This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2894322, IEEE Transactions on Medical Imaging

Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach

Jinming Duan, Ghalib Bello, Jo Schlemper, Wenjia Bai, Timothy J W Dawes, Carlo Biffi, Antonio de Marvao, Georgia Doumou, Declan P O'Regan, Daniel Rueckert, *Fellow, IEEE*

Deep learning approaches have achieved state-of-the-art performance in cardiac magnetic resonance (CMR) image segmentation. However, most approaches have focused on learning image intensity features for segmentation, whereas the incorporation of anatomical shape priors has received less attention. In this paper, we combine a multi-task deep learning approach with atlas propagation to develop a shape-refined bi-ventricular segmentation pipeline for short-axis CMR volumetric images. The pipeline first employs a fully convolutional network (FCN) that learns segmentation and landmark localisation tasks simultaneously. The architecture of the proposed FCN uses a 2.5D representation, thus combining the computational advantage of 2D FCNs networks and the capability of addressing 3D spatial consistency without compromising segmentation accuracy. Moreover, a refinement step is designed to explicitly impose shape prior knowledge and improve segmentation quality. This step is effective for overcoming image artefacts (e.g. due to different breath-hold positions and large slice thickness), which preclude the creation of anatomically meaningful 3D cardiac shapes. The pipeline is fully automated, due to network's ability to infer landmarks, which are then used downstream in the pipeline to initialise atlas propagation. We validate the pipeline on 1831 healthy subjects and 649 subjects with pulmonary hypertension. Extensive numerical experiments on the two datasets demonstrate that our proposed method is robust and capable of producing accurate, high-resolution and anatomically smooth bi-ventricular 3D models, despite the presence of artefacts in input CMR volumes.

***Index Terms*—Deep learning, bi-ventricular CMR segmentation, landmark localisation, non-rigid registration, label fusion, multi-atlas segmentation, shape prior, cardiac artefacts.**

I. Introduction

CARDIAC magnetic resonance (CMR) imaging is the gold standard for assessing cardiac chamber volume and mass for a wide range of cardiovascular diseases [1]. For decades, clinicians have been relying on manual segmentation approaches to derive quantitative measures such as left ventricle (LV) volume, mass and ejection fraction. However, manual expert segmentation of CMR images is tedious, time-consuming and prone to subjective errors. It becomes impractical when dealing with large-scale datasets. As such, there is a demand for automatic techniques for CMR image analysis that can handle the scale and variability associated with large imaging studies [2], [3]. Recently, automatic segmentation based on deep neural networks has achieved state-of-the-art performance in the CMR domain [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. For example, in the Automatic Cardiac Diagnosis Challenge (ACDC) [18] the 8 highest-ranked segmentation methods were all neural network-based methods.

Theoretically, 3D neural network-based segmentation methods may be designed with arbitrarily deep architectures. In practice however, the size of cardiac images, especially that of high-resolution volumetric images [11], often presents a computational bottleneck at the training stage. To deal with this, shallow 3D network architectures [11] or fewer feature/activation maps [5] are typically considered. Also, to reduce the computational burden, most methods extract the region of interest (ROI) containing the whole heart as a first step to reduce the volume size [8], [9], [10], [11], [14], [15], [17], or train a 2D network to separately segment each short-axis slice in the volume [12], [13], [14], [15], [16]. However, there are fundamental problems associated with each of these workarounds. For example, the use of shallow 3D network architectures or fewer feature maps is known to compromise segmentation accuracy. The ROI extraction approach is carried out using ROI detection algorithms, whose robustness remains questionable [8]. In addition, as no 3D context is taken into account, 2D network-based methods suffer from lack of 3D spatial consistency between the segmented slices (leading to lack of smoothness in the long-axis direction), and may result in a false positive prediction at an image slice containing non-ventricular tissues that are similar to target ventricles [8].

Due to the limitations of standard clinical acquisition protocols, raw volumetric CMR images acquired from standard scans often contain several artefacts [19], including inter-slice shift (i.e. respiratory motion), large slice thickness, and lack of slice coverage. Most deep learning methods do not routinely account for imaging artefacts [4], [5], [6], [7], [8], [9], [10], [12], [13], [14], [16]. As such, these artefacts are inevitably propagated onto the resulting segmentations. An example is given in Fig 1 e. The figure shows the segmentation of a 3D volume (whose short- and long-axis views are shown in Fig 1 a and b) using a state-of-the-art CNN approach [13]. As can be seen, the segmentation Fig 1 e inherits the misalignment and staircase artefacts present in the original volumetric image due to cardiac motion and large slice thickness. Further, holes exist at the apical region of the 3D model due to incomplete slice coverage of the whole heart. Different approaches have been proposed to tackle each artefact accordingly before building a smooth model. For example, misalignment was corrected using quadratic polynomials [15] or rigid registration [20]; Large slice thickness can be addressed by super-resolution techniques [21]. However, few studies have addressed different

JD, JS, WB, CB and DR are with the Biomedical Image Analysis Group, Imperial College London, London, UK. JD, GB, TJWD, A de M, GD and DPO'R are with MRC London Institute of Medical Sciences, Imperial College London, London, UK. TJWD is with the National Heart and Lung Institute, Imperial College London, London, UK.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2894322, IEEE Transactions on Medical Imaging

artefacts directly from an image segmentation perspective. To date, we are aware of only one deep learning segmentation method [11] that takes into account different cardiac artefacts, but the method was tested on only simulated images of the LV, whose anatomy is less complex than the bi-ventricular anatomy. It is thereby still an open problem as to how to build an artefact-free and smooth bi-ventricular segmentation model from real artefact-corrupted CMR volumes with novel image segmentation methods.

For clinical applications, segmentation algorithms need to maintain accuracy across diverse patient populations with varying disease phenotypes. In the existing literature, however, most methods [5], [6], [9], [11], [12], [13], [14], [15] have been developed and validated over normal (healthy) hearts or mildly abnormal hearts. Few studies have focused on hearts with very significant pathology with altered geometry and motion compared to healthy hearts. In addition, most methods [4], [5], [6], [7], [9], [10], [12], [14], [15] tend to use small image datasets. For example, four representative MICCAI challenges, namely the 2009 automatic LV segmentation challenge[1] (also known as Sunnybrook cardiac data), the 2011 LV segmentation challenge[2] (organized as part of the STACOM workshop), the 2015 RV segmentation challenge [22] and the 2017 ACDC, were tested on only 30, 100, 48 and 100 CMR datasets respectively. Given the small size of the datasets used for training and testing, whether the reported results can be generalised to larger cohorts remains questionable.

In this paper, we propose a segmentation pipeline to address the aforementioned limitations of current approaches. Specifically, we make the following contributions:

- We propose a multi-task deep learning network that simultaneously predicts segmentation labels and anatomical landmarks in CMR volumes. The network takes input volumetric images as multi-channel vector images (2.5D representation), requires no ROI extraction, and contains up to 15 convolutional layers. As such, the network has the computational advantage of 2D networks and is able to address 3D issues without compromising accuracy and spatial consistency. To our knowledge, this is the first work applying deep learning to CMR landmark localisation in a 3D context.
- We introduce anatomical shape prior knowledge to the network segmentation, which is a refinement step that is carried out using atlas propagation with a cohort of high-resolution atlases. As such, the pipeline is able to produce an accurate, smooth and clinically meaningful bi-ventricular segmentation model, despite the existing artefacts in the input volume. Moreover, due to the use of landmarks detected by the network, the proposed pipeline is entirely automatic.
- We demonstrate that the proposed pipeline can be readily generalised to segmenting volumetric CMR images from subjects with pulmonary hypertension (a cardiovascular disease). We thoroughly assess the effectiveness and robustness of the proposed pipeline using a large-scale dataset, comprising 2480 short-axis CMR volumetric images for training and testing. To our knowledge, this is one of the first CMR segmentation studies utilising a volumetric dataset of this size, and the technique introduced herein is the first automatic approach capable of producing a full high-resolution bi-ventricular model in 3D.

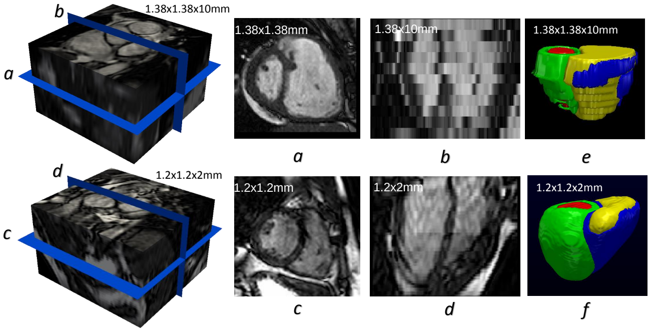

Figure 1. Illustrating the differences between a low-resolution CMR volume (top row) and a high-resolution CMR volume (bottom row). The images in the short-axis view are shown in a and c, while those in the long-axis view are in b and d. The corresponding segmentations are given in e and f.

II. Method

A. Overview

The proposed automatic segmentation pipeline handles two types of CMR volumetric inputs: low-resolution (LR) and high-resolution (HR) volumes. Fig 1 illustrates the differences between them. The LR volume has a large slice thickness (10 mm), giving rise to a staircase effect in the long-axis[3] view (Fig 1 b). Moreover, since the slices in Fig 1 b were acquired from multiple breath-holds, inconsistency of each breath-hold results in an inter-slice shift artefact. In contrast, the cross plane resolution of the HR volume is 2 mm, making its long-axis image Fig 1 d relatively smooth. In addition, HR imaging requires only one single 20-25 second breath-hold and therefore it introduces no inter-slice shift artefact. However, HR imaging may not be feasible for pathological subjects who are unable to hold their breath for 20-25s during each scan. Since HR imaging acquisition generates artefact-free cardiac volumes [23], it enables an accurate delineation of ventricular morphology, as shown in Fig 1 f. In comparison, Fig 1 e shows that the segmentation of an LR volume contains different cardiac artefacts [19] (e.g. inter-slice shift, large slice thickness, and lack of slice coverage). Note that the in-plane resolution of both HR and LR volumes is about 1.3×1.3 mm, so their corresponding short-axis views Fig 1 a and c are of relatively high quality.

The proposed pipeline has three main components: segmentation, landmark localisation and atlas propagation. We term the proposed network used in the pipeline as the Simultaneous Segmentation and Landmark Localisation Network (SSLLN).

[1] http://www.cardiacatlas.org/challenges/lv-segmentation-challenge/

[2] http://www.cardiacatlas.org/studies/sunnybrook-cardiac-data/

[3] In a standard CMR acquisition, short-axis and long-axis images are acquired separately, both of which have high in-plane resolution. However, in this paper, only CMR-acquired short-axis images are used, and a long-axis image denotes a vertical slice/cross-section of a stack of these short-axis images. Bigger gap between short-axis images would result in poorer resolution in the long-axis image. An example is given in Fig 1.

