared to other substructures; in MRI WHS, the segmentation of myocardium, AO and PA appears to be more difficult. One reason could be that these regions have much larger variation in terms of shapes and image appearance across different scans. Particularly, the diverse pathologies can result in heterogeneous intensity of the myocardium and blood fluctuations to the great vessels. The other reason could be the large variation of manual delineation of boundaries for these regions, which results in more ambiguity for the training of learning-based algorithms and the generation of the gold standard.

5.2. MAS versus DL-based segmentation

As Table 4 and Table 5 summarize, 9 out of the 11 benchmarked CT WHS methods and 8 out of the 10 MRI

Table 6: Summary of the advantages and limitations of the 12 benchmarked methods.

Method	Strengths	Limitations
GUT	- Combining localization and segmentation layers of the CNNs to reduce the requirements of memory and computation time. - Good segmentation performance for both CT and MRI.	- The cropping of the fixed physical size ROI is required.
UOL	- The discrete registration can capture large shape variations across scans. - The regularization is used to obtain smooth surfaces that are important for mesh generation and motion or electrophysiological modelling.	- Only tested on the MRI data. - The automatic cropping of ROI sometimes do not cover the whole heart.
KTH	- Combining shape context information with orthogonal U-Nets for more consistent segmentation in 3-D views. - Good segmentation performance, particularly for CT.	- Potential of overfitting because the U-Nets rely much on the shape context channels. - Weighting factors of the shape context generation are determined empirically.
CUHK1	- Pre-trained 3-D Network provides good initialization and reduces overfitting. - Auxiliary loss functions are used to promote gradient flow and ease the training procedure. - Tackling the class-imbalance problem using a multi-class Dice based metric.	- The introduced hyperparameters need determining empirically. - Relatively poor performance in MRI WHS.
UCF	- Multi-planar information reinforce the segmentation along the three orthogonal planes. - Multiple 3-D CNNs require less memory compared to a 3-D CNN.	- The softmax function in the last layer could cause information loss due to class normalization.
CUHK2	- Coupling the 3-D FCN with transfer learning and deep supervision mechanism to tackle potential training difficulties caused by overfitting and vanishing gradient. - Enhance local contrast and reduce the image inhomogeneity.	- Relatively poor performance in MRI WHS.
SEU	- Three-step multi-atlas image registration method is lightweight for computing resources. - The method can be easily deployed.	- Only tested on the CT data.
UT	- The proposed incremental segmentation method is based on local atlases and allows users to perform partial and incremental segmentation.	- The registration of MRI atlas can be inaccurate, and the evaluated segmentation accuracy is low.
SIAT	- Combining a 3-D U-Net with a ROI detection to alleviate the impact of surrounding tissues and reduce the computational complexity. - Fusing MRI and CT images to increase the training samples and take full advantage of multi-modality information so that features of different substructures can be better extracted.	- Poor segmentation performance, particularly for MRI data.
UB1*	- The focal loss and Dice loss are well encapsulated into a complementary learning objective to segment both hard and easy classes. - An iterative switch training strategy is introduced to alternatively optimize a binary segmentation task and a multi-class segmentation task for a further accuracy improvement.	- Late submission of the WHS results. - The clinical usage and usefulness of the uncertainty measurements are not clear.
UB2*	- Multi-scale context and multi-scale deep supervision are employed to enhance feature learning and to alleviate the potential gradient vanishing problem during training. - Reliable performance on the tested MR data.	- Late submission of the WHS results. - Only tested on the MRI data.
UOE*	- The proposed two-stage U-Net framework can directly segment the images with their original resolution.	- Late submission of the WHS results. - Poor performance, particularly for CT data.

WHS algorithms are based on deep neural networks. In general, the DL-based approaches can obtain good scores when the models have been successfully trained. However, tuning the parameters for a network to obtain the optimal performance can be difficult, as several DL-based methods reported poor results. This is also evident from Fig. 4 and Fig. 5 where some of the DL methods have very large interquartile ranges and outliers, and from the 3D visualization results presented in Fig. 6 and Fig. 7. In several cases, the shape of the heart from the segmentation results can be totally unrealistic, such as the worst CT case of UOE*, median and worst MRI cases of SIAT, worst MRI cases of CUHK1 and UCF.

