Medical Image Analysis 60 (2020) 101595

Contents lists available at ScienceDirect

Medical Image Analysis

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

Atrial scar quantification via multi-scale CNN in the graph-cuts framework

Lei Li[a,b], Fuping Wu[b,c], Guang Yang[d,e], Lingchao Xu[f], Tom Wong[e], Raad Mohiaddin[d,e], David Firmin[d,e], Jennifer Keegan[d,e], Xiahai Zhuang[b,g,*]

[a] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[b] School of Data Science, Fudan University, Shanghai, China
[c] Dept of Statistics, School of Management, Fudan University, Shanghai, China
[d] National Heart and Lung Institute, Imperial College London, London, UK
[e] Cardiovascular Research Center, Royal Brompton Hospital, London, UK
[f] School of NAOCE, Shanghai Jiao Tong University, Shanghai, China
[g] Fudan-Xinzailing Joint Research Center for Big Data, Fudan University, Shanghai, China

ARTICLE INFO

Article history:
Received 24 December 2018
Revised 5 June 2019
Accepted 26 October 2019
Available online 16 November 2019

Keywords:
Atrial fibrillation
Left atrium
LGE MRI
Scar segmentation
Graph learning
Multi-scale CNN

ABSTRACT

Late gadolinium enhancement magnetic resonance imaging (LGE MRI) appears to be a promising alternative for scar assessment in patients with atrial fibrillation (AF). Automating the quantification and analysis of atrial scars can be challenging due to the low image quality. In this work, we propose a fully automated method based on the graph-cuts framework, where the potentials of the graph are learned on a surface mesh of the left atrium (LA) using a multi-scale convolutional neural network (MS-CNN). For validation, we have included fifty-eight images with manual delineations. MS-CNN, which can efficiently incorporate both the local and global texture information of the images, has been shown to evidently improve the segmentation accuracy of the proposed graph-cuts based method. The segmentation could be further improved when the contribution between the t-link and n-link weights of the graph is balanced. The proposed method achieves a mean accuracy of 0.856 ± 0.033 and mean Dice score of 0.702 ± 0.071 for LA scar quantification. Compared to the conventional methods, which are based on the manual delineation of LA for initialization, our method is fully automatic and has demonstrated significantly better Dice score and accuracy ($p < 0.01$). The method is promising and can be potentially useful in diagnosis and prognosis of AF.

1. Introduction

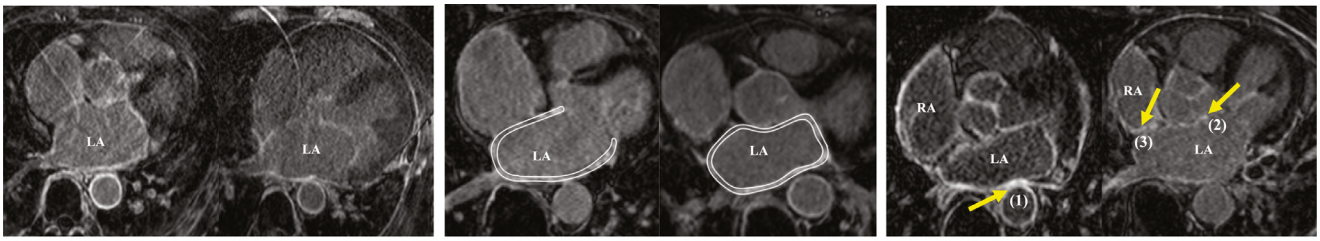
Atrial fibrillation (AF) is the most common arrhythmia observed in clinical practice, occurring in up to 1% of the population and rising fast with advancing age (Chugh et al., 2013). Radiofrequency catheter ablation using the pulmonary vein (PV) isolation technique has emerged as one of the most common methods for the treatment of AF patients (Wilber et al., 2010; Calkins et al., 2012). Quantification of atrial scars is potentially beneficial in selecting candidates and guiding ablation treatment. Late gadolinium enhancement magnetic resonance imaging (LGE MRI) is a promising technique to visualize and quantify the atrial scars (Vergara and Marrouche, 2011). Many clinical studies mainly focus on the location and extent of scarring areas of the left atrium (LA) myocardium (McGann et al., 2008; Vergara et al., 2011; Badger et al., 2010).

Automatic delineation of scars from LGE MRI is still challenging due to various reasons. First, the image quality of LGE MRI could be poor. Second, the prior model of scars is hard to construct on account of the various LA shapes, the thin wall (mean thickness of 1.89 ± 0.48 mm reported by Beinart et al., 2011), the surrounding enhanced regions and the complex patterns of scars in AF patients. Fig. 1 illustrates and explains the challenges in more details. To the best of our knowledge, little work has been reported in the literature to achieve the fully automatic quantification of LA scars from LGE MRI.

The most widespread methods for atrial scar segmentation are mainly based on thresholding (Badger et al., 2010; Karim et al., 2013). Pontecorboli et al. (2016) provided an overall review of scar

* Corresponding author.
E-mail address: zxh@fudan.edu.cn (X. Zhuang).

https://doi.org/10.1016/j.media.2019.101595



(a) (b) (c)

Fig. 1. The challenges of automatic delineation of scars from LGE MRI: (a) two typical LGE MRIs with poor quality; (b) thin atrial walls highlighted using bright white color in the figure; (c) surrounding enhanced regions pointed out by the arrows, where (1) and (2) respectively indicate the enhanced walls of descending and ascending aorta, (3) denotes the enhanced walls of right atrium.

segmentation using various threshold techniques. For these methods, an appropriate threshold value is decisive, but setting this value can be subjective, eventually limiting the applicability and reproducibility. Perry et al. (2012) proposed to use k-means clustering to classify the normal and fibrosis tissue from manually segmented LA walls. Karim et al. (2014) combined the scar intensity priors and Gaussian mixture model (GMM) to construct a cost function for scar segmentation, which was achieved by an optimization using the graph-cuts framework. Yang et al. (2018) employed the super-pixel method and support vector machine (SVM) to segment the atrial scars.

Most of the reported methods rely on manual segmentation of the LA or LA walls to provide an accurate initialization. In ISBI 2012 challenge (Karim et al., 2013), manual segmentation of LA was provided. There was large variance in terms of segmentation accuracy, especially for the pre-ablation cases, and the teams using manually delineated LA walls generally obtained much better performance than those using fully automatic approaches in the challenge. Their benchmark study emphasizes the importance of an accurate initialization.

For LA segmentation, Ravanelli et al. (2014) proposed a method using threshold for initialization, followed by the 3D fast marching for segmentation. They required manual correction from the clinicians to achieve reliable performance. Tao et al. (2016) combined LGE MRI with another MRI sequence with better anatomical information to segment the LA, and the combined segmentation achieved better results than the method solely using LGE MRI. Xiong et al. (2018) proposed a dual fully convolutional neural network for LA segmentation from LGE MRI with promising results. Later, they organized a LA segmentation challenge in MICCAI 2018 (Zhao and Xiong, 2018). For LA wall segmentation, Veni et al. (2017) proposed an algorithm named ShapeCut, combining a shape-based system and the graph-cuts approach to make a Bayesian dual surface estimation. Ji et al. (2018) applied the advanced two-layer level set with a soft distance constraint for dual surface segmentation of LA and left ventricle (LV) walls. Their method was 2D-based and required a manual initialization of the endocardial boundaries. Karim et al. (2018) provided a benchmark dataset for LA wall segmentation. For CT data, three algorithms were evaluated, including the marker-controlled geodesic active contours, level-set, and blood pool mesh vertex normal traversal method. For MRI data, the level-set, region growing, and watershed algorithms were studied.

Recently, deep-learning based algorithms have been successfully applied to the LV myocardial segmentation. Xu et al. (2018) proposed an end-to-end framework, named as OP-RNN, to segment myocardial infarction from cine cardiac MRI without contrast agents. Moccia et al. (2019) employed a fully convolutional neural network to segment scars of LV, assisted by an initialization of manually segmented myocardium. Lau et al. (2018) utilized a chained generative adversarial network, referred to as ScarGAN, to simulate scar tissues on LGE MRI of healthy patients for data augmentation.

In summary, in previous studies atrial scar quantification relies on an accurate segmentation of the LA or LA walls for initialization, but automating this segmentation is still an open question. In this work, we propose a fully automatic method for LA scar quantification and analysis, without the requirement of an accurate LA segmentation.

Firstly, we propose to perform scar quantification on a surface, onto which the LA endocardium is projected. We neglect the thickness of LA walls, because the clinical studies are generally performed by projecting the scars onto the LA endocardial surface for visualization (Peters et al., 2007; Knowles et al., 2010; Ravanelli et al., 2014). In this framework, we represent the surface using a graph, and formulate the classification as an energy minimization problem which can be solved by graph-cuts. We further propose to explicitly learn the edge weights of the graph, i.e., n-link and t-link potentials (Boykov and Jolly, 2001). This is achieved by a convolutional neural network (CNN), which learns features from the images (Krizhevsky et al., 2012). Here, we do not directly compute these weights solely based on the intensity similarity, as the conventional graph-cuts methods do. This is because the enhancement patterns in LGE MRI are complex and can vary greatly across different patients, leading to inconsistent intensity patterns. Also, currently the automatic methods could have a few millimeters under or over segmentation, leading to the estimated endocardial surface being misaligned to the ground truth. The proposed CNN scheme can exploit both image features and spatial context by means of neighborhood information, to provide more accurate estimation of the graph weights. We finally obtain the classification based on the graph-cuts framework, which mitigates the effect of misalignments of the endocardial surface due to automatic LA segmentation errors. Note that the graph-cuts in this work was not embedded into the network for an end-to-end training, which could be computational inefficient (please c.f. Section 4 for details).