Further, the related terms SSLLN-HR and SSLLN-LR will be used to refer to versions of SSLLN trained with HR and LR volumetric data, respectively. In Fig 2, we illustrate the pipeline schematically. For an HR volume input, the trained SSLLN-HR is deployed to predict its segmentation labels as well as landmark locations. Since the HR volume input is artefact-free, the resulting segmentation is an accurate and smooth bi-ventricular 3D model. Afterwards, the HR volume and its corresponding SSLLN-HR outputs (landmarks and segmentation) are used as part of an HR atlas. For an LR volume input, the pipeline consists of two steps: First, the trained SSLLN-LR predicts an initial segmentation of the LR volume. In order to guarantee an artefact-free smooth segmentation output, a further refinement is carried out (second step). In this step, multiple selected HR atlases derived from SSLLN-HR are propagated onto the initial LR segmentation to form a smooth segmentation. This step explicitly fits anatomical shapes and is fully automatic due to the use of landmarks predicted from SSLLN-HR and -LR. We detail each of the two steps in the next two subsections.

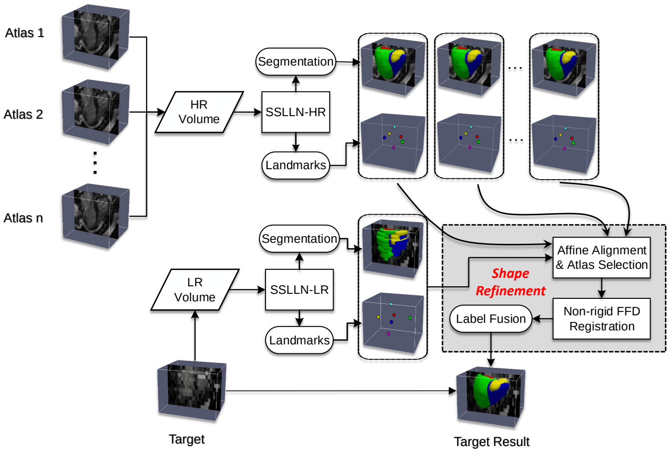

Figure 2. Pipeline for automatic bi-ventricular segmentation of low- and high-resolution volumetric images. The pipeline includes segmentation, landmark localisation and atlas propagation. It is capable of producing accurate, high-resolution and anatomically smooth bi-ventricular models, despite existing artefacts in input CMR volumes.

B. Learning segmentation labels and landmark locations

We treat the problem of predicting segmentation labels and landmark locations as a multi-class classification problem. First, let us formulate the learning problem as follows: we denote the input volumetric training dataset by $S = \{(U_i, R_i, L_i), i = 1, ..., N_t\}$, where $U_i = \{u_j^i, j = 1, ..., |U_i|\}$ is the raw input CMR volume (Fig 3 left), $R_i = \{r_j^i, j = 1, ..., |R_i|\}$, $r_j^i \in \{1, ..., N_r\}$ denotes the ground-truth segmentation labels for volume U_i ($N_r = 5$ representing 4 tissue types and a background as shown in Fig 3 right), $L_i = \{l_j^i, j = 1, ..., |L_i|\}$, $l_j^i \in \{1, ..., N_l\}$ stands for the ground-truth landmark labels for U_i ($N_l = 7$ representing 6 landmarks and a background as shown in Fig 3 middle), and N_t is the number of samples in the training data.

Note that $|U_i| = |R_i| = |L_i|$ is the total number of voxels in a CMR volume. We then define all network layer parameters as $\mathbf{W}$. In a supervised setting, we propose to solve the following minimisation problem via the standard (back-propagation) stochastic gradient descent

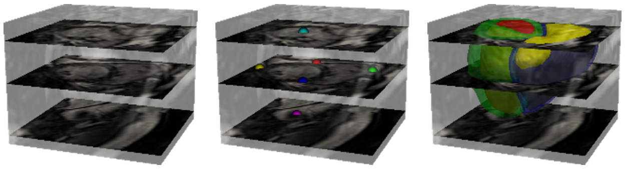

Figure 3. An exemplar raw volumetric CMR image, its ground-truth landmarks and segmentation labels, which are utilised as inputs to train the network in Fig 4. On the left, three short-axis slices in the volume are highlighted, corresponding to basal, mid-ventricular, and apical locations (from top to bottom) of the heart. In the middle, six landmarks are shown, coloured according to the following cardiac regions: the left ventricular lateral wall mid-point (yellow), two right ventricular insert points (red and blue), right ventricular lateral wall turning point (green), apex (pink) and centre of the mitral valve (cyan). Together, they reflect the size, pose and shape of the heart. On the right, a full anatomical bi-ventricular heart model is shown, coloured according to the left ventricular cavity (red), left ventricular wall (green), right ventricular cavity (yellow) and right ventricular wall (blue).

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}}(L_D(\mathbf{W}) + \alpha L_L(\mathbf{W}) + \beta\|\mathbf{W}\|_F^2), \quad (1)$$

where α and β are weight coefficients balancing the three terms. $L_D(\mathbf{W})$ is the segmentation loss that evaluates spatial overlap with ground-truth labels. $L_L(\mathbf{W})$ is the landmark associated loss for predicting landmark locations. $\|\mathbf{W}\|_F^2$, known as the weight decay term, represents the Frobenius norm on the weights $\mathbf{W}$. This term is used to prevent over-fitting in the network. The training problem is to estimate the parameters $\mathbf{W}$ associated with all the convolutional layers and by minimising (1) the network is able to simultaneously predict segmentation labels and landmark locations. The definition of $L_D(\mathbf{W})$ above is first given as follows

$$L_D(\mathbf{W}) = -\sum_i \frac{2\sum_k\sum_j \mathbb{1}_{\{r_j^i=k\}} \cdot P(r_j^i = k|U_i, \mathbf{W})}{\sum_k\sum_j\left(\mathbb{1}^2_{\{r_j^i=k\}} + P^2(r_j^i = k|U_i, \mathbf{W}) + \epsilon\right)}, \quad (2)$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function. ϵ is a small positive value used to avoid dividing by zero. i, k and j respectively denote the training sample index, the segmentation label index and the voxel index. $P(r_j^i = k|U_i, \mathbf{W})$ corresponds to the softmax probability estimated by the network for a specific voxel j (subject to the restriction $r_j^i = k$), given the training volume U_i and network weights $\mathbf{W}$. Note that (2) is known as the differentiable Dice loss [24], in which the summations are carried out over all voxels, labels and training samples.

For landmark localisation in a CMR volume, the primary challenge is the extreme imbalance between the proportion of voxels belonging to landmark regions and the proportion belonging to non-landmark regions (the 6 landmarks are represented by 6 voxels, while all the remaining voxels (numbering in the millions) represent background). To solve this highly imbalanced classification problem, we propose the class-balanced weighted categorical cross-entropy loss

$$L_L(\mathbf{W}) = -\sum_i \sum_k \left(w_k^i \sum_{j \in Y_k^i} \log P(l_j^i = k|U_i, \mathbf{W}) \right). \quad (3)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2894322, IEEE Transactions on Medical Imaging

Here k denotes the landmark label index, ranging from 1 to 7. Y_k^i represents the voxels in training sample i that belong to the region for which the value of landmark label index is k. To automatically balance landmark and non-landmark classes, we use a weight w_k^i for (3), where $w_k^i = 1 - \left|Y_k^i\right| / |Y_i|$, $k = 1, .., 7$. Here $|Y_k^i|$ denotes the number of voxels in Y_k^i, while $|Y_i|$ represents the total number of voxels in training sample i. Let us explain how the weighting process works intuitively. For the voxel falling in any one of the 6 landmark locations, $\left|Y_k^i\right|$ is 1 and $\left|Y_k^i\right| / |Y_i|$ is close to zero. Therefore, $1 - \left|Y_k^i\right| / |Y_i|$ is close to 1. On the other hand, $\sum \log P(l_j^i = k | U_i, \mathbf{W})$ in (3) is very small as only one voxel contributes to this term. Therefore, the product $w_k^i \sum \log P(l_j^i = k | U_i, \mathbf{W})$ ends up being a small value. In contrast, for a voxel falling in background area, $1 - \left|Y_k^i\right| / |Y_i|$ is a very small value close to zero. $\sum \log P(l_j^i = k | U_i, \mathbf{W})$ is however very large as almost all voxels (excluding the 6 landmark voxels) contribute to this term. Therefore, the product $w_k^i \sum \log P(l_j^i = k | U_i, \mathbf{W})$ becomes a small value. As such. the losses resulting from the landmark and non-landmark voxels are well balanced, which is crucial for successfully detecting merely 6 landmarks from a volume containing millions of voxels.

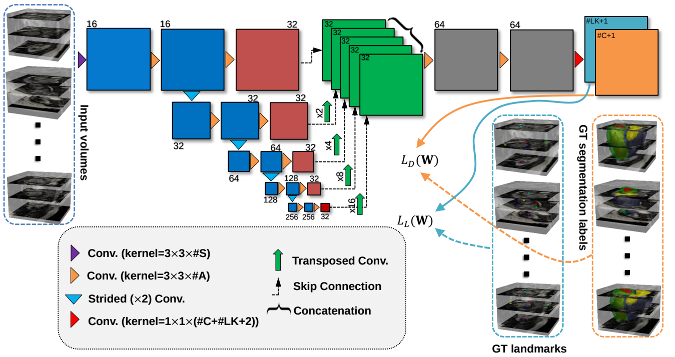

Figure 4. The architecture of the proposed SSLLN with 15 convolutional layers. The network takes different CMR volumes as input, applies a branch of convolutions, learns image features from fine to coarse levels, concatenates multi-scale features and finally predicts the probability maps of segmentation and landmarks simultaneously. These probability maps, together with the ground-truth segmentation labels and landmark locations, are then utilised in the loss function in (1) which is minimised via the stochastic gradient descent. Here #S, #A, #C, #LK and GT represent the number of volume slices, the number of activation maps, the number of anatomies, the number of landmarks, and ground truth, respectively.

