Fully Automated Segmentation of the Left Ventricle in Cine Cardiac MRI using Neural Network Regression

Abstract

Background: Left ventricle (LV) structure and functions are the primary assessment performed in most clinical cardiac magnetic resonance imaging (MRI) protocols. Fully automated LV segmentation might improve the efficiency and reproducibility of clinical assessment.

Purpose: To develop and validate a fully automated neural network regression based algorithm for segmentation of the LV in cardiac MRI, with full coverage from apex to base across all cardiac phases, utilizing both short axis (SA) and long axis (LA) scans.

Study Type: Cross-sectional survey; diagnostic accuracy.

Subjects: 200 subjects with coronary artery diseases and regional wall motion abnormalities from public 2011 Left Ventricle Segmentation Challenge (LVSC) database; 1140 subjects with mix of normal and abnormal cardiac functions from public Kaggle Second Annual Data Science Bowl database.

Field Strength / Sequence: 1.5T, steady-state free precession.

Assessment: Reference standard data generated by experienced cardiac radiologists. Quantitative measurement and comparison via Jaccard and Dice index, modified Hausdorff distance (MHD), and blood volume.

Statistical Tests: Paired *t*-tests comparing to previous work.

Results: Tested against the LVSC database, we obtain 0.77 ± 0.11 (Jaccard index) and 1.33 ± 0.71 mm (MHD), both metrics demonstrating statistically significant improvement ($p < 0.001$) compared to previous work. Tested against the Kaggle database, the signed difference in evaluated blood volume is +7.2 ± 13.0 mL and −19.8 ± 18.8 mL for the end-systolic (ES) and end-diastolic (ED) phases respectively, with a statistically significant improvement ($p < 0.001$) for the ED phase.

Data Conclusion: A fully automated LV segmentation algorithm was developed and validated against a diverse set of cardiac cine MRI data sourced from multiple imaging centers and scanner types. The strong performance overall is suggestive of practical clinical utility.

Keywords: cardiac MRI, LV segmentation, deep learning, cine MRI, automated segmentation

Introduction

Cardiovascular diseases (CVD) are the primary cause of death globally, accounting for approximately 30% of all deaths in 2013 (1). Ultrasound is the primary imaging modality for CVD diagnosis due to its portability and low cost. However, cardiac MR is recognized as the reference standard for the assessment of cardiac volumes and regional functions due to its greater accuracy and reproducibility (2, 3). Most standard cardiac MR protocols begin with assessing the left ventricle (LV) structure and functions via steady state free precession (SSFP) gated cine imaging due to its high signal-to-noise ratio (SNR) and excellent contrast between the myocardium and blood pool (4). Standard acquisitions include long axis (LA) images captured along the plane passing through the LV apex and mitral valve, and a stack of short axis (SA) images orthogonal to the LA plane, captured between the LV apex and mitral valve.

In standard clinical practice, quantification of LV function is performed via manual delineation of the LV myocardium (endo- and epicardium) within the SA images, for the end-diastole (ED) and end-systole (ES) cardiac phases. This allows the evaluation of standard clinical measurements such as LV ED and ES blood volumes, ejection fraction, and LV mass (5). Despite delineating only two cardiac phases, such manual tracing can take up to 20 minutes by a radiologist. Full delineation across all cardiac phases would enable useful quantification of motion parameters to identify regional LV dysfunction. However, the excessive effort required for manual full delineation makes it impractical for clinical adoption.

Many LV segmentation algorithms have been published to date, with an extensive review provided by Petitjean et al. (6). Techniques used include general image-processing based methods such as intensity distribution modelling of the LV tissue and blood pool (7); deformable models such as active contours targeting the myocardium boundaries (8);

statistical shape and appearance models(9); and anatomical atlas-based registration (10). The attributes of published algorithms vary, ranging from semi-automated (11) to fully automated (12); endocardium only (13) to complete myocardium (14); mid-slices only (9) to full coverage from apex to base (mitral valve) (7); and dual ED/ES phases only (13) to all cardiac phases (15). Fully automated algorithms are inherently superior in terms of convenience, as well as their elimination of subjective inter-observer variability. However, it is still difficult for these algorithms to provide comparable performance to semi-automated algorithms which utilize human expert input. In addition, most published segmentation algorithms – both fully and semi-automated – are only validated against non-public, single institution datasets (6). This makes comparison between such published results infeasible due to the differing test conditions: differing quantity of test images, the pathological status (or none) of the patients, as well as the imaging parameters and hardware.

In our previous work, convolutional neural network (CNN) regression was introduced for the segmentation of LV myocardium, including full coverage from LV apex to base, across all cardiac phases (16). However, the approach was only semi-automatic (i.e. required user input) and was solely trained on SA images; manual intervention was still required to identify the basal and apical SA slices to constrain the volumetric quantification within the LV. In addition, the algorithm did not combine segmentation results across neighboring phases, i.e. each 2D slice was processed in isolation, leading to inconsistent results in the contour from one phase to the next. In this paper, we aim to develop and validate a fully automated algorithm for segmentation of the LV in cardiac MRI. We improve over the previous work through: (i) redesigning the network architecture to incorporate LA images for localizing apex and base landmarks, which enables full automation of the SA segmentation task, (ii) utilizing LA landmarks to stabilize SA segmentation in challenging apical and basal slices by

restricting the input SA field-of-view (FOV), and (iii) implementing 2D+time post processing to improve segmentation consistency.

Material and Methods

Data

We utilized data from three sources – two sources for the training and primary assessment of the algorithm, and the third source solely for the assessment of scan-rescan reproducibility. The first was data from the 2011 Left Ventricle (LV) Segmentation Challenge (LVSC) – a segmentation competition initiated during the 2011 Statistical Atlases and Computational Modelling of the Heart (STACOM) workshop (17). The LVSC database consists of 200 publically accessible cardiac MR cine cases; 100 with ground truth myocardial contours in all slices and phases for training and cross-validation, and 100 unlabeled cases for test validation. The database includes short axis (SA) and long axis (LA) volumes. The LA volumes are typically a mix of two-chamber, four-chamber, and LV outflow tract (LVOT) single slice 2D+time views. The subjects were comprised of patients with coronary artery diseases and regional wall motion abnormalities due to prior myocardial infarction. Patient characteristics were 76.8%/23.2% male/female, with mean age 62.7 years (range 34–84 years). By convention, the ground truth contours included trabeculae and papillary muscles in the blood pool, excluding them from the LV mass.

The second data source was the Kaggle Second Annual Data Science Bowl – a competition held in 2016 to evaluate end-systolic (ES) and end-diastolic (ED) blood volumes (18). The Kaggle database consists of 1140 publically accessible cardiac MR cine cases; 700 with ground truth ES and ED blood volume measurements for training and cross-validation, and 440 unlabeled cases for test validation. The database includes SA and LA volumes for most subjects, the LA volumes having a mix of two-chamber and four-chamber single slice

2D+time views (<1% of subjects also included LVOT views, but these were omitted for ease of processing). The subjects were comprised of a mix of patients with both normal and abnormal cardiac functions. Patient characteristics were 58.8%/41.2% male/female, with mean age 42.1 years (range 2 weeks – 88 years). The Kaggle ground truth only contains clinical blood volume measurements, i.e., no myocardial contours, centerpoints, or LA landmarks were available. To address this limitation, we manually assessed the training dataset and manually labelled 104 LA views for additional LA landmark localization training, and 15 cases for additional SA centerpoint localization training.

The combined LVSC training data and the manually labelled Kaggle data was initially split 85:15 by subject for training and cross validation, respectively, during hyperparameter optimization. The entire training and cross-validation dataset was then combined for the final training run after all hyperparameters had been finalized. There are a total of 26,069 and 9,860 individual images for SA and LA training respectively (Table 1).