In general, the conventional methods, mainly based on MAS framework, can generate results with more realistic shapes, though their mean accuracies can be less compared to the well trained DL models. Particularly, in MRI WHS the MAS-based methods obtained better mean accuracies than the DL-based approaches, though only two MAS methods were submitted for comparisons. Notice that the WHS of MRI is generally considered more challenging compared to that of CT. Since the DL-based approaches performed much better in the CT WHS, one can expect the performance of MR WHS could be significantly improved by resorting to new DL technologies in the future.

5.3. CT WHS versus MRI WHS

The MRI WHS is generally more arduous than the CT WHS, which is confirmed by the results presented in this work. The mean generalized Dice score of CT WHS is evidently better than that of MRI WHS averaged from the benchmarked algorithms, namely 0.872 ± 0.087 (CT) versus 0.824 ± 0.102 (MRI). One can further confirm this by comparing the results for these two tasks in Table 4 and Table 5, as nine methods have been evaluated on both the CT and MRI test data, and the same algorithms generally obtain better accuracies for CT data. Similar conclusion can be also drawn for the individual substructures as well as for the whole heart, when one compares the boxplots of segmentation Dice scores between Fig. 4 and Fig. 5.

5.4. Progress and challenges

The MM-WHS challenge provides an open access dataset and ongoing evaluation framework for researchers, who can make full use of the open source data and evaluation platform to develop and compare their algorithms. Both the conventional methods and the new DL-based algorithms have made great progress shown in this paper. It is worth mentioning that the DL models with best performance have demonstrated potential of generating accurate and reliable WHS results, such as the methods from GUT, UB1* and UB2*, though they were trained using 40 training images (20 CT and 20 MRI). Nevertheless, there are limitations, particularly from the methodological point of view. Table 6 summarizes the advantages and potential limitations of the benchmarked works.

WHS of MRI is more arduous. The average performance of the MRI WHS methods is not as good as that of the CT methods, concluded from the submissions. The challenges could mainly come from the low image quality and inconsistent appearance of the images, as well as the large shape variation of the heart which CT WHS also suffers from. Enlarging the size of training data is a commonly pursued means for improving the learning-based segmentation algorithms. However, availability of whole heart training images can be as challenging as the task itself. One potential solution is to use artificial training data, such as by means of data augmentation or image synthesis using generative adversarial networks (Goodfellow et al., 2014). Alternately, shape constraints can be incorporated into the training and prediction framework, which is particularly useful for the DL-based methods to avoid generating results of unrealistic shapes.

6. Conclusion

Knowledge of the detailed anatomy of the heart structure is clinically important as it is closely related to cardiac function and patient symptoms. Manual WHS is labor-intensive and also suffers from poor reproducibility. A fully automated multi-modality WHS is therefore highly in demand. However, achieving this goal is still challenging, mainly due to the low quality of whole heart images, complex structure of the heart and large variation of the shape. This manuscript describes the MM-WHS challenge which provides 120 clinical MRI/ CT images, elaborates on the methodologies of twelve evaluated methods, and analyzes their evaluated results.

The challenge provides the same training data and test dataset for all the submitted methods. Note that these data are also open to researchers in future. The evaluation has been performed by the organizers, blind to the participants for a fair comparison. The results show that WHS of CT has been more successful than that of MRI from the twelve submissions. For segmentation of the substructures, the four chambers generally are easy to segment from the submitted results. By contrast, the great vessels, including aorta and pulmonary artery, still need more efforts to achieve good results. For different methodologies, the DL-based methods could achieve high accuracy for the cases they succeed. They could also generate poor results with unrealistic shape, namely the performance can vary a lot. The conventional atlas-based approaches, either using segmentation propagation or probabilistic atlases, however generally perform stably, though they are not as widely used as the DL technology now. The hybrid methods, combining deep learning with prior information from either the multi-modality atlas or shape information of the heart substructures, should have potential and be worthy of future exploration.

Authors contributions

XZ initialized the challenge event, provided the 60 CT images, 41 MRI images (with KR and SO) of the 60 MRI images, and the manual segmentations of all the 120 images. GY, RM, JK, and DF provided the other 19 MRI images. XZ, GY and LL organized the challenge event, and LL evaluated all the submitted segmentation results. GY generated the first draft, based on which XZ and LL restructured and rewrote the manuscript. CP, DS, MU, MPH, JO, CW, OS, CB, XY, PAH, AM, UB, JB, GYu, CS, GG, JYR, TB, QT, WS, and XL were the participates of the MM-WHS challenge and contributed equally. GZ, ZS, GZ, CW, TM and DN submitted their results after the deadline of the challenge. All of the participants provided their results for evaluation and the description of their algorithms. All authors have read and approved the publication of this work.