In Fig 4, we show the architecture of SSLLN. There are two major differences between our network architecture and existing 2D or 3D ones, which we highlight as novel contributions of this work. First, 2D networks [4], [5], [6], [7], [9], [10], [12], [13], [14], [15], [16] are often trained using 2D short-axis slices separately. Therefore, there is no 3D spatial consistency between the resulting segmented slices. 3D networks [5], [6], [7], [11], [24] often rely on 3D convolutions, which in practice leads to 5D tensors (e.g. batch size × [3D volume size] × classification categories) during forward and backward propagations and requires far more GPU memory than their 2D counterparts. Workarounds such as subsampling [25] or use of small batch size and fewer convolutional layers [5], [7], [11] are often considered when training 3D networks, but these either complicate the training process or cause loss of information and accuracy. Unlike 2D networks, our network treats each input CMR volume as a multi-channel vector image, known as '2.5D' representation. In this sense, 3D volumes rather than 2D short-axis slices are used to train our network. As such, our network accounts for the spatial consistency between slices. Retaining the 3D spatial relationship is crucial for landmark localisation as landmarks encode spatial information. Unlike 3D networks, our network only involves 4D tensors (excluding the last layer). After the input volume passes through the first convolutional layer, the subsequent convolutional operations (excluding the last layer) in our network function exactly the same as those in 2D methods. Hence, the proposed network has the computational advantage of 2D networks, and also handles the input explicitly as a 3D volume (rather than a series of 2D slices), thus retaining accuracy and spatial consistency. This will be demonstrated later in Section III-C. We also note that other network architecture, such as the multi-view CNN [26] that parses 3D data into different 2D components, may also suit our applications. Second, our network predicts segmentation labels and landmark locations simultaneously as we integrate the two problems into a unified image classification problem for which we tailored a novel loss function (1). We are not aware of any previous approach that detects cardiac landmarks using a deep learning-based classification method. This is also the first work that focuses on segmentation and landmark localisation simultaneously.

After the network is trained, given an unseen CMR volume $f : \Omega \to \mathbb{R}^{\#S}$ ($\#S$ is the number of short-axis slices in the volume) defined on the domain $\Omega \subset \mathbb{R}^2$, we deploy the network on it and obtain the probability maps of segmentation (P_S) and the probability maps of landmarks (P_L) from the last convolutional layer. The binary segmentation and landmark labels are the indices of the maximum values of their probability maps along the channel direction, i.e. $S = \arg\max_{k=1,\ldots,N_r} P_S$ and $\mathcal{L} = \arg\max_{k=1,\ldots,N_l} P_L$.

C. Introducing anatomical shape prior knowledge

Due to limitations of cardiac MR imaging, low-resolution (LR) volumetric training datasets often contain artefacts, such as inter-slice shift, large slice thickness, lack of slice coverage, etc. Inevitably, the deployment of SSLLN-LR trained from such a dataset causes the propagation of these artefacts to the resulting segmentation. An example can be found in Fig 5 d and f. In this section, we introduce shape prior knowledge through atlas propagation to overcome such artefacts in SSLLN-LR segmentation. In Fig 5, we outline the shape refinement framework, including initial affine alignment, atlas selection, deformable registration and label fusion. The framework involves using a cohort of high-resolution (HR) atlases produced from SSLLN-HR, each of which consists of an HR CMR volume ($1.25 \times 1.25 \times 2.0$ mm), and its corresponding landmarks and segmentation labels. Next, we detail the framework.

Due to individual differences, the scanned heart often shows marked variations in size, pose and shape (as shown in Fig 5 a and d). This poses difficulty for existing image registration algorithms due to their non-convex nature. For

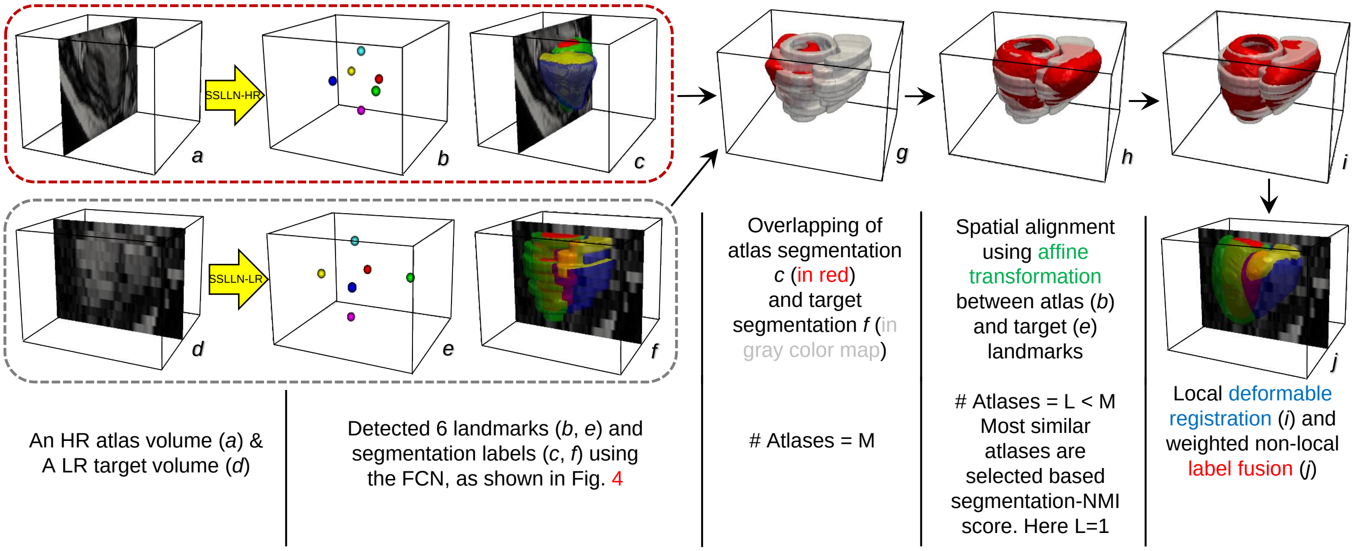
SSLLN-HR
a
b
c
SSLLN-LR
d
e
f
g
h
i
j
Overlapping of atlas segmentation c (in red) and target segmentation f (in gray color map)
Spatial alignment using affine transformation between atlas (b) and target (e) landmarks
An HR atlas volume (a) & A LR target volume (d)
Detected 6 landmarks (b, e) and segmentation labels (c, f) using the FCN, as shown in Fig. 4
# Atlases = M
# Atlases = L < M Most similar atlases are selected based segmentation-NMI score. Here L=1
Local deformable registration (i) and weighted non-local label fusion (j)

Figure 5. A block diagram illustrating how to explicitly introduce an anatomical shape refinement to the SSLLN-LR segmentation. As is evident in j, such a shape refinement enables an accurate, smooth and clinically meaningful bi-ventricular segmentation model, despite the artefacts in the LR input volume d. The framework is fully automated due to the use of the landmarks detected from SSLLN-HR and -LR.

this, the landmarks detected from SSLLN-HR and -LR were used to initialise the subsequent non-rigid algorithm between target and each atlas, which is similar to [27], [28]. An affine transformation with 12 degrees of freedom was first computed between the target landmarks (predicted by SSLLN-LR) and the atlas landmarks (predicted by SSLLN-HR). In addition to initialising the non-rigid image registration, the resulting affine transformations were used to warp segmentations in all atlases to the target space for atlas selection. According to the normalised mutual information (NMI) scores between the target segmentation and each of affinely warped atlas segmentations, L most similar atlases can be selected to save registration time and to remove dissimilar atlases for label fusion.

Since the correspondences of structures across both target and atlas volumes are explicitly encoded in their segmentations, we only use segmentations for the following non-rigid registration. Let S and l_n $(n = 1, ..., L)$ be the SSLLN-LR segmentation and the nth atlas segmentation, respectively. Let $P_{S,l_n}(i, j)$ be the joint probability of labels i and j in S and l_n, respectively. It is estimated as the number of voxels with label i in S and label j in l_n divided by the total number of voxels in the overlap region of both segmentations. We then maximise the overlap of structures denoted by the same label in both S and l_n by minimising the following objective function

$$\Phi_n^* = \arg\min \mathcal{C}\left(S, l_n(\Phi_n)\right) \tag{4}$$

where Φ_n is the transformation between S and l_n, which is modelled by a free-form deformation (FFD) based on B-splines [29]. $\mathcal{C}(S, l_n) = \sum_{i=1}^{N_r} P_{S,l_n}(i, i)$, representing the label consistency [30]. $\mathcal{C}$ in (4) is a similarity measure of how many labels, of all the labels in the atlas segmentation, are correctly mapped into the target segmentation. With the affine transformation as initialisation, a multi-scale gradient descent was then used to minimise the objective function (4). After the optimal Φ_n^* is found, the segmentations and volumes in the nth atlas are warped to the target space. The process is repeated until $n = L$.

Lastly, we perform non-local label fusion to generate an accurate and smooth bi-ventricular model $\tilde{S}$ for the imperfect SSLLN-LR segmentation S. Let us first denote the warped atlas volumes and segmentations as $\{(f_n, l'_n)|n = 1, ..., L\}$, respectively. Here, n denotes the warped atlas index and L is the number of selected atlases. For each voxel x in the target LR volume f, a patch f_x centred at x can be constructed. The aim of the label fusion task is to determine the label at x in f using $\{(f_n, l'_n)|n = 1, ..., L\}$. For each voxel x in f_n, we define $\{(f_{n,y}, l_{n,y})|n = 1, ..., L, y \in \mathcal{N}(x)\}$, where y denotes a voxel in the search window $\mathcal{N}(x)$, $f_{n,y}$ denotes the patch centred at voxel y in the nth warped atlas, and $l_{n,y}$ denotes the corresponding label for voxel y. The resulting label at voxel x in the target volume f can be calculated as

$$S_x = \arg\max_{k=1,..,N_r} \sum_{n} \sum_{y \in \mathcal{N}(x)} e^{-\frac{\|f_x - f_{n,y}\|_F^2}{h}} \cdot \delta_{l_{n,y},k} \tag{5}$$

where h denotes the bandwidth for the Gaussian kernel function and $\delta_{l_{n,y},k}$ denotes the Kronecker delta, which is equal to one when $l_{n,y} = k$ and equal to zero otherwise. The equation (5) can be understood as a form of weighted voting, where each of the patches from each of the atlases contributes a vote for the label. It is a non-local method because it uses patch similarity formulation (i.e. Gaussian kernel function), which is inspired by the non-local methods used in image denoising [31], inpainting [32], [33] and reconstruction [34]. It has been shown in [35] that, in a Bayesian framework, (5) is essentially a weighted K nearest neighbours (*KNN*) classifier, which determines the label by maximum likelihood estimation. By aggregating high-resolution atlas shapes in this way, an explicit anatomical shape prior can be inferred. The artefacts in the SSLLN-LR segmentation can thus be resolved, as shown in Fig. 5 j.

III. Experiments

In this section, we cover extensive experiments to evaluate (both qualitatively and quantitatively) the performance of the proposed pipeline on short-axis CMR volumetric images. Dice index and Hausdorff distance [13] were employed for evaluating segmentation accuracy. Dice varies from 0-1, with high values corresponding to a better results. The Hausdorff distance is computed on an open-ended scale, with smaller values implying a better match. We also validate the performance using clinical measures (ventricular volume and mass) derived from the segmentations. In the following experiments, each component in the pipeline is studied separately.