Our third source of data consists of in-house scan-rescan data from a previous study on LV motion correction (19). Ten healthy subjects were recruited and scanned three times during the same session, and their respective ES and ED blood volumes quantified via a manually delineated, motion-corrected 3D surface model. All subjects had SA and LA scans, the LA volumes being two-chamber, four-chamber, and LVOT single slice 2D+time views. Subject characteristics were 30%/70% male/female, with mean age 48.4 years (range 39−61 years)

Patient information from the public LVSC and Kaggle databases were anonymized by their respective providers; data usage agreements were obtained for use in this study. This study also received Institutional Review Board (IRB) approval (ref: 989.75) for the reproducibility analysis.

MR protocol

All MR images from the three databases were acquired using a gated steady-state free precession (SSFP) pulse sequence. The LVSC and Kaggle databases were sourced from a variety of imaging centers and scanner types, leading to a heterogenous mix of imaging protocols and parameters. For the LVSC database, the range (mode) of imaging parameters were: echo time [TE] 0.96 – 2.98 (1.13) ms, repetition time [TR] 2.50 – 79.1 (59.2) ms, flip angle 25° – 90° (45°), in-plane (x/y) resolution 0.68 – 2.14 (1.56) mm, slice thickness 6 – 10 (8) mm, number of slices 5 – 17 (11), number of cardiac phases 18 – 35 (25). For the Kaggle database, the range (mode) of imaging parameters were: echo time [TE] 1.04 – 1.54 (1.19) ms, repetition time [TR] 14 – 54.7 (34.2) ms, flip angle 35° – 79° (50°), in-plane (x/y) resolution 0.59 – 1.95 (1.41) mm, slice thickness 5 – 11 (8) mm, number of slices 2 – 21 (10), number of cardiac phases 25 – 30 (30). The in-house scan-rescan data was acquired on a single 1.5T MRI system (Signa HDxt 1.5T, GE Healthcare, WI). The imaging parameters were: echo time [TE] 1.6 ms, repetition time [TR] 3.7 ms, flip angle 55°, in-plane (x/y) resolution 1.37 mm, slice thickness 8 mm, number of slices 10 – 15, number of cardiac phases 20.

Automated segmentation

Neural networks

Artificial neural networks are a family of mathematical functions with numerous recent successes in tackling artificial intelligence problems, including image processing and recognition (20). Though originated in the 1960s, neural networks have seen a strong resurgence in recent years, thanks largely to improvements in computing hardware and the availability of large quantities of training data (21).

Neural networks can be understood as a chain of linear operations interspersed with various nonlinear *activation* functions. Each group in the chain is more commonly known as a *layer*, which consists of a matrix of weights, *W*, and a vector of biases, *b*. For each individual layer the input vector is multiplied and summed against *W* and *b* respectively. An element-wise nonlinear *activation* function (e.g. a hyperbolic tangent function) is then applied and the resulting output is used as the input to the subsequent layer and the general series of operation is repeated in further layers.

Traditional neural networks are also known as fully connected networks (FCON), and are typically used with unstructured vector input. For inputs with regular structure (e.g. a 2D image), convolutional neural networks (CNN) are a more suitable variant. Here, *W* and *b* are applied repeatedly in a sliding window fashion analogous to the standard convolution operation in signal processing.

The *W* and *b* values of all layers are referred to as the network parameters. Starting from a random initialization, the parameters are iteratively updated by calculating a loss function (e.g. mean squared error) and back-propagating the result via an optimization function such as gradient descent, until convergence. Further information may be found in (21).

Segmentation System Overview

Our system primarily operates on 2D intensity images, as well as 2D first harmonic magnitude images, $H1_{mag}$, obtained by applying a 1D Fourier transform across the temporal dimension of a 2D+time slice. In our previous work we found $H1_{mag}$ to efficiently incorporate temporal information for short axis (SA) volumes (16).

Three separate neural networks were trained (Figure 1): (i) LV Landmarks (LM) network, where the LV base plane (mid of mitral valve) and the LV apex tip were localized in LA images. (ii) Centerpoint (CTR) network, where the LV centerpoint was localized in SA

images, (iii) Myocardial Boundaries (MB) network, where the myocardial boundaries were delineated in SA images. In all three networks, the inferred result is obtained through neural network regression. This is particularly notable for the MB network, where the myocardial boundaries are delineated as individual radial points inferred from a polar transform of the input image centered on the LV centroid, as opposed to the more common technique of myocardial segmentation by per-pixel classification. In our previous work, we found this regression technique superior to other state-of-the-art per-pixel classification networks; it implicitly enforces useful physiological constraints in the model, such as there being only a single connected object, and that the endo- and epicardium contours share a common centerpoint (16).

For both SA and LA images, an initial estimate of the location of the LV centerpoint, C_0, was first determined by calculating the intersection point between the SA and LA images. If insufficient LA images were available, C_0 was initialized at the center of each image instead.

In the LM network, individual LA intensity images were first resampled to a standard resolution of 2mm/pixel, with intensity values normalized to zero mean and a standard deviation (SD) of 1, and a 96×96 pixel crop was performed centered on C_0 (Figure 1a – red x-mark) to include the entire LV. The cropped images were input to the LM network, which outputs four values via regression: the evaluated (x, y) coordinates of the LV base and LV apex tip (Figure 1b – magenta dots). These coordinates were used to determine LV longitudinal coverage within the stack of corresponding SA images. This is primarily used to identify which SA slices to be included when evaluating clinical measurements such as ES and ED blood volume, as well as to identify basal and apical SA slices for additional processing, as described in the Pre- and Post-processing section.

For the input of the CTR network, the same process of resampling (2mm/pixel), intensity-normalization (zero mean, SD of 1) and cropping (96x96 pixels) was applied to both the SA intensity and $H1_{mag}$ images with center at C_0 (Figure 1c – red x-mark). Given this input, the CTR network outputs two values via regression: the evaluated (x, y) coordinates of the LV centerpoint, C_1 (Figure 1d – red crosshair). These coordinates were used for the polar remapping of the SA images for input into the MB network.

For the input of the MB network, individual SA intensity and $H1_{mag}$ images were first scaled to 1mm/pixel, normalized to zero mean and SD of 1, and the images were remapped to polar coordinate space centered on C_1, with radius 80 pixels and 96 angular sections (Figure 1e – circular dot pattern represents the polar coordinate space and bounds, red dots indicate zero θ position and orange dots indicate positive angular direction). The finer resampling was chosen to improve accuracy of delineation of the myocardial boundary. The resulting remapped images were 80×96 in size with cyclical buffering on both ends of the angular dimension. From this, individual 80×64 crops were taken along the angular dimension and input to the MB network, in the manner of a sliding window operation with unit step size (Figure 1f & 1g – dotted blue box indicates the size of sliding window). Each individual pass outputs two values via regression: the evaluated radius of the endo- and epicardial wall. Thus, for a single SA slice, the MB network was evaluated 96 times, resulting in 96 endo- and epicardial radius values (Figure 1g – red and green lines respectively). Finally, these values were remapped back to Cartesian coordinate space to form the myocardial boundaries.

Real-time random augmentation

Where there are small numbers of training datasets, random augmentation has been shown to improve generalization of the network by artificially increasing the number of training datasets (21). Specifically, this involves distorting the original training input data and target output result to create a new training sample, e.g. by displacing the image slightly and

calculating the new corresponding target centerpoint. Random augmentation is only performed during network training.

