Medical Image Analysis 51 (2019) 21–45

Contents lists available at ScienceDirect

Medical Image Analysis

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

Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers

Mahendra Khened [a], Varghese Alex Kollerathu [a], Ganapathy Krishnamurthi [a,*]

Department of Engineering Design, Indian Institute of Technology Madras, Chennai, India

ARTICLE INFO

Article history:
Received 16 January 2018
Revised 11 October 2018
Accepted 18 October 2018
Available online 19 October 2018

Keywords:
Deep learning
Ensemble classifier
Cardiac MRI
Segmentation
Automated diagnosis
Fully convolutional densenets.

ABSTRACT

Deep fully convolutional neural network (FCN) based architectures have shown great potential in medical image segmentation. However, such architectures usually have millions of parameters and inadequate number of training samples leading to over-fitting and poor generalization. In this paper, we present a novel DenseNet based FCN architecture for cardiac segmentation which is parameter and memory efficient. We propose a novel up-sampling path which incorporates long skip and short-cut connections to overcome the feature map explosion in conventional FCN based architectures. In order to process the input images at multiple scales and view points simultaneously, we propose to incorporate Inception module's parallel structures. We propose a novel dual loss function whose weighting scheme allows to combine advantages of cross-entropy and Dice loss leading to qualitative improvements in segmentation. We demonstrate computational efficacy of incorporating conventional computer vision techniques for region of interest detection in an end-to-end deep learning based segmentation framework. From the segmentation maps we extract clinically relevant cardiac parameters and hand-craft features which reflect the clinical diagnostic analysis and train an ensemble system for cardiac disease classification. We validate our proposed network architecture on three publicly available datasets, namely: (i) Automated Cardiac Diagnosis Challenge (ACDC-2017), (ii) Left Ventricular segmentation challenge (LV-2011), (iii) 2015 Kaggle Data Science Bowl cardiac challenge data. Our approach in ACDC-2017 challenge stood second place for segmentation and first place in automated cardiac disease diagnosis tasks with an accuracy of 100% on a limited testing set (n=50). In the LV-2011 challenge our approach attained 0.74 Jaccard index, which is so far the highest published result in fully automated algorithms. In the Kaggle challenge our approach for LV volume gave a Continuous Ranked Probability Score (CRPS) of 0.0127, which would have placed us tenth in the original challenge. Our approach combined both cardiac segmentation and disease diagnosis into a fully automated framework which is computationally efficient and hence has the potential to be incorporated in computer-aided diagnosis (CAD) tools for clinical application.

1. Introduction

Cardiac cine Magnetic Resonance (MR) Imaging is primarily used for assessment of cardiac function and diagnosis of Cardiovascular diseases (CVDs). Cardiac MRI is considered the most accurate method for the estimation of clinical parameters such as ejection fraction, ventricular volumes, stroke volume and myocardial mass. Delineating important organs and structures from volumetric medical images, such as MR and computed tomography (CT) images, is usually considered the primary step for estimating clinical parameters, disease diagnosis, prediction of prognosis and surgical planning. In a clinical setup, a radiologist delineates the region of interest from the surrounding tissues/ organs by manually drawing contours encompassing the structure of interest. However, this approach becomes infeasible in a hospital with high throughput as it is time-consuming, tedious and also introduces intra and inter-rater variability (Petitjean and Dacher, 2011; Miller et al., 2013; Tavakoli and Amini, 2013; Suinesiaputra et al., 2014). Hence, a fully automatic method for segmentation and indication of clinical diagnosis is desirable.

The organization of the paper is as follows. Discussion of prior art on short-axis cardiac cine MR segmentation and our contributions is summarized in the remainder of Section 1. In Section 2 we introduce the datasets employed in our experiments and discuss the proposed methodology for segmentation and disease classifi-

* Corresponding author.
E-mail address: gankrish@iitm.ac.in (G. Krishnamurthi).

https://doi.org/10.1016/j.media.2018.10.004
1361-8415/

Table 1
Summary of the methods for (semi-)/automatic cardiac segmentation in cine MRI.

Method	References
Image-based classification	Jolly (2006), Katouzian et al. (2006), Üzümcü et al. (2006), Lu et al. (2009), Cousty et al. (2010)
Pixel classification	Lynch et al. (2006), Pednekar et al. (2006), Nambakhsh et al. (2013)
Variational and level sets	Paragios (2003), Fradkin et al. (2008), Lynch et al. (2008), Ayed et al. (2008)
Graph cuts and image-driven approaches	Boykov and Jolly (2000), Lin et al. (2006b), Cocosco et al. (2008)
Cardiac atlases based registration	Lorenzo-Valdés et al. (2004), Lötjönen et al. (2004), Bai et al. (2015)
Statistical shape & active appearance models	Ordas et al. (2007), Zhu et al. (2010), Zhu et al. (2010), Zhang et al. (2010), Albá et al. (2018)
Learning-based approaches	Margeta et al. (2011)

cation. In Section 3, we experimentally analyze the effectiveness of our proposed segmentation and disease classification models. In 4, we shall comprehensively present the results on three publicly available cardiac MR challenge datasets. In Section 5, we discuss our results and conclude with future direction of work.

1.1. Related work

1.1.1. Segmentation and automated diagnosis from cardiac cine MR images

Segmentation of left ventricular (LV) endocardium and epicardium as well as right ventricular (RV) endocardium from multi-slice cine MR datasets has received significant research attention over past few years and several grand challenges (Radau et al., 2009; Suinesiaputra et al., 2014; Petitjean et al., 2015; Booz Allen Hamilton Inc and Kaggle, 2015; Bernard et al., 2018) have been organized for advancing the state of art methods in (semi-)/automated cardiac segmentation. These challenges usually provide expert-ground truth contours and provide set of evaluation metrics to benchmark various approaches. Petitjean and Dacher (2011), Frangi et al. (2001), Tavakoli and Amini (2013), and Peng et al. (2016) provide a comprehensive survey on cardiac segmentation using semi-automated and fully automated approaches. These approaches can be categorized into three levels based on the type of prior information incorporated during segmentation: (i) no prior information is used but manual input is required, (ii) weak prior such as anatomical assumptions on the spatial intensity or ventricle shape, (iii) strong prior such as statistical models, manually annotated images are used to construct such models. A summary of these approaches can be found in Table 1. For automating cardiac diagnosis from cine MR images statistical shape based methods (Sonka et al., 2003; Suinesiaputra et al., 2009; Zhao et al., 2009; Suinesiaputra et al., 2018) have been proposed over the years.

1.1.2. Fully convolutional neural networks for medical image segmentation

In the field of medical image analysis, considerable amount of work has been done in the lines of automating segmentation of various anatomical structures, detection and delineation of lesions and tumors. Non-learning based algorithms such as statistical shape modeling, level sets, active contours, multi-atlas and graphical models have shown promising results on limited dataset, but they usually tend to perform poorly on data originating from a database outside the training data. Some of these techniques heavily relied on engineering hand-crafted features and hence required domain knowledge and expert inputs. Moreover, hand-crafted features have limited representational ability to deal with the large variations in appearance and shapes of anatomical organs. In order to overcome these limitations, learning based methods have been explored to seek more powerful features. A variant of artificial neural networks for image related tasks were first introduced as Neocognitron (Fukushima, 1979). These learning based networks were later on popularized as convolution neural networks (CNNs) by LeCun et al. (1998). CNNs have been adopted for a variety of computer vision and pattern recognition tasks. Most common

Table 2
Summary of deep learning based methods for cardiac cine MR image analysis.

Reference	Method
Deep learning for feature extraction	
Emad et al. (2015)	Patch-based CNN for LV localization
Kong et al. (2016)	CNN+LSTM for detection of systole & diastole phases from cardiac MRI
Zhang et al. (2016)	CNN to detect missing slices (apical & basal) in cardiac MRI
Deep learning combined with classical segmentation	
Rupprecht et al. (2016)	Patch-based CNN + Active contour framework for boundary extraction
Avendi et al. (2016)	CNN+Auto-encoders +Deformable models for LV segmentation
Ngo et al. (2017)	Deep belief network to initialize and guide level-set to segment LV
Yang et al. (2016)	Deep network for label-fusion operation in multi-atlas LV segmentation
End-to-end deep learning based segmentation	
Tran (2016)	FCN to segment LV and RV
Poudel et al. (2016)	Recurrent FCN to segment LV leveraging inter-slice spatial dependency
Tan et al. (2017)	CNN based regression for LV contour delineation in polar space domain
Oktay et al. (2018)	Incorporated anatomical prior in CNNs for cardiac image segmentation, enhancement and pathology classification
Zheng et al. (2018)	FCN with spatial propagation for 3-D consistent segmentation
Bai et al. (2018)	FCN to segment LV and RV on short-axis CMR images and the left atrium (LA) and right atrium (RA) on long-axis CMR images.

ones are image classification (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; He et al., 2016) and semantic segmentation using fully convolutional networks (FCN) (Shelhamer et al., 2017). In all these applications, CNNs have demonstrated greater representational and hierarchical learning ability. Recently in medical image analysis, FCN and its popular extensions like U-NET (Ronneberger et al., 2015) have achieved remarkable success in segmentation of various structures in heart (Dou et al., 2016), brain lesions (Havaei et al., 2017; Kamnitsas et al., 2017; Pereira et al., 2016), liver lesions (Christ et al., 2016; Dou et al., 2016; Ben-Cohen et al., 2015; Deng and Du, 2008) from medical volumes. Also with availability of huge amounts of labelled data and increase in the computational capability of general purpose graphics processors units (GPUs), CNN based methods have the potential for application in daily clinical practice. A summary of CNN/ deep learning based methods for cardiac cine MR image analysis can be found in Table 2.

Deep learning based architectures like FCNs have shown larger model capacity and the ability to learn highly discriminative features. However, in the context scarcity of labelled data, training such deep networks leads to over-fitting and poor generalization. Unlike natural images, medical images are inherently 3D data and segmentation of 3D structures using FCNs demands higher memory and computational requirements. When training FCNs for multi-class segmentation tasks, severe class imbalance is alleviated by employing manually calibrated weighted loss-functions. Medical image datasets are acquired with different scanners with varying imaging resolution, for better generalization across multi-centric data, the deep architectures need to be scale-invariant and viewpoint independent.

1.1.3. Contributions

In this paper, we present a novel 2D fully convolutional neural network (FCN) architecture for medical image segmentation which is parameter and memory efficient. We also develop a fully automated framework which incorporates cardiac structures segmentation and cardiac disease diagnosis. Our contributions are summarized as follows:

- The proposed network connectivity pattern was based on densely connected convolutional neural networks (DenseNets) (Huang et al., 2017). DenseNets facilitates multi-path flow for gradients between layers during training by back-propagation and hence does implicit deep-supervision. DenseNets encourage feature reuse and thus substantially reduces the number of parameters while maintaining good performance, which is ideal in scenarios with limited data. In addition, we incorporated multi-scale processing in the initial layers of the network by performing convolutions on the input with different kernel sizes in parallel paths and later fusing them as in Inception architectures. We proposed a novel long skip and short-cut (residual) connections in the up-sampling path which was computationally and memory efficient when compared to standard skip connections. We introduce a weight calibration scheme for both cross-entropy and Dice loss. We propose to combine both the loss functions to yield higher segmentation accuracy alongside with qualitative improvements in segmentation.
- The proposed methodology for region of interest (ROI) extraction from the cine MRI sequences utilizes spatio-temporal variation statistics of cardiac structures and circular Hough Transform. The approach of training the network on ROI patches aided in the reduction of computational and GPU memory requirement.
- For the automated cardiac disease classification, the predicted segmentation labels were used to estimate cardiac physiological parameters and hand-crafted features. Random Forest based feature importance analysis was performed to identify most relevant features. We developed an ensemble classifier system which processed the features in two-stages for prediction of the cardiac disease.

We extensively validated our proposed network on two cardiac segmentation tasks: (i) segmentation of left ventricle (LV), right ventricle (RV) and myocardium (MYO) from multi-slice cine MR images for both end-diastolic (ED) and end-systolic (ES) phase instances and, (ii) segmentation of myocardium for the whole cardiac frames and slices in multi-slice cine MR images, by participating in two challenges organized at Statistical Atlases and Computational Modeling of the Heart (STACOM) workshops: (i) Automated cardiac diagnosis challenge: ACDC-2017 (Bernard et al., 2018), and (ii) Left ventricular segmentation challenge: LV-2011 (Suinesiaputra et al., 2014) respectively. The proposed segmentation network's generalization across different data distributions was assessed by evaluating segmentation performance of model trained on ACDC-2017 training dataset and testing on LV-2011 challenge test set, and vice versa. We evaluated our segmentation network trained on ACDC dataset on the Kaggle Data Science Bowl Cardiac Challenge (Booz Allen Hamilton Inc and Kaggle, 2015) testing set (n=440) for the task of estimating LV volume at ED and ES phases. We achieved competitive segmentation results to state-of-the-art approaches in both the challenges and demonstrated the effectiveness and generalization capability of the proposed network architecture. To facilitate further research and improvements, we have made our implementations publicly available[1]

1.1.4. Connections with prior related work

A preliminary version of our work (Khened et al., 2017) was presented at STACOM-2017 workshop held in conjunction with Medical Image Computing and Computer Assisted Interventions (MICCAI-2017). In this paper, we improved the network architecture and these changes in the design improved segmentations alongside reduction of GPU memory foot-print and number of trainable parameters. The changes made to the methodology used for cardiac disease classification model gave an accuracy of 100% on the ACDC challenge test dataset (n=50). The main modifications in this paper include:- (i) Elaborating proposed methods, analyzing underlying network connectivity pattern, loss functions and adding experiments on LV-2011 and Kaggle challenge datasets. (ii) Design of 2-stage classifier ensemble for cardiac diagnosis and engineering features based on myocardial wall thickness variation for characterizing myocardial infarction.

While our work was being reviewed, DRINet (Chen et al., 2018) was published incorporating connectivity pattern of DenseNets, residual networks and Inception networks in a FCN for biomedical image segmentation. Our proposed architecture differed from DRINet in the following manner:- (i) In the down-sampling path, our network had Inception module in the first layer for processing input image at multiple scales and subsequent layers comprised of dense blocks. Whereas in DRINet the down-sampling path was composed of dense blocks only. (ii) Our network had residual long skip connections, which was 1×1 convolution for dimension reduction and followed by addition of features maps in the up-sampling path at a given resolution. DRINet did not have skip-connections. (iii) In the up-sampling path, our network had dense blocks with residual connections and transposed convolution blocks. In DRINet it comprised of residual inception blocks and unpooling blocks.

[1] Please contact the authors.

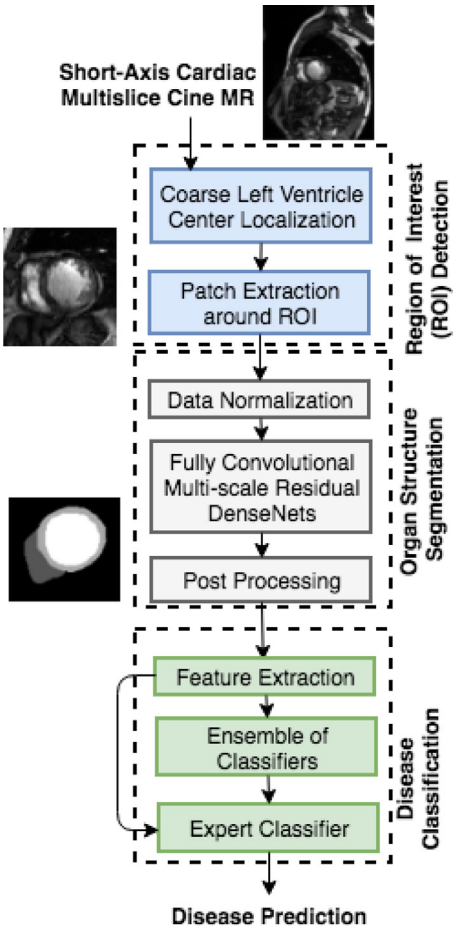


Fig. 1. Proposed pipeline for automated cardiac segmentation & cardiac disease diagnosis.