A. Clinical datasets

UK Digital Heart Project Dataset: This dataset[4] (henceforth referred to as Dataset 1) is composed of 1831 cine HR CMR volumetric images from healthy volunteers, with corresponding dense segmentation annotations at the end-diastolic (ED) and end-systolic (ES) frames. The ground-truth segmentation labels were manually annotated by a pair of clinical experts working together, and each volume was only annotated by one expert at a time. For each volume at ED, 6 landmarks, as shown in Fig 3 middle, were manually annotated by a clinician (inter-user 1). The raw volumes were derived from healthy subjects, scanned at Hammersmith Hospital, Imperial College London using a 3D cine balanced steady-state free precession (b-SSFP) sequence [23] and has a resolution of $1.25 \times 1.25 \times 2$ mm. As introduced in Section II-A, HR imaging technique does not produce cardiac artefacts which are often seen in LR imaging acquisition [19].

Pulmonary Hypertension Dataset: This dataset (henceforth referred to as Dataset 2) was acquired at Hammersmith Hospital National Pulmonary Hypertension Centre, and composed of 649 subjects with pulmonary hypertension (PH) - a cardiovascular disease characterised by changes in bi-ventricular volume and geometry. PH subjects often have breathing difficulties, therefore HR imaging was impractical for the majority of patients in this cohort due to the relatively long breath-hold time required. Within the cohort, 629 of the 649 patients were scanned using conventional LR image acquisition, and this manner of image acquisition (over multiple short breath-holds) often leads to lower-resolution volumes and inter-slice shift artefacts. In contrast, the remaining 20 subjects managed to perform a single breath-hold, and therefore HR volumes could be acquired for these subjects. Coupled with these HR volumes, LR volumes were also acquired during scanning, forming 20 pairs of LR and HR cine CMR volumes. The resolutions for LR and HR volumes are $1.38 \times 1.38 \times 10$ mm and $1.25 \times 1.25 \times 2$ mm, respectively. For all 649 subjects, the manual ground-truth segmentation labels at ED and ES were generated, and 6 landmarks at ED were also annotated.

B. Preprocessing and augmentation

Image preprocessing was carried out to ensure: 1) the size of each volumetric image fits the network architecture; 2) the intensity distribution of each volume was in a comparable range so that each input could be treated equally importantly. As such, each of the HR volumes in Dataset 1 was reshaped to common dimensions of $192 \times 192 \times 80$ with zero-padding if necessary, while each of LR volumes in Dataset 2 was interpolated to $1.25 \times 1.25 \times 2$ mm and then reshaped to $192 \times 192 \times 80$. For the best visual effect, the figures shown in experiments may be cropped manually. However, no ROI detection algorithm (for localisation of the heart) was used in image preprocessing. The intensity redistribution processes for both HR and LR volumes are the same. After reshaping, we first clipped the extreme voxel values (i.e. outliers) in each HR/LR volume. We defined outliers as voxel values lying outside of the 1st to 99th percentile range of original intensity values. Finally, the resulting voxel intensities of each volume were scaled to the $[0, 1]$ range.

Since our network takes volumetric images as inputs, we performed 3D data augmentation on-the-fly during training. At each iteration, augmentation included rescaling of voxel intensities in the input volume, and a 3D random affine transformation of the volume and corresponding labels and landmarks. For simplicity, the affine transformation only involved in-plane translation, isotropic scaling and rotation along one random direction (x-, y- or z-axis) at the central voxel of the volume. Neither shearing nor volume flipping was used. Data augmentation enables the network to see a large and diverse array of inputs by the end of training, and was implemented using the *SimpleITK* library in *Python*. With an Nvidia Titan XP GPU, training (50 epochs) took approximately 20 and 10 hours for Datasets 1 and 2, respectively. For inference, segmentation (without shape refinement) of an HR/LR volume for a single subject at ED took $< 1s$.

C. Segmentation of high-resolution volumes

First, we conducted experiments using Dataset 1, which includes 1831 HR CMR volumes. We randomly split the dataset into two disjoint subsets of 1000/831. The first subset was used to train SSLLN-HR, and the second subset was used for testing the accuracy of segmentation and landmark localisation, respectively. During training, we only used ED instances (volumes, landmarks and segmentation labels). Note that the proposed SSLLN-HR is a multi-task network that simultaneously outputs labels and landmarks. Next we segmented a cardiac volume into 5 regions: the left ventricular cavity (LVC), right ventricular cavity (RVC), left ventricular wall (LVW), right ventricular wall (RVW) and background. Our method is the first one capable of producing a full HR bi-ventricular segmentation (LVC+LVW+RVC+RVW) in 3D.

In Fig 6, we compare SSLLN-HR with two baseline methods for segmentation. The first one is the 2D FCN proposed in [13], where the network[5] was trained using each short-axis slice in the volume separately. The second one is the 3D FCN, whose architecture is similar as in Fig 4. To make the 3D FCN fit GPU memory, we halved the number of activation maps in each layer (excluding last one) and cropped the original image to a size of $112 \times 112 \times 64$. To focus exclusively on segmentation accuracy, we removed the landmark localisation activation maps in the last layer of the 3D FCN. As Fig 6 shows, 2D FCN produces a jagged appearance as shown in

[4] https://digital-heart.org/

[5] Code is publicly available at https://github.com/baiwenjia/ukbb_cardiac

the long-axis view image Fig 6 d, and there are ‘cracks’ in the corresponding 3D model as shown in Fig 6 e. This problem is due to the fact that the 2D method does not consider 3D context of the volumetric image, leading to a lack of spatial consistency between segmented slices. In contrast, both SSLLN-HR and 3D FCN account for the spatial consistency between slices, enabling smooth results. Visually, SSLLN-HR is comparable to 3D FCN. However, SSLLN-HR is less memory demanding and therefore can be directly implemented on non-cropped volumes with a faster training speed.

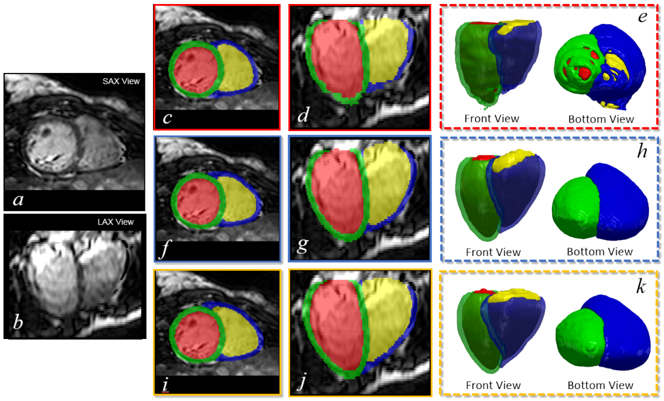

Figure 6. Visual comparison of segmentation results by 2D slice-by-slice FCN, 3D FCN and SSLLN-HR. a and b: two views of a high-resolution volume; c, d and e: results by 2D FCN; f, g and h: results by 3D FCN; i, j and k: SSLLN-HR. SAX and LAX denote short-axis and long-axis, respectively.

Table I
DICE INDEX AND HAUSDORFF DISTANCE DERIVED FROM 2D FCN, 3D FCN, AND SSLLN-HR FOR SEGMENTING 831 HIGH-RESOLUTION SHORT-AXIS VOLUMETRIC IMAGES. THE MEAN ± STANDARD DEVIATION ARE REPORTED.

	Dice Index (%)			Hausdorff Dist. (mm)		
	2D FCN	3D FCN	SSLLN-HR	2D FCN	3D FCN	SSLLN-HR
LVC	0.950±0.022	**0.963±0.010**	0.962±0.015	2.584±1.108	**2.037±0.413**	2.203±0.922
LVW	0.836±0.060	**0.888±0.024**	0.873±0.034	3.927±1.712	**3.028±1.062**	3.242±0.992
RVC	0.887±0.061	0.917±0.025	**0.929±0.026**	6.614±4.032	4.748±1.253	**4.171±1.527**
RVW	0.633±0.132	0.732±0.073	**0.755±0.068**	8.252±3.644	6.184±1.403	**5.996±1.424**

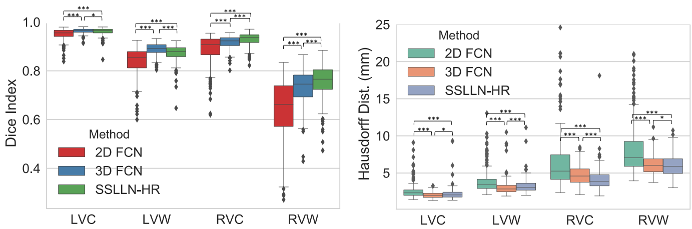

Figure 7. Boxplot comparison of segmentation accuracy between 2D FCN, 3D FCN and SSLLN-HR on 831 high-resolution short-axis volumetric images. The symbol ‘***’ denotes $p \ll 0.001$, and ‘*’ denotes $p < 0.1$.

Table I provides a summary of quantitative comparisons between 2D FCN, 3D FCN and SSLLN-HR, with statistics derived from 831 subjects. Statistical significance of the observed differences in the evaluation metrics (Dice index and Hausdorff distance) between each pair of methods is assessed via the Wilcoxon signed-rank test. The results in the table demonstrate the high consistency between automated and manual segmentations. In terms of Dice and Hausdorff distance, SSLLN-HR and 3D FCN outperformed 2D FCN, and SSLLN-HR achieved comparable performance to 3D FCN. Of note, all three methods achieved a relative low Dice score on the RVW anatomy. This is due to the thinness of RVW and the fact that the Dice index is more sensitive to errors in this structure. In Fig 7, boxplots visually depicting the results of Table are presented. As these plots show, the 2D method produced large variation across different segmentations for the four anatomies, resulting in a inferior accuracy than the 2.5D and 3D methods. SSLLN-SR achieved similar results to 3D FCN, with the segmentation accuracy of RVC and RVW slightly higher than that of 3D FCN.

In Fig 8, we further compare the proposed SSLLN-HR with the 2D FCN. We selected batches of k consecutive short-axis slices in a volumetric image, with multiple settings of k (=5,13, and 20). In each case, we set intensities in the selected slices to zero, as shown in the 1st column. The two methods under comparison were then applied to these partially zero-filled volumes, and the results are given in 2nd-5th columns. As is evident, 2D FCN fails to segment these zero-filled slices, thus leaving gaps in the resulting 3D segmentations. In contrast, SSLLN-HR demonstrates robustness to missing slices and has the capability of ‘inpainting’ these gap regions. However, as the gap (number of zero-filled slices) increases (from k=5 to k=20), the segmentation performance becomes worse. These results further illustrate that the proposed network retains 3D spatial consistency, which the 2D FCN is unable to achieve. Our method thus outperforms the 2D approach in this regard.