For the LM and CTR networks, input images were augmented by random rotation (by ±180°), flipping, displacement (by 35 mm) and scaling (by ±15%). For the MB network, input images were randomly rotated by ±180°, flipped, centerpoint perturbed up to 75% of the minimum endocardial radius, and scaled by ±15%. In addition, for the CTR and MB networks, 10% of the time a circular crop mask was applied to mask out arbitrary non-LV portions of the image, emulating a post-processing field-of-view (FOV) reduction operation. Finally, for all networks we added random Gaussian noise (0 mean, 0.15 SD). The aforementioned random FOV reduction was implemented to train the CTR and MB networks to handle the associated FOV reduction task during inference, as described later in the Pre- and Post-processing – CTR and MB network section. Interestingly, during training we noted a small improvement in cross-validation loss with the addition of the random FOV reduction, even though the cross-validation inference did not apply any corresponding FOV reduction (Figure 2). We hypothesize that the random FOV reduction influenced the network to de-emphasize non-LV information.

We performed real-time random augmentation as opposed to pre-generating a fixed number of augmented samples, i.e., the random distortions were generated and applied continuously during training.

Network architecture

Our earlier work utilized two networks (CTR and MB) with differing architecture (16). In particular, the earlier MB network utilized a specific “coarse and fine” dual sub-network design with varying input windows; the total number of parameters for both networks was around three million. In this paper, we significantly simplified the network architectures and

made them consistent across all three networks, while maintaining the total number of parameters at three million despite the inclusion of an additional new network (LM) for processing the long axis views.

All three networks (LM, CTR, MB) now use a single architecture: 8 CNN layers + 4 FCON layers, including the final output layer. Each network consists of approximately 1 million parameters. All intermediate layers had parameter quantities of comparable orders of magnitude (right column of Table 2).

Exponential linear units (ELU) (22) were used as activation functions for all layers, except the final output layer (no activation function used) and layer nine (maxout activation (23) with three units used). We opted not to utilize any pooling; striding of size two produced similar accuracy at lower operational cost. Standard CNNs were used in all convolutional layers except layer nine, where a separable convolution was used instead; this is an operation where the spatial convolution (depth-wise) is performed independently from the channel convolution (pointwise). The use of maxout activation and separable convolution in layer nine was primarily motivated by the need to control the number of parameters in this layer to be of comparable magnitude to the other layers.

Although the quantity of training data here is relatively small, we opted not to use any regularization techniques such as dropout, as we saw little benefit in the observed cross-validation loss. The real-time augmentation used during training appears to provide sufficient regularization for the task.

We used the Adam stochastic optimizer (24) to minimize a standard mean squared error loss function, using a mini-batch size of 64. The initial learning rate was set to 0.001, and annealed by half every ten thousand training runs. The network was designed using the TensorFlow r1.0 machine learning framework (Google Inc., California, U.S.), and executed on a 3.4GHz

Intel processor based workstation with a single NVIDIA GTX980 graphics processing unit (GPU). Each network took approximately two hours to complete training. During inference, complete execution of all three networks took approximately 12s per study, including SA and LA volumes.

Adjustment for pediatric cases

There were at least 99 subjects in the Kaggle database with age below 12 years, including 17 subjects below one year old. The median spatial resolution for subjects ≤12 years and >12 years of age were 0.7 and 1.4 mm respectively. The standard 2mm (LM & CTR) and 1mm (MB) rescaling used during data preparation would result in a loss of spatial resolution for images captured at finer resolutions. Despite being acceptable for adult-sized hearts, such loss is detrimental for pediatric subjects, particularly infants.

We utilize acquisition field of view (FOV) as a surrogate measurement for heart size, calculated as $\sqrt{m_x \times m_y} \times \sqrt{s_x \times s_y}$, where $m_{x,y}$ is the matrix size, and $s_{x,y}$ is the pixel spacing. From analysis of the Kaggle database, we empirically determined a median reference FOV value of 310mm, and a "small heart" threshold of 250mm. For datasets with FOV metric below the threshold, the image is scaled up to match the reference metric. E.g., a dataset with a 200mm FOV metric would be scaled up by 1.55× during data preparation.

Pre- and Post-processing

LM network

If a single subject had multiple LA volumes, the multiple evaluated landmark points were consolidated via the arithmetic mean. If an individual set of points were separated from their counterparts by >15 mm Euclidean distance, they were assumed to be errors and discarded. If only two LA volumes were available but their landmark points disagreed by >15 mm, preference was applied in this order: LA two-chamber (LA2C) → LV outflow tract (LVOT)

→ LA four-chamber (LA4C). We chose this order as the LA2C acquisition is the most straightforward longitudinal view for LV landmark localization, whereas the LVOT and LA4C acquisitions are more complicated to process as they may have the aorta and the right ventricle, respectively, in view.

We obtained a single set of (x, y) coordinates for each LA landmark by calculating the median across the cardiac phase. This provides a representative position of the LV apex tip and base (mid of mitral valve). These coordinates were then mapped to their corresponding positions in each SA slice to estimate z-dimension proximity to the LV apex and base position, the z-dimension here being the perpendicular dimension with respect to the SA plane. We expect at least 20% LV coverage (i.e., at least 20% of SA slices classified as between apex and base positions). If not, the points were assumed invalid and discarded. In all cases, the threshold constants were determined from analysis of the LVSC training data.

CTR and MB network

Using results from the LM network, each SA slice was categorized as apex, mid-level, or basal, using the criteria <0.2, 0.2 – 0.9, and >0.9 fractional z-position respectively. The mid-level slices were assumed to produce more reliable results and were processed through the CTR and MB networks as-is. From these results, we obtain the various LV mid-level centerpoints and calculate the global 95th percentile of the epicardial radius, r_{95}, to be used as contextual information for processing the apical and basal slices.

In the CTR network: for apical and basal slices, each initial centerpoint estimate C_0 (the intersection between SA and LA images) is replaced by its evaluated C_1 counterpart (the output of the CTR network) from the neighboring medial slice, with the assumption that the neighboring medial C_1 is a better starting estimate of the current centerpoint under evaluation, particularly for apical slices. We also applied a FOV reduction (i.e., a circular crop mask)

using r_{95} (the mid-level 95th percentile epicardial radius) as the base value, with 1.2× for basal slices, and 0.5× – 1× for apex-tip to apex-mid slices. The LV is especially small relative to the full image when close to the apex, potentially causing the network to be confused by other high intensity objects. This FOV reduction can be thought of as a conservative crop to exclude non-LV objects, which forces the network to only consider image data within the reduced FOV.

For the MB network, directly applying a similar reduced FOV tended to reduce the segmentation accuracy of good quality images due to the elimination of surrounding contextual data. Instead, a two-pass run is performed: in the first pass the uncropped image is processed. A second-pass with reduced FOV is applied for images with >15% outlier points. Outlier points were determined by the filtering and smoothing process described below.

For post-processing, CTR and MB results were filtered and smoothed using a periodic cubic spline filter. MB results were smoothed across spatial and temporal dimensions (i.e. 2D+time contour smoothing), while CTR results were smoothed temporally only. For the CTR network: outliers were identified by analyzing point-to-point Euclidean distances between neighboring phases; points >5.9 mm distance (which correspond to 99.99 percentile of LVSC training data) were filtered out. For the MB network: outliers for each slice and time frame were identified by analyzing the point-to-point Euclidean distance between neighboring radial points; the threshold was determined using a standard sigmoid function, $d/\left(1+\exp\left(-\left(b\times(x+a)\right)\right)\right)+c$, with the median radial distance as input, x. Points exceeding the threshold were filtered out. In all cases, the threshold, categorization, and multiplier constants were empirically determined from analysis of the LVSC training data.