Acknowledgement

This work was funded in part by the Chinese NSFC research fund, the Science and Technology Commission of Shanghai Municipality (17JC1401600) and the British Heart Foundation Project Grant (PG/16/78/32402).

References

Avendi, M., Kheradvar, A., Jafarkhani, H., 2016. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. Medical image analysis 30, 108–119.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging .

Cai, K., Yang, R., Chen, H., Li, L., Zhou, J., Ou, S., Liu, F., 2017. A framework combining window width-level adjustment and Gaussian filter-based multi-resolution for automatic whole heart segmentation. Neurocomputing 220, 138–150.

Crum, W.R., Camara, O., Hill, D.L., 2006. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE transactions on medical imaging 25, 1451–1461.

Dong, C., Loy, C.C., He, K., Tang, X., 2016. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence 38, 295–307.

Galisot, G., Brouard, T., Ramel, J.Y., 2017. Local probabilistic atlases and a posteriori correction for the segmentation of heart images, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 207–214.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

Heinrich, M.P., Blendowski, M., 2016. Multi-organ segmentation using vantage point forests and binary context features, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 598–606.

Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A., 2013a. MRF-based deformable registration and ventilation estimation of lung CT. IEEE transactions on medical imaging 32, 1239–1248.

Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, M., Schnabel, J.A., 2013b. Towards realtime multimodal fusion for image-guided interventions using self-similarities, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 187–194.

Heinrich, M.P., Oster, J., 2017. MRI whole heart segmentation using discrete nonlinear registration and fast non-local fusion, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 233–241.

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat 37, 547–579.

Kang, D., Woo, J., Kuo, C.J., Slomka, P.J., Dey, D., Germano, G., 2012. Heart chambers and whole heart segmentation techniques: a review. Journal of Electronic Imaging 21, 010901.

Karim, R., Blake, L.E., Inoue, J., Tao, Q., Jia, S., Housden, R.J., Bhagirath, P., Duval, J.L., Varela, M., Behar, J., et al., 2018. Algorithms for left atrial wall segmentation and thickness–evaluation on an open-source CT and MRI image database. Medical image analysis 50, 36–53.

Kim, J., Kwon Lee, J., Mu Lee, K., 2016. Deeply-recursive convolutional network for image super-resolution, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1637–1645.

Kittler, J., Hatef, M., Duin, R.P., Matas, J., 1998. On combining classifiers. IEEE transactions on pattern analysis and machine intelligence 20, 226–239.

Mahbod, A., Chowdhury, M., Smedby, Ö., Wang, C., 2018. Automatic brain segmentation using artificial neural networks with shape context. Pattern Recognition Letters 101, 74–79.

Mendis, S., Puska, P., Norrving, B., et al., 2011. Global atlas on cardiovascular disease prevention and control. World Health Organization.

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

Moghari, M.H., Pace, D.F., Akhondi-Asl, A., Powell, A.J., 2016. HVSMR 2016: MICCAI workshop on whole-heart and great vessel segmentation from 3D cardiovascular MRI in congenital heart disease. `http://segchd.csail.mit.edu/index.html`.

Mortazi, A., Burt, J., Bagci, U., 2017a. Multi-planar deep segmentation networks for cardiac substructures from MRI and CT, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 199–206.

Mortazi, A., Karim, R., Rhode, K., Burt, J., Bagci, U., 2017b. CardiacNET: Segmentation of left atrium and proximal pulmonary veins from MRI using multi-view CNN, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 377–385.

Ngo, T.A., Lu, Z., Carneiro, G., 2017. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Medical image analysis 35, 159–171.

Nikolaou, K., Alkadhi, H., Bamberg, F., Leschka, S., Wintersperger, B.J., 2011. MRI and CT in the diagnosis of coronary artery disease: indications and applications. Insights into imaging 2, 9–24.

Pace, D.F., Dalca, A.V., Geva, T., Powell, A.J., Moghari, M.H., Golland, P., 2015. Interactive whole-heart segmentation in congenital heart disease, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 80–88.

Payer, C., Štern, D., Bischof, H., Urschler, M., 2016. Regressing heatmaps for multiple landmark localization using CNNs, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 230–238.