2. Material and methods

2.1. Overview

Fig. 1 illustrates our automated cardiac segmentation and disease diagnosis framework. The pipeline involved: (i) Fourier analysis and circular Hough-transform for region of interest (ROI) cropping, (ii) FCN for cardiac structures segmentation and (iii) an ensemble of classifiers for disease diagnosis based on features extracted from the segmentation.

2.2. Experimental datasets and materials

2.2.1. ACDC-2017 dataset

The automated cardiac diagnosis challenge dataset comprised of 150 exams of different patients. Based on cardiac physiological parameters in the medical reports, the patients were grouped into 5 classes namely- (i) normal- NOR, (ii) patients with previous myocardial infarction- MINF, (iii) patients with dilated cardiomyopathy- DCM, (iv) patients with hypertrophic cardiomyopathy- HCM, (v) patients with abnormal right ventricle- ARV. The preparation of the dataset ensured even distribution of patient groups in all the classes. The cine MR images were acquired in breath hold with a retrospective or prospective gating and with a SSFP sequence in short axis orientation. A series of short axis slices cover the LV from the base to the apex, with a slice thickness of 5–8 mm and an inter-slice gap of 5 or 10 mm. The spatial resolution goes from 1.37 to 1.68 mm^2/pixel and 28 to 40 images cover completely or partially the cardiac cycle. For each patient, the weight, height and the diastolic and systolic phase instants were provided. The challenge organizers had evenly divided the patient database based on the pathological condition and was made available in two phases, 100 for training and 50 for testing. The manual annotations of LV, RV and MYO were done by clinical experts at systolic and diastolic phase instances only. The clinical diagnosis and manual annotations were provided for the training set and those of testing set were held out by the challenge organizers for their independent evaluation.

2.2.2. LV-2011 dataset

LV segmentation challenge dataset was made publicly available as part of the STACOM 2011 challenge on automated LV myocardium segmentation from short-axis cine MRI. The dataset comprised of 200 patients with coronary artery disease and myocardial infarction. The dataset provided by the organizers was divided into two sets of 100 cases each: training and validation. The spatial resolutions of the images varied from 0.7 to 2.1 mm/pixel and the matrix sizes varied from 156×192 to 512×512, and cardiac cycle comprised of 18 to 35 frames. The training set was provided with expert-guided semi-automated segmentation contours for the myocardium. The ground truths for the validation set were generated using the STAPLE algorithm (Warfield et al., 2004) from the results of STACOM 2011 challenge (Suinesiaputra et al., 2014).

2.2.3. 2015 Kaggle data science bowl cardiac challenge data

The Kaggle database comprised of cardiac MRI images in DICOM format. The organizers gave 500 patients for training, 200 for validation and 440 for final testing and scoring. For each patient, 2D cine MR images contained approximately 30 images across the cardiac cycle. Each slice was acquired on a separate breath hold. These images were acquired in different planes which included multiple short-axis views covering the entire heart (the number of slices varied from 1 to 23), a 2-chamber view and a 4-chamber view. Unlike the ACDC and LV-2011 datasets, the Kaggle dataset had no ground truth segmentations for LV, but the reference volumes at end systole and end diastole phases were provided.

2.3. Region of interest (ROI) detection

The cardiac MR images of the patient comprise of the heart and the surrounding chest cavity like the lungs and diaphragm. The proposed ROI detection step was specifically designed to get an approximate localization of heart region (LV center) in cine MR images. ROI detection involved spatio-temporal statistical analysis of cardiac phases and circular Hough transform (Duda and Hart, 1972; Korshunova et al., 2016) to delineate the heart structures from the surrounding tissues. The ROI extraction involved finding an approximate LV center and extracting a patch of size 128×128 centered around it. The extracted ROI patch was used for training and inference of the FCN models. This approach alleviated the class-imbalance problem associated with labels for heart structures seen in the full sized cardiac MR images. Appendix B gives a detailed overview of ROI detection steps.

2.4. Normalization

In both ACDC-2017 and LV-2011 dataset, the acquisition of images was done using multi-slice cine MRI. The training set for both the challenges were prepared by extracting 2D MR slices (with annotations) from all the patient's cine MR sequences and the normalization was performed for every time frame of the cine sequences separately. Assuming that the batch of image samples drawn from the training set were independent and identical distributed, slice-wise normalization was considered appropriate for training the 2D segmentation network. The multi-slice cine MR

cardiac datasets employed slice-wise normalization of voxel intensities using Eq. (1).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X is voxel intensity. $X_{\min}$ and $X_{\max}$ are minimum and maximum of the voxel intensities in a slice respectively.

2.5. Network architecture

The proposed network's connectivity pattern was inspired by DenseNet for semantic segmentation (Jégou et al., 2017). Appendix C contains a brief overview on DenseNets, Residual networks and Inception architectures.

2.5.1. Fully convolutional multi-scale residual DenseNets for semantic segmentation

Figs. 2 & 3 illustrate the building blocks and the schematic diagram of the proposed network architectures for segmentation respectively. A typical semantic segmentation architecture comprises of a down-sampling path (contracting) and an up-sampling path (expanding). The down-sampling path of the network was similar to the DenseNet architecture (C.1). The last layer of the down-sampling path was referred to as bottleneck. The input spatial resolution was recovered in the up-sampling path by transposed convolutions, dense blocks and skip connections coming from the down-sampling path. The up-sampling operation was referred to as transition up (TU). The up-sampled feature maps were added element-wise with skip-connections. The feature maps of the hindmost up-sampling component were convolved with a 1×1 convolution layer followed by a soft-max layer to generate the final label map of the segmentation.

We discuss and compared 3 different architectural variants of densely connected fully convolutional networks (DFCN). We refer to the architecture introduced by Jégou et al. (2017) as DFCN-A and other two proposed variants as DFCN-B & DFCN-C (Fig. 3). The following changes in the proposed network's connectivity pattern lead to further improvement in terms of parameter efficiency, rate of convergence and GPU memory footprint required:

- The GPU memory footprint increases with the number of feature maps of larger spatial resolution. In order to mitigate feature map explosion in the up-sampling path, the skip connections from down-sampling path to up-sampling path used element-wise addition operation instead of concatenation operation. In order to match the channel dimensions, a projection operation was done on the skip connection path using BN-ELU-1×1-convolution-dropout. This operation when compared to concatenation of feature maps helps in reduction of the parameters and memory footprint without affecting the quality of the segmentation output. The proposed projection operation does dimension reduction and also allows complex and learnable interactions of cross channel information (Lin et al., 2013). Replacing the activation function from rectified linear units (ReLUs) to exponential linear units (ELUs) manifested in faster convergence.
- In DenseNets, without pooling layers the spatial resolution of feature maps increases with depth and hence leads to memory explosion. So, Jégou et al. (2017) overcame this limitation in the up-sampling path by not concatenating the input to a dense block with its output (only exception was at the last dense block, see Fig. 3 (a)). Hence, the transposed convolution was applied only to the feature maps obtained by the last layer of a dense block and not to all feature maps concatenated so far. However, we observed that by introducing shortcut (residual) connections (He et al., 2016) in the dense blocks of up-sampling path by element-wise addition of dense blocks input with its output would better aggregate a set of previous transformations rather than completely discarding the inputs of dense blocks. In order to match dimensions of dense blocks input and output a projection operation was done using BN-ELU-1×1-convolution-dropout. We found that this was effective in addressing the memory limitations found in DenseNet based architecture for semantic segmentation. We also observed faster convergence and improved segmentation metrics due to shortcut connections. The DFCN-B refers to architecture got by the above modifications to DFCN-A.
- All of the above proposed modifications were incorporated in DFCN-C (Fig. 3(b)) and additionally its initial layer included parallel CNN branches similar to inception module (Szegedy et al., 2015). Incorporating multiple kernels of varying receptive fields in each of the parallel paths would help in capturing view-point dependent object variability and learning relations between image structures at multiple-scales (See Appendix C and C.18 for more details).

2.5.2. Loss function

In medical images, there exists an acute class imbalance between the region of interest and the surrounding background. This issue was addressed by different loss functions such as Dice loss (Milletari et al., 2016) and weighted cross-entropy loss.

For training the network, a dual loss function which incorporated both cross-entropy and Dice loss was proposed. Additionally, two different weighting mechanisms for both the loss functions were introduced. The cross-entropy loss measures accumulated error across all voxels by calculating voxel-wise error probability between the predicted output class and the target class. Spatial weight maps generated from the ground-truth images (Fig. 4) were used for weighting the loss computed at each voxel in the cross-entropy loss Eq. (2).

Let $W = (w_1, w_2, \cdots, w_l)$ be the set of learn-able weights, where w_l is weight matrix corresponding to the *lth* layer of the deep network, $p(t_i|x_i; W)$ represent the probability prediction of a voxel x_i after the soft-max function in the last output layer, the spatially weighted cross-entropy loss was formulated as:

$$L_{CE}(X; W) = -\sum_{x_i \in X} w_{map}(x_i) \log(p(t_i|x_i; W)) \tag{2}$$

where X represents the training samples and t_i is the target class label corresponding to voxel $x_i \in X$ and $w_{map}(x_i)$ is the weight estimated at each voxel x_i.

Let L be the set of all ground truth classes in the training set. For each ground-truth image, let N be the set of all voxels, T_l be the set of voxels corresponding to each class $l \in L$ and C_l be the set of contour voxels corresponding to each class $l \in L$.

$$w_{map}(x_i) = \sum_{l \in L} \frac{|N| * \mathbb{1}_{T_l}(x_i)}{|T_l|} + \sum_{l \in L} \frac{|N| * \mathbb{1}_{C_l}(x_i)}{|C_l|} \tag{3}$$

where $|.|$ denotes the cardinality of the set and $\mathbb{1}$ represents the indicator function defined on the subsets of N, i.e. $C_l \subset T_l \subset N$, $\forall\, l \in L$.

For using Dice overlap coefficient score as loss function an approximate value of Dice-score was estimated by replacing the predicted label with its posterior probability $p(t_i|x_i; W)$. Since, Dice-coefficient needs to be maximized for better segmentation output, the optimization was done to minimize its complement, i.e. $(1 - \widehat{DICE})$.

For multi-class segmentation, the Dice loss was computed using weighted mean of $\widehat{DICE_l}$ for each class $l \in L$. The weights were estimated for every mini-batch instance from the training set. The Dice loss for multi-class segmentation problem is given in Eq. (5):

$$l_{DICE}(X; W) = \frac{\sum_{x_i \in X} p(t_i|x_i; W) g(x_i) + \epsilon}{\sum_{x_i \in X} (p(t_i|x_i; W)^2 + g(x_i)^2) + \epsilon} \tag{4}$$

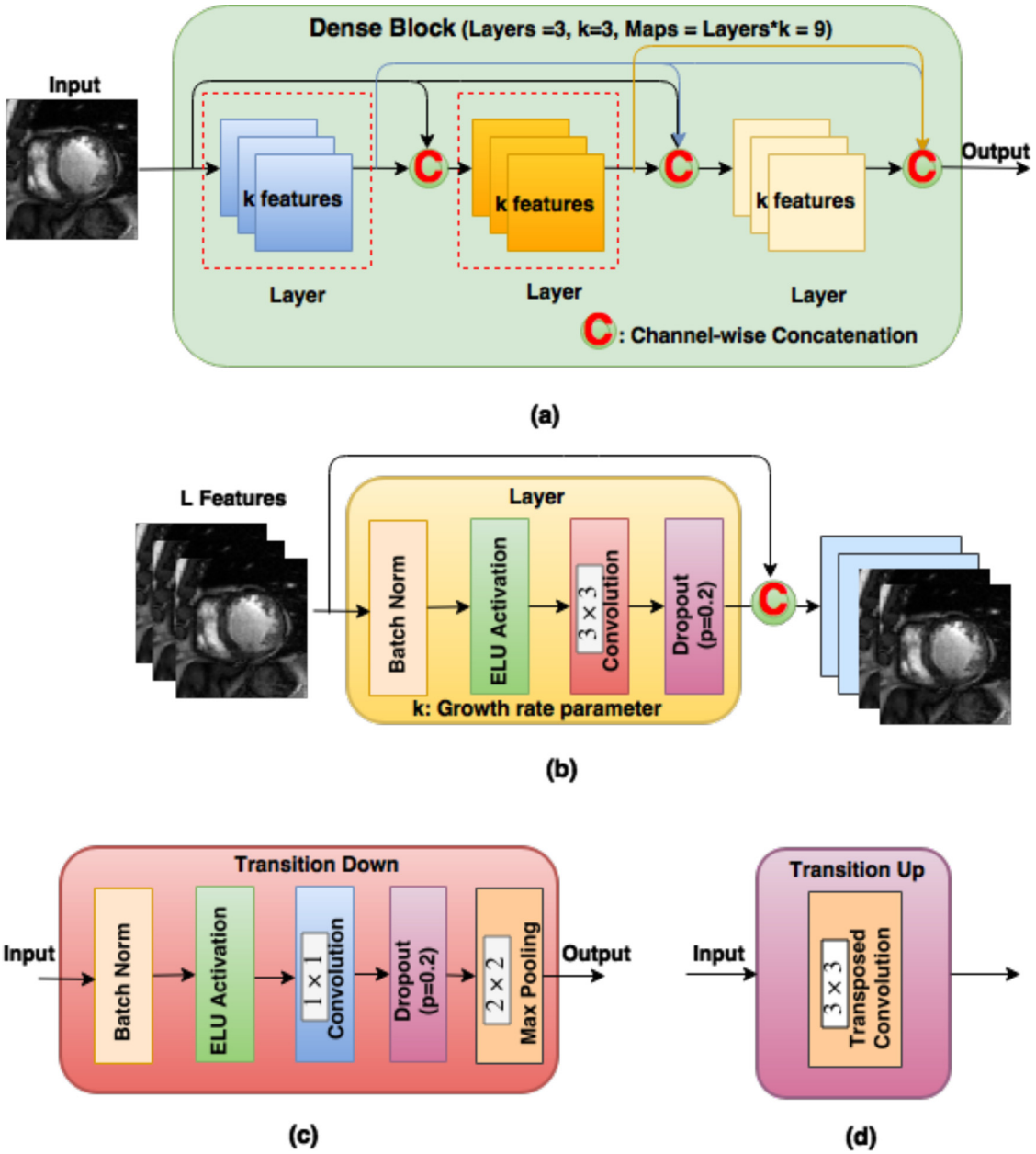


Fig. 2. The proposed architecture for semantic segmentation using densely connected fully convolutional network (DFCN) was made up of several modular blocks which are explained as follows: (a) An example of dense block (DB) with 3 Layers. In a DB the input was fed to the first layer to create k new feature maps. These new feature maps were concatenated with the input and fed to the second layer to create another set of k new feature maps. This operation was repeated for 3 times and the final output of the DB was a concatenation of the outputs of all the 3 layers and thus contains 3*k feature maps, (b) A layer in a DB was a composition of batch normalization (BN), exponential linear unit (ELU), 3×3 convolution and a dropout layer with a drop-out rate of $p = .2$, (c) A transition down (TD) block reduces the spatial resolution of the feature maps as the depth of the network increases. TD block was composed of BN, ELU, 1×1 convolution, dropout ($p = .2$) and 2×2 max-pooling layers, (d) A transition up (TU) block increases the spatial resolution of the feature maps by performing 3×3 transposed convolution with a stride of 2.

$$L_{DICE} = \frac{\sum_{l \in L} w_l l_{DICE}}{\sum_{l \in L} w_l} \tag{5}$$

where w_l was the estimated weight for each class $l \in L$ and ϵ was a small value added to both numerator and denominator for numerical stability. Let M be the set of pixels in the mini-batch, M_l be the set of pixels corresponding to each class $l \in L$ and $M_l \subset M$, then the weight estimate for the current mini-batch is given by:

$$w_l = \frac{|M|}{|M_l|} \quad \forall l \in L \tag{6}$$

The parameters of the network were optimized to minimize both the loss functions in tandem. In addition, an $L2$ weight-decay penalty was added to the loss function as regularizer. The total loss function was given in Eq. (7).

$$LOSS = \lambda(L_{CE}) + \gamma(1 - L_{DICE}) + \eta||W||^2 \tag{7}$$

where λ, γ and η are weights to individual losses. The $L2$ loss decay factor was set to $\eta = 5 \times 10^{-4}$.