Furthermore, we propose to employ the multi-scale patch (MSP) strategy (Zhuang and Shen, 2016), and combine it with the CNN for graph potential learning. This is because distinguishing scars can be challenging solely based on local texture information, particularly in the area where the LA wall is surrounded by other enhanced regions, such as the fibrosis of mitral valve, aortic wall and right atrial (RA) wall. The MSP method is developed from the image scale space theory, which can handle information at different levels within a limited window and has been widely applied to the tasks of feature extraction, detection and image matching (Lorensen and Cline, 1987). The MSP strategy can incorporate both the local fine texture features and the global structural information into the CNN architecture. We refer to such MSP-based CNN as multi-scale CNN, i.e. MS-CNN. In addition, the MSPs are extracted with random offsets along the perpendicular direction of the LA endocardial surface, simulating the misalignments between the au-

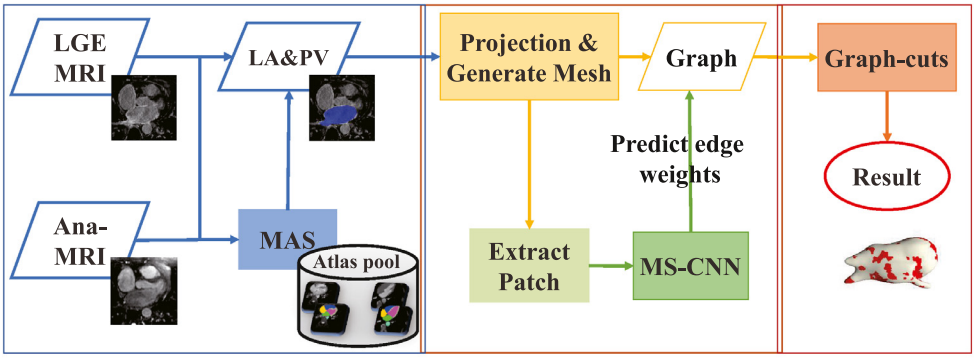


Fig. 2. Flowchart of the proposed framework for LA scar quantification and analysis.

tomatically segmented LA surface and the ground truth. Therefore, such patches not only can model the multi-scale texture patterns of the images, but also can further improve the robustness of the proposed method against the LA segmentation errors.

The remainder of the paper is organized as follows: the detailed framework of the proposed algorithm is presented in Section 2. Section 3 presents the experiments and results. Discussion and conclusion are given in Section 4.

2. Method

Fig. 2 provides an overview of the proposed framework. First, we use a well-developed multi-atlas whole heart segmentation (MA-WHS) to obtain an initial segmentation of the LA (see Section 2.1). Then, we project the LA endocardium to generate a surface mesh, where the quantification is performed (see Section 2.2). The labeling of scars is achieved by optimizing a cost function based on the graph-cuts framework (see Section 2.3), whose potentials for edge weights are explicitly learned by the proposed MS-CNN (see Section 2.4). Note that the graph-cuts based classification is performed on the surface mesh. This can avoid the challenging segmentation of thin LA wall and can also greatly reduce the computational cost. At the same time, both the texture and anatomical features of the LA myocardium can be adequately extracted by employing the MSP strategy. Thus, the features of the nodes in the graph are represented by a set of MSPs, and the potentials are learned and predicted by the MS-CNN.

2.1. Initialization of atrial endocardium and pulmonary veins

We use MA-WHS, which is based on multi-atlas segmentation (MAS), to obtain the geometrical information of the LA. This is because the LGE MRI covers the whole heart, and MA-WHS has been well developed and applied in recent years (Zhuang and Shen, 2016; Yang et al., 2018). MAS algorithm segments an unknown target image by propagating and fusing the labels from multiple annotated atlases using registration. As the LGE MRI could have relatively poor image quality, we first apply MA-WHS on the anatomical MRI (Ana-MRI), and then propagate the segmentation using affine registration from the Ana-MRI to LGE MRI. The Ana-MRI is normally acquired in the same MRI examination as LGE MRI, using the b-SSFP sequence, which generates higher quality images for atlas-based segmentation.

Having finished the WHS for LA and PV delineation, the marching cubes algorithm (Lorensen and Cline, 1987) is then used to obtain a surface mesh of the LA endocardium which excludes the mitral valve. Note that the LA segmentation is generally reliable, but still contains errors leading to misalignments between the extracted surface mesh and the ground truth. For example, the mean Dice score of our MA-WHS for LA is 0.898 $\pm$ 0.044 (please c.f. Section 3.3 for details). However, the effect of inaccurate LA segmentation can be minimized by using the projection strategy and the MS-CNN learning coupled with the randomly shifted MSP sampling strategy. The reader is referred to Fig. 3 for illustration and following methodology sections for details.

2.2. Projection of the atrial endocardium

We project the LA endocardium onto a surface mesh, and then the atrial scars can be classified on a graphical surface. This is because the clinical demands for scar quantification in AF patients mainly concern the location and extent of scarring areas (Ravanelli et al., 2014). Williams et al. (2017) proposed a method to simultaneously represent multiple parameters on a surface model based on the template of an average LA mesh. By projection, both the errors due to LA wall thickness and misregistration of the WHS can be mitigated. At the same time, the computational complexity of the algorithm can be reduced dramatically.

The endocardial surface is generated from the volumetric binary segmentation result of the LA cavity using the marching cubes algorithm (Lorensen and Cline, 1987). The resolution of the surface mesh is denser than the resolution of the image, which protects the small scars. The projection from the LA geometry to the surface mesh can preserve the geodesic distances between two nodes. This equidistant projection is required due to the definition of n-link weights in the proposed graph-cuts framework. In this formulation, each vertex on the surface, i.e., node of the graph, should include a profile that represents the texture information of the corresponding location in the LGE MRI. Here, we represent this profile using MSPs, which can incorporate both global structural features and local texture information.

2.3. Graph formulation for scar segmentation

Classification and quantification of scars on the LA surface can be formulated as an energy minimization problem solved via graph-cuts. The weights of the graph come from two parts, i.e., the regional term E_R and the boundary term E_B (Boykov and Jolly, 2001). The regional term encodes the intensity distributions of different classes, and the boundary term maintains the continuity between neighbors.

Let $G = \{\mathcal{X}, \mathcal{N}\}$ denotes a graph, where $\mathcal{X} = \{x_i\}$ indicates the set of graph nodes, and $\mathcal{N} = \{< x_i, x_j >\}$ is the set of edges. Here, the weights of edges connecting graph nodes to the terminals are known as t-link weight, and the weights of edges connecting neighboring nodes are referred to as n-link weight (Boykov and Jolly, 2001). The two terminals respectively denote the scars and normal myocardium in our problem, analogous to the foreground and background of the general image segmentation task. Let $l_{x_i} \in \{0, 1\}$ be the label assigned to x_i, and $l = \{l_{x_i} | x_i \in \mathcal{X}\}$ be the label

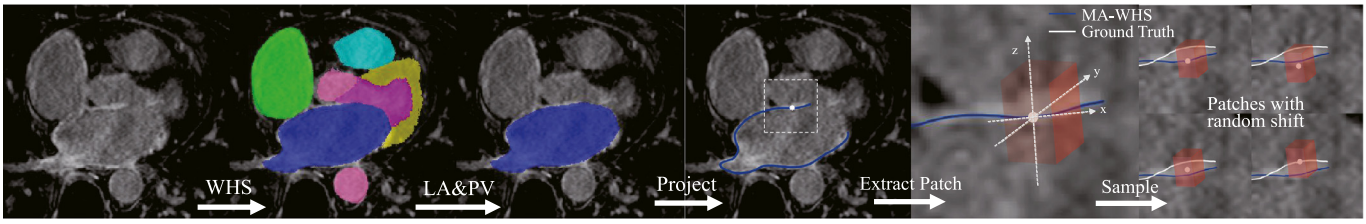


Fig. 3. Pipeline of the projection and patch extraction phases: (a) target LGE MRI; (b) WHS result of LGE MRI propagated from Ana-MRI; (c) extracted LA and PV from WHS; (d) LA endocardial surface after projection; (e) patch along the normal direction of the LA endocardial surface; (f) patches with random shift in the training phase.

vector that defines a segmentation. The segmentation energy is defined as follows,

$$E(l) = E_R(l) + \lambda E_B(l) = \sum_{x_i \in \mathcal{X}} W_{x_i}^{t\text{-}link}(l_{x_i}) + \lambda \sum_{(x_i,x_j)\in\mathcal{N}} W_{\{x_i,x_j\}}^{n\text{-}link}(l_{x_i}, l_{x_j}), \quad (1)$$

where $W_{x_i}^{t\text{-}link}$ and $W_{\{x_i,x_j\}}^{n\text{-}link}$ are respectively the t-link and n-link weight, and λ is a balancing parameter.

In conventional graph-based segmentation, the regional term is generally obtained by optimizing based on a manual defined initial model. For example, Boykov and Jolly (2001) manually selected a number of seed points to construct such model, referred to as graph cuts method, and Rother et al. (2004) manually defined a bounding box for interactive segmentation, known as GrabCut approach. The boundary term in these works was normally defined according to the dissimilarity of intensity and distance between two connected nodes. Veni et al. (2017) designed a regional term based on a generative image model incorporating both local and global shape priors. The boundary term was defined for regularizing the smoothness of the estimated surface, i.e., minimizing the squared difference of the offsets between neighboring vertices. Lu et al. (2017) estimated a regional term combining three maps, including a probability map, a thresholding map and a local appearance map. The boundary term they defined was related to the intensity difference and distance of two connected nodes.

In this work, we propose to directly learn and predict the t/n-link potentials for the regional and boundary terms. This is different from the conventional means, where the profile of a graph node is commonly represented by the intensity of a single pixel or its local texture, which consists of limited information. In contrast, we combine the profile representation of graph nodes with the MSP strategy, and learn the potentials using the proposed MS-CNN. Fig. 4 illustrates the flowchart of constructing the graph.

2.4. Explicit learning of graph potentials using MS-CNN

Fig. 4 illustrates the computation of graph potentials for the graph-cuts based classification of LA scars.

2.4.1. Multi-scale patch and patch extraction

We propose to extract MSPs from LGE MRI to represent the profile of the graph node, and to feed the MS-CNN for training and prediction. MSP can represent different levels of structural information at a location in an image, with low scale capturing local fine details and high scale providing global structural information of the image (Zhuang and Shen, 2016).