Validation and Testing

For the LVSC database, the validation ground truth was based on a merged, consensus dataset (identified as CS*) built from multiple automatic and semi-automatic raters (17). We benchmarked our results to the LVSC ground truth in two ways: on an averaged point-by-point basis via the modified Hausdorff distance (MHD) (25); and in the form of binary myocardium images via the Jaccard index and Dice index. In addition, we subdivided the images to apical, mid-level, and basal slice locations, and analyzed their MHD metrics separately (the Dice and Jaccard index can be unreliable metrics for apical slices due to the small size of the LV binary image).

For the Kaggle database, only the ground truth ED and ES blood volumes are provided. We calculated blood volume via trapezoidal rule integration across the identified LV slices, adjusting the result to compensate for LV slice coverage, i.e. in cases where the SA slices did not cover the full extent of the LV, the integrated volume result was adjusted to compensate via a truncated ellipsoid function simulating the generic shape of the endocardium. The Kaggle challenge utilizes a continuous ranked probability score (CRPS) for evaluation, which necessitates building a cumulative distribution function (CDF) for the LV volume as opposed to a single value prediction (18). We fit a linear regression model against the Kaggle training set, with the predicted LV volumes and subject age and gender as regressors. The CDF was built as a Gaussian distribution with mean and standard deviation obtained from the regression model. For the LVSC and Kaggle evaluations, we performed a paired *t*-test comparing the previous work (16) to the current results.

For the scan-rescan reproducibility experiment using in-house datasets from 10 healthy subjects, we calculated the ED and ES blood volumes in a similar manner to the Kaggle evaluation, then calculated the standard deviation (SD) of the volumes across the three scans

for each subject, and obtained the overall mean SD across subjects. A paired *t*-test was utilized to compare the automated and manual results.

By design, the CTR and MB networks are tightly coupled. The results from the CTR network directly feed into the MB network; whereby the CTR network infers the LV centerpoints, and the MB network infers the myocardium as radial distances from the inferred centerpoint to the myocardium boundary. To test the independent errors of both networks, we generated gold standard LV centerpoints from the LVSC CS* reference binary images. For the CTR network, we calculated the independent error as the Euclidean distance between the evaluated C_1 result and the gold standard centerpoints. For context, we also converted this to a fractional result normalized by the average radius of the endocardium (in images with no blood pool, we used the epicardium radius instead). For the MB network, we repeated the LVSC validation test and metrics, but centered on the gold standard centerpoints (as opposed to the C_1 centerpoints). Unfortunately, the LVSC CS* dataset does not include reference delineations for the LA volumes. Thus we were unable to perform similar independent error analysis for the LM network.

Finally, we tested the added effect of the pre- and post-processing. The LVSC validation test and metrics were repeated, but with all pre- and post-processing disabled for both the previous (16) and current work. Specifically, we disabled the special processing for apical and basal slices (FOV reduction), as well as all filtering and smoothing functions as described in the Pre- and Post-processing section. The results reflect the raw, unfiltered output from the three networks.

Results

Evaluated against the LVSC database, there is a small but statistically significant improvement in all metrics when compared to our earlier, semi-automated work (16)

(Table 3). Notably, there is an approximately 2% improvement in the modified Hausdorff distance (MHD) despite the change to a fully-automated algorithm. To place this metric in perspective, over 95% of slices have MHD values ≤2× the in-plane resolution (i.e. ≤2 pixels). To the best of our knowledge, this is the highest overall performance to date for a fully-automated algorithm tested against the LVSC database (Table 4). Figure 3 illustrates results of a representative case from apex to base, diastole to systole.

Comparing the previous (16) semi-automated results and current automated results visually, the perceived improvement appears to come from the adoption of 2D+time contour smoothing, which results in higher phase-to-phase consistency. We measure this by calculating the standard deviation of the Jaccard index, Dice index, and MHD metrics across all cardiac phases for each individual slice, demonstrating approximately 4% to 7% average reduced variation in performance on a phase-to-phase basis (Table 3).

Another notable change is the additional pre- and post-processing applied to slices at the base and apex. In particular, apical slices are very challenging due to the relative small size of the blood pool (including disappearance during systolic phases). We did not find a statistically significant difference between the previous (16) and current work in the MHD metric for apical slices ($p = 0.027$). However, we suspect this is because the LVSC consensus validation dataset (CS*) does not include ground truth data for many apical slices. Consensus images were only generated for slices with valid results from at least three contributing raters; since apical slices are problematic for many algorithms, this likely resulted in invalid results for many raters. Around 47% of slices categorized as apical by the LM network were not included in the LVSC consensus validation dataset (Figure 4 – low result density at both ends). This included slices that demonstrated clear visual improvement when compared to the previous work (Figure 5). To explore this further, we calculated the MHD metric directly between the previous (16) and current work for all apical slices (i.e., as opposed to calculating

MHD against the CS* reference). We found that slices missing from CS* had significantly higher mean MHD compared to slices in CS* (2.59 vs. 1.23 mm, $p < 0.001$), i.e., slices missing from SC* showed larger differences in results between the previous and current work. This strongly suggests that apical slices missing from CS* may be more challenging (and are thus affected by the additional pre- and post-processing in the current work).

Evaluated against the Kaggle database, there is a similar statistically significant improvement in blood volume estimation for the end-diastole (ED) phase, though the end-systole (ES) results are more mixed (Table 3). Notably, these results are comparable to reported inter-reader variability values for multiple independent expert readers (27): bias (mean signed difference) up to ±13 / ±19 mL for ES/ED, and precision (standard deviation of signed difference) up to 13 / 13 mL for ES/ED. In comparison, our bias is +7.2 / −19.8 mL for ES/ED, and our precision is 13.0 / 18.8 mL for ES/ED. Despite this, we did not find a statistically significant difference for the continuous ranked probability score (CRPS) ($p = 0.67$). This is likely because our CRPS score is strongly affected by the separate linear regression used for its calculation; the regression technique used was unchanged between our previous (16) and current work.

Evaluating the scan-rescan reproducibility, we found no statistically significant difference between the automated and manual methods. The mean variability (standard deviation across three scans, averaged across all subjects) for the ES phase was 2.43 ± 1.10 mL and 3.35 ± 3.63 mL, $p = 0.41$ for the automated and manual methods respectively. The mean variability for the ED phase was 3.20 ± 2.26 mL and 4.40 ± 3.17 mL, $p = 0.32$ for the automated and manual methods, respectively.

For the independent error analysis, we found the CTR network to perform well, with a mean error of around 1.8 mm or 8% of the endocardial radius (Table 5). As expected, apical slices

performed the worst and mid-level slices the best. The performance of basal slices was almost similar to apical slices in terms of absolute error, though we suspect this is partially due to limitations of the gold standard centerpoints used; some of the LVSC CS* reference basal slices included only partial binary coverage of the myocardium, affecting calculation of the blood pool centerpoint. Given the low error of the CTR network, we found the independent error of the MB network to be largely similar to its end-to-end performance shown in Table 3. The independent error is around 1% better for all metrics, the biggest increase being the MHD for apical slices (around 8% improvement).

Finally, we tested the added effect of pre- and post-processing (PPP), and found its overall effect to be modest. For example, the mean Jaccard, Dice, and MHD metrics for the current work worsened from 0.769 / 0.864 / 1.329 mm to 0.767 / 0.863 / 1.338 mm with PPP disabled. There was still a statistically significant difference ($p < 0.001$) for all three metrics comparing the previous (16) to current work, demonstrating improvements in network training and architecture independent from the additional PPP. In contrast, with PPP disabled, there were no statistically significant differences for all three metrics when testing phase-to-phase consistency (i.e. standard deviation between phases) between the previous (16) to current work, further demonstrating the real effect of PPP 2D+time contour smoothing.