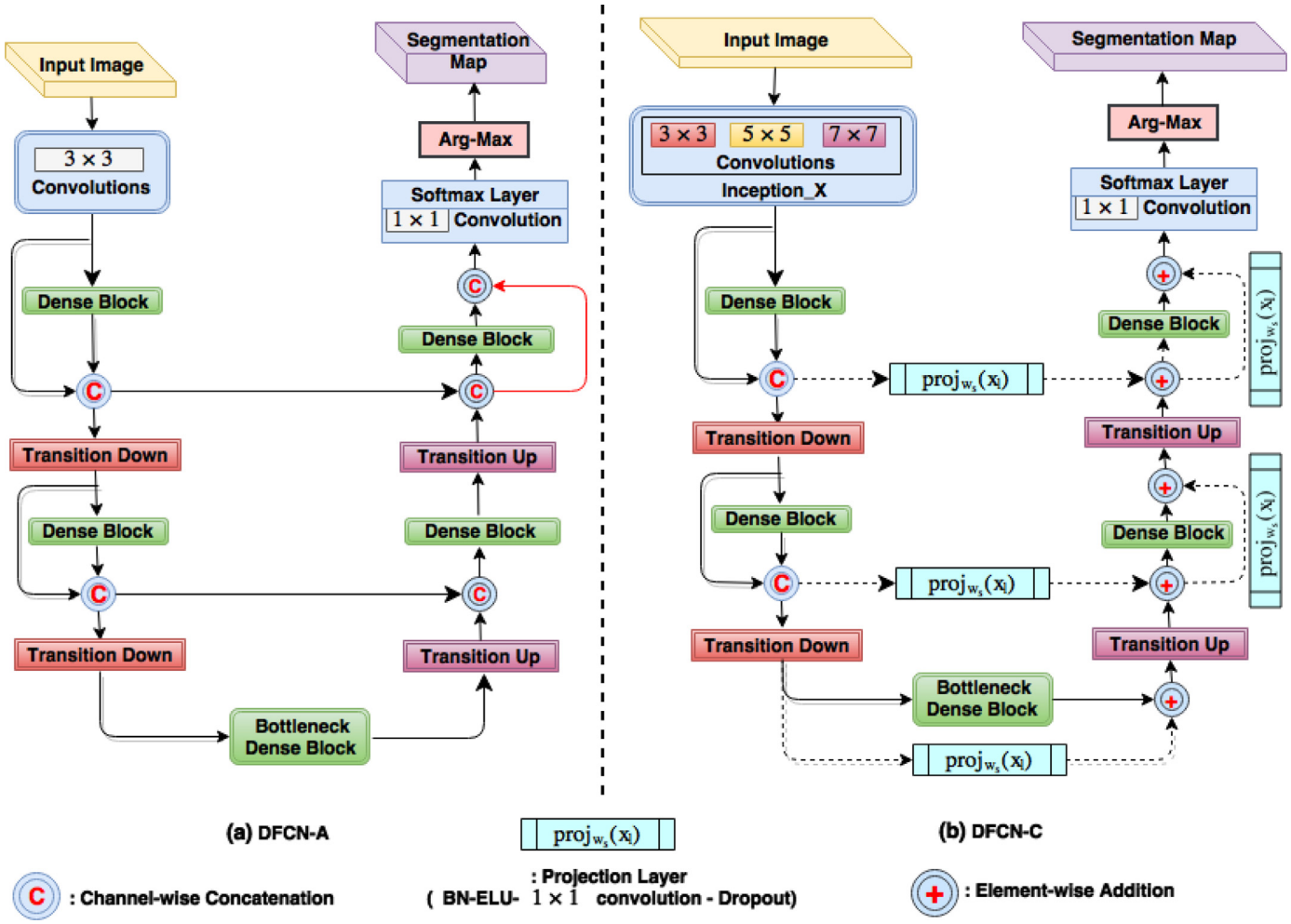


Fig. 3. The figures shows the modifications introduced to DenseNets to mitigate the feature map explosion when extending to FCN. (a) DFCN-A (Jégou et al., 2017), (b) The proposed architecture was referred to as DFCN-C. The main modifications include:- (i) replacing the standard copy and concatenation of skip connections from down-sampling path to up-sampling path with a projection layer and an element-wise addition of feature-maps respectively, aiding in reduction of parameters and GPU memory footprint. (ii) introduction of short-cut connections (residual) in the up-sampling path, (iii) parallel pathways in the initial layer.

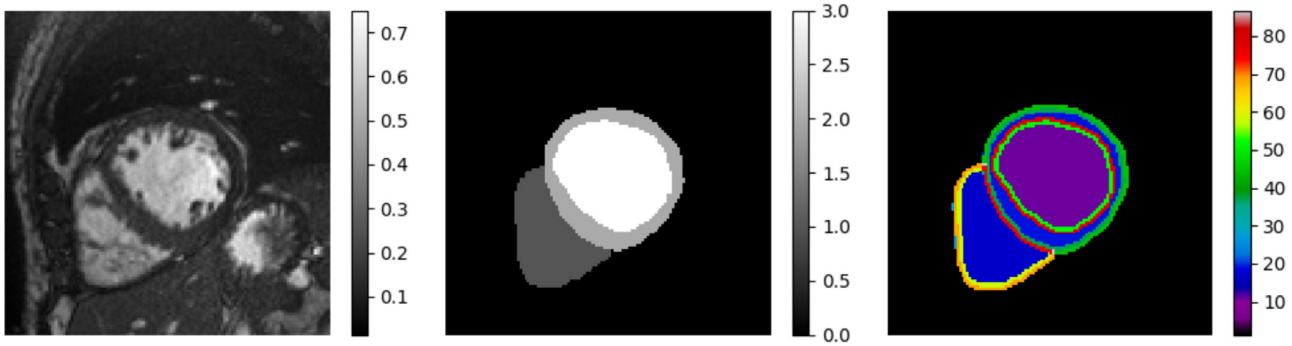


Fig. 4. The figure shows the spatial weight-map generated from the ground truth image. The spatial weight map was used with voxel-wise cross-entropy loss. The contour voxels for each class were identified using Canny edge detector with $\sigma = 1$ and was followed by morphological dilation. The colors in spatial weight-map indicate weight distribution based on their relative class frequency. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.6. Post-processing

The results of segmentation were subjected to 3-D connected component analysis followed by slice-wise 2-D connected component analysis and morphological operations such as binary hole-filling inside the ventricular cavity.

2.7. Automated cardiac disease diagnosis

The goal of the automated cardiac disease diagnosis challenge was to classify the cine MRI-scans of the heart into one of the five groups, namely:- (i) DCM, (ii) HCM, (iii) MINF, (iv) ARV and (v) NOR.

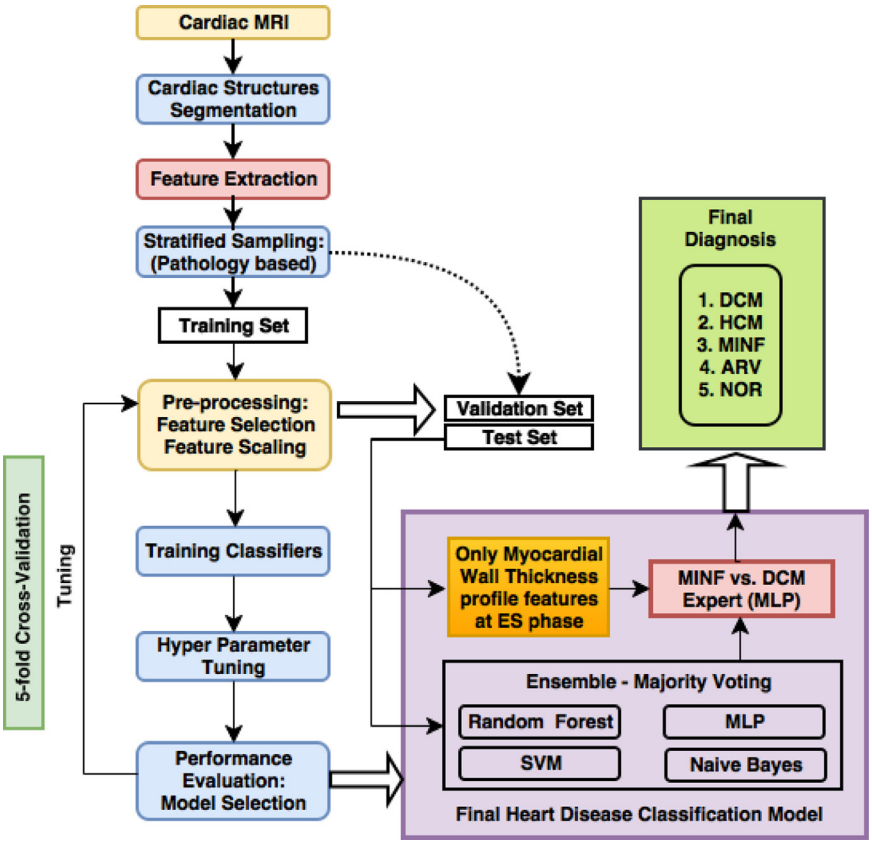


Fig. 5. Automated cardiac diagnosis using ensemble of classifiers and expert classifier approach.

2.7.1. Feature extraction

From the ground truth segmentations of training data, several cardiac features pertaining to left ventricle (LV), right ventricle (RV) and myocardium (MYO) were extracted. The cardiac features were grouped into two categories, namely primary feature & derived features. Primary features were calculated directly from the segmentations and DICOM tags pixel spacing and slice-thickness.

1. Volume of LV, RV and MYO at end diastole (ED) and end systole (ES) phases.
2. Myocardial wall thickness measures at each slice.

Derived features were a combination of primary features:

1. Ejection Fraction (EF) of LV and RV.
2. Ratio of Primary features: *LV*: *RV, MYO*: *LV* at ED and ES phases.
3. Variation profile of myocardial wall thickness (MWT) in Short-Axis (SA) and Long-Axis (LA). In cases of infarction the myocardial walls were irregular and the wall thickness variation was not smooth. It was observed that the features such as standard deviation of myocardial wall thickness measures within a slice or across slices had the potential to characterize myocardial infarction. Appendix D gives the methodology in arriving at these set of features.

2.7.2. Two-stage ensemble approach for cardiac disease classification

The cardiac disease diagnosis was approached as two stage classification using an ensemble system. Fig. 5 illustrates the methodology adopted for cardiac disease classification. Ensemble classification is a process in which multiple classifiers are created and strategically combined to solve a particular a classification problem. Combining multiple-classifiers need not always guarantee better performance than the best individual classifier in the ensemble. The ensemble ensures that overall risk due to poor model selection is minimized. The accuracy of the classifiers were estimated based on 5-fold cross-validation scores. Based on the cross-validation scores only the top performing classifiers were selected for combining in the ensemble-based system. The first stage of the ensemble comprised of four classifiers namely- (i) Support Vector Machine (SVM) with radial basis function kernel, (ii) Multi-layer Perceptron (MLP) with 2 hidden layers with 100 neurons each, (iii) Gaussian Naive Bayes (GNB) and (iv) Random Forest (RF) with 1000 trees. All the classifiers were independently trained to classify the patient's cine MR scan into five groups by extracting all the features listed in Table D.16. In the first-stage of the ensemble a voting classifier finalized the disease prediction based on majority vote.

In some of the cases, the first stage of the ensemble had difficulty in distinguishing between MINF and DCM groups. In-order to eliminate such misclassifications, a two class "expert" classifier trained only on myocardial wall thickness variation profile features at ES phase was proposed. The expert classifier re-assessed only those cases for which the first stage's predictions were MINF or DCM. The expert classifier used was a MLP with 2 hidden layers with 100 neurons each.

3. Experimental analysis of segmentation and classification models

In this section, we experimentally analyze the effectiveness of our proposed network architecture, loss function, data-augmentation scheme, effect of ROI cropping, post-processing and disease classification ensemble. To conduct the experiments, ACDC training dataset with ground truth masks were used in-order to analyze the learning process and compare the segmentation results. The metrics of evaluation were Dice score and Hausdorff Distance (HD) in mm (Appendix A). The neural network architectures

were designed using TensorFlow (Abadi et al., 2016) software. We ran our experiments on a desktop computer with NVIDIA-Titan-X GPU, Intel Core i7-4930K 12-core CPUs @ 3.40GHz and 64GB RAM.

3.1. ACDC-2017 training dataset preparation

The training dataset comprising of 100 patient cases ($\approx 1.8k$ 2D images) were split into 70: 15: 15 for training, validation and testing subsets. Stratified sampling was done so as to ensure each split comprised of equal number of cases from different cardiac disease groups. Each patient scan had approximately 20 $2D$ images with ground truth annotations for left ventricle (LV), right ventricle (RV) and myocardium (MYO) at the end diastole (ED) and end systole (ES) phases.

3.2. Segmentation network architecture and hyper-parameter tuning

Fig. 6 describes the proposed network architecture used for segmentation. Based on exhaustive hyper parameter tuning experiments, the following network configuration were found to be optimal:- (i) Number of max-pooling operations were limited to three ($P = 3$), (ii) Growth-rate of dense locks (DBs) was set to ($k = 12$), (iii) Number of initial feature maps (F) generated by the first convolution layers was ensured to be at-most 3 times the growth-rate ($F \approx 3k$).

3.3. Training settings

The network was trained by minimizing the proposed loss function (Eq. 7) using ADAM optimizer (Kingma and Ba, 2014) with a learning rate set to 10^{-3}. The network weights were initialized using He normal initializer (He et al., 2015) and trained for 200 epochs with data augmentation scheme as described in Section 3.7. The training batch comprised of 16 ROI cropped 2D MR images of dimension 128×128. After every epoch the model was evaluated on the validation set and the final best model selected for evaluating on test set was ensured to have highest Dice score for MYO class on the validation set.

3.4. Evaluating the effect of growth rate

Table 3 shows the DFCN-C performance with varying growth-rate (k) parameter. For the same architecture the segmentation performance steadily improved with increasing value of k. Also, the network exhibited potential to work with extremely small number of trainable parameters.

3.5. Evaluating the effect of different loss functions

Table 4 compares the segmentation performance of six different loss functions. Because of ROI cropping, the heavy class-imbalance was already mitigated and hence the standard cross-entropy loss showed optimal performance in terms of both Dice score and Hausdorff measures. In terms of Hausdorff distance metric alone, the spatially weighted cross-entropy loss showed best performance suggesting heavy weight penalization for contour voxels aided in learning the contours precisely. The performance of standard Dice loss was better than the mini-batch weighted Dice loss, indicating weighting was not necessary when using only Dice loss. The simple combination of Dice and cross-entropy losses showed slight dip in the performance. It was observed that cross-entropy loss optimized for voxel-level accuracy whereas the Dice loss helped in improving the segmentation metrics. Weighted Dice loss alone caused over segmentation at the boundaries whereas the weighted cross-entropy loss alone led to very sharp contours with minor under-segmentation. So, in order to balance between these trade-offs and combine the advantages of both the losses, we empirical estimated that by setting $\gamma = 1$ and $\lambda = 1$ in the proposed loss (Eq. (7)) gave optimal performance in terms of faster convergence and better segmentation quality. Statistical t-tests on the limited test set (n=15) indicated model trained on proposed loss function had no statistical significance in terms of Dice and Hausdorff metrics (Table 4). However, it was observed that the loss function altered the segmentation network's output behavior. If the metric of evaluation was Dice score then standard Dice loss would appear to perform best. However, visual inspection of segmentation revealed that models trained on Dice loss over-segmented leading to more false positives, a similar observation was made by Pawlowski et al. (2018). The proposed loss function in contrast to other loss function showed qualitative improvements as illustrated with an example in Fig. 7.

3.6. Evaluating the effect of ROI cropping and post-processing

For evaluating the effect of not using ROI cropped images, the proposed network was trained on the input images resized to 256×256 (zero-padding or center-cropping was done to ensure this image dimension). Table 5 compares the effect of using ROI-cropping and post-processing on segmentation results. Even though the non-ROI based technique resulted in better Dice-score but it had higher Hausdorff distance, this was mainly because of false positives at basal and apical slices. As shown in Fig. 8 post-processing steps aided in removing false positives and outliers. Table 5 indicates that the performance of ROI vs. non-ROI was comparable after post-processing steps. However, in the interest of reducing the GPU memory footprint and time required for training and inference, ROI based method was proposed.