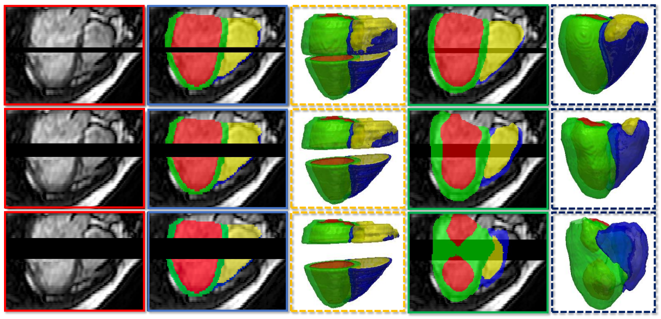

Figure 8. Testing 3D spatial consistency of the 2D FCN and SSLLN-HR methods. 1st column: target segmentation volumes with zero-filled gaps of different sizes; 2nd and 3rd columns: 2D FCN results; 4th and 5th columns: SSLLN-HR results.

D. Landmark localisation

To enable automatic alignment for subsequent non-rigid registration, we also predicted landmark locations (together with segmentation) for each input volume using SSLLN-HR. Same as above, we used the split subsets 1000/831 for training and testing. Note that SSLLN-HR was trained with manual landmarks carried out by inter-user 1 on each of the 1000 subjects. For the 831 unseen test subjects, the automatically detected landmarks were compared with the manual ones from inter-user 1 using the point-to-point Euclidean distance. Also, to study inter-user variability of landmarking, a second expert (inter-use 2) was recruited to manually annotate landmarks for each of 831 test subjects. The annotations were then compared with those of inter-user 1.

Fig 9 first shows a visual comparison of automated and manual (inter-user 1) landmarks. Fig 9 b shows the landmark locations predicted by our SSLLN-HR. As is evident, each landmark is represented by a few locally clustered voxels. The central gravity (represented by a single voxel) of each landmark in Fig 9 b can be computed by averaging the positions of the voxels forming the true landmark. The corresponding results are shown in Fig 9 c, where the two type of landmarks are superimposed. The respective colour-coded single-voxel landmarks are shown in Fig 9 d, which were used for initial point-to-point affine registration, as shown in Fig 5. In Fig 9 f, we superimposed the automated detected landmarks and manual landmarks (Fig 9 e). As can be seen, Fig 9 f demonstrates very good consistency between the automated and manual landmarks.

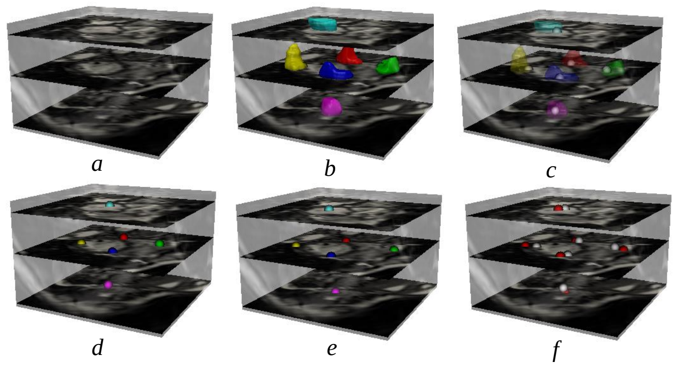

Figure 9. Landmark localisation using the proposed network. a: input volume; b: landmarks detected by SSLLN-HR directly; c: single-voxel landmark extraction from each clustered landmark in b; d: colour coded single-voxel landmarks; e: ground-truth landmarks annotated by inter-user 1; f: superimposed ground-truth (red) and automated (white) landmarks.

Table II
POINT-TO-POINT (P2P) DISTANCE ERROR STATISTICS IN LANDMARK LOCALISATION OVER 831 VOLUMES. THE SECOND COLUMN SHOWS THE ERRORS BETWEEN AUTOMATED (SSLLN-HR) AND MANUAL (INTER-USER 1) LANDMARK LOCALISATIONS. THE THIRD COLUMN SHOWS THE ERRORS BETWEEN TWO MANUAL (INTER-USER 1 AND 2) ANNOTATIONS. THE MEAN ± STANDARD DEVIATION IN MM ARE REPORTED. THE DESCRIPTION OF THE 6 LANDMARKS IS GIVEN IN FIG 3.

Region	Auto vs Man 1	Man 1 vs Man 2	p-value
Landmark-I (blue)	3.67±3.20	9.16±4.37	$p \ll 0.001$
Landmark-II (green)	8.18±6.91	17.4±9.27	$p \ll 0.001$
Landmark-III (red)	3.99±3.54	9.69±5.39	$p \ll 0.001$
Landmark-IV (yellow)	5.77±4.15	11.7±5.87	$p \ll 0.001$
Landmark-V (pink)	3.86±2.74	5.61±2.62	$p \ll 0.001$
Landmark-VI (cyan)	5.15±2.82	14.6±5.44	$p \ll 0.001$

In Table II, we compare the landmark localisation errors between automated and manual methods, as well as between the two manual methods on 831 test volumes. Using inter-user 1 as a baseline for comparison, we observe that the SSLLN-HR detections are more accurate than the annotations of inter-user 2. The point-to-point distance errors between SSLLN-HR and inter-user 1 vary only from 3.67±3.20 mm for Landmark-I to 8.18±6.91 mm for Landmark-II. In contrast, the errors between the two inter-users vary from 5.61±2.62 mm for Landmark-V to 17.4±9.27 mm for Landmark-II. This confirms that computer-human difference can be smaller than human-human difference.

Fig 10 provides a simple visualisation of the relative error distribution in the test sample of 831 volumes. The two plots show the cumulative distribution of point-to-point distance error for each landmark. As can be seen, the curves in the right are more clustered and stacked vertically than those in the left, indicating superior accuracy of landmark localisations by SSLLN-HR. For example, from the right plot we see that for all landmarks, about 92% of test volumes had point-to-point distance error of <20 mm. In contrast, only 60% of test volumes reached point-to-point distance error of ∼20 mm, as shown in the left plot.

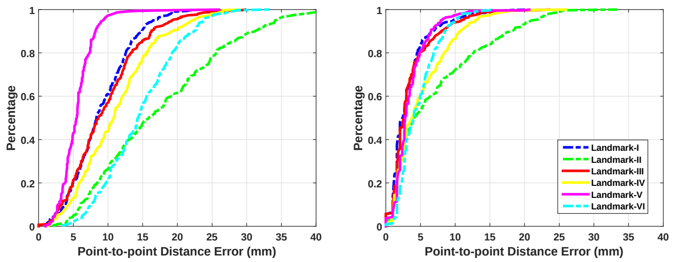

Figure 10. Cumulative error distribution curves of landmark localisation errors. The left curves are derived from manual landmarks of inter-user 1 and inter-user 2, and the right curves are plotted based on automated (SSLLN-HR) and manual (inter-user 1) landmark localisations.

Fig 11 provides a 3D visualisation of landmarks for all the 831 volumes. These landmarks were acquired from inter-user 1, inter-user 2 and SSLLN-HR. This figure further illustrates that SSLLN-HR, trained from manual annotations of a human, excellently matches the performance of that human on an unseen test set. On the other hand, the discrepancy between human-human performance could be very large. Fig 11, together with Fig 9, Table II and Fig 10, provide an ample evidence that the proposed SSLLN-HR has the capability of detecting landmarks robustly and accurately, and that it tends to produce less variability in predictions relative to variability among human experts.

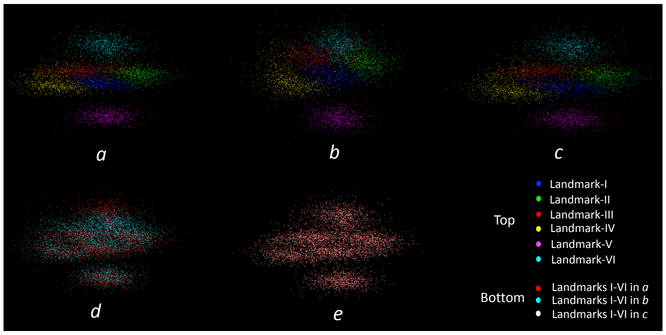

Figure 11. Visualisation of landmarks in 3D. a: manual landmarks by inter-user 1; b: manual landmarks by inter-user 2; c: landmarks localised by the network, trained on the manual annotations of inter-user 1; d: superimposed a and b; e: superimposed a and c. It is evident that a and c overlap to a greater degree than b and c.

E. Impact of landmarks

In this section, we show that landmarks localisation is a ne-

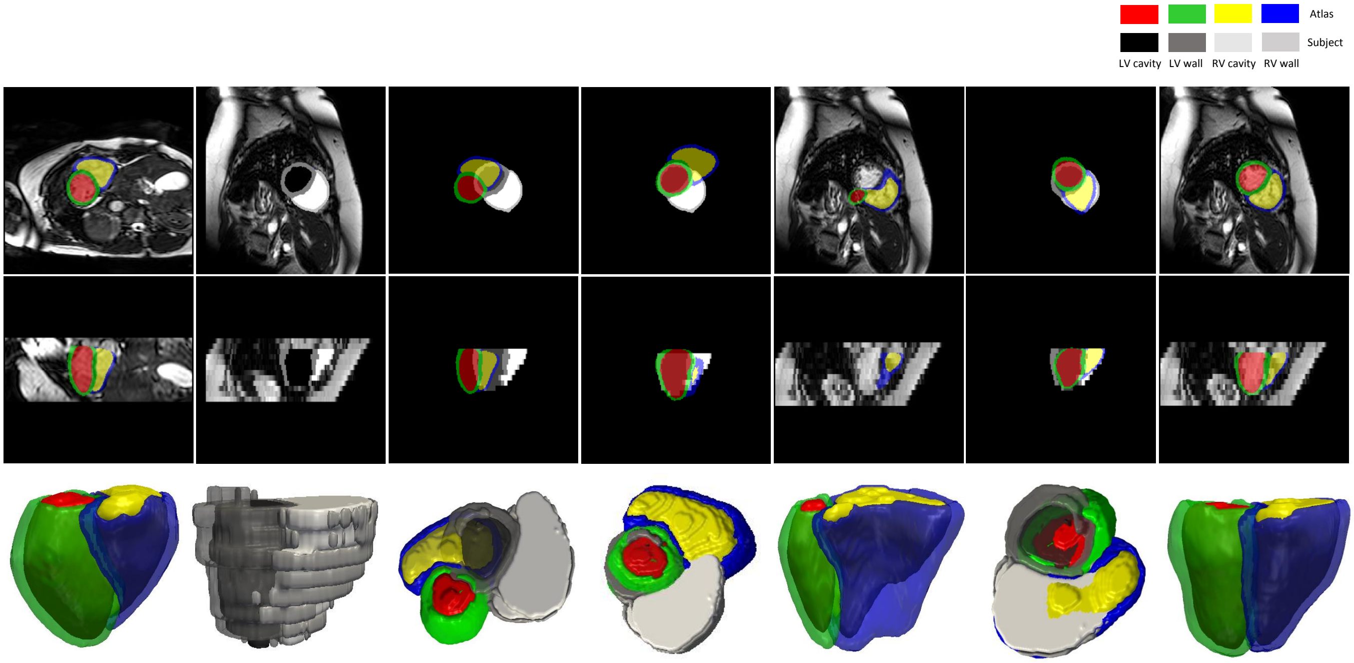

Figure 12. Impact of using landmarks in the proposed pipeline. 1st column: SSLLN-HR segmentation of an HR atlas volume; 2nd column: SSLLN-LR segmentation of an LR volume; 3rd column: superimposed SSLLN-HR and SSLLN-LR segmentation labels; 4th column: SSLLN-HR segmentation affinely warped to the SSLLN-LR segmentation based on their labels; 5th column: final SSLLN-LR segmentation refined by the non-rigid registration initialised with the label-based affine transform; 6th column: SSLLN-HR segmentation affinely warped to the SSLLN-LR segmentation based on their landmarks localised by the network in Fig 4; 7th column: final SSLLN-LR segmentation refined by the non-rigid registration initialised with the landmark-based affine transform. The 1st, 2nd and 3rd rows respectively show the short-axis view, long-axis view and corresponding 3D visualisation of segmentation.