Discussion

In this paper we have presented a fully automated algorithm utilizing both short axis (SA) and long axis (LA) information concomitantly for the segmentation of left ventricular (LV) myocardium in SA cardiac MR images, with full coverage from apex to base, for all cardiac phases. Despite being a fully-automated operation, we show a small but statistically significant improvement in segmentation performance as compared to our previous semi-

automated approach (16), while significantly simplifying and making consistent the network architecture.

To the best of our knowledge, our mean 0.77 Jaccard index represents the best performance to date for a fully automated algorithm as evaluated against the public Left Ventricle (LV) Segmentation Challenge (LVSC) database. The only approach exceeding our performance is the semi-automated AU rater (Table 4), which requires significant manual input through the interactive placement of guide points in 4D (14). Additionally, the performance scores for AU are slightly advantaged due to its results being used to build the consensus validation dataset; i.e., the results for AU are not fully independent of the consensus ground truth.

In our evaluation against the Kaggle Second Annual Data Science Bowl challenge, we demonstrated a small but significant improvement in end-diastole (ED) blood volume estimation compared to our previous, semi-automated approach (16). Notably, the performance of the current algorithm is comparable to reported variability values for human raters (27), despite being tested against an order of magnitude more studies. Our Kaggle continuous ranked probability score (CRPS) of 0.0122 would have placed us in tenth position out of the 192 original challengers, a respectable outcome considering the MB network used for segmentation was not trained against any of the Kaggle data. Notably, the top three competitors in the original challenge all utilized convolutional neural networks (CNN) in some way or form: the champion (28) and second runner up (29) utilized per-pixel CNN segmentation of the blood pool (endocardium only), whereas the first runner up (30) utilized direct CNN regression of blood volume (no delineations).

LV delineation is inferred through the use of neural network regression. Our design necessitates the use of the polar transform so that the myocardium contour can be parameterized as radial distances from the LV centroid. The polar transform may introduce

errors where the blood pool is small; the endocardium contour approaches the LV centroid and may lead to significant interpolation artefacts. This is most apparent in apical slices, and is likely a significant reason for the reduced performance there. Nevertheless, the proposed approach has been shown to be effective overall.

Our approach is dependent on three notable assumptions. First, to handle pediatric cases, we assume that small fields of view (FOV) are a reliable proxy for small hearts, where we can then trigger an extra zoom factor to compensate. This assumption can fail in cases where an inappropriate large FOV was used during acquisition (i.e., excessive inclusion of empty space), in patients with a small heart but a large body (would not trigger the small FOV threshold), or in images that have been cropped beforehand (smaller FOV than expected, leading to inadvertent trigger of the zoom factor). Nevertheless, we did not see any of these situations occur in our extensive collection of test data.

The second assumption is a reliance on consistent patient positioning metadata in the DICOM tags; these are used to correlate the SA and LA scans together, enabling the initial centerpoint estimate, C_0, as well as for mapping the localized LA landmarks to the SA images, defining the LV apex-to-base extent and thus the proper calculation of the clinical measurements, and enabling the FOV reduction in apical and basal slices. Unlike the first assumption of FOV size, we did identify a small number of cases in the Kaggle database where the patient positioning metadata between the SA and LA volumes were inconsistent (e.g. where SA and LA volumes in the same study appeared to have different frames of reference). However, these situations were always easily detectable via out-of-bounds C_0 or landmark coordinates, allowing for straightforward flagging for manual intervention.

Finally, the majority of our approach is purely based on 2D or 2D+time. This allows the algorithm as a whole to be insensitive to inter-slice shifts due to patient movement between

slice acquisitions. The FOV reduction for apical and basal slices are exceptions to this, where they depend on consistent inter-slice positioning. The FOV reduction is based on a relatively conservative value – the 95th percentile of the mid-level epicardium radiuses – nevertheless it may fail in situations of extreme inter-slice shift, though we saw no evidence of that in the independent test data.

There were some limitations in our study design. The study design was only retrospective in nature; no new datasets were collected. In addition, we lack reference data for some aspects of our evaluation: the Kaggle dataset contains reference clinical volume measurements but no myocardium delineations, while the LVSC dataset is lacking gold standard delineations in a significant fraction of apical slices.

In conclusion, we have presented a fully automated algorithm utilizing SA and LA information for the segmentation of LV myocardium in SA cardiac MR images, with full coverage from apex to base, for all cardiac phases. This is the best performing fully automated algorithm to date as evaluated by the public LVSC challenge, while demonstrating performance comparable to human readers in both absolute variability of clinical parameters, as well as in scan-rescan reproducibility. This overall performance is a strong indicator of practical clinical utility.

References

1. Roth GA, Huffman MD, Moran AE, et al.: Global and Regional Patterns in Cardiovascular Mortality From 1990 to 2013. *Circulation* 2015; 132:1667–1678.

2. Gardner BI, Bingham SE, Allen MR, Blatter DD, Anderson JL: Cardiac magnetic resonance versus transthoracic echocardiography for the assessment of cardiac volumes and regional function after myocardial infarction: an intrasubject comparison using simultaneous intrasubject recordings. *Cardiovascular Ultrasound* 2009; 7:38.

3. Faridah Abdul Aziz Y, Fadzli F, Rizal Azman R, Mohamed Sani F, Vijayananthan A, Nazri M: State of the Heart: CMR in Coronary Artery Disease. *Current Medical Imaging Reviews* 2013; 9:201–213.

4. Kramer CM, Barkhausen J, Flamm SD, Kim RJ, Nagel E: Standardized cardiovascular magnetic resonance (CMR) protocols 2013 update. *J Cardiovasc Magn Reson* 2013; 15:1–10.

5. Schulz-Menger J, Bluemke DA, Bremerich J, et al.: Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (SCMR) Board of Trustees Task Force on Standardized Post Processing. *Cardiovasc Magn Reson* 2013; 15:35.

6. Petitjean C, Dacher J-N: A review of segmentation methods in short axis cardiac MR images. *Med Image Anal* 2011; 15:169–184.

7. Jolly M-P, Guetter C, Lu X, Xue H, Guehring J: Automatic Segmentation of the Myocardium in Cine MR Images Using Deformable Registration. In *Statistical Atlases and Computational Models of the Heart Imaging and Modelling Challenges*. Edited by Camara O,

Konukoglu E, Pop M, Rhode K, Sermesant M, Young A. Springer Berlin Heidelberg; 2012:98–108. [*Lecture Notes in Computer Science*, vol. 7085]

8. Berbari RE, Bloch I, Redheuil A, et al.: An automated myocardial segmentation in cardiac MRI. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; 2007:4508–4511.

9. Mitchell SC, Lelieveldt BPF, Geest RJ van der, Bosch HG, Reiver JHC, Sonka M: Multistage hybrid active appearance model matching: segmentation of left and right ventricles in cardiac MR images. *IEEE Transactions on Medical Imaging* 2001; 20:415–423.

10. Zhuang X, Rhode KS, Razavi RS, Hawkes DJ, Ourselin S: A Registration-Based Propagation Framework for Automatic Whole Heart Segmentation of Cardiac MRI. *IEEE Transactions on Medical Imaging* 2010; 29:1612–1625.

11. Bricq S, Frandon J, Bernard M, et al.: Semiautomatic detection of myocardial contours in order to investigate normal values of the left ventricular trabeculated mass using MRI. *J Magn Reson Imaging* 2016; 43:1398–1406.