Each graph node x_i has its associated MSPs, denoted as $\mathcal{P}_i = \{p_{x_i}^0, p_{x_i}^1, \ldots, p_{x_i}^{N_s-1}\}$, where N_s indicates the number of scales. These patches are extracted from the corresponding volumetric region in the LGE MRI, by back projecting the node to the position in the image. They are elongate-shaped and are defined along the normal direction of the LA endocardial surface, as Fig. 3(e) shows, and their local orientations are maximally aligned to the common world coordinate system of the LGE MRI. The multi-scale strategy is implemented by adjusting the sample spacing to generate patches with different scales, corresponding to different resolutions of the LGE MRI. We employ parallel convolutional pathways for multi-scale processing, to feed the different scale information of images to the neural network simultaneously, as Fig. 5(a) shows.

2.4.2. Multi-scale convolutional neural network

We have two neural networks, i.e., *T*-NET and *N*-NET. *T*-NET learns and predicts the t-link potentials, i.e., the probabilities of a node belonging to scars and normal walls respectively, as Fig. 5(a) shows. *N*-NET calculates the n-link potential between two connected nodes, as Fig. 5(b) shows. The sub-network for extracting patch features in *T*-NET and *N*-NET is referred to as *Patch*-NET, as Fig. 5(c) shows.

For training of the t-link potentials, we define a sample for each node of a graph constructed from LGE MRI. The sample is composed of the MSPs associated to the node x_i, and its label probability L_i generated from the ground truth label. As Fig. 5(a) shows, the training data of *T*-NET can be represented as $\mathcal{D}^{\mathbf{T}} = [(\mathcal{P}_1, L_1), \ldots, (\mathcal{P}_N, L_N)]$, i.e., N nodes with corresponding labels. Thus, the *T*-NET can be parameterized by $\theta^{\mathbf{T}}$ as follows,

$$\hat{\theta}^{\mathbf{T}} = \arg\min_{\theta^{\mathbf{T}}} \sum_{i=1}^{N} (\hat{L}(\mathcal{P}_i; \theta^{\mathbf{T}}) - L_i)^2, \quad (2)$$

where $\mathcal{P}_i = \{p_{x_i}^0, p_{x_i}^1, p_{x_i}^2\}$, and $\hat{L}$ is the estimated t-link weight.

For training of the n-link, we define a sample for each pair of two neighboring nodes $\{x_i, x_j\}$, consisting of three elements, i.e., (1) the pair of the two sets of MSPs associated with the two nodes, i.e. $\{\mathcal{P}_i, \mathcal{P}_j\}$, (2) the geodesic distance between them, denoted as d_{ij}, which is computed using the length of the shortest path along the edges connecting them, and (3) their ground truth label similarity M_{ij}, defined as $L_i \times L_j + (1 - L_i) \times (1 - L_j)$.

As Fig. 5(b) shows, the training data of *N*-NET can be represented as $\mathcal{D}^{\mathbf{N}} = [(\mathcal{P}_i, \mathcal{P}_j, d_{ij}, M_{ij})]_{i,j=1}^{i,j=N}$. The distance d_{ij} is viewed as an additional similarity feature, namely the labeling of two nodes can be more similar if they are closer. To this end, we design a sub-network, denoted as $\vec{\mathbb{F}}$, to extract high-level and dense features, i.e. $\vec{\mathbb{F}}(\mathcal{P})$. We then obtain a new feature vector from $\mathcal{P}_i$ and $\mathcal{P}_j$, as follows,

$$\vec{\mathbb{G}}_{ij} = \vec{\mathbb{F}}(\mathcal{P}_i) \times \vec{\mathbb{F}}(\mathcal{P}_j) + (1 - \vec{\mathbb{F}}(\mathcal{P}_i)) \times (1 - \vec{\mathbb{F}}(\mathcal{P}_j)). \quad (3)$$

Each element of $\vec{\mathbb{G}}_{ij}$ can be considered as a similarity metric in the feature space. Finally, we combine $\vec{\mathbb{G}}_{ij}$ and d_{ij}, and feed them to another sub-network for computing the label similarity, i.e., the n-link weight. Thus, the *N*-NET can be parameterized by $\theta^{\mathbf{N}}$ as follows,

$$\hat{\theta}^{\mathbf{N}} = \arg\min_{\theta^{\mathbf{N}}} \sum_{i,j=1}^{N} (\hat{M}(\vec{\mathbb{G}}_{ij}, d_{ij}; \theta^{\mathbf{N}}) - M_{ij})^2, \quad (4)$$

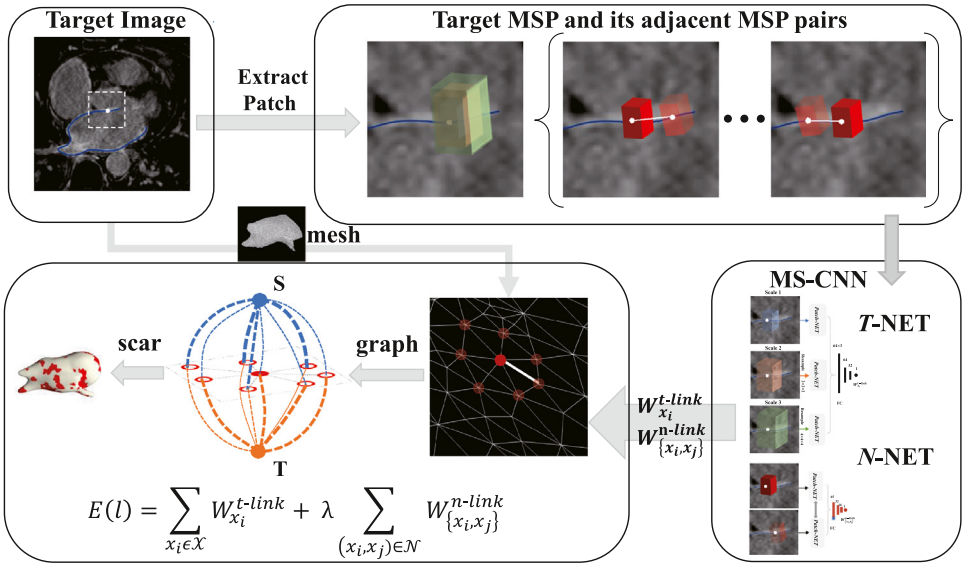


Fig. 4. Construction of the graph and the explicit learning of the graph potentials by MS-CNN (MSPs are integrated and represented using red cuboid in this work). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

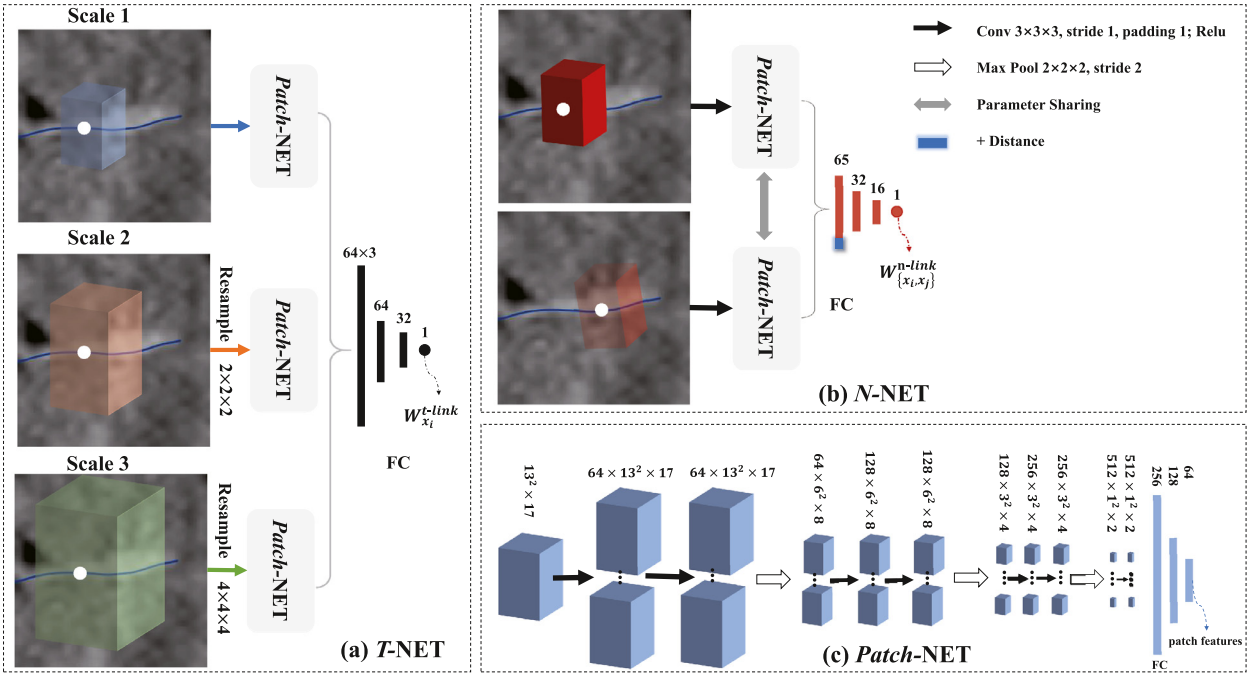


Fig. 5. The hierarchical architecture of the networks: (a) *T*-NET: the input is the MSPs of the target node, and the output is the predicted t-link weight; (b) *N*-NET: the input is the patch pair of two neighbor nodes, and the output is the predicted n-link weight; (c) *Patch*-NET: the input is a patch, and the output is the patch features. Note that the diagram takes 13 × 13 × 17 patch as an example.

where $\hat{M}$ is the estimated n-link weight.

As Fig. 5(c) shows, in *Patch*-NET the convolution layers with kernel size 3 × 3 × 3 use 1 pixel stride and 1 pixel width zero padding, and the max pooling layers use a stride equal to the pooling size, i.e., 2. The number of parameters is about 7.27M in the *Patch*-NET, 21.83M in the *T*-NET, and 21.82M in the *N*-NET.