cessary step in our pipeline. In Fig 12, we compare the SSLLN-LR segmentation results refined by the non-rigid deformation with different initialisations of affine transformation. As shown in the 5th column of Fig 12, the non-rigid refinement failed completely if the affine transform is initialised from the tissue classes. In contrast, initialising it directly on the landmarks resulted in an accurate refinement, as shown in the last column of Fig 12. We propose two reasons for this observation: 1) the six anatomical landmarks defined in the study effectively reflect the underlying pose, size and shape of a heart. As such, warping a heart with landmark-based affine transformation produces a very robust initialisation for the subsequent non-rigid registration; 2) Computing an affine transformation from a pair of landmarks is a convex least squares problem, a unique solution to which exits. In contrast, initialising an affine transform directly on the tissue classes is a non-convex problem. As such, the warped result is sometimes sub-optimal, which may negatively impact the non-rigid registration and increase uncertainty of the registration method. Moreover, label-based affine registration is much more computational expensive than landmark-based affine registration as it needs to deal with millions of voxels in the 3D volumes. In the *IRTK* implementation[6], it took ∼0.005s to compute an affine transformation on a pair of landmarks, whilst it took ∼5s to perform an affine registration using the 3D segmentation labels with size $256 \times 256 \times 56$.

We note that it may also be possible to detect landmarks automatically from segmentation labels. In this case, the accuracy of landmarks will be conditioned on the accuracy of segmentation. On the other hand, it may not be straightforward to determine which landmarks should be detected from segmentation labels for robust registration. As such, directly detecting the six landmarks defined in the study using the proposed network is neater and better.

F. Experiments on simulated low-resolution volumes

To quantitatively assess the performance of SSLLN-LR and shape refinement (SSLLN-LR+SR) in the pipeline (bottom path in Fig 2), we developed a method to simulate different types of artefacts seen in LR cardiac volumes. Specifically, in Fig 13 an HR volume and its manual segmentation were first downsampled from $1.25\times1.25\times2$ mm to $1.25\times1.25\times10$ mm, as shown in the 1st and 2nd columns. The downsampling produces a staircase artefact due to reduction in long-axis resolution. Moreover, the segmentation (Fig 13 d) around the apical region is now incomplete due to the lack of coverage of the whole heart. We further simulated inter-slice shift artefact by randomly translating each 2D short-axis slice horizontally. This step produced misalignment in the cardiac volume and its segmentation, as shown in the 3rd column.

Next, for training the SSLLN-LR, the LR volume Fig 13 e and its segmentation Fig 13 f were used as inputs. Note that our method is capable of producing an HR smooth segmentation model even from misaligned inputs such as the example in Fig 13 f. Since we have the smooth ground truth Fig 13 b for the simulated Fig 13 e, we can quantitatively assess the ability of our method to recover the original smooth shape. For these simulation experiments, we split Dataset 1 into subsets (1000/600/231). The first two subsets were corrupted

[6] Code is publicly available at https://github.com/BioMedIA/IRTK

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2894322, IEEE Transactions on Medical Imaging

with the simulated artefacts described above, which were used for training the SSLLN-LR and testing the proposed shape refinement (SC) component of the pipeline. The HR atlas shapes (segmented by SSLLN-HR) in the last cohort (n=231) were used to refine SSLLN-LR segmentations.

Here we highlight three reasons why we used SSLLN-HR network results as a reference atlas set for shape-refinement: 1) our SSLLN-HR is able to produce results that are very similar to the corresponding ground truth, as confirmed from Section III-C and III-D; 2) Once the SSLLN-HR is trained, it can be readily deployed on an external dataset (where HR atlases are not available) to create new HR atlases so as to facilitate the running of our pipeline; 3) The atlas set can be enriched by adding more SSLLN-HR results, which will increase the possibility to select better atlases for the sequential registration-based refinement.

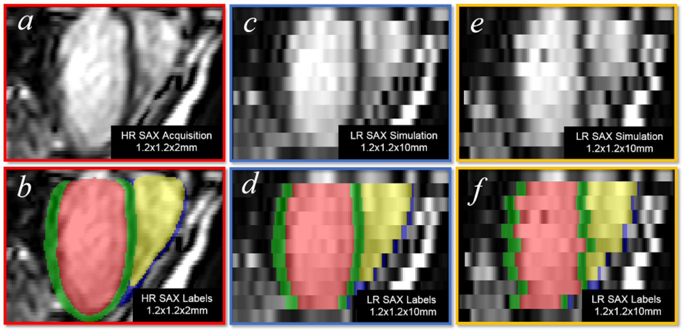

Figure 13. Simulating cardiac artefacts in real scenarios. 1st column: artefact-free high-resolution cardiac volume and ground-truth labels. 2nd column: downsampled versions of volumes in the 1st column. 3rd column: inter-slice shift is added to the downsampled volumes in the 2nd column.

Table III
Comparison of Dice index and Hausdorff distance between SSLLN-HR and SSLLN-LR+SR (shape refinement). SSLLN-HR was validated on 600 high-resolution short-axis volumetric images from Dataset 1, whilst SSLLN-LR+SR was validated on 600 low-resolution volumes, simulated from the corresponding high-resolution volumes.

	Dice Index (%)			Hausdorff Dist. (mm)		
	SSLLN-HR	SSLLN-LR+SR	p-value	SSLLN-HR	SSLLN-LR+SR	p-value
LVC	**0.960±0.015**	0.940±0.024	$p \ll 0.001$	**3.396±0.505**	4.045±0.675	$p \ll 0.001$
LVW	**0.879±0.030**	0.863±0.049	$p \ll 0.001$	**3.868±1.306**	4.394±0.841	$p \ll 0.001$
RVC	**0.929±0.025**	0.914±0.033	$p \ll 0.001$	**4.560±1.040**	5.039±1.218	$p \ll 0.001$
RVW	**0.662±0.103**	0.557±0.121	$p \ll 0.001$	**5.664±2.701**	6.119±2.956	$p \ll 0.001$

In Table III, we compare the Dice index and Hausdorff distance between the SSLLN-HR and SSLLN-LR+SR results. SSLLN-HR was directly evaluated on 600 artefact-free HR volumes at ED in Section III-C, while SSLN-LR+SC was tested on the 600 corresponding simulated LR volumes where cardiac artefacts exist, as shown in Fig 13. Although SSLLN-HR performs better than SSLLN-LR+SR, the performance gap between two approaches is minor. For LVC, LVW and RVC, the Dice index of SSLLN-HR is only about 0.2 higher than that of SSLLN-LR+SR. The Hausdorff distance of SSLLN-HR is about 0.5 mm smaller than that of SSLLN-LR+SR for all 4 regions. Again due to the thin structure of RVW, the mean Dice values of the two methods are relatively low: 0.662 and 0.557, respectively. This table shows that SSLLN-LR+SR achieves good segmentation results for imperfect LR input volumes, and the results are comparable to direct segmentation of artefact-free HR results.

In Table IV, we report the mean and standard deviation of the measurements derived from the two automated methods and manual segmentation. The table further demonstrates SSLLN-LR+SR results are comparable to SSLLN-HR results, proving that our proposed method can produce results comparable to direct segmentation of artefact-free HR volumes, even though target segmentation volumes are of low resolution and contain artefacts. Moreover, the RVM measurement derived from the two methods is consistent with the manual RVM measurement, confirming adequate segmentation of RVW using the two methods despite relatively lower Dice scores, as shown in Table III

Table IV
Comparison of clinical measures between SSLLN-HR, SSLLN-LR+SR and manual measurements on 600 volumetric cardiac images. SSLLN-HR was validated on high-resolution volumes from Dataset 1, whilst SSLLN-LR+SR was validated on 600 low-resolution volumes, simulated from the corresponding high-resolution volumes. The 4th/5th columns show absolute difference between automated and manual measures.

	SSLLN-HR*	SSLLN-LR+SR†	Manual	$*$ vs Manual	† vs Manual
LVV (ml)	148.392±34.352	151.048±35.016	147.638±34.711	4.623±4.014	6.066±5.782
LVM (gram)	123.028±24.123	124.240±24.383	119.278±25.685	5.551±4.308	6.737±5.244
RVV (ml)	168.638±37.144	174.383±39.480	171.553±38.622	8.547±7.540	9.299±8.551
RVM (gram)	35.466±8.121	32.290±7.381	33.704±7.261	3.571±2.803	2.956±2.899

Table V
Comparison of Dice index and Hausdorff distance between the proposed SSLLN-LR+SR and 5 state-of-the-art 3D approaches. These methods were tested on 20 simulated LR volumes (∼200 CMR images). The ground-truth labels were obtained from high-resolution volumes acquired from same subjects, which do not contain cardiac artefacts.