12. Margeta J, Geremia E, Criminisi A, Ayache N: Layered Spatio-temporal Forests for Left Ventricle Segmentation from 4D Cardiac MRI Data. In *Statistical Atlases and Computational Models of the Heart Imaging and Modelling Challenges*. Edited by Camara O, Konukoglu E, Pop M, Rhode K, Sermesant M, Young A. Springer Berlin Heidelberg; 2012:109–119. [*Lecture Notes in Computer Science*, vol. 7085]

13. Avendi MR, Kheradvar A, Jafarkhani H: A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Medical Image Analysis* 2016; 30:108–119.

14. Li B, Liu Y, Occleshaw CJ, Cowan BR, Young AA: In-line Automated Tracking for Ventricular Function With Magnetic Resonance Imaging. *JACC: Cardiovascular Imaging* 2010; 3:860–866.

15. Tran PV: A Fully Convolutional Neural Network for Cardiac Segmentation in Short-Axis MRI. *arXiv:160400494 [cs]* 2016.

16. Tan LK, Liew YM, Lim E, McLaughlin RA: Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine MR sequences. *Medical Image Analysis* 2017; 39:78–86.

17. Suinesiaputra A, Cowan BR, Al-Agamy AO, et al.: A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images. *Medical Image Analysis* 2014; 18:50–62.

18. Second Annual Data Science Bowl | Kaggle [https://www.kaggle.com/c/second-annual-data-science-bowl]

19. Liew YM, McLaughlin RA, Chan BT, et al.: Motion corrected LV quantification based on 3D modelling for improved functional assessment in cardiac MRI. *Phys Med Biol* 2015; 60:2715.

20. LeCun Y, Bengio Y, Hinton G: Deep learning. *Nature* 2015; 521:436–444.

21. Goodfellow I, Bengio Y, Courville A: *Deep Learning*. 2016.

22. Clevert D-A, Unterthiner T, Hochreiter S: Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). 2016.

23. Goodfellow I, Warde-farley D, Mirza M, Courville A, Bengio Y: Maxout Networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13). Volume 28*. Edited by Dasgupta S, Mcallester D. JMLR Workshop and Conference Proceedings; 2013:1319–1327.

24. Kingma D, Ba J: Adam: A Method for Stochastic Optimization. *arXiv:14126980 [cs]* 2014.

25. Dubuisson MP, Jain AK: A modified Hausdorff distance for object matching. In *Proceedings of 12th International Conference on Pattern Recognition*. *Volume 1*; 1994:566–568 vol.1.

26. Fahmy AS, Al-Agamy AO, Khalifa A: Myocardial Segmentation Using Contour-Constrained Optical Flow Tracking. In *Statistical Atlases and Computational Models of the Heart Imaging and Modelling Challenges*. Edited by Camara O, Konukoglu E, Pop M, Rhode K, Sermesant M, Young A. Springer Berlin Heidelberg; 2012:120–128. [*Lecture Notes in Computer Science*, vol. 7085]

27. Suinesiaputra A, Bluemke DA, Cowan BR, et al.: Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. *Journal of Cardiovascular Magnetic Resonance* 2015; 17:63.

28. Automatic Left ventricle volume calculation in cardiac MRI using Convolutional Neural Network [https://github.com/woshialex/diagnose-heart]

29. 3rd place solution for the second national datascience bowl [http://juliandewit.github.io/kaggle-ndsb/]

30. Diagnosing Heart Diseases with Deep Neural Networks

[http://irakorshunova.github.io/2016/03/15/heart.html]

Tables

TABLE 1. Datasets used for training, cross-validation, and test validation. Descriptions of the three networks (LM, CTR, MB) are in the "Segmentation System Overview" section

	Training & cross-validation		**Test validation**	
Networks	**LVSC**	**Kaggle**	**LVSC**	**Kaggle**
LM network * − LA2c	98 (2,275)	52 (1,560)	98 (2,310)	434 (13,020)
− LA4c	99 (2,295)	52 (1,560)	98 (2,315)	429 (12,870)
− LVOT	93 (2,170)	None	93 (2,210)	None
CTR network †	100 (22,259)	15 (3,810)	100 (28,115)	440 (136,620)
MB network †	100 (22,259)	None	100 (28,115)	440 (136,620)

* Values shown are: number of 2D+time long axis (LA) views (number of images);
LA2c – two chamber, LA4c – four chamber, LVOT – outflow tract

† Values shown are: number of 3D+time short axis (SA) volumes (number of images)

TABLE 2. Basic architecture of all three networks (LM, CTR, MB)

Layer	Size	Parameters (000's)
1. CNN *	5×5×64	3
2. CNN	5×5×64	102
3. CNN *	5×5×64	102
4. CNN	5×5×64	102
5. CNN *	3×3×96	55
6. CNN	3×3×96	83
7. CNN *	3×3×96	83
8. CNN	3×3×96	83
9. CNN → FCON †	6×6×960	95
10. FCON	320	103
11. FCON	320	103
12. FCON → OUT	4 (LM) or 2 (CTR, MB)	0.6

* x = 2, y = 2 stride applied on input data

† separable (separate depthwise and pointwise convolution) CNN. Maxout activation with 3 units used.

TABLE 3. Comparison of results between the previous semi-automated algorithm of (16), and the fully-automated algorithm presented here. ES = end-systole, ED = end-diastole.

	Semi-automated (16)	**Fully-automated (this paper)**	**Difference ‡**
LVSC database			
Jaccard index (JI) *	0.765 ± 0.111	0.769 ± 0.109	0.003 ± 0.053
Dice index (DI) *	0.862 ± 0.083	0.864 ± 0.080	0.003 ± 0.041
Modified Hausdorff distance (MHD) (mm) *	1.355 ± 0.718	1.329 ± 0.710	0.026 ± 0.495
Apical	1.720 ± 1.002	1.769 ± 1.126	-0.049 ± 1.103
Mid-level *	1.250 ± 0.557	1.212 ± 0.493	0.037 ± 0.287
Basal *	1.963 ± 1.192	1.913 ± 1.168	0.050 ± 0.352
Std. deviation of JI between phases †	0.051	0.049	0.002
Std. deviation of DI between phases *	0.035	0.034	0.002
Std. deviation of MHD between phases (mm) †	0.352	0.329	0.023
Kaggle database			
ES blood volume (absolute diff.) (mL) *	9.9 ± 9.1	11.4 ± 9.5	−1.5 ± 9.1
ES blood volume (signed diff.) (mL) *	+1.9 ± 13.4	+7.2 ± 13.0	−5.3 ± 9.8

ED blood volume (absolute diff.) (mL) *	26.8 ± 16.3	21.7 ± 16.5	5.1 ± 13.2
ED blood volume (signed diff.) (mL) *	−25.0 ± 19.0	−19.8 ± 18.8	5.2 ± 15.1
Continuous ranked probability score	0.0124	0.0122	0.0002

* Statistically significant difference at $p<0.001$ using paired *t*-test

† Statistically significant difference at $p<0.01$ using paired *t*-test

‡ Direction of comparison chosen such that positive values indicate improvement

TABLE 4. Comparison of results between our proposed algorithm and other published techniques. AU (14), AO (26), SCR (7), DS, and INR (12) values are taken from Table 2 of (17). FCN values are taken from Table 3 of (15). Values are of mean (standard deviation).

Method	**Manual input**	**Jaccard Index**
AU	Interactive 4D guide point placement	.84 (.17)
CNR (this paper)	**None**	**.77 (.11)**
FCN	None	.74 (.13)
AO	Delineate first frame	.74 (.16)
SCR	None	.69 (.23)
DS	Delineate first frame	.64 (.18)
INR	None	.43 (.10)

TABLE 5. Independent error analysis of the CTR and MB networks. For the CTR network, the fractional error was determined by normalizing the absolute error against the average endocardium radius. For the MB network, the rightmost column is the end-to-end network error analysis reproduced from Table 3 for convenience of comparison.