3.7. Evaluating the effect of data augmentation scheme

Data augmentation was done to artificially increase the training set and to prevent the network from over-fitting on the training set. For analyzing the effect of data augmentation during the learning process of the proposed network, two separate models were trained with and without data-augmentation. For the model which incorporated data-augmentation scheme the training batch comprised of the mixture of original dataset and on the fly randomly generated augmented data which included: (i) rotation: random angle between -5 and 5 degrees, (ii) translation x-axis: random shift between -5 and 5 mm, (iii) translation y-axis: random shift between -5 and 5 mm, (iv) rescaling: random zoom factor between 0.8 and 1.2, (v) adding Gaussian noise with zero mean and 0.01 standard deviation and (vi) elastic deformations using a dense deformation field obtained through a 2×2 grid of control-points and B-spline interpolation. Fig. 9 summarizes the learning curves. Both the models validation loss decreased consistently as the training loss went down, indicating less over-fitting on the training data. Close observation on the validation curves with model trained on data-augmentation revealed minor improvement in the segmentation performance on the validation set which was also corroborated on the held-out test-set.

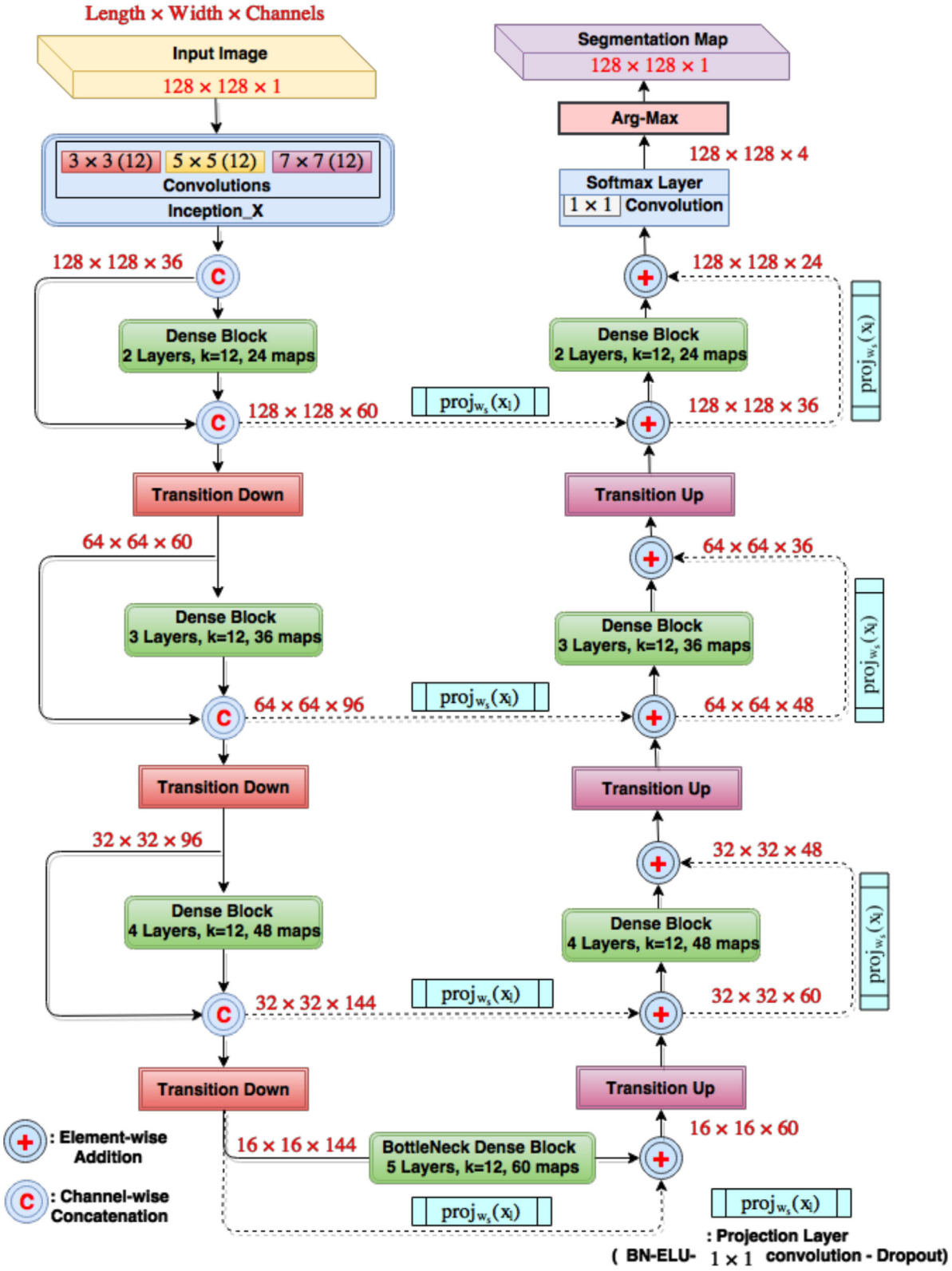
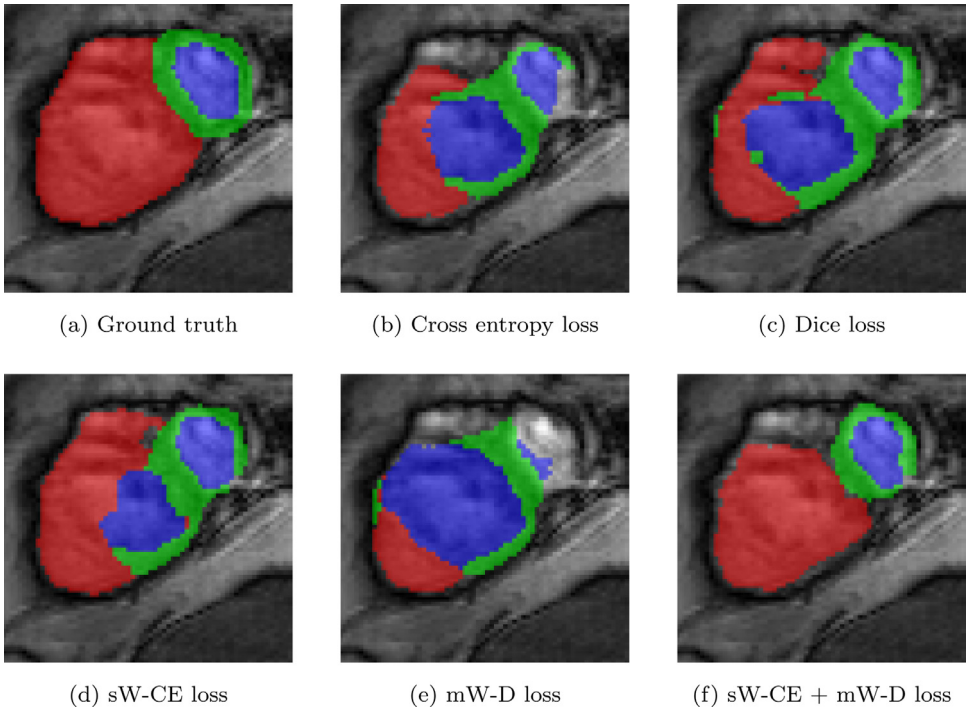


Fig. 6. The proposed network architecture (DFCN-C) comprises of a contracting path (down-sampling path) and an expanding path (up-sampling path). The arrows indicate network's information flow pathways. The dotted horizontal arrows represent residual skip connections where the feature maps from the down-sampling path were added in an element-wise manner with the corresponding feature maps in the up-sampling path. In order to match the channel dimensions a linear projection was done using BN-ELU-1 × 1-convolution-dropout in the residual connection path. In the down-sampling path, the input to a dense block was concatenated with its output, leading to a linear growth in the number of feature maps. Whereas in the bottleneck and up-sampling path the features were added element-wise to enable learning a residual function.



(a) Ground truth (b) Cross entropy loss (c) Dice loss

(d) sW-CE loss (e) mW-D loss (f) sW-CE + mW-D loss

Fig. 7. The figure compares the ground truth segmentation with the automated segmentation results on a typical cardiac MR short axis slice exhibiting intensity inhomogeneity. The proposed network architecture was trained using different loss functions as listed in Table 4. The model trained on proposed loss function exhibited minor under segmentations of the RV region ((f)). Whereas other models gave anatomical impossible segmentations such as two disjoint LV regions, incomplete/ incorrect MYO and RV contours ((b), (c), (d)). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Evaluation of segmentation results for various growth rates. The values are provided as mean (standard deviation). Numbers with * indicate significant difference compared to the $k = 12$ model, according to Student's t-test for two independent samples on the Dice and Hausdorff metrics ($p < 5 \cdot 10^{-2}$).

	Growth Rate							
DICE	***k=2***	***k=4***	***k=6***	***k=8***	***k=10***	***k=12***	***k=14***	***k=16***
LV	0.86 (0.08)	0.90 (0.06)	0.88 (0.09)	0.93 (0.05)	0.93 (0.04)	0.93 (0.05)	0.93 (0.05)	0.93 (0.05)
RV	0.76 (0.13)*	0.82 (0.14)	0.85 (0.17)	0.87 (0.13)	0.91 (0.04)	0.91 (0.05)	0.91 (0.05)	0.92 (0.05)
MYO	0.78 (0.06)*	0.80 (0.08)*	0.82 (0.09)*	0.88 (0.04)	0.88 (0.03)	0.89 (0.03)	0.90 (0.02)	0.90 (0.03)
Mean	0.80 (0.09)*	0.84 (0.09)*	0.85 (0.12)*	0.89 (0.07)	0.91 (0.04)	0.91 (0.04)	0.92 (0.04)	0.92 (0.04)
HD	***k=2***	***k=4***	***k=6***	***k=8***	***k=10***	***k=12***	***k=14***	***k=16***
LV	16.89 (10.51)*	13.04 (9.51)	17.64 (11.33)	7.26 (8.97)	4.96 (5.15)	5.46 (6.39)	3.95 (4.21)	4.16 (4.26)
RV	17.96 (10.86)*	9.41 (4.74)*	7.49 (4.20)*	6.01 (2.65)	5.60 (2.58)	5.65 (2.38)	5.49 (2.82)	4.79 (2.05)
MYO	14.27 (8.47)*	14.72 (9.57)*	16.71 (8.37)	5.58 (4.08)	7.72 (7.80)	5.18 (4.44)	4.32 (3.20)	4.25 (4.12)
Mean	16.37 (9.95)*	12.39 (7.94)*	13.95 (7.97)*	6.29 (5.24)	6.09 (5.18)	5.43 (4.40)	4.59 (3.41)	4.40 (3.48)
Parameters	11,452	43,036	94,756	166,612	258,604	370,732	502,996	655,396

3.8. Comparing performance and learning curves with other baseline models

For analyzing and benchmarking the proposed network DFCN-C (F=36, P=3, k=12), we constructed 3 baseline models, namely:- (i) Modified U-Net (Ronneberger et al., 2015) starting with 32 initial feature maps and 3 max-pooling layers (F=32, P=3), (ii) DFCN-A (Jégou et al., 2017) (F=32, P=3, k=12), (iii) DFCN-B (F=32, P=3, k=12). The architecture of DFCN-B was similar to DFCN-C except the initial layer. Its initial layer had only one CNN branch which learnt 32 3×3 filters. All the models were trained in the same manner.

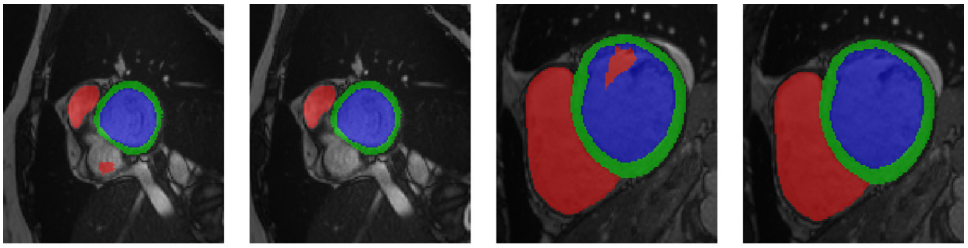
Fig. 10 summarizes and compares the learning process of DFCN-C with other three baseline models. In U-Net, it was observed that the loss associated with training and validation decreased and increased respectively as the training progressed. Such patterns indicate the possibility of the network to over-fit on smaller datasets. Moreover, the number of parameters and GPU memory usage were highest for U-Net. For all the DFCN based architectures, the validation loss consistently decreased as the training loss decreased, hence showed least tendency to over-fit. When comparing amongst the DFCN variants, the validation curves of DFCN-C showed faster convergence and better segmentation scores when compared DFCN-B. Hence, corroborating the effectiveness of proposed methodology of multi-scale feature extraction and feature fusion. When compared to DFCN-A, the number of parameters and GPU memory usage were relatively low for both DFCN-B and DFCN-C. Table 6 compares the results of the proposed DFCN-C architecture against the baseline models. Both DFCN-A and DFCN-C were almost on-par with respect to most of the metrics of evaluation. It was observed that only for the task of RV segmentation at ES phase, Hausdorff distance (HD) metric was significantly different with $p < 5 \cdot 10^{-2}$. Manual inspection of segmentation results on held-out test set indicated both the networks had difficulty in

Table 4
Evaluation of segmentation results for different loss functions.
Note:- CE: cross-entropy loss, D: Dice loss, sW-CE : cross-entropy loss with weighting scheme based on spatial weight map, mW-D: Dice loss with weighting scheme based on mini-batch. The values are provided as mean (standard deviation). Numbers with * indicate significant difference compared to the loss function **sW-CE + mW-D**, according to Student's *t*-test for two independent samples on the Dice and Hausdorff metrics ($p < 5 \cdot 10^{-2}$).

Method	DICE LV		DICE RV		DICE MYO		Mean Dice
	ED	ES	ED	ES	ED	ES	
CE	0.96 (0.02)	0.91 (0.07)	0.94 (0.02)	0.88 (0.06)	0.87 (0.03)	0.88 (0.03)	0.91 (0.04)
sW-CE	0.96 (0.02)	0.91 (0.06)	0.92 (0.09)	0.85 (0.17)	0.89 (0.03)	0.89 (0.03)	0.90 (0.07)
D	0.96 (0.02)	0.91 (0.09)	0.95 (0.02)	0.88 (0.07)	0.89 (0.03)	0.90 (0.03)	0.91 (0.04)
mW-D	0.96 (0.02)	0.90 (0.08)	0.95 (0.01)	0.87 (0.06)	0.87 (0.02)	0.89 (0.03)	0.90 (0.04)
CE+D	0.96 (0.02)	0.90 (0.08)	0.93 (0.04)	0.86 (0.11)	0.88 (0.03)	0.88 (0.04)	0.90 (0.05)
sW-CE + mW-D	0.96 (0.02)	0.90 (0.08)	0.95 (0.02)	0.87 (0.08)	0.89 (0.03)	0.89 (0.03)	0.91 (0.04)
	HD LV		**HD RV**		**HD MYO**		**Mean HD**
	ED	ES	ED	ES	ED	ES	
CE	2.98 (2.93)	4.73 (3.78)	4.81 (1.94)	6.73 (3.51)	3.57 (2.55)	7.90 (7.12)	5.12 (3.64)
sW-CE	3.82 (4.01)	4.04 (2.51)	5.09 (2.62)	6.19 (3.93)	4.46 (2.82)	4.85 (2.83)	4.74 (3.12)
D	4.70 (6.58)	7.82 (9.72)	4.82 (1.79)	6.41 (3.26)	4.59 (4.93)	6.16 (5.44)	5.75 (5.29)
mW-D	4.49 (8.35)	7.45 (9.01)	10.14 (8.1)*	9.62 (6.65)	6.47 (7.38)	9.54 (10.37)	7.95 (8.31)*
CE+D	7.09 (10.24)	9.47 (10.86)	4.85 (2.08)	6.93 (2.88)	7.67 (8.90)	11.56 (9.27)*	7.93 (7.37)*
sW-CE + mW-D	4.42 (6.39)	6.51 (6.40)	4.20 (1.81)	7.10 (2.95)	4.48 (4.52)	5.87 (4.35)	5.43 (4.40)

Table 5
Evaluation results with and without region of interest (ROI) cropping as a pre-processing step and largest connected Component (LCC) as a post-processing step. The values are provided as mean (standard deviation). Numbers with * indicate significant difference compared to the model employing ROI-Crop and LCC steps, according to a two-sided, paired *t*-test on the Dice and Hausdorff metrics ($p < 5 \cdot 10^{-2}$).