2.4.3. Training and testing strategy

In the training phase, we use weighted sampling to mitigate the problem of class imbalance in the training set, where the number of the nodes belonging to normal myocardium in a subject could be tens or even hundreds times more than that of scars. Hence, instead of extracting the patches of all nodes for training, we first count the total number of the nodes on scars and scar boundaries, and then randomly select the similar number of nodes from the background to deal with the imbalanced training problem. In addition, we add a random shift, along with the normal direction, to the center of the MSPs, to mitigate the effects from the inaccurate delineation of the LA boundaries due to over or under segmentation. This is illustrated in Fig. 3(d)–(f). This shift should be large enough to overcome potential segmentation errors, while at the same time be small enough to avoid being too distant and cannot capture the texture profile of the LA wall. We propose to assign this random value in a given range, i.e. $\gamma \in (-\mathcal{R}, +\mathcal{R})$, to a node

in the training phase, where γ is the shift value, - and + represent being inside and outside of the LA blood cavity, respectively. Note that this random shift strategy is not needed in the testing phase.

In the testing phase, one can compute the t-link and n-link potentials of the graph, and the classification of scars on the LA surface can be achieved by embedding these estimated weights into the graph-cuts framework, i.e.,

$$W_{x_i}^{t\text{-}link} = \hat{L}(\mathcal{P}_i^{\gamma}; \theta^{\mathbf{T}}), \tag{5}$$

and,

$$W_{\{x_i,x_j\}}^{n\text{-}link} = \hat{M}(\mathcal{P}_i^{\gamma}, \mathcal{P}_j^{\gamma}, d_{ij}; \theta^{\mathbf{N}}) = \hat{M}(\vec{\mathbb{G}}_{ij}, d_{ij}; \theta^{\mathbf{N}}). \tag{6}$$

Note that the two normalized t-link weights of a node, respectively indicating the potentials to the foreground and background, can also be viewed as the probabilities of this node belonging to scars or normal tissues.

3. Experiments and results

3.1. Data acquisition and experimental setup

We collected fifty-eight post-ablation LGE MRI data from patients with longstanding persistent AF for experiments. Transverse navigator-gated 3D LGE MRI was performed on a 1.5T Siemens Magnetom Avanto scanner (Siemens Medical Systems, Erlangen, Germany), which used an inversion prepared segmented gradient echo sequence (TE/TR 2.2 ms/5.2 ms) 15 min after gadolinium administration. The LGE MRI data were acquired at resolution of (1.4-1.5) × (1.4-1.5) × 4 mm, and reconstructed to (0.7-0.75) × (0.7-0.75) × 2 mm. For each patient, prior to contrast agent administration, coronal navigator-gated 3D b-SSFP (TE/TR 1 ms/2.3 ms) data were acquired, with acquisition resolution of (1.6-1.8) × (1.6-1.8) × 3.2 mm, and reconstructed to (0.8-0.9) × (0.8-0.9) × 1.6 mm. Both LGE MRI and b-SSFP data were acquired during free breathing with respiratory motion control (Keegan et al., 2014).

The available data were randomly divided into two sets, one for training (31 images) and the other for testing (27 images). *T*-NET was trained using stochastic gradient descent optimizer, with following hyper-parameters: momentum = 0.9, batch size = 50, weight decay = 10^{-4}, number of epochs = 15. The learning rate was initially set to 0.01, and had a stepped decay rate of 0.8 every 1000 iterations. The same configuration was used for *N*-NET except that its number of epochs was 10.

We first evaluated the accuracy of automatic segmentation of LA in Section 3.3. Then, we performed four parameter studies to verify the effects of the parameters and explore their optimal values. In Section 3.4.1, we investigated the influence of different patch sizes to the proposed framework using the single-scale CNN, and then compared the results with that of MS-CNN. In Section 3.4.2, we studied the proposed method with different values of the balancing parameter λ. Section 3.4.3 and 3.4.4 present the studies of random shift and multi-scale learning, respectively. The optimal parameters concluded from these studies were used for the proposed method, in comparisons with other methods, in Section 3.5. Finally, Section 3.6 reports the performance of the proposed method and results of the inter-observer study.

3.2. Gold standard and evaluation

All the LGE MRIs were manually segmented by an experienced physicist specialized in cardiac MRI, to label the enhanced atrial scarring regions, which are considered as ground truth in this work. To assess the scar classification results, we generated the ground truth reference by projecting the manually segmented scars onto the LA surface. With regard to different initializations, i.e., the manual (abbreviated as LA_M) and automatic (abbreviated as LA_{auto}) delineation of LA, two different ground truths, respectively referred to as GT_M and GT_{auto}, were generated for evaluation. In the comparison studies in Section 3.5, the fully automatic methods were evaluated using GT_{auto}, while the semi-automatic algorithms based on LA_M were evaluated using GT_M.

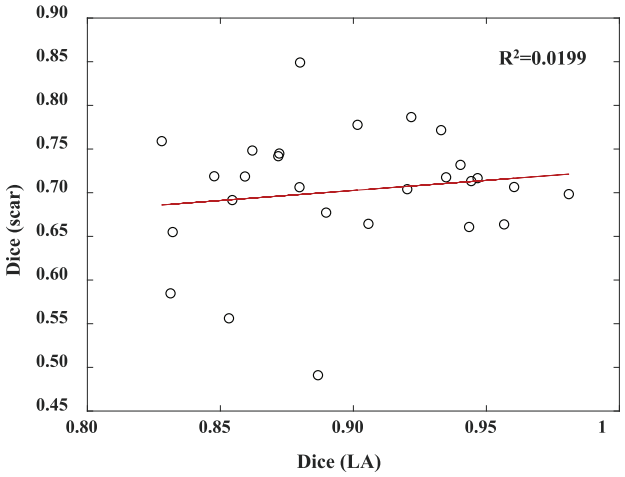


Fig. 6. Scatter point plot for analyzing the correlation between the LA segmentation performance and scar quantification accuracy, both indicated by Dice scores. The Pearson coefficient and Spearman's rank coefficient are respectively 0.1412 and 0.0110.

For evaluation, we computed the statistical measures, Dice score of scars, referred to as Dice (scar), and the generalized Dice score, denoted as *G*Dice. The statistical measures include accuracy, sensitivity and specificity. *G*Dice is a weighted Dice score by evaluating the segmentation of all labels (Crum et al., 2006; Zhuang, 2013), and is formulated as follows,

$$GDice = \frac{2\sum_{k=0}^{N_k-1}\left|S_k^{auto} \cap S_k^{manual}\right|}{\sum_{k=0}^{N_k-1}\left(\left|S_k^{auto}\right|\right) + \left(\left|S_k^{manual}\right|\right)}, \tag{7}$$

where S_k^{auto} and S_k^{manual} indicate the segmentation results of label k from the automatic method and manual delineation, respectively, and N_k is the number of labels. All the metrics are computed on the projected LA surface.

3.3. Automatic segmentation of LA and correlation analysis

To obtain an initialization of LA for scar segmentation, we developed the MA-WHS method using 30 b-SSFP MRI atlases. The 30 high resolution atlases were constructed from the Left Atrial Segmentation Challenge (STACOM 2013) (Tobon-Gomez et al., 2015). The manual delineation of LA was regarded as the gold standard for this experiment. The MA-WHS results of Ana-MRI were mapped to LGE MRI from the same subject, and then generated the initial LA labels. The average Dice score of this LA segmentation to the manual delineation was 0.898 ± 0.044.

To analyze the relation between the LA segmentation error and the scar quantification accuracy by the proposed method, we plotted these two values for each of the 27 test subjects as two dimension scatter points in Fig. 6. One can see that the plot shows little direct relationship between them. We further performed linear regression, Pearson correlation and Spearman's rank correlation. The R^2 and Pearson coefficient were respectively 0.0199 and 0.1412, indicating low linear correlation between Dice (scar) and Dice (LA); and the rank correlation coefficient was 0.0110, meaning hardly monotonic relationship between them either. To conclude, the result illustrates the low correlation between the scar