Payer, C., Štern, D., Bischof, H., Urschler, M., 2017. Multi-label whole heart segmentation using CNNs and anatomical label configurations, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. 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. Magnetic Resonance Materials in Physics, Biology and Medicine 29, 155–195.

Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.C., et al., 2015. Right ventricle segmentation from cardiac MRI: a collation study. Medical image analysis 19, 187–202.

Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A., Wright, G., Huang, S., Liu, J., Lee, L., Venkatesh, S., et al., 2008. Cardiac MR left ventricle segmentation challenge. insight-journal. org .

Roberts, W., Bax, J., Davies, L., 2008. Cardiac CT and CT coronary angiography: technology and application. Heart 94, 781–792.

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

Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 520–527.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Shi, Z., Zeng, G., Zhang, L., Zhuang, X., Li, L., Yang, G., Zheng, G., 2018. Bayesian VoxDRN: A probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3D MR images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 569–577.

Suinesiaputra, A., Cowan, B.R., Finn, J.P., Fonseca, C.G., Kadish, A.H., Lee, D.C., Medrano-Gracia, P., Warfield, S.K., Tao, W., Young, A.A., 2011. Left ventricular segmentation challenge from cardiac MRI: a collation study, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer.

pp. 88–97.

Tan, L.K., McLaughlin, R.A., Lim, E., Abdul Aziz, Y.F., Liew, Y.M., 2018. Fully automated segmentation of the left ventricle in cine cardiac MRI using neural network regression. Journal of Magnetic Resonance Imaging 48.

Tobon-Gomez, C., Geers, A.J., Peters, J., Weese, J., Pinto, K., Karim, R., Ammar, M., Daoudi, A., Margeta, J., Sandoval, Z., et al., 2015. Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets. IEEE transactions on medical imaging 34, 1460–1473.

Tong, Q., Ning, M., Si, W., Liao, X., Qin, J., 2017. 3D deeply-supervised U-Net based whole heart segmentation, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 224–232.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3D convolutional networks, in: Proceedings of the IEEE international conference on computer vision, pp. 4489–4497.

Wang, C., Smedby, Ö., 2014. Automatic multi-organ segmentation in non-enhanced CT datasets using hierarchical shape priors, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE. pp. 3327–3332.

Wang, C., Smedby, Ö., 2017. Automatic whole heart segmentation using deep learning and shape context, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 242–249.

Xu, Z., Lee, C.P., Heinrich, M.P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R.G., Landman, B.A., 2016. Evaluation of six registration methods for the human abdomen on clinically acquired CT. IEEE Transactions on Biomedical Engineering 63, 1563–1572.

Yang, G., Sun, C., Chen, Y., Tang, L., Shu, H., Dillenseger, J.l., 2017a. Automatic whole heart segmentation in CT images based on multi-atlas image registration, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 250–257.

Yang, X., Bian, C., Yu, L., Ni, D., Heng, P.A., 2017b. 3D convolutional networks for fully automatic fine-grained whole heart partition, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 181–189.

Yang, X., Bian, C., Yu, L., Ni, D., Heng, P.A., 2017c. Hybrid loss guided convolutional networks for whole heart parsing, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 215–223.

Yu, F., Koltun, V., Funkhouser, T., 2017. Dilated residual networks, in: Computer Vision and Pattern Recognition, p. 2.

Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31, 1116–1128.

Zhao, J., Xiong, Z., 2018. 2018 atrial segmentation challenge. `http://atriaseg2018.cardiacatlas.org/`.

Zhou, R., Liao, Z., Pan, T., Milgrom, S.A., Pinnix, C.C., Shi, A., Tang, L., Yang, J., Liu, Y., Gomez, D., et al., 2017. Cardiac atlas development and validation for automatic segmentation of cardiac substructures. Radiotherapy and Oncology 122, 66–71.

Zhuang, X., 2013. Challenges and methodologies of fully automatic whole heart segmentation: A review. Journal of Healthcare Engineering 4, 371–407.

Zhuang, X., Bai, W., Song, J., Zhan, S., Qian, X., Shi, W., Lian, Y., Rueckert, D., 2015. Multiatlas whole heart segmentation of CT data using conditional entropy for atlas ranking and selection. Medical physics 42, 3822–3833.

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

Zuluaga, M.A., Cardoso, M.J., Modat, M., Ourselin, S., 2013. Multi-atlas propagation whole heart segmentation from MRI and CTA using a local normalised correlation coefficient criterion, in: International Conference on Functional Imaging and Modeling of the Heart, Springer. pp. 174–181.