Pre-Process	Post-Process	DICE LV		DICE RV		DICE MYO		Mean Dice
		ED	ES	ED	ES	ED	ES	
ROI-Crop	LCC	0.96 (0.02)	0.90 (0.08)	0.95 (0.02)	0.87 (0.08)	0.89 (0.03)	0.89 (0.03)	0.91 (0.04)
ROI-Crop	-	0.96 (0.02)	0.90 (0.08)	0.94 (0.04)	0.84 (0.13)	0.88 (0.03)	0.88 (0.03)	0.90 (0.05)*
-	-	0.96 (0.02)	0.91 (0.07)	0.93 (0.03)*	0.84 (0.10)	0.89 (0.02)*	0.90 (0.03)*	0.91 (0.04)
-	LCC	0.96 (0.02)	0.91 (0.07)	0.94 (0.03)*	0.86 (0.08)	0.90 (0.02)*	0.90 (0.02)*	0.91 (0.04)
		HD LV		**HD RV**		**HD MYO**		**Mean HD**
		ED	ES	ED	ES	ED	ES	
ROI-Crop	LCC	4.42 (6.39)	6.51 (6.40)	4.20 (1.81)	7.10 (2.95)	4.48 (4.52)	5.87 (4.35)	5.43 (4.40)
ROI-Crop	-	8.04 (11.26)	15.28 (13.45)*	11.80 (19.17)	13.34 (11.92)*	11.31 (12.84)	17.52 (14.10)*	12.88 (13.79)*
-	-	2.97 (3.15)	9.69 (16.88)	26.90 (42.44)	21.49 (35.86)	22.11 (34.04)	34.55 (44.34)*	19.62 (29.42)*
-	LCC	2.97 (3.15)	3.92 (2.71)	4.67 (1.96)	7.05 (4.29)	3.72 (2.58)	6.09 (6.22)	4.74 (3.48)



(a) Before LCCA (b) After LCCA (c) Before MBHF (d) After MBHF

Fig. 8. The figure shows the sequence of post-processing steps applied to eliminate false-positives and outliers in the predictions. Largest connected component analysis (LCCA) retained only the largest common structure and discarded the rest as seen in (a) and (b). Morphological binary hole filling (MBHF) operation eliminated outliers as seen in (c) and (d). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

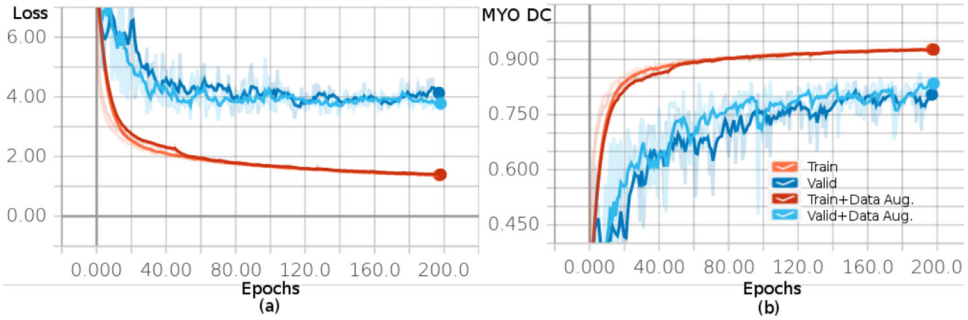


Fig. 9. Comparison of the learning curves of the DFCN-C with and without data augmentation. (a) Loss curves, (b) Dice score curves for MYO class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

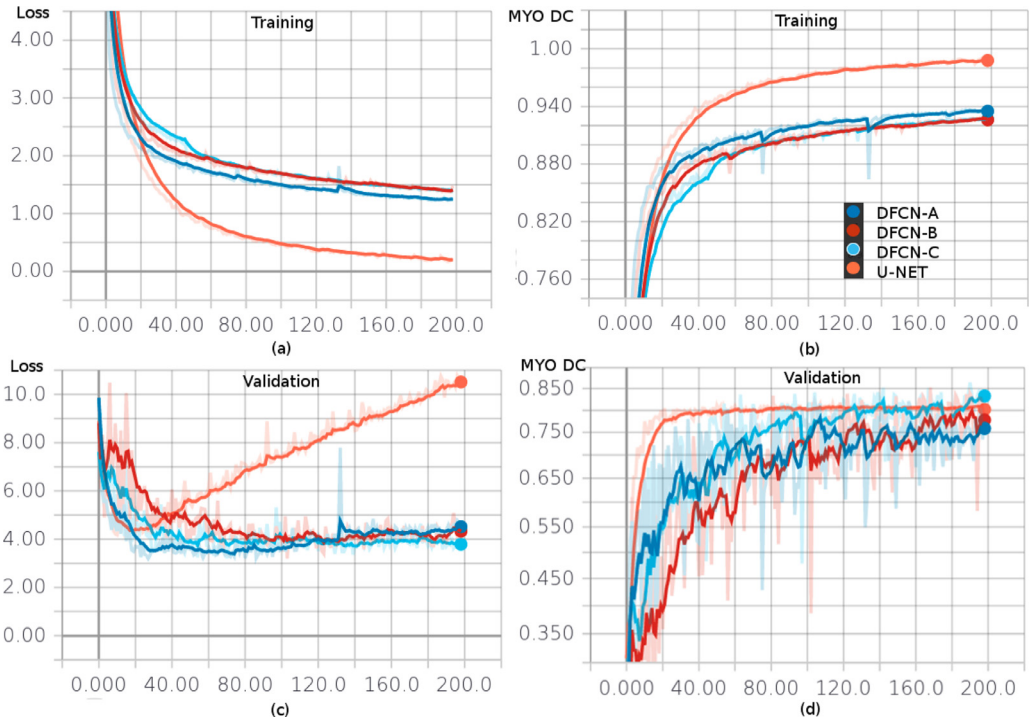


Fig. 10. Comparison of the learning curves of the DFCN-A, DFCN-B, DFCN-C (proposed) and U-Net. (a) Training loss curves, (b) Dice score curves for MYO class during training, (c) Validation loss curves, (d) Dice score curves for MYO class during validation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Evaluation results for baseline models and DFCN-C. The values are provided as mean (standard deviation). The GPU memory usage shown in the table was when with input images were of dimension 128×128 and the batch-size was 1. Numbers with * indicate significant difference compared to the DFCN-C, according to Student's t-test for two independent samples on the Dice and Hausdorff metrics ($p < 5 \cdot 10^{-2}$).

Method	DICE LV		DICE RV		DICE MYO		Mean Dice	
	ED	ES	ED	ES	ED	ES		Params.
U-Net	0.94 (0.02)	0.90 (0.06)	0.90 (0.04)*	0.82 (0.07)	0.85 (0.05)*	0.86 (0.04)	0.88 (0.05)*	1,551k
DFCN-A	0.96 (0.02)	0.91 (0.08)	0.94 (0.02)	0.88 (0.08)	0.89 (0.03)	0.90 (0.03)	0.91 (0.04)	435k
DFCN-B	0.96 (0.02)	0.90 (0.08)	0.92 (0.05)	0.84 (0.17)	0.88 (0.04)	0.88 (0.05)	0.90 (0.07)	360k
DFCN-C	0.96 (0.02)	0.90 (0.08)	0.95 (0.02)	0.87 (0.08)	0.89 (0.03)	0.89 (0.03)	0.91 (0.04)	371k
	HD LV		HD RV		HD MYO		Mean HD	
	ED	ES	ED	ES	ED	ES		GPU M.
U-Net	10.99 (12.29)	11.90 (10.56)	12.56 (8.39)*	12.67 (9.42)*	9.64 (6.85)*	12.22 (8.08)*	11.66 (9.26)*	3GB
DFCN-A	6.85 (10.45)	4.00 (2.73)	4.59 (2.09)	5.24 (1.67)*	8.16 (8.95)	4.65 (3.17)	5.58 (4.84)	2GB
DFCN-B	8.28 (10.24)	6.68 (7.37)	5.19 (2.12)	7.88 (4.91)	8.15 (7.51)	9.18 (7.19)	7.56 (6.56)*	1GB
DFCN-C	4.42 (6.39)	6.51 (6.40)	4.20 (1.81)	7.10 (2.95)	4.48 (4.52)	5.87 (4.35)	5.43 (4.40)	1GB

segmenting out relevant cardiac structures from MR images at end systole phase (ES) when compared to end diastole (ED) phase. The possible reason for larger HD metric could be attributed to difficulty in segmenting the right atrium from right ventricle at basal slices during end systole phase. In some of the test cases, the right atrium was confused to be RV which lead to generation of false positives. In the interest of minimizing the GPU memory foot-print and the number of trainable parameters, DFCN-C was preferred.

3.9. Classifier selection and ensemble evaluation for cardiac disease diagnosis

Table 7 tabulates the performance of various classifiers for the task of cardiac disease diagnosis. Classifiers which showcased a mean accuracy score greater than or equal to 95% in 5-fold cross-validation study were retained to form the ensemble. Fig. 11 shows the confusion matrices of the classification ensemble predictions from 5-fold cross-validation on the training set. The confusion matrix of the first stage classifier indicated misclassifications were prominently due to complexities in distinguishing between MINF and DCM as they visually appear to be similar. DCM cases lack global myocardial contraction whereas in MINF cases it is limited to few segments of myocardium. Hence, a second stage 2-class MINF vs. DCM expert classifier was designed for refining the predictions from first stage. The confusion matrix of the expert classifier indicated reduced misclassification rate.

4. Challenge results

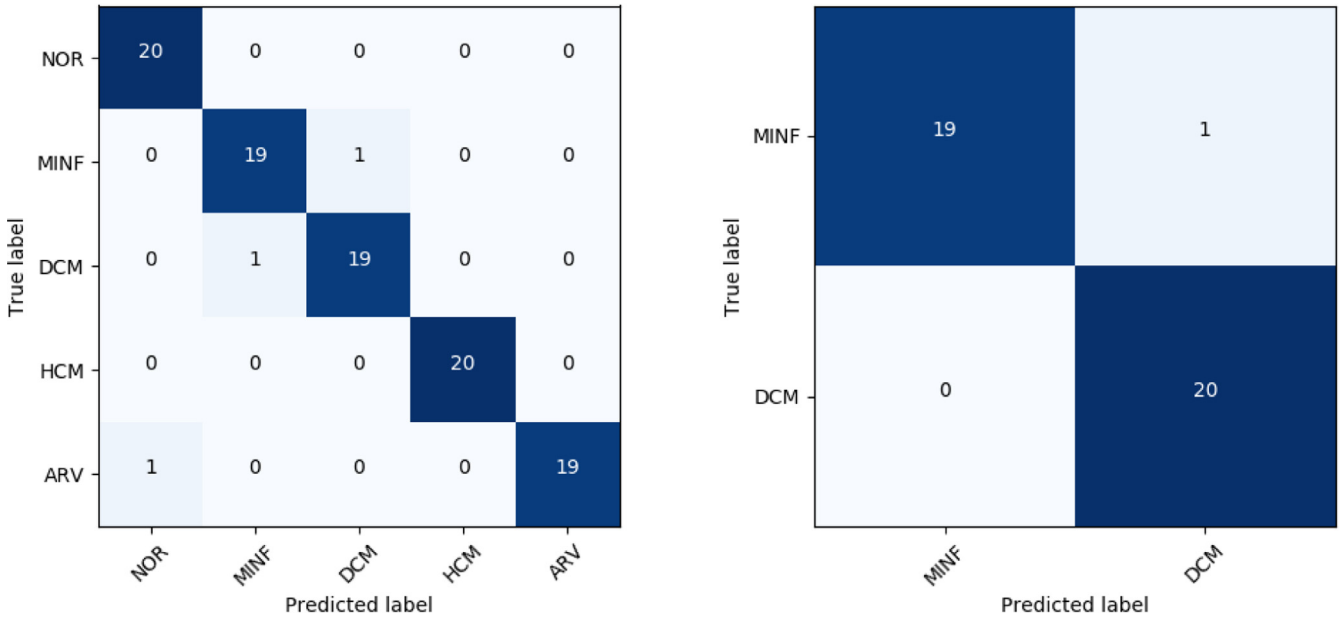
4.1. Performance evaluation on ACDC challenge

4.1.1. Network architecture and training settings

The network architecture and training settings were as discussed in Fig. 6 and Section 3.2.

4.1.2. Segmentation evaluation metrics

The challenge organizers provided an online platform for evaluation of segmentation results. On this platform, the performance of submitted methods were compared and ranked on the basis of geometrical and clinical perspective. **Geometric metrics** measured



(a) Stage 1: Voting classifier (b) Stage 2: Expert MLP classifier

Fig. 11. The figure shows the confusion matrices of the classification ensemble predictions from 5-fold cross-validation on the training set (n=100). (a) Stage 1 voting classifier and (b) Stage 2 Expert MLP classifier. The misclassifications seen in stage 1 were in distinguishing (i) between MINF and DCM, and (ii) ARV from NOR.

Table 7
The table lists the classifiers evaluated and their corresponding five-fold cross-validation accuracy scores. The classifiers were evaluated for both first stage of the ensemble (5-class) and expert discrimination (2-class). The values are provided as mean (standard deviation). LR- Logistic Regression, RF- Random Forest with 1000 trees, GNB- Gaussian Naive Bayes, XGB- Extreme Gradient Boosting with 1000 trees, SVM- Support Vector Machine with Radial Basis kernel, MLP- Multi-layer perceptron with 2-hidden layers with 100 neurons each, K-NN- 5-Nearest Neighbors, Vote- Voting classifier based on SVM, MLP, NB & RF.

Classifier	5-class	2-class
LR	0.94 (0.06)	0.82 (0.06)
RF	0.96 (0.02	0.85 (0.09)
GNB	0.96 (0.04)	0.82 (0.06)
XGB	0.93 (0.04)	0.88 (0.11)
SVM	0.95 (0.04)	0.85 (0.09)
MLP	**0.97 (0.02)**	**0.97 (0.05)**
K-NN	0.91 (0.04)	0.85 (0.09)
Vote	0.97 (0.04)	0.93 (0.06)

segmentation accuracy of LV, MYO and RV structures at ED and ES phases using Dice and 3D Hausdorff distance. **Clinical metrics** included correlation, bias and standard deviation values. These three metrics were computed from the measurements of: i) the ED volumes of LV and RV; ii) the ejection fractions of LV and RV; iii) the MYO mass at ED.

4.1.3. Segmentation results

Table 8 presents our segmentation results on the ACDC-2017 challenge test dataset (n=50). The top, middle and bottom part list the segmentation results of LV, RV and MYO structures respectively. Our method was ranked 2nd for LV and MYO and 3rd for RV. We compared our method with the top 4 performing deep learning approaches from other participating teams in the challenge. The overall top performing method (Isensee et al., 2017) developed an ensemble comprising of 10 models based on 2D and 3D U-Net architectures trained with Dice loss. This approach was closely followed by other methods with Jang et al. (2017) implementing 2D M-Net architecture with weighted cross-entropy loss function; Baumgartner et al. (2017) implementing 2D U-Net with cross-entropy loss; Zotti et al. (2018) implementing 2D Grid-Net architecture with an automatically-registered shape prior; Patravali et al. (2018) implementing 2D U-Net with Dice loss. During inference, our ROI pre-processing method cropped a patch of size 128×128 from the predicted LV center. Sometimes, the predicted ROI partially missed RV borders due to larger image matrix size/ resolution. This caused minor under-segmentation of RV structure, which was reflected in lower Dice score & higher Hausdorff distance for our RV segmentation.