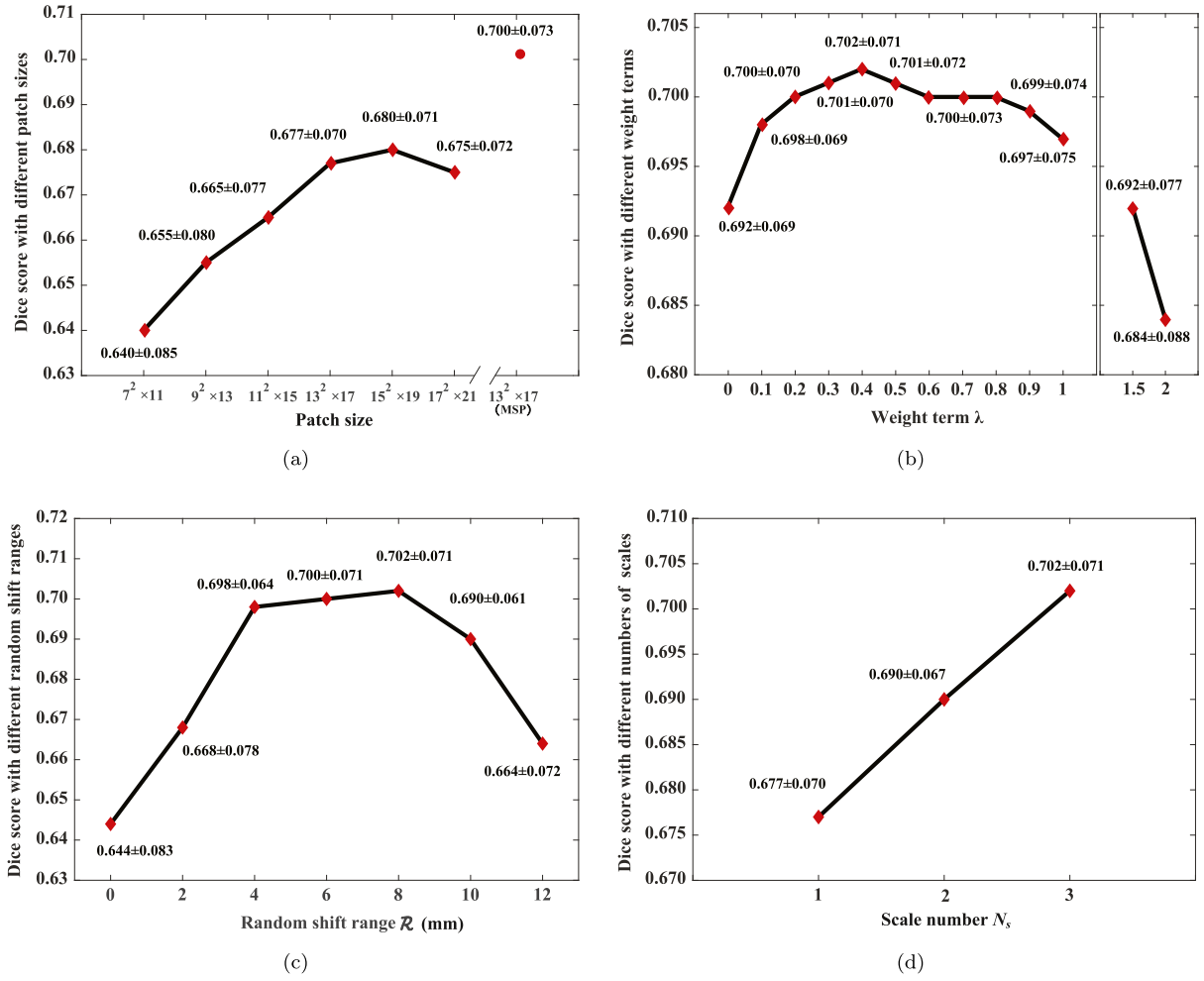


Fig. 7. Dice scores of the proposed method with different parameterizations: (a) performance against different patch sizes (λ=0.6); (b) performance against different values of the balancing parameter λ to weight the t-link and n-link terms in the graph-cuts framework; (c) performance against different random shift ranges $\mathcal{R}$; (d) performance against different numbers of scales N_s.

quantification accuracy and the LA segmentation accuracy by the proposed method.

3.4. Parameter studies

3.4.1. Study of the patch sizes

We used one-scale MSP, namely only the original image (scale 0) was used and the CNN was a single-scale network, for studying the proposed method with different sizes of patches. The patch sizes ranged from 7 × 7 × 11 to 17 × 17 × 21 voxel, where the voxel size is 1 × 1 × 1 mm. Then, we implemented the three-scale MSP and CNN with patch size 13 × 13 × 17 voxel, for comparisons with the single-scale CNNs. The balancing parameter λ in this study was set to 0.6, and the random shift range $\mathcal{R}$ was set to half of the patch length.

Fig. 7(a) shows that the average Dice score increases dramatically at first with respect to the increased sizes of patches, then starts to converge after the patch size reaching 13 × 13 × 17 voxel. This is reasonable, as the larger size is used, the richer intensity profile is included for feature training and detection. However, the increase of patch size generally requires more complex networks, either more kernels or more convolutional layers, which increases computation load and memory requirements. This also rationalizes our proposal to use MSP and MS-CNN. As Fig. 7(a) presents, our MS-CNN obviously increases the accuracy of the classification results, thanks to the usage of the MSP strategy which incorporates both local and global information of the images. In the following experiments, we adopted this three-scale setting (except for Section 3.4.4) and patch size of 13 × 13 × 17 voxel.

3.4.2. Study of the balancing parameter

In this study, we compared the results of the proposed scheme using different values for the balancing parameter λ, $\lambda \in [0, \infty)$, to demonstrate the effect of graph-cuts. Here, we set the values ranging from 0 to 2. The patch strategy was as follows, number of scales was three, patch size was 13 × 13 × 17 voxel, and the random shift range $\mathcal{R}$ was set to a maximum of 8 mm.

Fig. 7(b) presents the results. One can see that the best performance in terms of Dice score is obtained when λ is set to 0.4. This indicates that the inter-node relation (n-link) is important, and the weighting between the t-link and n-link terms should be balanced to achieve optimal performance. In the following experiments, λ was set to 0.4 for the proposed method.

3.4.3. Study of the random shift range

To demonstrate the effect of random shift, we compared the performance of the proposed method with different random shift ranges $\mathcal{R}$, for $\gamma \in (-\mathcal{R}, +\mathcal{R})$. Here, we set $\mathcal{R}$ ranging from 0 to 12 mm. The patch size was 13 × 13 × 17 voxel, and λ was set to 0.4.

Fig. 7(c) provides the results of this study. The best Dice score is obtained when $\mathcal{R}$ is set to 8 mm, i.e., half of the patch length in the long-axis direction, and the performance of the proposed method deteriorates when the shift range becomes larger than 8 mm. This is rational, because the shift range should cover all the potential misalignments of the constructed surface to the ground truth. When the random shift range $\mathcal{R}$ is greater than 8 mm, the patch may not cover the regions which include the important features for training and classification.

3.4.4. Study of the scales

To study the effect of multi-scale learning, we compared the results using different numbers of scales, i.e. $N_s = \{1, 2, 3\}$. The patch size of MSP was set to 13 × 13 × 17 voxel, λ was set to 0.4, and the random shift range $\mathcal{R}$ was set to a maximum of 8 mm.

Fig. 7(d) presents the mean Dice scores of the method. This study demonstrates that the effectiveness of the multi-scale learning. It indicates that the more scales we used the better accuracy we obtained. It should be noted that when we tried to use more scales, the training session failed, due to the limited computation capacity of our computer.

3.5. Comparison with other methods

In this study, we implemented nine segmentation approaches, including the proposed method, for comparisons. Here, LA_M indicates the methods adopt the manual segmentation of LA for initialization, and LA_{auto} denotes the methods employ the automatic segmentation from the MA-WHS approach described in Section 2.1.

(1) $LA_M + 2SD$: This is one of the most widespread thresholding methods to detect atrial scars. It calculates a specific number of standard deviation (SD) above a reference value. The reference value is generally set to the mean intensity from the blood pool or LA wall. It is however generally patient-specific and slice-specific, and different numbers of SD have been used (Karim et al., 2013). In our study, we obtained the optimal performance by setting the threshold value to 2 SD above the mean intensity of LA walls. Here, we constructed the LA wall from a manual segmentation of the LA with a morphological dilation, which was also used for the following experiments when the LA wall was needed from LA_M.

(2) $LA_M + Otsu$: This method uses the Otsu algorithm (Otsu, 1979) for automatic thresholding of the scarring tissues from the LA wall obtained from LA_M.

(3) $LA_M + MGMM$: This method adopts the multi-component Gaussian mixture model (MGMM) for scar segmentation from the LA wall (Liu et al., 2017). MGMM can deal with the intensity heterogeneity of myocardium caused by the infarcts, and has been proven to be effective in myocardium segmentation.

(4) $LA_M + MGMM + GC$: This method further regularizes the spatial continuity using the graph-cuts framework, based on the result of MGMM. Here, we defined the boundary weight using the intensity difference between neighboring points, and the regional weight was computed from the posterior probability map of scars generated from MGMM.

(5) $LA_M + MS\text{-}CNN$: This method employs a 2D U-Net architecture (Ronneberger et al., 2015) for scar segmentation. The U-Net is trained using a stochastic gradient descent optimizer, with following hyper-parameters: batch size=25, weight decay=10^{-4}, number of epochs=100. The learning rate is initially set to 0.01, and has a stepped decay rate of 0.95 every 1000 iterations. The input image is a 2D slice of LGE MRI cropped into 128 × 128 centered on LA_M.

(6) $LA_M + MS\text{-}CNN^0$: This learning based method only uses the two t-link weights estimated from *T*-NET to classify scars. The two weights, i.e. respectively linked to the foreground scar and background normal tissue, are normalized and considered as the posterior probability of the two labels. Here, both training data and test data were initialized using manually segmented LA, so the random shift in the training phase was set to zero, i.e. γ=0.

(7) $LA_{auto} + MS\text{-}CNN^0$: This method uses the estimated t-link weights from *T*-NET, similar to $LA_M + MS\text{-}CNN^0$, to classify scars. However, the LA here was automatically segmented using MA-WHS. For comparisons with $LA_M + MS\text{-}CNN^0$, here we also set the random shift to zero (γ=0).

(8) $LA_{auto} + MS\text{-}CNN$: Similarly, this method uses the estimated t-link weights from *T*-NET to classify scars, and the LA was automatically segmented using MA-WHS. However, in the training phase we set the random shift accordingly based on the parameter study in Section 3.4.3.

(9) $LA_{auto} + LearnGC$: This is the proposed method in which the LA was initialized by MA-WHS and the weights of the graph were learned and predicted using MS-CNN. Here, the balancing parameter λ was set to 0.4. Noted that when $\lambda = 0$, $LA_{auto} + LearnGC$ becomes $LA_{auto} + MS\text{-}CNN$.

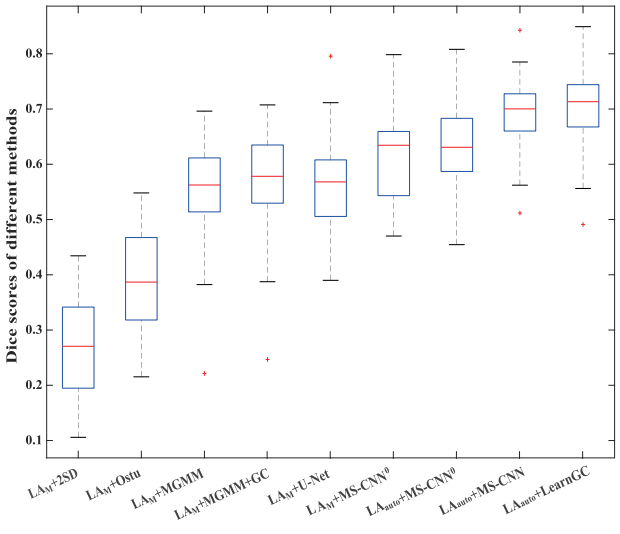


Fig. 8. Boxplots of the Dice scores of scar segmentation by the nine methods.