	Endocardium		Myocardium	
	Dice Index (%)	Hausdorff Dist. (mm)	Dice Index (%)	Hausdorff Dist. (mm)
3D-Seg [11]	0.923±0.019	10.28±8.25	0.773±0.038	10.15±10.58
3D-UNet [36]	0.923±0.019	9.94±9.92	0.764±0.045	9.81±11.77
3D-AE [37]	0.926±0.019	8.42±3.64	0.779±0.033	8.52±2.72
3D-ACNN [11]	0.939±0.017	7.89±3.83	0.811±0.027	7.31±3.59
MAM [38]	0.87±0.029	6.65±1.74	0.711±0.064	8.89±2.07
SSLLN-LR+SR	**0.943±0.020**	**4.09±0.69**	**0.854±0.042**	**4.37±1.04**

Next, we compare SSLLN-LR+SR with the 3D-seg model [11], 3D-UNet model [36], cascaded 3D-UNet and convolutional auto-encoder model (3D-AE) [37], 3D anatomically constrained neural network model (3D-ACNN) [11] as well as multi-atlas method[7] (MAM) [38]. To ensure a fair comparison, we used the same 20 CMR volumes as in [11] and the quantitative results are summarised in Table V. Since 3D-ACNN only segments the left ventricle (LV), the table only shows the results for the endocardium and myocardium of LV. Among the methods compared, 3D-seg and 3D-UNet do not use shape information, while 3D-AE and 3D-ACNN infer shape constraints using an auto-encoder during network training. As Dice shows, MAM is inferior to deep learning-based methods, shape-based models outperform those without shape priors, and our SSLLN-LR+SR achieved the best performance. We propose three main reasons for this: 1): SSLLN-LR+SR uses atlas propagation to impose a shape refinement explicitly while 3D-AE and 3D-ACNN impose shape constraints in an implicit fashion. When the initial segmentation by SSLLN-LR is of

[7] Code is publicly available at https://github.com/baiwenjia/CIMAS

sufficiently adequate quality, such an explicit shape refinement is able to produce more accurate segmentation. 2): SSLLN-LR+SR is a 2.5D-based method which allows the use of deeper network architectures than the 3D-based methods (e.g. ACNN-seg only uses 7 convolutional layers while SSLLN-LR+SR has 15), leading to improved segmentation accuracy. 3): SSLLN-LR+SR uses label-based non-rigid registration (4), which may be more accurate for segmentation purpose than the intensity-based non-rigid registration used in MAM.

G. Experiments on pathological low-resolution volumes

In Section III-F, we have quantitatively studied the performance of the proposed SSLLN-LR+SR using simulated LR cardiac volumes. In this section, we will use real LR volumes. In particular, we test SSLLN-LR+SR on volumetric data in patients with pulmonary hypertension (PH) in Dataset 2. PH leads to a progressive deterioration in cardiac function and ultimately death, due to RV failure. As such, it is critical to accurately segment different functional regions of the heart in PH so as to study PH patients quantitatively. Fig 14 shows the difference in two CMR volumes from a representative healthy subject and a PH subject. In health, the RV is crescentic in short-axis views and triangular in long-axis views, wrapping around the thicker-walled LV. In PH, the dilated RV pushes onto the LV causing deformation and loss of its circular shape. The abnormal cardiac morphology of PH heart poses challenges for existing segmentation algorithms.

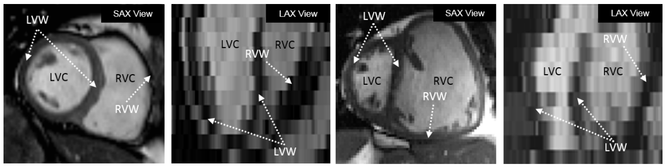

Figure 14. Illustrating the difference between a healthy subject (first two) and a PH subject (last two) from short- and long-axis views. Both subjects were scanned using low-resolution acquisition.

For training and testing, we use Dataset 2 introduced in Section III-A. This dataset includes 629 LR PH volumes and 20 pairs of LR and HR PH volumes. We randomly split the 629 volumes into two disjoint subset of 429/200. The first subset is used to train SSLLN-LR, while the second subset is used for visually testing the accuracy of SSLLN-LR+SR segmentations (due to lack of corresponding HR ground truths). The 20 LR volumes are also used to quantitatively evaluate SSLLN-LR+SR using their HR volumes as ground-truth references. 231 HR atlases appearing in Section III-C are used to refine SSLLN-LR segmentations.

200 greyscale PH volumes (1.38 × 1.38 × 10 mm) were segmented by SSLLN-LR+SR into HR smooth models (1.25 × 1.25 × 2 mm). Results were visually assessed by one clinician with over five years' experience in CMR imaging and judged satisfactory in all cases. We propose three reasons why the shape refinement works for PH cases: 1) the landmark-based affine and non-rigid registrations are collectively able to capture both global and local deformations between subjects; 2) for the non-rigid registration, we used label consistency as a loss function (4). It is based on segmentation masks, which can provide stronger regional and edge information for an accurate registration; 3) multiple atlases (i.e. the most similar to the subject) were selected for registration and fusion, and these selected atlases together vote for the final result, which further prevents diseased cases producing healthy results.

In Fig 15 a-h and Fig 16, we show an exemplary bi-ventricular segmentation of a cardiac volume in PH. We visually compare SSLLN-LR+SR with 2D FCN [13] and two approaches (nearest neighbour interpolation (NNI) and shape-based interpolation (SBI) [30], [39]) that interpolate the 2D FCN results. Both 2D FCN and interpolation methods do not use anatomical shape information, so they performed worse than SSLLN-LR+SR in the long-axis view, as confirmed in Fig 15 f-h. Due to the high in-plane resolution, similar results in the short-axis view were achieved by different methods, as shown in Fig 15 b-d. Moreover, we observed from Fig 16 that SSLLN-LR+SR gives a better 3D phenotype result which is smooth, accurate and artefact-free.

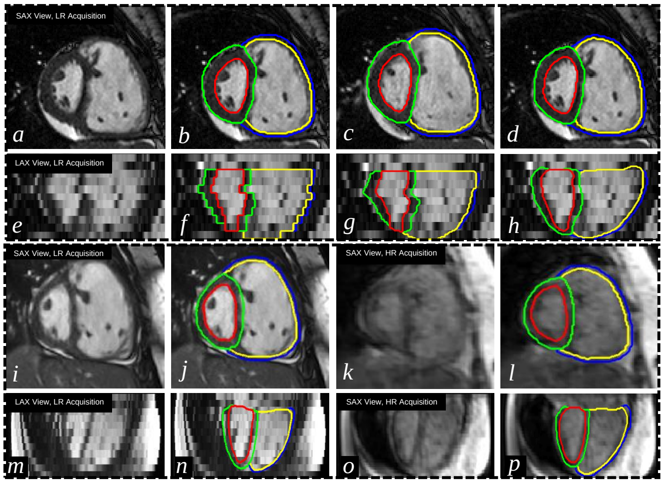

Figure 15. Bi-ventricular segmentation of volumetric images from two PH patients. a and e: original low-resolution volume (two views) from patient I; b and f: 2D FCN+NNI results; c and g: 2D FCN+SBI results; d and h: SSLLR-LR+SC results. i and m: original low-resolution volume from patient II; j and n: SSLLN-LR+SR results; k and o: original high-resolution volume from patient II; l and p: ground truth. The proposed SSLLN-LR+SR is not only insensitive to cardiac artefacts (inter-slice shift, large slice thickness, and lack of slice coverage), but also robust against pathology-induced morphological changes.

Next, we test SSLLN-LR+SR using 20 pairs of LR and HR cardiac volumetric images. In Fig 15 i-p, we first demonstrate a segmentation example on a pair of LR and HR volumes acquired from the same patient with PH. The original low-resolution volume (1.38 × 1.38 × 10 mm) was segmented by SSLLN-LR+SR into a HR smooth model (1.25 × 1.25 × 2 mm). The smooth segmentation is then visually compared with the ground truth, obtained directly from segmenting the corresponding HR volume of the patient. As is evident, the paired segmentation results show a very good agreement in terms of their cardiac morphology. Further, Table VI is provided, which shows a quantitative comparison between the SSLLN-LR+SR results and the ground-truth segmentations. The automated measurements are quantitatively consistent with the manual measurements. Comparing Table VI with Table IV, we observed that PH patients have a bigger RVC and a smaller

LVC than healthy subjects, and that the RVW of PH patients is thicker than that of healthy subjects. Note that the Dice scores computed from the paired LR and HR volumes are not applicable here due to the fact that they were acquired from subjects scanned at different positions with different breath-holds. We also note that p values in Table VI are relatively large. This is likely due to the relatively low sample size of the dataset used in this experiment, in addition to the fact that automatic and manual measurements are not substantially different.

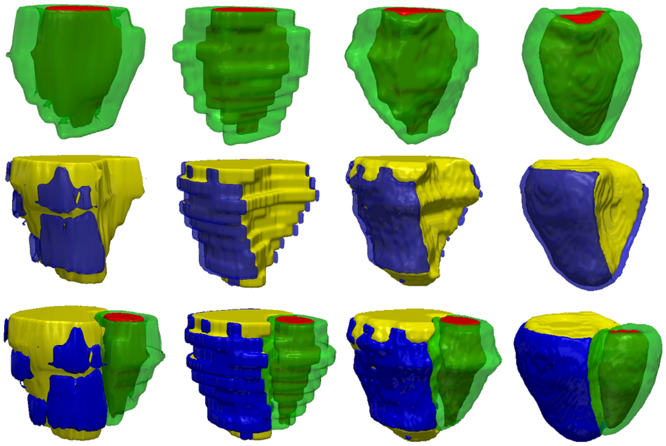

Figure 16. Visualisation of a 3D bi-ventricular model obtained through segmenting the volumetric image from a PH patient. 1st column: 2D FCN results; 2nd column: 2D FCN+NNI results; 3rd column: 2D FCN+SBI results; 4th colunm: SSLLR-LR+SC results. The proposed approach is capable of producing accurate, high-resolution and anatomically smooth bi-ventricular models for pathological subjects.

Table VI
COMPARISON OF CLINICAL MEASURES DERIVED FROM SSLLN-LR+SR AND MANUAL SEGMENTATIONS ON 20 PAIRS OF LOW-RESOLUTION AND HIGH-RESOLUTION VOLUMETRIC IMAGES FROM DATASET 2. SSLLN-LR+SR SEGMENTED 20 LOW-RESOLUTION VOLUMES INTO HIGH-RESOLUTION MODELS, WHILST MANUAL SEGMENTATION WAS PERFORMED ON 20 HIGH-RESOLUTION CARDIAC VOLUMES DIRECTLY. THE 4TH COLUMN SHOWS ABSOLUTE DIFFERENCE BETWEEN SSLLN-LR+SR AND MANUAL MEASURES.

	SSLLN-LR+SR†	Manual	† vs Manual	p-value
LVV (ml)	120.098±20.822	114.815±25.099	10.42±9.378	$p \approx 0.119$
LVM (gram)	125.989±34.639	124.237±25.271	8.032±9.614	$p \approx 0.855$
RVV (ml)	221.514±64.534	204.293±58.534	18.69±13.55	$p \approx 0.001$
RVM (gram)	51.621±14.938	49.877±14.166	3.857±2.558	$p \approx 0.501$

IV. DISCUSSION AND CONCLUSION

In this paper, we developed a fully automatic pipeline for shape-refined bi-ventricular segmentation of short-axis CMR volumes. In the pipeline, we proposed a network that learns segmentation and landmark localisation tasks simultaneously. The proposed network combines the computational advantage of 2D networks and the capability of addressing 3D spatial consistency issues without loss of segmentation accuracy. The pipeline also induces an explicit shape prior information, thus allowing accurate, smooth and anatomically meaningful bi-ventricular segmentations despite artefacts in the cardiac volumes. Extensive experiments were conducted to validate the effectiveness of the proposed pipeline for both healthy and pathological cases.