In terms of clinical metrics, all the methods listed in the Table 8 gave correlation scores above 0.97 for EF of LV and also for volumes of LV, RV and MYO at ED. The EF of RV was the most difficult clinical metric to estimate and even the best performing method had a correlation score of 0.9. All the methods had difficulty in RV segmentation, and hence gave highest Hausdorff distances and lowest Dice scores at ES. When compared to other structures, Dice score of MYO was the lowest for all the methods, as MYO segmentation required precise delineation of two walls. Fig. 12 shows the results of segmentation produced by our method at ED and ES phase of the cardiac cycle on held-out training dataset reserved for testing. Our approach gave accurate results on most of the mid-level slices, but sometimes gave erroneous segmentation at basal and apical slices. Table 9 compares segmentation results with and without post-processing. The post-processing of the results removed small erroneous segments, which was reflected in improved Hausdorff distance metric.

4.1.4. Classification results

Table 10 presents our results on automated disease classification challenge. The organizers calculated accuracy on the entire challenge test dataset, and also per disease group precision and recall. Confusion matrix was provided to highlight the results. The challenge test set comprised of a small number of patients (n=50) and a misclassification lead to an accuracy drop of 2%. Our proposed approach gave an accuracy score of 1. Our scores

Table 8
Comparisons with different approaches on cardiac segmentation. The evaluations of left ventricle, right ventricle and myocardium are listed in top, middle and bottom, respectively. Our proposed method had an overall ranking of 2. Note: DC- Dice score, HD- Hausdorff distance, cor- correlation. The mean values provided for DC and HD metrics.

Rank	Method	DC ED	DC ES	HD ED	HD ES	EF cor.	EF bias	EF std.	Vol. ED corr.	Vol. ED bias	Vol. ED std.
LV											
1	Isensee et al. (2017)	0.968	0.931	7.384	6.905	0.991	0.178	3.058	0.997	2.668	5.726
2	**Ours**	0.964	0.917	8.129	8.968	0.989	-0.548	3.422	**0.997**	0.576	**5.501**
3	Jang et al. (2017)	0.959	0.921	7.737	7.116	0.989	-0.330	3.281	0.993	-0.440	8.701
4	Baumgartner et al. (2017)	0.963	0.911	6.526	9.170	0.988	0.568	3.398	0.995	1.436	7.610
RV											
1	Isensee et al. (2017)	0.946	0.899	10.123	12.146	0.901	-2.724	6.203	0.988	4.404	10.823
2	Zotti et al. (2018)	0.941	0.882	10.318	14.053	0.872	-2.228	6.847	0.991	-3.722	9.255
3	**Ours**	0.935	0.879	13.994	13.930	0.858	-2.246	6.953	0.982	-2.896	12.650
4	Baumgartner et al. (2017)	0.932	0.883	12.670	14.691	0.851	1.218	7.314	0.977	-2.290	15.153
MYO		**DC ED**	**DC ES**	**HD ED**	**HD ES**	**Vol. ES corr.**	**Vol. ES bias**	**Vol. ES std.**	**Mass ED corr.**	**Mass ED bias**	**Mass ED std.**
1	Isensee et al. (2017)	0.902	0.919	8.720	8.672	0.985	-3.842	9.153	0.989	-4.834	7.576
2	**Ours**	0.889	0.898	9.841	12.582	0.979	-2.572	11.037	**0.990**	-2.873	**7.463**
3	Baumgartner et al. (2017)	0.892	0.901	8.703	10.637	0.983	-9.602	9.932	0.982	-6.861	9.818
4	Patravali et al. (2018)	0.882	0.897	9.757	11.256	0.986	-4.464	9.067	0.989	-11.586	8.093

Table 9
Automated segmentation results on ACDC challenge dataset (n=50) with and without the application of proposed post-processing scheme. The values are provided as mean (standard deviation). Numbers with * indicate significant difference compared to the model without post-processing step, according to a two-sided, paired *t*-test on the Dice and Hausdorff metrics (* $p < 5 \cdot 10^{-2}$). PP:- Post-Processing.

PP	DICE LV		DICE RV		DICE MYO	
	ED	**ES**	**ED**	**ES**	**ED**	**ES**
√	0.96 (0.01)	0.92 (0.06)	0.94 (0.04)	0.88 (0.07)	0.89 (0.03)	0.90 (0.03)
✗	0.96 (0.01)	0.92 (0.06)	0.94 (0.04)	0.88 (0.07)	0.89 (0.03)	0.90 (0.03)
	HD LV		**HD RV**		**HD MYO**	
	ED	**ES**	**ED**	**ES**	**ED**	**ES**
√	8.13 (6.70)	8.97 (5.16)	13.99 (9.40)*	13.93 (8.17)	9.84 (6.60)	12.58 (10.18)
✗	8.22 (6.76)	9.59 (6.01)	17.37 (17.66)	14.44 (9.35)	9.75 (6.76)	12.97 (10.18)

Table 10
Comparisons with different approaches on automated cardiac disease diagnosis. Our accuracy score reported in the table is from post-MICCAI leader board.

Rank	Method	Accuracy
1	**Ours**	**1**
2	Isensee et al. (2017) and Cetin et al. (2017)	0.92
3	Wolterink et al. (2018)	0.86

were boosted due to the incorporation of the second stage 2-class MINF versus DCM classifier. The classification metrics with only first stage classifier are provided in Table 11. The confusion matrix of the first stage indicated difficulties in distinguishing DCM from MINF cases. These misclassifications were resolved with the inclusion of second stage classifier.

Table 10 compares our method with other approaches which used automated and semi-automated segmentation methods for feature extraction. The best performing automated segmentation method (Isensee et al., 2017) extracted a series of instant and dynamic features from segmentation maps and used an ensemble of 50 MLPs and a 1000-trees RF classifiers. Wolterink et al. (2018) had developed an automated segmentation approach using dilated CNN, from segmentation maps they extracted 14 features and used a 1000-trees RF classifier. Cetin et al. (2017) used semi-automatic segmentation method, from the segmentation maps they extracted 567 radiomic (shape, intensity and texture) and physiological features and used a SVM classifier.

Even though our segmentation method was judged as second best performing, our disease classification model (Section 2.7.2)

Table 11
Classification metrics for automated cardiac disease diagnosis on ACDC challenge dataset (n=50) without second stage 2-class DCM vs. MINF classifier.

(a)

Group →	NOR	DCM	HCM	MINF	ARV
Recall	1	0.8	1	0.8	1
Precision	1	0.8	1	0.8	1
Overall Accuracy	0.92				

(b)

	NOR	DCM	HCM	MINF	RV
NOR	10	0	0	0	0
DCM	0	8	0	2	0
HCM	0	0	10	0	0
MINF	0	2	0	8	0
RV	0	0	0	0	10

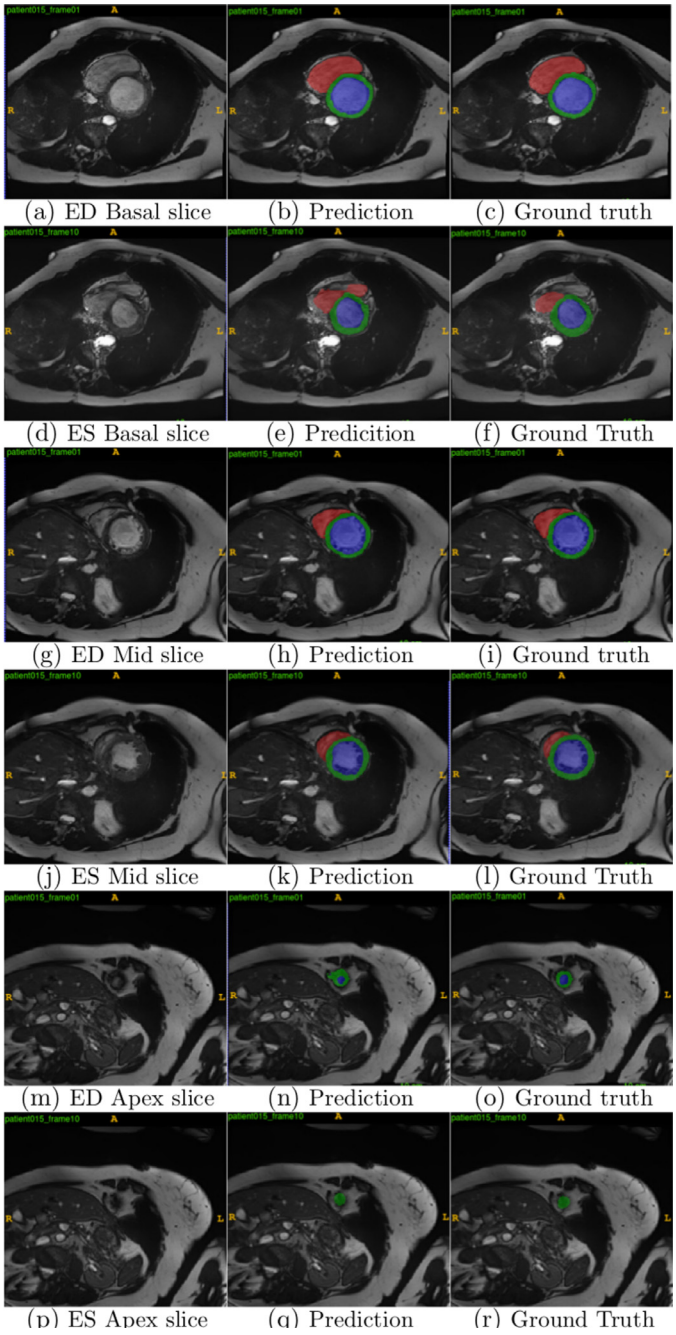


Fig. 12. Segmentation results at both ED & ES phases of cardiac cycles on a subset of ACDC training set reserved for testing. The columns from left to right indicate: the input images, segmentations generated by the model and their associated ground-truths. The rows from top to bottom indicate: short axis slices of the heart at basal, mid and apex. In all figures the colors red, green and blue indicate RV, MYO and LV respectively. The model did erroneous segmentation for RV in the basal slice (e) and the myocardium was over-segmented in the apical slice (n). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

could perfectly classify all the 50 patients. It was observed that most of the segmentation errors came from our network's inability to segment cardiac structures in extremely difficult slices of the heart such as apex and basal regions. Segmentation errors on apical slices had very minor impact on the overall volume computation and henceforth the features derived from it. Also, our model was tested on a very limited challenge test dataset (n=50) where certain disease-related abnormalities located in apical regions were most likely not present (e.g. patient cases with infarction at Apical regions). Even the segmentation results in Table 8 reveal that our approach gave very accurate correlation scores on the automatically extracted clinical indices with low bias and standard deviation on the LV EDV, MYO ED mass and LV EF, which are commonly used cardiac physiological parameters to diagnose heart disease.

4.2. Performance evaluation on LV-2011 segmentation challenge

4.2.1. Network architecture and training settings

The LV-2011 network's final classification layer was modified to have 2 feature maps (myocardium class plus background). The rest of the configuration remained as per Section 3.2 and Fig. 6. The LV-2011 training dataset comprised of 100 patients and annotations for myocardium were provided in all cardiac phases ($\approx 29k$ 2D annotated images). The patient cases were randomly split into 70: 20: 10 for training, validation and internal testing. The model was trained for 50 epochs as described in Section 3.3. Additionally the data-augmentation scheme included random vertical and horizontal flips.

4.2.2. Segmentation evaluation metrics

The challenge organizers evaluated our segmentation results on those images of final validation set for which reference consensus contours CS* Suinesiaputra et al. (2014) were available. The organizers categorized individual images of the final validation set into apex, mid and basal slices. The reported metrics were Jaccard index, Dice score, Sensitivity, Specificity, Accuracy, Positive Predictive Value and Negative Predictive Value.

4.2.3. Segmentation results

Table 12 compares the results between our proposed approach and other published results using LV-2011 validation dataset (n=100). Our Jaccard index (mean ± standard deviation) was 0.68 ± 0.16, 0.78 ± 0.13, 0.74 ± 0.18 for the apex, mid and base slices respectively. The errors were mostly concentrated in apical slices. Fig. 13 illustrates the myocardium prediction by our approach at three different slice levels on a subset of training set reserved for internal testing. For most of the evaluation measures including Jaccard index, our approach was on par with other fully automated published methods. The AU method used manual guide-point modeling and required human expert approval for all slices and frames. The CNR (Tan et al., 2017) method used manual input for identifying basal and apical slices and used convolutional regression approach to trace endocardium and epicardium contours in polar space domain. They also used a CNN for predicting LV center and their networks had about 3 million parameters. Our network performance was on par with FCN (Tran, 2016). But, in contrast to our network architecture and training settings, FCN had about 11 million parameters and used 95% of the dataset for training.

4.3. Model generalization across different data distributions

In this section we evaluated our trained segmentation model's generalization on other challenge test sets.

4.3.1. ACDC model on LV-2011 validation dataset

Comparison of models trained on LV-2011 and ACDC-2017 dataset for the task of myocardium segmentation on LV-2011 final validation set (n=100) is provided in Table 13. The ACDC-2017 model used a smaller training set (n=1.4k images) with annotations for LV, RV and MYO at ED and ES phases only. The trained ACDC-2017 model was used to segment myocardium at all cardiac phases of the LV-2011 validation set, and gave a comparable Jaccard score of 0.71. These results corroborates the ACDC-2017 model's effectiveness even with sparse annotations across cardiac phases and its generalization across different datasets.

Table 12
Comparison of segmentation performance with other published approaches on LV2011 validation set using the consensus contours. AU (Li et al., 2010), AO (Fahmy et al., 2011), SCR (Jolly et al., 2011), DS, and INR (Margeta et al., 2011) values were taken from Table 2 of (Suinesiaputra et al., 2014). FCN values are taken from Table 3 of Tran (2016) and CNR regression was taken from Table 3 of Tan et al. (2017). Values are provided as mean (standard deviation) and in descending order by Jaccard index. FA- Fully Automated.

Method	FA	Jaccard	Sensitivity	Specificity	PPV	NPV
AU	✗	0.84 (0.17)	0.89 (0.13)	0.96 (0.06)	0.91 (0.13)	0.95 (0.06)
CNR	✗	0.77 (0.11)	0.88 (0.09)	0.95 (0.04)	0.86 (0.11)	0.96 (0.02)
FCN	✓	0.74 (0.13)	0.83 (0.12)	0.96 (0.03)	0.86 (0.10)	0.95 (0.03)
Ours	✓	0.74 (0.15)	0.84 (0.16)	0.96 (0.03)	0.87 (0.10)	0.95 (0.03)
AO	✗	0.74 (0.16)	0.88 (0.15)	0.91 (0.06)	0.82 (0.12)	0.94 (0.06)
SCR	✓	0.69 (0.23)	0.74 (0.23)	0.96 (0.05)	0.87 (0.16)	0.89 (0.09)
DS	✗	0.64 (0.18)	0.80 (0.17)	0.86 (0.08)	0.74 (0.15)	0.90 (0.08)
INR	✓	0.43 (0.10)	0.89 (0.17)	0.56 (0.15)	0.50 (0.10)	0.93 (0.09)

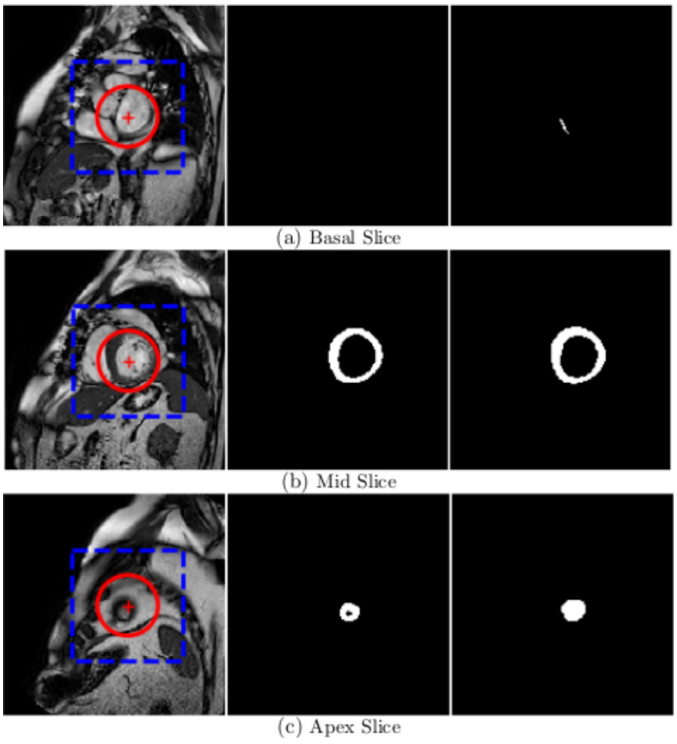


Fig. 13. From Right to Left: Input, Prediction and Ground Truth for basal, mid and apical slices. The segmentation results are for a subset of LV-2011 training set reserved for testing. The ⊕ indicates the ROI center detected by Fourier-Circular Hough Transform approach. The red circle represents the best fitting circle across all slices (myocardium contours of the mid-slices are more circular in shape and hence the best fitting Hough circle radii mostly encompass them). The blue bounding box indicates the 128×128 patch cropped around ROI center for feeding the segmentation network. Some of the basal slices of the training data had ground truths with partial myocardium. Even though we trained our model with such slices, our model showed the ability to generalize well.