Table 1 presents all the quantitative results of the nine methods, and Fig. 8 provides their boxplots of Dice scores of scars. The proposed learning graph-cuts method, i.e. $LA_{auto} + LearnGC$, obtained evidently better scar segmentation (Dice of scars) than the conventional methods based on LA_M. It also performed statistically better than all the other eight methods in terms of Dice scores of scars ($p < 0.01$). Compared to the conventional methods, $LA_M + U\text{-}Net$ performed better than the two threshold-based methods ($p < 0.001$), and achieved similar results to the two MGMM-based methods ($p > 0.1$). Note that $LA_{auto} + MS\text{-}CNN^0$ has a slightly better Dice (scar) than $LA_M + MS\text{-}CNN^0$ but without statistical significance ($p = 0.255$), even though the former is based on automatic segmentation of LA and the latter uses manual segmentations. When combined with the random shift strategy, $LA_{auto} + LearnGC$ and $LA_{auto} + MS\text{-}CNN$ obtained evidently and statistical better Dice (scar) than the other methods ($p < 0.01$). For them, $LA_{auto} + LearnGC$ is generally better, but the gain is marginal, due to the fact that the graph-cuts is considered as a built-in smoothness constraint to generate less patchy results. In this study, $LA_{auto} + LearnGC$ did not obtain the best figures in *sensitivity* or *specificity* metrics. Sensitivity measures the proportion of actual

Table 1
Summary of the quantitative evaluation results. *GDice* denotes the generalized Dice score. Here, the asterisk (*) in column Dice (scar) indicates the methods obtained statistically poorer ($p < 0.01$) results compared to the proposed $LA_{auto} + LearnGC$. The p value of the Dice (scar) between $LA_M + MS\text{-}CNN^0$ and $LA_{auto} + MS\text{-}CNN^0$ is 0.225.

Method	Accuracy	Sensitivity	Specificity	Dice (scar)	GDice
$LA_M + 2SD$	0.809 ± 0.074	0.168 ± 0.067	**0.994 ± 0.005**	0.275 ± 0.091*	0.758 ± 0.098
$LA_M + Otsu$	0.763 ± 0.188	0.346 ± 0.214	0.880 ± 0.289	0.396 ± 0.090*	0.726 ± 0.207
$LA_M + MGMM$	0.708 ± 0.160	0.781 ± 0.127	0.690 ± 0.236	0.545 ± 0.101*	0.716 ± 0.190
$LA_M + MGMM + GC$	0.716 ± 0.162	**0.799 ± 0.124**	0.694 ± 0.240	0.562 ± 0.102*	0.721 ± 0.192
$LA_M + U\text{-}Net$	0.832 ± 0.046	0.540 ± 0.149	0.920 ± 0.035	0.568 ± 0.083*	0.826 ± 0.052
$LA_M + MS\text{-}CNN^0$	0.798 ± 0.051	0.775 ± 0.099	0.805 ± 0.078	0.615 ± 0.083*	0.811 ± 0.047
$LA_{auto} + MS\text{-}CNN^0$	0.806 ± 0.052	0.743 ± 0.126	0.824 ± 0.088	0.631 ± 0.080*	0.814 ± 0.047
$LA_{auto} + MS\text{-}CNN$	0.846 ± 0.032	0.786 ± 0.118	0.886 ± 0.057	0.692 ± 0.069*	0.851 ± 0.030
$LA_{auto} + LearnGC$	**0.856 ± 0.033**	0.773 ± 0.132	0.883 ± 0.058	**0.702 ± 0.071**	**0.859 ± 0.031**

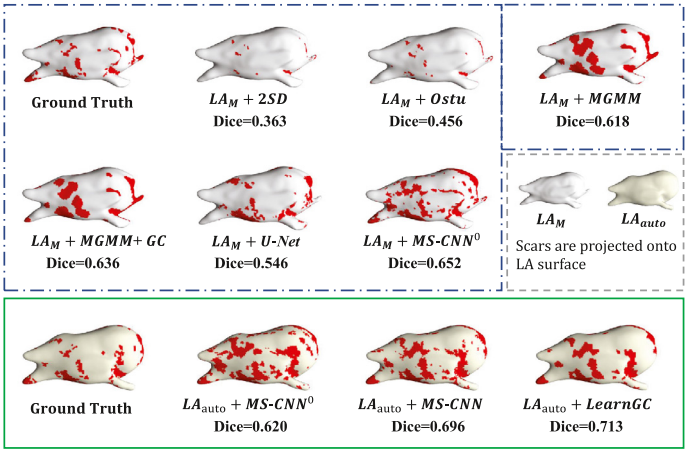


Fig. 9. 3D visualization of the LA scar classification results using the nine methods. This is the median case selected from the test set in terms of Dice score of scars by the proposed method. The scarring areas are red-colored on the LA surface mesh, which can be constructed either from LA_M (LA surface in white) or from LA_{auto} (LA surface in light yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

scarring regions that are correctly identified, and specificity measures the proportion of actual normal wall regions that are correctly segmented. One can see the misleading of these two metrics in evaluating the performance of a method from Table 1, where $LA_M + 2SD$ and $LA_M + MGMM + GC$ achieved the best specificity or sensitivity, though their performance was actually poor in our visual assessment.

In addition, we chose a representative case, the median in terms of Dice (scar) from the test set by the proposed $LA_{auto} + LearnGC$. Fig. 9 visualizes the 3D results by the nine methods. One can observe that the 3D visualization agrees well with the quantitative analysis result using Dice (scar). Though the manually segmented scars in LA_M and LA_{auto} are projected onto two different reference surfaces, GT_M and GT_{auto} visually appear similar when we compare the location and extent of scars. Both the two threshold algorithms, 2SD and Otsu, tended to under estimate (segment) the scars, though Otsu generally performed better. The results of $LA_M + MGMM$ and $LA_M + MGMM + GC$ were acceptable, but the accuracy and automation needed improving. $LA_M + U\text{-}Net$ performed reasonably well, but it tended to misclassify the scars with small areas, especially around the PV. The learning-based methods, from $LA_M + MS\text{-}CNN^0$, $LA_{auto} + MS\text{-}CNN^0$ and $LA_{auto} + MS\text{-}CNN$, to $LA_{auto} + LearnGC$, improved the performance when the new methodologies were introduced. Particularly, $LA_{auto} + LearnGC$ further reduced the noise and patchy segmentation results, and it obtained full automation and best Dice score of scar quantification.

3.6. Performance of the proposed method and inter-observer study

This study analyzes the performance of the proposed method in detail. To provide a reference for the quantitative evaluation metrics, we conducted a study of inter-observation variation from two manual delineations. We randomly selected ten cases from the available data, and asked two experts to manually label the scars separately. For each case, the two labelling results of scars were projected onto the LA_M surface. The Dice (scar), generalized Dice, and accuracy of inter-observer variation were respectively 0.695 ± 0.049, 0.868 ± 0.027 and 0.867 ± 0.026.

Table 1 summarizes the quantitative evaluation results of the proposed method, i.e. $LA_{auto} + LearnGC$. The average Dice of scar is 0.702 ± 0.071, which is comparable to the inter-observer variation (0.695 ± 0.049), and the difference is not significant (p=0.7783). This conclusion also applies when we compare them using accuracy and *GDice* evaluation metrics. We have repeated this experiment another 3 times, by randomly selecting 31 subjects for training and the remaining 27 for test. The mean Dice scores of the three experiments are respectively 0.703 ± 0.082, 0.695 ± 0.094 and 0.698 ± 0.083, which are generally stable.

Fig. 10 provides 2D visualization of the axial view from three examples. These three cases were the first quarter, median and third quarter cases from the test set in terms of Dice (scar) by the proposed method. This illustrates that the method could provide promising performance for localizing and quantifying atrial scars of LA. In the median and third quarter cases, we highlight the errors, particularly due to the enhanced adjacent regions, pointed out by arrow (1), (2) and (3). These mis-classifications, representing the main challenges of this task, contributed to the major errors of scar quantification by the proposed method. Another type of error was caused by the misalignments of the automatic LA segmentation, as arrow (4) pointed out. This happened in some local areas where the errors occurred because of the different shapes of LA after reconstruction from the automatic segmentation. One can also see that even when it existed large LA segmentation errors, indicated by arrow (5) in Fig. 10, the proposed method still could identify the scars at the corresponding location of the projected surface. This is mainly attributed to the effective training of the MS-CNN, which assigns random shifts along the perpendicular direction of the surface when extracting the training patches. The multi-scale learning also contributes to the less demanding of accuracy from the automatic LA segmentation, thus enables to achieve fully automated LA scar quantification.