However, there still exist limitations in the pipeline. For example, the pipeline is a 2-stage approach, which is not end-to-end learning. In such a case, the network parameters learned in stage 1 might not be optimal to generate high-resolution smooth segmentations in stage 2. In addition, although the deployment of a trained network (SSLLN-HR or SSLLN-LR) in stage 1 took less than 1s, the shape refinement (SR) in stage 2 is relatively computationally expensive, which is a big disadvantage. SR combines the computational costs from atlas selection, target-to-atlas non-rigid image registration, and non-local label fusion. In our implementation, SR was performed in parallel for 5 selected atlases using multiple CPUs of a workstation and it took 15-20 mins per subject at ED.

In future work, we will investigate how to train a single network to compute smooth shapes from artefact-corrupted low-resolution cardiac volumes. A simple solution would be training an end-to-end super-resolution network, as in [21], but with the segmentation labels acquired from our pipeline as the ground truth inputs. We will also investigate how to improve the computational speed of Stage 2 in our pipeline. For example, a GPU-based non-rigid image registration toolbox[8] could be utilised. Besides the GPU-based implementation, deep hashing [40], [41] may be explored to select relevant atlas subjects instead of the brute force search technique (i.e. nearest neighbour) currently used in our atlas selection process. Another direction will be to investigate how to adapt the proposed network architecture for different tasks. For example, a fully connected layer may be concatenated for classification of subjects into healthy versus pathological groups, which will be carried out simultaneously with segmentation and landmark localisation tasks. Our pipeline treats landmarks as voxels and classifies them. In future work, we will explore an alternative approach that treats landmarks as points and regresses their coordinates, which could be implemented with a fully connected layer.

V. ACKNOWLEDGEMENTS

The research was supported by the British Heart Foundation (NH/17/1/32725, RE/13/4/30184); National Institute for Health Research (NIHR) Biomedical Research Centre based at Imperial College Healthcare NHS Trust and Imperial College London; and the Medical Research Council, UK. We would like to thank Dr Simon Gibbs, Dr Luke Howard and Prof Martin Wilkins for providing the CMR image data. The TITAN Xp GPU used for this research was kindly donated by the NVIDIA Corporation.

REFERENCES

[1] D. Ripley, T. Musa, L. Dobson, S. Plein, and J. Greenwood, "Cardiovascular magn. reson. imaging: what the general cardiologist should know," *Heart*, pp. heartjnl–2015, 2016.

[2] P. Gracia, B. Cowan, A. Suinesiaputra, and A. Young, "Challenges of cardiac image analysis in large-scale population-based studies," *Curr. Cardiol. Rev.*, vol. 17, no. 3, p. 9, 2015.

[3] D. Rueckert, B. Glocker, and B. Kainz, "Learning clinically useful information from images: Past, present and future," *Med. Image Anal.*, vol. 33, pp. 13–18, 2016.

[8] http://cmictig.cs.ucl.ac.uk/research/software/software-nifty/niftyreg

[4] H. Winther, C. Hundt, B. Schmidt, C. Czerner, J. Bauersachs, F. Wacker, and J. Vogel, "V-net: Deep learning for generalized biventricular cardiac mass and function parameters," *ArXiv Preprint ArXiv:1706.04397*, 2017.

[5] J. Patravali, S. Jain, and S. Chilamkurthy, "2d-3d fully convolutional neural networks for cardiac mr segmentation," *ArXiv Preprint ArXiv:1707.09813*, 2017.

[6] C. Baumgartner, L. Koch, M. Pollefeys, and E. Konukoglu, "An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation," *ArXiv Preprint ArXiv:1709.04496*, 2017.

[7] F. Isensee, P. Jaeger, P. Full, I. Wolf, S. Engelhardt, and K. Maier, "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features," *ArXiv Preprint ArXiv:1707.00587*, 2017.

[8] Q. Zheng, N. Delingette, Hand Duchateau, and N. Ayache, "3d consistent & robust segmentation of cardiac images by deep learning with spatial propagation," *IEEE T. Med. Imaging*, 2018.

[9] M. Nasr, M. Mohrekesh, M. Akbari, S. Soroushmehr, E. Nasr, N. Karimi, S. Samavi, and K. Najarian, "Left ventricle segmentation in cardiac mr images using fully convolutional network," *ArXiv Preprint ArXiv:1802.07778*, 2018.

[10] M. Khened, V. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *ArXiv Preprint ArXiv:1801.05173*, 2018.

[11] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. Cook, A. Marvao, T. Dawes, D. ORegan, *et al.*, "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE T. Med. Imaging*, vol. 37, no. 2, pp. 384–395, 2018.

[12] P. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis mri," *ArXiv Preprint ArXiv:1604.00494*, 2016.

[13] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. Lee, N. Aung, E. Lukaschuk, M. Sanghvi, *et al.*, "Human-level cmr image analysis with deep fully convolutional networks," *J. Cardiov. Magn. Reson.*, 2018.

[14] T. Ngo, Z. Lu, and G. Carneiro, "Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance," *Med. Image Anal.*, vol. 35, pp. 159–171, 2017.

[15] M. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri," *Med. Image Anal.*, vol. 30, pp. 108–119, 2016.

[16] J. Duan, J. Schlemper, W. Bai, T. J. Dawes, G. Bello, G. Doumou, A. De Marvao, D. P. ORegan, and D. Rueckert, "Deep nested level sets: Fully automated segmentation of cardiac mr images in patients with pulmonary hypertension," in *MICCAI*, pp. 595–603, Springer, 2018.

[17] J. Schlemper, O. Oktay, W. Bai, D. C. Castro, J. Duan, C. Qin, J. V. Hajnal, and D. Rueckert, "Cardiac mr segmentation from undersampled k-space using deep latent representation learning," in *MICCAI*, pp. 259–267, Springer, 2018.

[18] O. Bernard, A. Lalande, C. Zotti, *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE T. Med. Imaging*, 2018.

[19] S. Petersen, P. Matthews, J. Francis, M. Robson, F. Zemrak, R. Boubertakh, A. Young, S. Hudson, P. Weale, S. Garratt, *et al.*, "Uk biobanks cardiovascular magnetic resonance protocol," *J. Cardiov. Magn. Reson.*, vol. 18, no. 1, p. 8, 2015.

[20] G. Tarroni, O. Oktay, W. Bai, A. Schuh, *et al.*, "Learning-based quality control for cardiac mr images," *arXiv preprint arXiv:1803.09354*, 2018.

[21] O. Oktay, W. Bai, M. Lee, *et al.*, "Multi-input cardiac image super-resolution using convolutional neural networks," in *MICCAI*, pp. 246–254, Springer, 2016.

[22] C. Petitjean and J. Dacher, "A review of segmentation methods in short axis cardiac mr images," *Med. Image Anal.*, vol. 15, no. 2, pp. 169–184, 2011.

[23] A. Marvao, T. Dawes, W. Shi, C. Minas, N. Keenan, T. Diamond, G. Durighel, G. Montana, D. Rueckert, and S. o. Cook, "Population-based studies of myocardial hypertrophy: high resolution cardiovascular magnetic resonance atlases improve statistical power," *J. Cardiov. Magn. Reson.*, vol. 16, no. 1, p. 16, 2014.

[24] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision*, pp. 565–571, IEEE, 2016.

[25] K. Kamnitsas, C. Ledig, V. F. Newcombe, *et al.*, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, 2017.

[26] A. Mortazi, R. Karim, K. Rhode, J. Burt, and U. Bagci, "Cardiacnet: Segmentation of left atrium and proximal pulmonary veins from mri using multi-view cnn," in *MICCAI*, pp. 377–385, Springer, 2017.

[27] S. Grbic, R. Ionasec, D. Vitanovski, I. Voigt, Y. Wang, B. Georgescu, N. Navab, and D. Comaniciu, "Complete valvular heart apparatus model from 4d cardiac ct," *Med. Image Anal.*, vol. 16, no. 5, pp. 1003–1014, 2012.

[28] J. Duan, J. Schlemper, W. Bai, T. J. Dawes, G. Bello, C. Biffi, G. Doumou, A. De Marvao, D. P. ORegan, and D. Rueckert, "Combining deep learning and shape priors for bi-ventricular segmentation of volumetric cardiac magnetic resonance images," in *International Workshop on Shape in Medical Imaging*, pp. 258–267, Springer, 2018.

[29] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes, "Nonrigid registration using free-form deformations: application to breast mr images," *IEEE T. Med. Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[30] A. Frangi, D. Rueckert, J. Schnabel, and W. Niessen, "Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling," *IEEE T. Med. Imaging*, vol. 21, no. 9, pp. 1151–1166, 2002.

[31] A. Buades, B. Coll, and J. Morel, "A non-local algorithm for image denoising," in *CVPR*, vol. 2, pp. 60–65, IEEE, 2005.

[32] J. Duan, Z. Pan, B. Zhang, W. Liu, and X.-C. Tai, "Fast algorithm for color texture image inpainting using the non-local ctv model," *J. Global Optim.*, vol. 62, no. 4, pp. 853–876, 2015.

[33] J. Duan, Z. Pan, W. Liu, and X.-C. Tai, "Color texture image inpainting using the non local ctv model," *Journal of Signal and Information Processing*, vol. 4, no. 03, p. 43, 2013.

[34] W. Lu, J. Duan, D. Orive-Miguel, L. Herve, and I. B. Styles, "Graph-and finite element-based total variation models for the inverse problem in diffuse optical tomography," *arXiv preprint arXiv:1901.01969*, 2019.

[35] W. Bai, W. Shi, C. Ledig, and D. Rueckert, "Multi-atlas segmentation with augmented features for cardiac mr images," *Med. Image Anal.*, vol. 19, no. 1, pp. 98–109, 2015.

[36] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *MICCAI*, pp. 424–432, Springer, 2016.

[37] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *MICCAI*, pp. 203–211, Springer, 2017.

[38] W. Bai, W. Shi, and othersl, "A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac mr images," *IEEE T. Med. Imaging*, vol. 32, no. 7, pp. 1302–1315, 2013.

[39] S. Raya and J. Udupa, "Shape-based interpolation of multidimensional objects," *IEEE T. Med. Imaging*, vol. 9, no. 1, pp. 32–42, 1990.

[40] V. Erin, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *CVPR*, pp. 2475–2483, 2015.

[41] Y. Gu and J. Yang, "Densely-connected multi-magnification hashing for histopathological image retrieval," *IEEE J. Biomed. Health Inform.*, 2018.