4.3.2. LV-2011 model on ACDC testing dataset

Comparison of models trained on LV-2011 and ACDC-2017 dataset for the task of myocardium segmentation on ACDC challenge test dataset (n=50) is provided in Table 14. The LV-2011 model used a relatively larger training set (n=20k) with MYO annotations at all cardiac frames, but still showed a drop in performance on ACDC challenge test set when compared to ACDC-2017 model, this could be possibly attributed to multi-centric data heterogeneity due to differences in scanner type and acquisition protocols. Another possible reason could be due to differing heart pathologies in two different training datasets.

4.3.3. ACDC model on Kaggle testing dataset

The Kaggle challenge utilized continuous ranked probability score (CRPS) (Booz Allen Hamilton Inc and Kaggle, 2015) for evaluation of LV volumes. The segmentation model trained on ACDC dataset was used to segment LV from the short-axis views of the cardiac MRI. From the segmentation, the LV volumes at the systolic and diastolic phases were estimated based on the minimum and maximum volume in the cardiac phases. A linear model was used to fit Kaggle training dataset against the calculated LV volume, patient's age and sex. In order to predict the cumulative probability distribution of LV volumes at both the phases, a Gaussian distribution was employed and its mean and standard deviation were obtained from the fitted linear model. On the final testing set (n=440), this model gave an CRPS score of 0.0127, which would have placed 10th out of 192 participating teams.

5. Discussion and conclusion

In this paper, we demonstrate the utility and efficacy of fully convolutional multi-scale residual DenseNets for automated cardiac segmentation. We also demonstrate that on a limited dataset, perfect cardiac pathology classification was achieved by employing machine learning techniques and engineering features from the segmentation maps. We propose a network design and training methodology that is parameter and memory efficient for the task of segmenting left ventricle, right ventricle and myocardium from short-axis cine MR images. We demonstrate the generalization ability of our segmentation approach by achieving near state-of-the-art segmentation results on three publicly available benchmark cardiac MR datasets. These datasets were acquired across multiple clinical institutions, scanners, populations and heart pathologies, thereby exhibiting variability in cardiac anatomical and functional characteristics. Comprehensive evaluations based

Table 13
The table compares the myocardium segmentation performance evaluated on LV-2011 final validation set (n=100). The values indicate mean (standard deviation).

Training Dataset (No. of images)	Jaccard	Dice	Accuracy	Sensitivity	Specificity	PPV	NPV
LV-2011 (20,360)	0.74 (0.15)	0.84 (0.14)	0.93 (0.04)	0.84 (0.16)	0.96 (0.03)	0.87 (0.10)	0.95 (0.03)
ACDC-2017 (1,352)	0.71 (0.13)	0.82 (0.11)	0.92 (0.04)	0.81 (0.15)	0.91 (0.06)	0.82 (0.12)	0.94 (0.06)

Table 14
The table compares the myocardium segmentation performance evaluated on ACDC-2017 final test set (n=50). Note: DC- Dice score, HD- Hausdorff distance, cor- correlation. The mean values provided for DC and HD metrics.

Training Dataset (No. of images)	DC ED	DC ES	HD ED	HD ES	Vol. ES corr.	Vol. ES bias	Vol. ES std.	Mass ED corr.	Mass ED bias	Mass ED std.
ACDC-2017 (1,352)	0.889	0.898	9.841	12.582	0.979	−2.57	11.04	0.990	−2.873	7.463
LV-2011 (20,360)	0.85	0.86	11.78	11.98	0.913	−7.19	22.29	0.974	−19.22	15.60

on multiple metrics revealed that our entire pipeline starting from classical computer vision based techniques for ROI extraction to CNN based segmentation was accurate, robust and efficient. For a typical 4D MR sequence with data dimension of $256 \times 256 \times 10 \times 30$ ($Height \times Width \times Slices \times Phases$) our approach took approximately 3 seconds for ROI detection and 7 seconds for segmentation. The proposed approach without ROI extraction step gave similar performance, however this necessitates higher computational resources for training and inference.

When compared to other CNN based approaches for cardiac segmentation (Lieman-Sifry et al., 2017; Tran, 2016; Tan et al., 2017), our network required the least number of trainable parameters (0.4 million, an order of 10 fold reduction when compared to standard U-Net based architectures). Based on experimental studies, it was observed that in the absence of proper training strategy like data-augmentation, FCN / U-Net based architectures showed higher tendencies to over-fit on smaller datasets. Our network's connectivity pattern ensured lower model complexity, which in turn led to better generalization when trained on a limited dataset. Memory explosion in DenseNet based FCN is due to concatenation of feature maps in the up-sampling path. The memory bottleneck in FCN due to concatenation was circumvented by incorporating residual type long skip and short-cut connections in the proposed up-sampling path. For multi-scale processing of images, an inception layer was introduced in the segmentation network. Though we limit the inception structure only to the first layer, this could be extended to deeper layers in a computationally efficient manner. The network trained on the proposed dual loss function generated anatomically plausible segmentation maps and qualitatively improved segmentations when compared to networks trained on standard loss functions such as cross-entropy loss or Dice loss. The generic nature of our segmentation framework allows direct application on different segmentation tasks without major adaptations. The network architecture can be extended for processing 3-D medical volumes by employing 3D convolution operation.

Currently, most of the CNN based techniques produce erroneous segmentation at basal and apical slices due to the following two major issues, namely:- (i) Uncertainties in the ground truth at valve level due to limited long-axis resolution of MRI, and (ii) Difficulty in exactly defining the apex and also presence of trabeculations near apex. The erroneous segmentations at apex and basal slices lead to higher Hausdorff distance and this leaves scope for further research and improvement from a segmentation point of view. Dice and Jaccard scores are the often used metrics to evaluate the performance of segmentation method. However, in Fig .7, we illustrate networks that generate anatomically impossible segmentation which are seldom detected by the currently used metrics. This clearly necessitates the need for coming up with better evaluation metrics, which penalizes abnormalities in segmentation.

Automated cardiac disease classification results on a limited dataset indicate that feature engineering based on domain knowledge of various cardiac abnormalities could yield machine learning techniques near perfect classification scores. However, these techniques need to be validated on a larger cohort comprising of various other cardiac pathologies so as to gauge the generalization capabilities. In a clinical setup, cardiac diagnosis is based upon cardiac physiological parameters estimated via segmentation of cardiac structures in MR images. An alternative method to avoid feature engineering from segmentation maps would be to device deep learning based techniques to predict the pathology directly from the MR sequences. Apart from requiring a larger annotated dataset, this calls for further research and development of techniques capable of learning spatial and temporal dynamics of heart in a cardiac cycle. In this direction, some of the recent works (Kong et al., 2016; Xue et al., 2018) on CNN-LSTM architectures for direct estimation of cardiac parameters (such as cardiac phases, regional wall thickness and LV areas) from MR sequences can be explored for disease classification.

Appendix A. Evaluation metrics

A brief overview of the main metrics reported in the literature used for comparative purposes are listed in this section. Let P and G be the set of voxels enclosed by the predicted and ground truth contours delineating the object class in a medical volume respectively. The following evaluation metrics were used to assess the quality of automated segmentation methods using the ground truth as reference:

1. **Dice overlap coefficient** is a metric used for assessing the quality of segmentation maps. It basically measures how similar predicted label maps are with respect to ground truth. The Dice score varies from zero to one (in-case of perfect overlap).

$$DICE = \frac{2|P \cap G|}{|P| + |G|} \tag{A.1}$$

2. **Hausdorff distance** is a symmetric measure of distance between two contours and is defined as:

$$H(P, G) = \max(h(P, G), h(G, P)) \tag{A.2}$$

$$h(P, G) = \max_{p_i \in P} \min_{g_i \in G} ||p_i - g_i|| \tag{A.3}$$

A high Hausdorff value implies that the two contours do not closely match. The Hausdorff distance is computed in millimeter with spatial resolution obtained from the DICOM tag Pixel Spacing.

Appendix B. Region of interest detection

B1. Fourier and statistical analysis methods for cardiac cine MR

The discrete Fourier transform Y of an N-D array X is defined as

$$Y_{K_1,K_2,\ldots,K_N} = \sum_{n_1=0}^{m_1-1} \omega_{m_1}^{K_1 n_1} \sum_{n_2=0}^{m_2-1} \omega_{m_2}^{K_2 n_2} \cdots \sum_{n_N=0}^{m_N-1} \omega_{m_N}^{K_N n_N} X_{n_1,n_2,\ldots,n_N} \tag{B.1}$$

Each dimension has length m_k for $k = 1, 2, \cdots, N$, and $\omega_{m_k} = e^{\frac{2\pi i}{m_k}}$ are complex roots of unity where i is the imaginary unit.

It can be seen that the N-D Fourier transform Eq. (B.1) of an N-D array was equivalent to computing the 1-D transform along each dimension of the N-D array. The short-axis cardiac MR images of a slice were taken across entire cardiac cycle and these

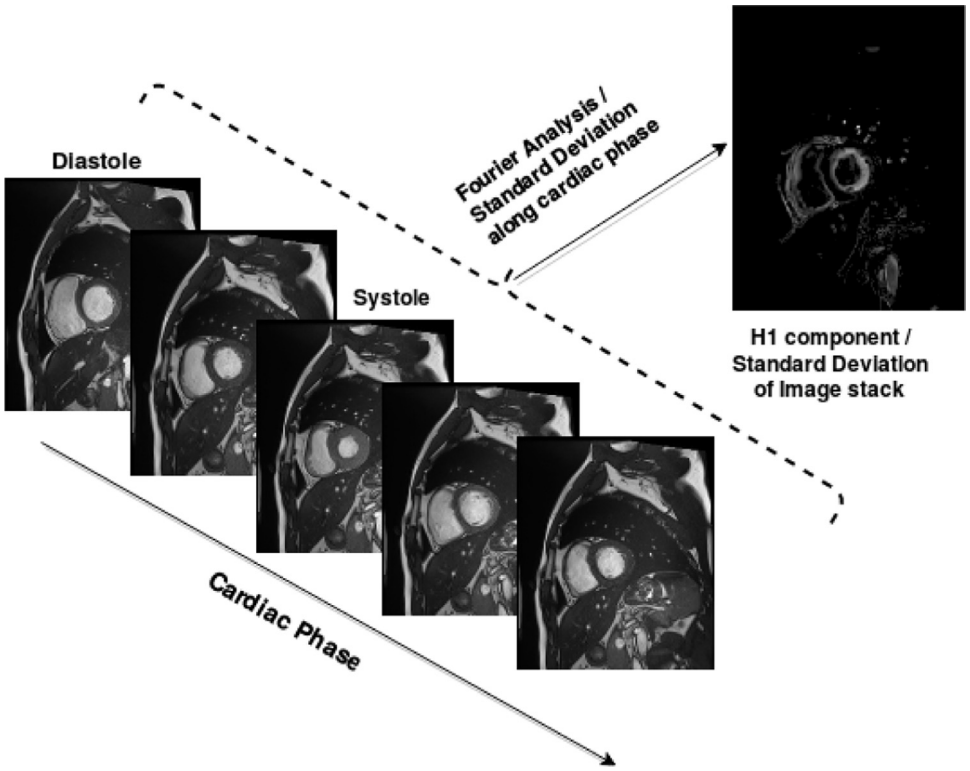


Fig. B.14. The figure shows an example of H1 component image got from a series of MR images of a cardiac short axis slice. It can be seen that most of the chest cavity excluding heart and some adjacent structures have disappeared. If we consider a pixel just outside the LV blood pool at the end diastole, it goes from being bright when it was inside the blood pool to dark when it was in myocardium at end systole, because the region containing blood was contracting inwards. The pixel gets brighter again as the heart approaches end diastole. The said pixel's intensity variation will resemble a waveform having frequency same as the heartbeat, hence the H1 component captures those structures of the heart which were responsible for heart beat. The standard deviation of image stack computed along the cardiac phase yielded an image similar to H1 component.

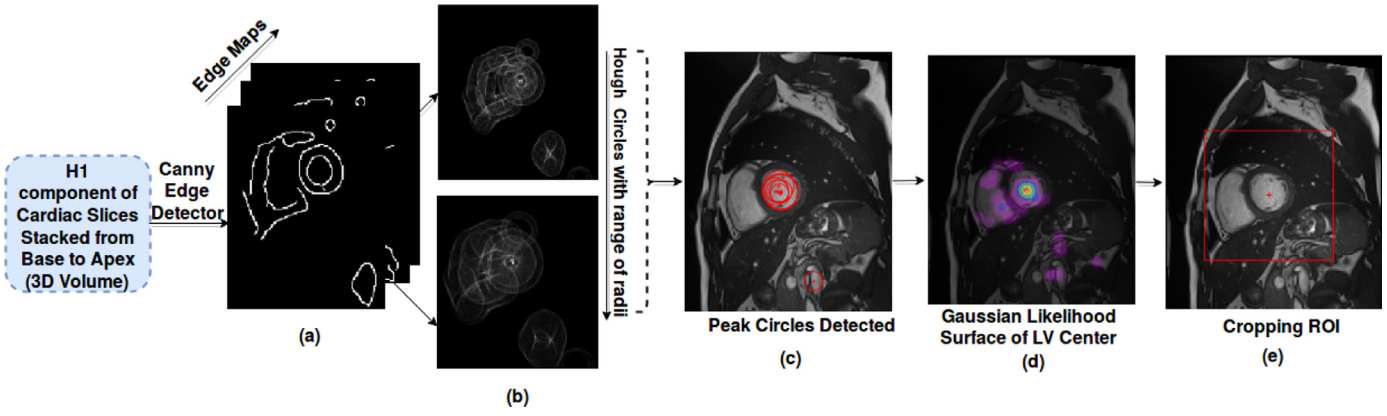


Fig. B.15. Gaussian kernel-based circular Hough transform approach was used for left ventricle (LV) localization. The steps involved in ROI detection were (a) Canny edge detection was done on each short axis slice's H1 component image, (b) For each of the edge maps the Hough circles for a range of radii were found, (c) For each of the edge maps, only P highest scoring Hough circles were retained, where P was a hyper-parameter, (d) For each of the retained circles, votes were cast using an Gaussian kernel that models the uncertainty associated with the circle's center. This approach makes the transform more robust to the detection of spurious circles (in the figure LV center's likelihood surface is overlayed on a slice, the red and purple regions indicates high and low likelihood of LV center respectively), (e) The maximum across LV likelihood surface was selected as the center of the ROI and a square patch of fixed size (128×128) was cropped. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sequence of images can be treated as 2-dimensional signal varying over time (2D+T signal - $Height \times Width \times Time$). The structures pertaining to heart like the myocardium and ventricles show significant changes due to heart beat motion. Hence, by taking 3-D Fourier Transform along time axis and analyzing the fundamental Harmonic (also called H1 component, where H stands for Hilbert Space) it was possible to determine pixel regions corresponding to ROI which show strongest response to cardiac frequency.

The first harmonic of the 3-D FFT was transformed back into original signal's domain (spatial) using 2-D inverse FFT. The result of the previous transformation lead to Complex valued signal. Since, the original signal was Real, the phase component was ignored and only the magnitude of the H1 component (see Fig. B.14) was retained. The H1 components were estimated for all the slices of the heart starting from base to apex and stacked to form a 3-D volume. The noise present throughout this whole volume was reduced by discarding pixel values which were less than 1% of

the maximum pixel intensity in the whole volume. A simpler and faster alternative to H1 component was to compute standard deviation of 3D image stack along time axis. The results obtained using standard deviation approach were similar to Fourier analysis method (Lin et al., 2006a).