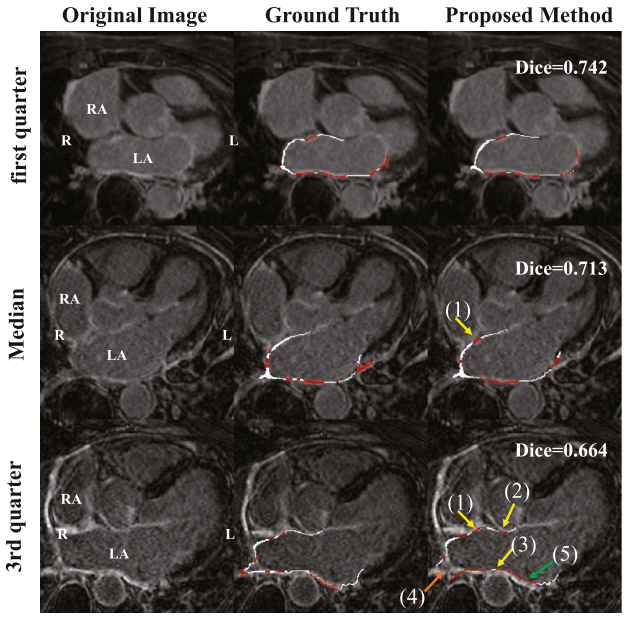


Fig. 10. Axial view of the images, the ground truth scar segmentation and the results by the proposed method. The red and white color labels represent the scar and normal wall, respectively. Arrow (1), (2) and (3) indicate the major classification errors of the proposed method caused by the surrounding enhanced regions, respectively from the right atrium wall, ascending aorta wall and descending aorta wall; arrow (4) shows an error from the misalignment between the automatic LA segmentation and the ground truth; arrow (5) illustrates that the proposed method can still perform well, even though the automatic LA segmentation contains obvious errors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.7. Computational time and computer systems

The pre-processing includes LA segmentation and patch extraction. The LA segmentation, based on a multi-atlas segmentation scheme, took about 18.22 min using a Lenovo D30 Workstation with 32 cores to parallel run the image registration (Zhuang and Shen, 2016). The patch extraction took approximately 1.2 min for one training image and 4.1 min for one test image. Note that the random sample strategy in training phase reduced the time for patch extraction. The inference of the *T*-NET and *N*-NET, with a patch size of 13 × 13 × 17 pixel, required about 47 seconds to process one test image, and the average computation time for the graph-cut algorithm was about 12 seconds.

4. Discussion and conclusion

In this work, we have proposed a fully automatic framework for segmentation and quantification of LA scars. Two major methodological contributions have been introduced. One is the formulation of quantifying the LA scarring based on a surface mesh. The classification and quantification are achieved via the surface projection and graph-cuts framework. The other is the adoption of the multi-scale learning combined with CNN, i.e. MS-CNN. The multi-scale learning is implemented using the MSP strategy, which extracts the features from both the local and global intensity profiles of LGE MRI. The MS-CNN learns both the label probability of each node and the relations between connected nodes in the graph. The surface projection in the proposed framework avoids the difficulty of providing an accurate and demanding LA wall segmentation, and the multi-scale patch-based learning, with the random shift training strategy, further mitigates the effect of less accurate LA initialization from a fully automatic approach, as demonstrated in Section 3.4.3. We employed fifty-eight images with manual delineation for experiments. The proposed method performs better when the size of extracted patches increases, but the performance converges when the size is larger than a certain value (see Section 3.4.1). The multi-scaling learning further improves the performance compared to the method with single-scale learning, as demonstrated in Section 3.4.4. Finally, the proposed learning graph-cuts based method demonstrates evidently better performance compared to the conventional approaches, and the mean accuracy and Dice (scar) for quantifying LA scars are respectively 0.856 and 0.702, which are comparable to those of inter-observer variation (accuracy=0.867, Dice =0.695).

Table 2 summarizes the related works from literature. Perry et al. (2012) evaluated their method on a dataset consisting of 34 images. The mean Dice score was 0.807 ± 0.106, and the inter-observation Dice was 0.786 ± 0.072. Their method required an accurate initialization of LA walls from manual segmentation, followed by a k-mean classification. Karim et al. (2014) employed GMM to model the enhancement of scar region, and used the graph-cuts method to consider neighbouring regions. This method used LA segmentation for initialization, which was achieved from a semi-automatic method with manual correction. They evaluated the method using numerical phantoms as well as using 15 *in vivo* images. They obtained more than 0.8 Dice scores on the two datasets. Ravanelli et al. (2014) adopted a threshold based approach, where the normalized voxel intensity (NVI) of LA walls was applied. The threshold value, $NVI = 4$, was assigned according to previous studies and visual validation by experts, base on which they used a 2-D skeletonization algorithm to quantify the atrial fibrosis. The authors evaluated both the fully automatic method and the semi-automatic approach with manual correction. The mean Dice scores of LA scar quantification increased from 0.60 ± 0.21 to 0.85 ± 0.07 when the manual correction was included. Wu et al. (2018) proposed a fully automatic method for LA fibrosis quantification. They formulated the joint distribution of images based on the multivariate mixture model, and optimized model parameters using the iterated conditional mode algorithm. They tested the method on 36 cases and reported a mean Dice score of 0.556 ± 0.187 and average accuracy of 0.809 ± 0.150. Chen et al. (2018a) developed a multi-view two-task recursive attention model for simultaneous segmentation of LA and scars. The mean Dice score of LA segmentation was 0.908 ± 0.031, which was similar to the result (Dice=0.898 ± 0.044) from our study, though their average Dice score of scar quantification was 0.776 ± 0.146. Yang et al. (2018) employed the super-pixel algorithm and SVM to segment the scars on 37 subjects. They obtained 0.790 ± 0.050 Dice score, 0.87 segmentation accuracy, 0.89 sensitivity and 0.79 specificity by using the leave-one-out cross-validation strategy. This study yielded better Dice score than ours in this work, but there was no evident difference in terms of the accuracy, sensitivity and specificity between these two works. It should be noted that among these six works, only one, i.e., Perry et al. (2012), reported the details of inter-observer variation. Also note that it can be difficult to pursue an objective cross-study comparison due to the difference of datasets, initialization methods, and evaluation metrics.

One of the challenges of LA scar quantification is to distinguish artifacts from the boundary regions, such as from the RA wall and aorta wall, as we discussed above and showed in Figs. 1 and 10. Conventionally, providing accurate LA walls is the crucial step (Karim et al., 2013; Perry et al., 2012). In this work, we propose to use multi-scale deep learning technology, with specifically designed training strategy, to tackle this challenge. However, due to the limited training data, the errors caused by this problem could still happen. Secondly, the quantification of scars in our work is performed on the surface mesh projected from the LA endocardium. Karim et al. (2018) discussed the importance of wall

Table 2
Overview of previous methods for scar quantification and segmentation in LA. Abbreviations: segmentation (seg); inter-observer variation in terms of Dice (Inter-ob); Society of Photo-Optical Instrumentation Engineers (SPIE), IEEE Journal of Translational Engineering in Health and Medicine (TEHM), IEEE transactions on medical imaging (TMI), Medical physics (MP), Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Work	No. subjects	LA (wall) seg	Scar seg method	Result (Dice)	Inter-ob
Perry et al. (2012), SPIE	34	manual	k-means	0.807 ± 0.106	0.786 ± 0.072
Karim et al. (2014), TEHM	15	semi-auto	GMM + Graph-cuts	>0.8	N/A
Ravanelli et al. (2014), TMI	10	semi-auto	NVI + Manual correction	0.850 ± 0.070	N/A
	10	auto	NVI	0.600 ± 0.210	N/A
Wu et al. (2018), MICCAI	36	auto	Multivariate mixture model	0.556 ± 0.187	N/A
Chen et al. (2018a), MICCAI	100	auto	Dilated attention network	0.776 ± 0.146	N/A
Yang et al. (2018), MP	37	auto	Super-pixels + SVM	0.790 ± 0.050	N/A

thickness, particularly considering the potential that the ectopic activity can prevail in scars that are non-transmural. However, they also emphasized that the relationship between the AF and the changes in wall thickness was not clear, and the thickness was difficult to measure based on current MRI data. In clinical practice, the location and extent of scarring areas are considered to have greater clinical significance, which is however arduous to represent and to perform quantitative cross-subject comparisons. In the future work, visual assessment will be considered. Besides, a limitation of this work is that the gold standard was constructed from the manual segmentation of only one physicist. In the future, we can combine the delineations from multiple experts to obtain an average and consensus gold standard.

A major limitation of this work is the lack of an end-to-end training scheme. Specifically, the framework is split into two subtasks, i.e., the MS-CNN and graph-cut, resulting in offline post-processing. Conditional random field (CRF), a probabilistic graphical model, has been broadly used in semantic segmentation to remove isolated false positives and to improve the localization of object boundaries. Chen et al. (2018b) employed a fully connected CRF as a post-processing step of deep CNN to capture fine edge details. Kamnitsas et al. (2017) used a 3D fully connected CRF network for post-processing to refine the output of a segmentation network. Zheng et al. (2015) achieve an end-to-end CRF network, known as CRFasRNN, by formulating the inference of CRF as recurrent neural networks. Additionally, the graph convolution network (Kipf and Welling, 2016) and PointNet (Qi et al., 2017) can cooperate with the neighbor information of nodes in an end-to-end style. In these studies, only the low dimensional features in the corresponding positions of the nodes are used. By contrast, the features associated with the nodes in the proposed LearnGC framework come from the image patches. The dimension of the patches could be thousands, e.g. the typical size of the patch is $13 \times 13 \times 17$, whose dimension is 2873. Therefore, an end-to-end training scheme of our method, with integral optimization based on the whole graph, could be infeasible in practice, due to its expensive time and space complexity. A detailed discussion of this issue can be found in the supplementary material document associated with this manuscript, and an efficient end-to-end network is considered as our future work.

Declaration of Competing Interest

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

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61971142) and the Science and Technology Commission of Shanghai Municipality (17JC14 01600). This study was also funded by the British Heart Foundation Project grant (PG/16/78/32402).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2019.101595.

References

Badger, T.J., Daccarett, M., Akoum, N.W., Adjei-Poku, Y.A., Burgon, N.S., Haslam, T.S., Kalvaitis, S., Kuppahally, S., Vergara, G., McMullen, L., et al., 2010. Evaluation of left atrial lesions after initial and repeat atrial fibrillation ablation: lessons learned from delayed-enhancement MRI in repeat ablation procedures. Circul.: Arrhyth. Electrophysiol. 3 (3), 249–259.

Beinart, R., Abbara, S., Blum, A., Ferencik, M., Heist, K., Ruskin, J., Mansour, M., 2011. Left atrial wall thickness variability measured by ct scans in patients undergoing pulmonary vein isolation. J. Cardiovasc. Electrophysiol. 22 (11), 1232–1236.