CTR network	**Independent Error**	**End-to-End**
Absolute error (mm) (fractional error)	1.782 ± 1.271 (0.083)	
Apical	2.108 ± 1.987 (0.150)	
Mid-level	1.704 ± 1.053 (0.071)	
Basal	2.086 ± 1.622 (0.092)	
MB network		
Jaccard index (JI)	0.771 ± 0.108	0.769 ± 0.109
Dice index (DI)	0.866 ± 0.079	0.864 ± 0.080
Modified Hausdorff distance (MHD) (mm)	1.310 ± 0.621	1.329 ± 0.710
Apical	1.692 ± 0.710	1.769 ± 1.126
Mid-level	1.201 ± 0.489	1.212 ± 0.493
Basal	1.930 ± 1.142	1.913 ± 1.168

Figure Legends

FIGURE 1. Overview of the segmentation system. Each row represents one of the three independent networks (LM, CTR, and MB). Columns illustrate the left-to-right sequential flow of the system: initial source image → pre-processed input images → neural network inference → evaluated output. The network imagery in the third column is representative for the purpose of illustration, refer to Table 2 for the detailed network architecture.

FIGURE 2. A small improvement in cross-validation loss is seen with the addition of random FOV reduction when training the CTR network. Dotted and solid lines are the absolute error averaged over 500 iterations. Each error bar indicates the 10^{th} and 90^{th} percentiles.

FIGURE 3. Representative segmentation result from the LVSC validation dataset. (Top to bottom) Representative slices from apex to base. (Left to Right) Representative cardiac phases from diastole to systole. Red and green contours are endo- and epicardium results from the MB network respectively.

FIGURE 4. Segmentation quality as a function of fractional slice position along LV apex (zero) to base (one). Performance is strongest in the mid-LV, with noticeable drop-offs

towards the apex and base ends. In addition, towards the apex and base ends there is a lack of data in the consensus gold standard for evaluation.

FIGURE 5. Sample images from LVSC validation dataset demonstrating improved stability due to field-of-view (FOV) reduction for apical slices. (Top row) Results from previous work (16), (bottom row) results for current paper. Red and green contours are endo- and epicardial contours resulted from MB network. Bottom row dark blue tint areas indicate the FOV reduction crop masks.

Figures

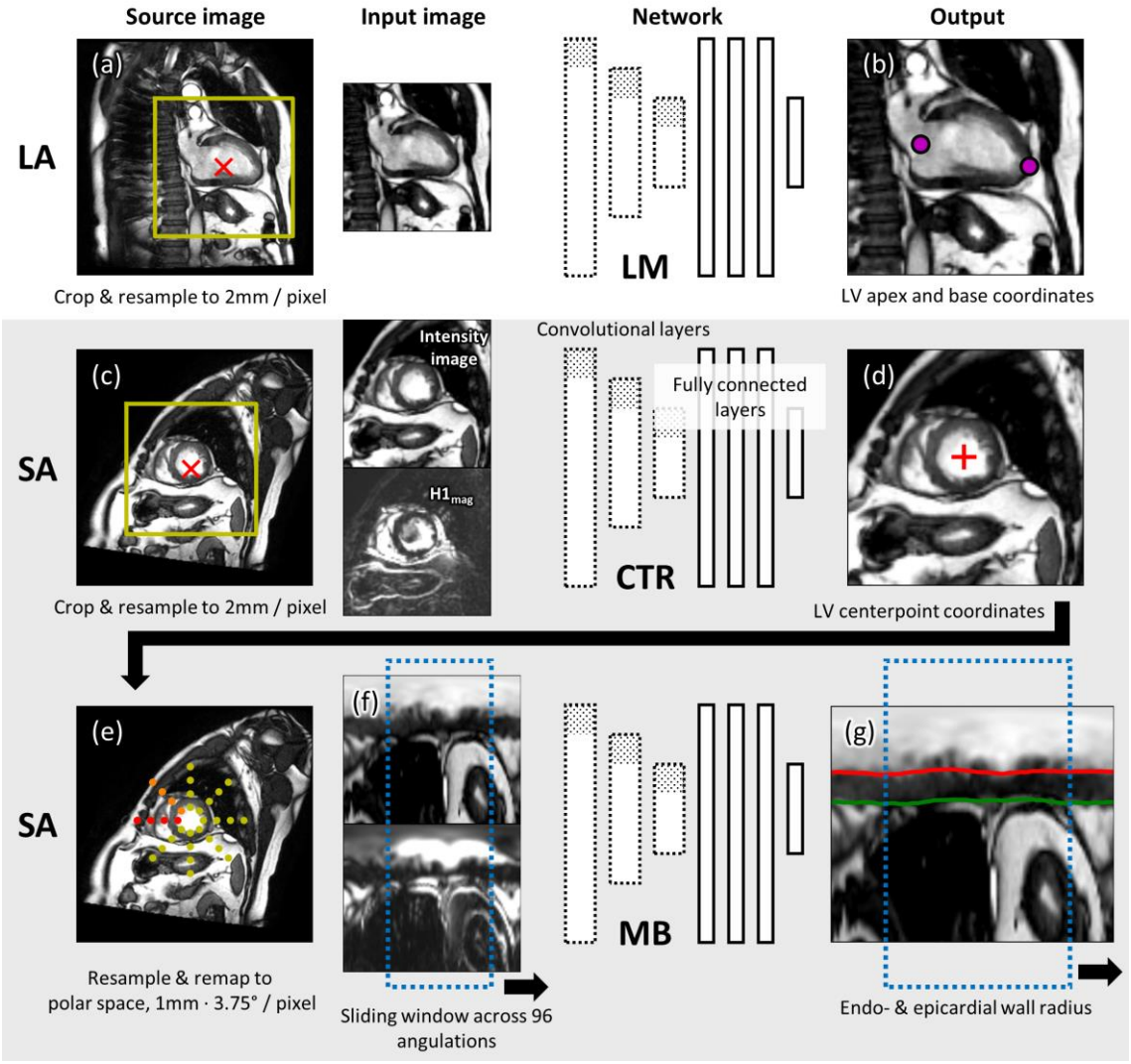

FIGURE 1. Overview of the segmentation system. Each row represents one of the three independent networks (LM, CTR, and MB). Columns illustrate the left-to-right sequential flow of the system: initial source image → pre-processed input images → neural network inference → evaluated output. The network imagery in the third column is representative for the purpose of illustration, refer to Table 2 for the detailed network architecture.

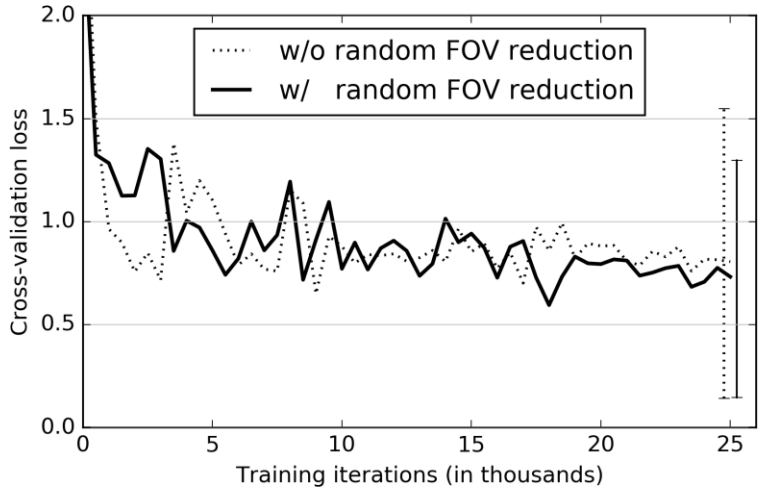

FIGURE 2. A small improvement in cross-validation loss is seen with the addition of random FOV reduction when training the CTR network. Dotted and solid lines are the absolute error averaged over 500 iterations. Each error bar indicates the 10^{th} and 90^{th} percentiles.