Boykov, Y.Y., Jolly, M.-P., 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In: IEEE International Conference on Computer Vision, 1, pp. 105–112.

Calkins, H., Kuck, K.H., Cappato, R., Camm, A.J., Chen, S.A., Crijns, H.J.G., Jr, R.J.D., Davies, D.W., Dimarco, J., Edgerton, J., 2012. 2012 HRS/EHRA/ECAS Expert consensus statement on catheter and surgical ablation of atrial fibrillation: recommendations for patient selection, procedural techniques, patient management and follow-up, definitions, endpoints, and research trial design. Heart Rhythm 9 (4), 632–696.

Chen, J., Yang, G., Gao, Z., Ni, H., Angelini, E., Mohiaddin, R., Wong, T., Zhang, Y., Du, X., Zhang, H., et al., 2018. Multiview two-task recursive attention model for left atrium and atrial scars segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chugh, S.S., Havmoeller, R., Narayanan, K., Singh, D., Rienstra, M., Benjamin, E.J., Gillum, R.F., Kim, Y.-H., McAnulty, J.H., Zheng, Z.-J., et al., 2013. Worldwide epidemiology of atrial fibrillation: a global burden of disease 2010 study. Circulation 837–847.

Crum, W.R., Camara, O., Hill, D.L., 2006. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imaging 25 (11), 1451–1461.

Ji, Y., van der Geest, R.J., Nazarian, S., Lelieveldt, B.P., Tao, Q., 2018. Advanced two-layer level set with a soft distance constraint for dual surfaces segmentation in medical images. In: Medical Imaging 2018: Image Processing, 10574, p. 105743B.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.

Karim, R., Arujuna, A., Housden, R.J., Gill, J., Cliffe, H., Matharu, K., Gill, J., Rindaldi, C.A., ÓNeill, M., Rueckert, D., Razavi, R., 2014. A method to standardize quantification of left atrial scar from delayed-enhancement MR images. IEEE J. Transl. Eng. Health Med. 2 (1), 1–15.

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

Karim, R., Housden, R.J., Balasubramaniam, M., Chen, Z., Perry, D., Uddin, A., Al-Beyatti, Y., Palkhi, E., Acheampong, P., Obom, S., et al., 2013. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. J. Cardiovasc. Magn. Reson. 15 (1), 105.

Keegan, J., Jhooti, P., Babu-Narayan, S.V., Drivas, P., Ernst, S., Firmin, D.N., 2014. Improved respiratory efficiency of 3D late gadolinium enhancement imaging using the continuously adaptive windowing strategy (CLAWS). Magn. Reson. Med. 71 (3), 1064–1074.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
Knowles, B.R., Caulfield, D., Cooklin, M., Rinaldi, C.A., Gill, J., Bostock, J., Razavi, R., Schaeffter, T., Rhode, K.S., 2010. 3-D Visualization of acute RF ablation lesions using MRI for the simultaneous determination of the patterns of necrosis and edema. IEEE Trans. Biomed. Eng. 57 (6), 1467–1475.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105.
Lau, F., Hendriks, T., Lieman-Sifry, J., Sall, S., Golden, D., 2018. Scargan: Chained Generative Adversarial Networks to Simulate Pathological Tissue on Cardiovascular MR Scans. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 343–350.
Liu, J., Zhuang, X., Wu, L., An, D., Xu, J., Peters, T., Gu, L., 2017. Myocardium segmentation from DE MRI using multicomponent gaussian mixture model and coupled level set. IEEE Trans. Biomed. Eng. 64 (11), 2650–2661.
Lorensen, W.E., Cline, H.E., 1987. Marching cubes: A high resolution 3D surface construction algorithm. In: ACM siggraph computer graphics, 21, pp. 163–169.
Lu, F., Wu, F., Hu, P., Peng, Z., Kong, D., 2017. Automatic 3D liver location and segmentation via convolutional neural network and graph cut. Int. J. Comput. Assist. Radiol. Surg. 12 (2), 171–182.
McGann, C.J., Kholmovski, E.G., Oakes, R.S., Blauer, J.J., Daccarett, M., Segerson, N., Airey, K.J., Akoum, N., Fish, E., Badger, T.J., et al., 2008. New magnetic resonance imaging-based method for defining the extent of left atrial wall injury after the ablation of atrial fibrillation. J. Am. Coll. Cardiol. 52 (15), 1263–1271.
Moccia, S., Banali, R., Martini, C., Muscogiuri, G., Pontone, G., Pepi, M., Caiani, E.G., 2019. Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images. Magn. Reson. Mater. Phys., Biol. Med. 32 (2), 187–195.
Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 9 (1), 62–66.
Perry, D., Morris, A., Burgon, N., McGann, C., MacLeod, R., Cates, J., 2012. Automatic classification of scar tissue in late gadolinium enhancement cardiac MRI for the assessment of left-atrial wall injury after radiofrequency ablation. In: Medical Imaging 2012: Computer-Aided Diagnosis, 8315. International Society for Optics and Photonics, p. 83151D.
Peters, D.C., Wylie, J.V., Hauser, T.H., Kissinger, K.V., Botnar, R.M., Essebag, V., Josephson, M.E., Manning, W.J., 2007. Detection of pulmonary vein and left atrial scar after catheter ablation with three-dimensional navigator-gated delayed enhancement MR imaging: initial experience. Radiology 243 (3), 690–695.
Pontecorboli, G., Figueras i Ventura, R.M., Carlosena, A., Benito, E., Prat-Gonzales, S., Padeletti, L., Mont, L., 2016. Use of delayed-enhancement magnetic resonance imaging for fibrosis detection in the atria: a review. EP Europace 19 (2), 180–189.
Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. Pointnet: deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660.
Ravanelli, D., dal Piaz, E.C., Centonze, M., Casagranda, G., Marini, M., Del Greco, M., Karim, R., Rhode, K., Valentini, A., 2014. A novel skeleton based quantification and 3-D volumetric visualization of left atrium fibrosis using late gadolinium enhancement magnetic resonance imaging. IEEE Trans. Med. Imaging 33 (2), 566–576.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.
Rother, C., Kolmogorov, V., Blake, A., 2004. Grabcut: interactive foreground extraction using iterated graph cuts. In: ACM transactions on graphics (TOG), 23, pp. 309–314.
Tao, Q., Ipek, E.G., Shahzad, R., Berendsen, F.F., Nazarian, S., van der Geest, R.J., 2016. Fully automatic segmentation of left atrium and pulmonary veins in late gadolinium-enhanced MRI: towards objective atrial scar assessment. J. Magn. Reson. Imaging 44 (2), 346–354.
Tobon-Gomez, C., Geers, A.J., Peters, J., Weese, J., Pinto, K., Karim, R., Ammar, M., Daoudi, A., Margeta, J., Sandoval, Z., et al., 2015. Benchmark for algorithms segmenting the left atrium from 3d ct and mri datasets. IEEE Trans. Med. Imaging 34 (7), 1460–1473.
Veni, G., Elhabian, S.Y., Whitaker, R.T., 2017. Shapecut: Bayesian surface estimation using shape-driven graph. Med. Image Anal. 40, 11–29.
Vergara, G.R., Marrouche, N.F., 2011. Tailored management of atrial fibrillation using a LGE-MRI based model: From the clinic to the electrophysiology laboratory. J. Cardiovasc. Electrophysiol. 22 (4), 481–487.
Vergara, G.R., Vijayakumar, S., Kholmovski, E.G., Blauer, J.J., Guttman, M.A., Gloschat, C., Payne, G., Vij, K., Akoum, N.W., Daccarett, M., et al., 2011. Real-time magnetic resonance imaging–guided radiofrequency atrial ablation and visualization of lesion formation at 3 Tesla. Heart Rhythm 8 (2), 295–303.
Wilber, D.J., Pappone, C., Neuzil, P., De Paola, A., Marchlinski, F., Natale, A., Macle, L., Daoud, E.G., Calkins, H., Hall, B., et al., 2010. Comparison of antiarrhythmic drug therapy and radiofrequency catheter ablation in patients with paroxysmal atrial fibrillation: a randomized controlled trial. JAMA 303 (4), 333–340.
Williams, S.E., Tobon-Gomez, C., Zuluaga, M.A., Chubb, H., Butakoff, C., Karim, R., Ahmed, E., Camara, O., Rhode, K.S., 2017. Standardized unfold mapping: a technique to permit left atrial regional data display and analysis. J. Interv. Cardiac Electrophysiol. 50 (1), 125–131.
Wu, F., Li, L., Yang, G., Wong, T., Mohiaddin, R., Firmin, D., Keegan, J., Xu, L., Zhuang, X., 2018. Atrial fibrosis quantification based on maximum likelihood estimator of multivariate images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 604–612.
Xiong, Z., Fedorov, V.V., Fu, X., Cheng, E., Macleod, R., Zhao, J., 2018. Fully automatic left atrium segmentation from late gadolinium enhanced magnetic resonance imaging using a dual fully convolutional neural network. IEEE Trans. Med. Imaging PP (99), 1–10.
Xu, C., Xu, L., Gao, Z., Zhao, S., Zhang, H., Zhang, Y., Du, X., Zhao, S., Ghista, D., Liu, H., et al., 2018. Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture. Med. Image Anal. 50, 82–94.
Yang, G., Zhuang, X., Khan, H., Haldar, S., Nyktari, E., Li, L., Wage, R., Ye, X., Slabaugh, G., Mohiaddin, R., et al., 2018. Fully automatic segmentation and objective assessment of atrial scars for long-standing persistent atrial fibrillation patients using late gadolinium-enhanced MRI. Med. Phys. 45 (4), 1562–1576.
Zhao, J., Xiong, Z., 2018. 2018 atrial segmentation challenge. http://atriaseg2018.cardiacatlas.org/.
Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H., 2015. Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision, pp. 1529–1537.
Zhuang, X., 2013. Challenges and methodologies of fully automatic whole heart segmentation: a review. J. Healthc. Eng. 4 (3), 371–407.
Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med. Image Anal. 31, 77–87.