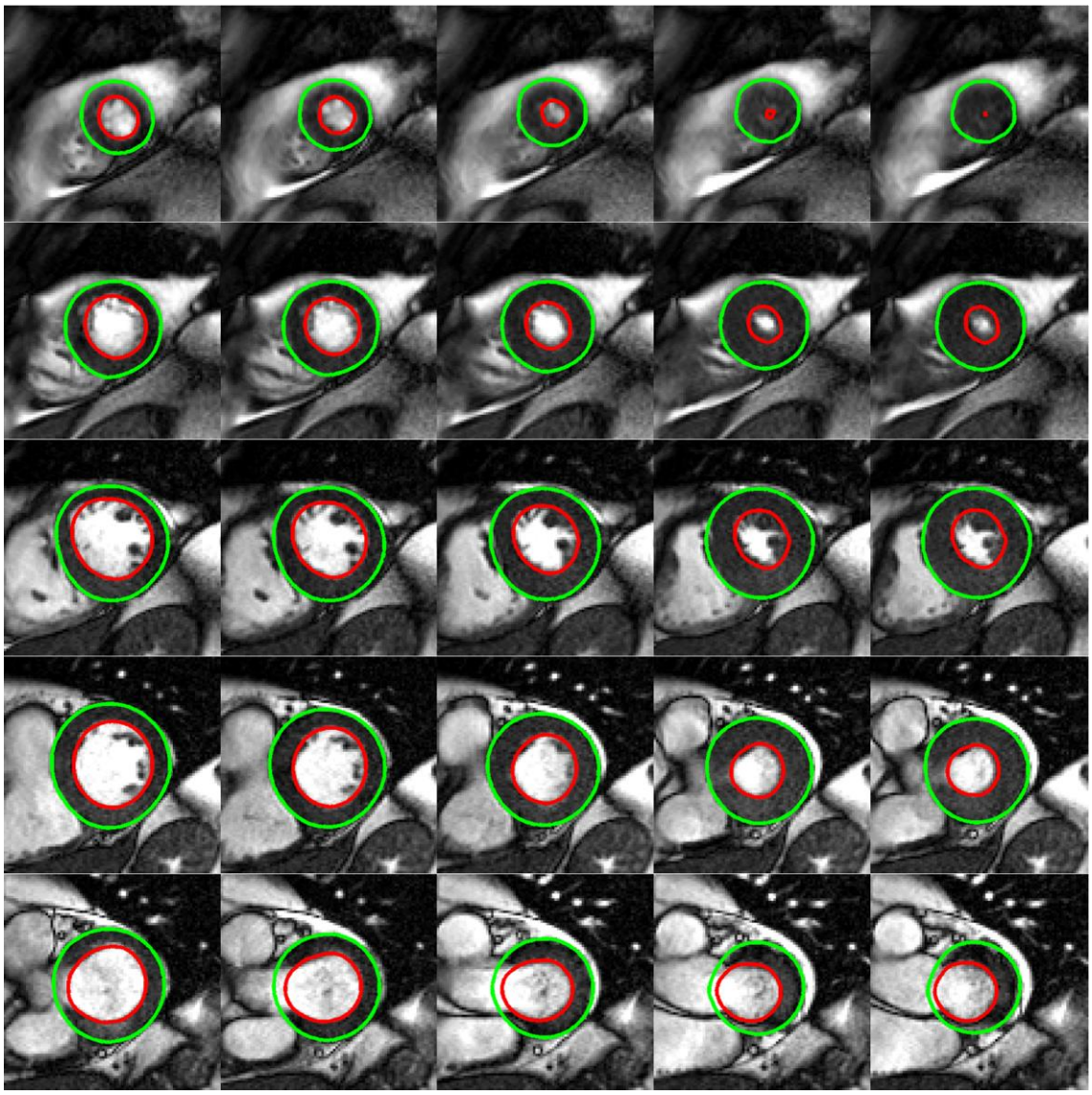

FIGURE 3. Representative segmentation result from the LVSC validation dataset. (Top to bottom) Representative slices from apex to base. (Left to Right) Representative cardiac phases from diastole to systole. Red and green contours are endo- and epicardium results from the MB network respectively.

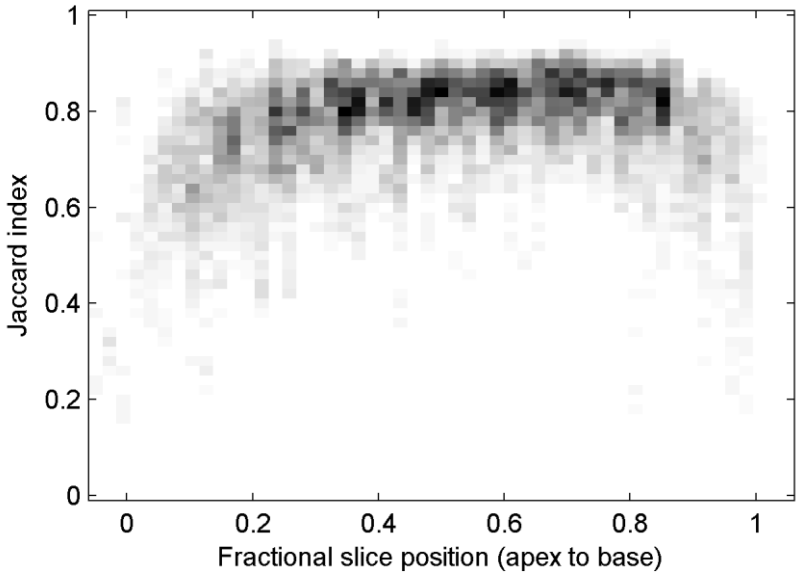

FIGURE 4. Segmentation quality as a function of fractional slice position along LV apex (zero) to base (one). Performance is strongest in the mid-LV, with noticeable drop-offs towards the apex and base ends. In addition, towards the apex and base ends there is a lack of data in the consensus gold standard for evaluation.

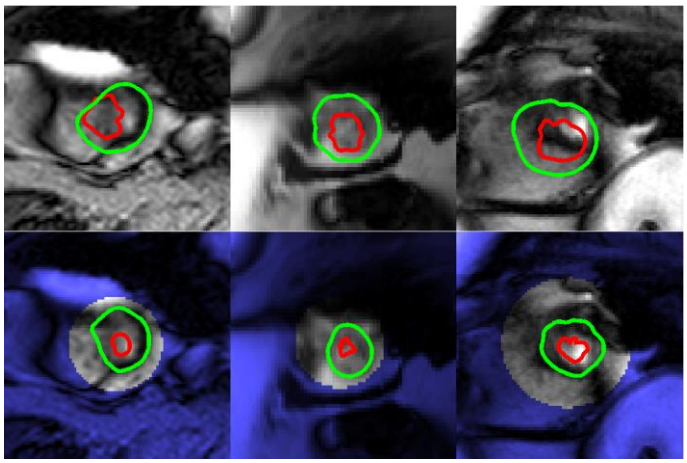

FIGURE 5. Sample images from LVSC validation dataset demonstrating improved stability due to field-of-view (FOV) reduction for apical slices. (Top row) Results from previous work (16), (bottom row) results for current paper. Red and green contours are endo- and epicardial contours resulted from MB network. Bottom row dark blue tint areas indicate the FOV reduction crop masks.