Table B.15
The table reports the statistics of the euclidean distance between the predicted LV center and the LV center derived from the ground truth on the ACDC dataset comprising of 1902 annotated images.

Mean	Standard deviation	Maximum
4.0	3.83	36.24

B2. Circular Hough transform

The LV myocardium wall resembles circular ring and this contracts and expands during the cardiac cycle. The pixel regions whose intensity varied because of this movement were captured by the H1 component (seen as bright regions in the image). On applying Canny edge detection on these H1 component images, two concentric circles were seen which approximate the myocardial wall boundaries at end diastole and end systole phases. Henceforth, the localization of the left ventricle center was done using Gaussian kernel-based circular Hough transform approach (See Fig. B.15).

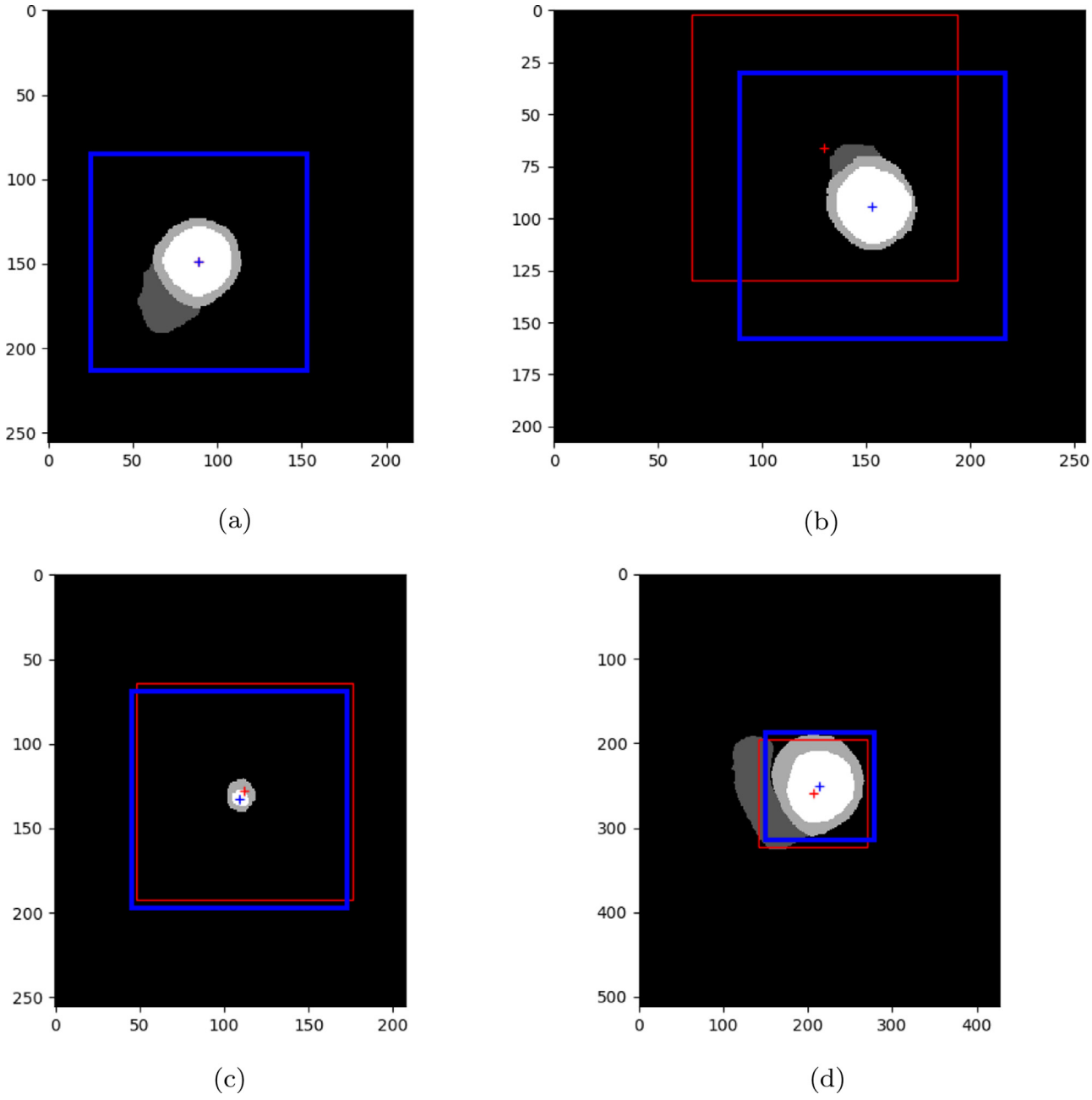


Fig. B.16. The figure shows the LV centers and its ROI bounding boxes of size = 128×128. The colors blue and red are used for the proposed ROI extraction method and ground truth reference respectively. (a) Euclidean distance was zero, (b) Euclidean distance was the maximum, (c) LV center predicted incorrectly on MYO region, and (d) ROI bounding box partially failed to capture heart's RV region owing to larger image matrix size (512×428). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

B3. Performance of ROI detection

On the ACDC dataset the statistics of the predicted LV center versus the LV center derived from the ground-truth are reported in Table B.15. The Fig. B.16 illustrates the performance of the proposed ROI extraction method on 4 different images. It was observed that even in cases where the predicted LV center was completely away from the ground truth LV center, the ROI bounding box of size 128×128 was sufficient to capture the complete heart region. In few MR images (as shown in Fig. B.16 (d)), the predicted ROI partially missed the heart due to larger image matrix size. Such images included an additional pre-processing step of re-sampling to lower resolution and a post-processing step of re-sampling the predicted segmentations to their original resolution.

Appendix C. Overview of CNN connectivity pattern variants

C1. Overview of DenseNets

DenseNets are built from dense blocks and pooling operations, where each dense block (DB) is an iterative concatenation of previous feature maps whose sizes match. A layer in dense block is composition of batch normalization (BN) (Ioffe and Szegedy, 2015), non-linearity (activation function), convolution and dropout (Srivastava et al., 2014). The output dimension of each layer has k feature maps where k, is referred to growth rate parameter, is typically set to a small value (e.g. k = 8). Thus, the number of feature maps in DenseNets grows linearly with the depth. For each layer in a DenseNet, the feature-maps of all preceding layers of matching spatial resolution are used as inputs, and its own feature-maps are passed onto subsequent layers. The output of the l^{th} layer is defined as:

$$x_l = H_l([x_{l-1}, x_{l-2}, \cdots, x_0]) \tag{C.1}$$

where x_l represents the feature maps at the l^{th} layer and $[\cdots]$ represents the concatenation operation. In our case, H is the layer comprising of batch normalization (BN), followed by exponential linear unit (ELU) (Clevert et al., 2015), a convolution and dropout rate of 0.2. A transition down (TD) layer is introduced for reducing spatial dimension of feature maps which is accomplished by using a 1×1 convolution (depth preserving) followed by a 2×2 max-pooling operation. This kind of connectivity pattern has the following advantages:

- It ensures that the error signal can be easily back-propagated to earlier layers more directly so this kind of implicit deep supervision, as earlier layers can get more direct supervision from the final classification layer.
- Higher parameter and computation efficiency is achieved than a normal ConvNet. In a normal ConvNet the number of parameters is proportional to square of the number of channels (C) produced at output of each layer (i.e. $\mathcal{O}(C \times C)$), however in DenseNets the number of parameters is proportional to $\mathcal{O}(l_{th} \times k \times k)$ where l_{th} is the layer index and we usually set k much smaller than C so the number of parameters in each layer of the DenseNet is much fewer than that in normal ConvNet.
- It ensures that there is maximum feature reuse as the features fed to each layer is a consolidation of the features from all the preceding layers and this leads to learning features which are more diversified and pattern rich.
- DenseNets have shown to be well suited even when the training data is minimal this is because the connectivity pattern in DenseNets ensures that both low and high complexity features are maintained across the network. Hence, the final classification layer uses features from all complexity levels and thereby ensures smooth decision boundaries. Whereas in normal ConvNet the final classification layers builds on top of the last convolution layers which are mostly complex high level features composed of many non-linear transformations.

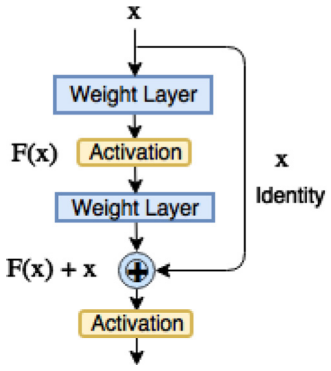


Fig. C.17. Residual learning: A building block.

C2. Overview of residual networks

Residual networks (ResNets) are designed to ease the training of very deep networks by introducing a identity mapping (shortcut connections) between input (x) and output of a layer ($H(x)$) by performing element-wise summation of the non-linear transformation introduced by the layer and its input. Referring to Fig. C.17, the residual block is reformulated as $H(x) = F(x) + x$, which consists of the residual function $F(x)$ and input x. The idea here is that if the non-linear transformation can approximate the complicated function $H(x)$, then it is possible for it to learn the approximate residual function $F(x)$.

C3. Overview of Inception architectures

The Inception modules were a parallel sub-networks (Fig. C.18 (a)) introduced in GoogLeNet architecture for ILSVRC 2014 competition, these modules are stacked upon each other, with occasional max-pooling layers with stride 2 to reduce the spatial dimension. The 1×1 convolution allowed dimension reduction, thereby making the architecture computationally efficient. These modules were used only in higher layers, whereas the lower layers maintained traditional convolution architecture because of technical constraints (Szegedy et al., 2015).

For the task of semantic segmentation we proposed to use modified version of the inception module (Fig. C.18(b)) only in the first layer, however this could be extended to higher layers. The inception architecture design allows the visual information to be processed at various scales and then aggregated so that the next stage can abstract features from the different scales simultaneously. The ratio of 3×3: 5×5: 7×7 convolutions could be skewed (like 2: 1: 1) as larger kernels have larger spatial coverage and can capture higher abstractions.

Appendix D. Cardiac disease classification

Table D.16 lists all the features extracted from the ground truth segmentations for cardiac disease classification task.

D1. Myocardial wall thickness variation profile features

Clinically, the myocardial wall thickness and its variation profile were the key discriminators in distinguishing between MINF and DCM (Karamitsos et al., 2009). The myocardial wall thickness peaks during end systole (ES) phase and is minimal during end

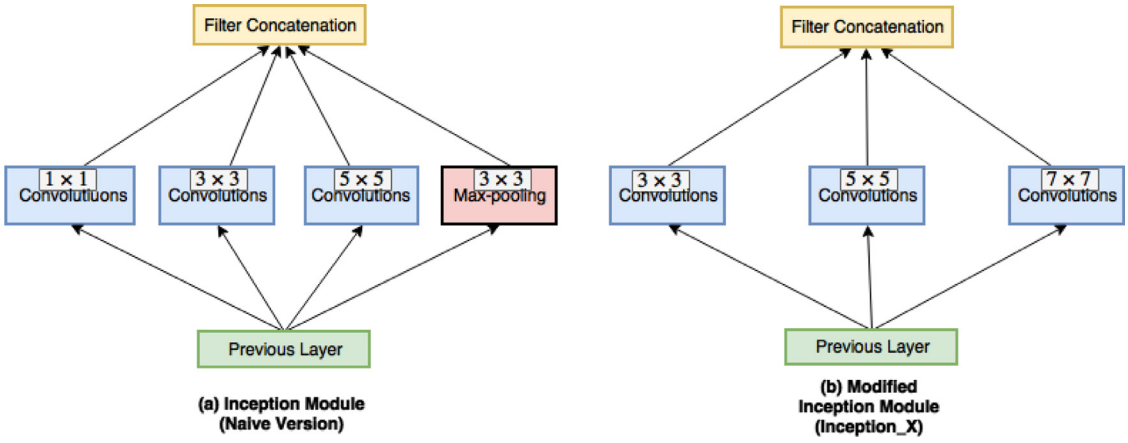


Fig. C.18. Inception module introduces parallel paths with different receptive field sizes by making use of multiple filters with different sizes, e.g. : 1×1, 3×3, 5×5 convolutions and 3×3 max-pooling layer. These feature maps are concatenated at the end. These operations are meant to capture sparse patterns of correlations in the stack of feature maps. The figures: (a) shows the naive version of the inception module, (b) modified version of inception module by excluding max-pooling and introducing a larger kernel (7×7) to increase the receptive field. For the task of semantic segmentation a small kernel helps in detecting small target regions whereas a larger kernel contributes to not only detecting larger target regions but also effectively aids in eliminating false positive regions that have similar properties as the target of interest.

Table D.16
The table lists all the features extracted from the predicted segmentation labels. All the 20 features were used for training the classifiers in the first stage of the ensemble. The expert classifier was trained with only a subset of the listed features indicated by *. *MWT|SA*:- Set of myocardial wall thickness measures (in mm) per short-axis slice. *Statistic(MWT|SA)|LA*:- Set of statistic (like mean, standard deviation) for all short axis slices when seen along long-axis at a particular cardiac phase.

Features	LV	RV	MYO
Cardiac volumetric features			
Volume at ED	✓	✓	✓
Volume at ES	✓	✓	✓
Ejection fraction	✓	✓	
Volume ratio: *ED*[*vol*(*LV*)/*vol*(*RV*)]	✓	✓	
Volume ratio: *ES*[*vol*(*LV*)/*vol*(*RV*)]	✓	✓	
Volume ratio: *ES*[*vol*(*MYO*)/*vol*(*LV*)]	✓		✓
Volume ratio: *ED*[*vol*(*MYO*)/*vol*(*LV*)]	✓		✓
Myocardial wall thickness variation profile			
ED[*max*(*mean*(*MWT*|*SA*)|*LA*)]			✓
ED[*stdev*(*mean*(*MWT*|*SA*)|*LA*)]			✓
ED[*mean*(*stdev*(*MWT*|*SA*)|*LA*)]			✓
ED[*stdev*(*stdev*(*MWT*|*SA*)|*LA*)]			✓
ES[*max*(*mean*(*MWT*|*SA*)|*LA*)]*			✓
ES[*stdev*(*mean*(*MWT*|*SA*)|*LA*)]*			✓
ES[*mean*(*stdev*(*MWT*|*SA*)|*LA*)]*			✓
ES[*stdev*(*stdev*(*MWT*|*SA*)|*LA*)]*			✓

diastole (ED) phase of the cardiac cycle. The myocardial wall thickness variation is smooth in normal cases. In patients with MINF the wall thickness variation profile is not smooth in both short-axis (SA) and long-axis (LA) views. Whereas, with DCM cases the wall thickness is extremely thin. Fig. D.19 illustrates the myocardial wall thickness variation in normal, DCM and MINF cases. We adopted the following procedure for estimating the myocardial wall thickness variation profile features:

1. The myocardial segmentation mask was subject to canny edge detection to detect the interior and exterior contours. Binary morphological operations like hole-filling and erosion was done to ensure contour thickness to be one pixel width. The Fig. D.20 illustrates the procedure adopted for finding contours.
2. Let I and E be the set of pixels corresponding to the interior I and exterior E contours. Then the myocardial wall thickness (MWT) was the set of shortest euclidean distance (d) measures from a pixel in interior contour I to any pixel in the exterior contour E. Formally, the MWT for a short-axis (SA) slice was given by:
$$MWT|SA = \{\min_{e \in E} d(i, e) : i \in I\}$$
3. The mean and standard deviation of the MWT was estimated for each SA slice of the heart in ED and ES phases.
4. From the above measurements, 8 features were derived for quantifying the MWT variation profile at ED and ES phases. The idea was to mathematically quantify how smooth was the variation of average MWT when seen across long-axis (LA). Also, how uniform was the MWT in SA slices and to check whether this uniformity was preserved across the slices in LA. The below hand-crafted features were estimated from MWT per SA slices at each cardiac phase:
   - *ES*[*max*(*mean*(*MWT*|*SA*)|*LA*)]: Maximum of the mean myocardial wall thickness seen across slices in LA at ES phase. The left ventricular myocardial wall thickness at the site of infarction was significantly less than adjacent non-infarcted myocardium. Due to remodeling of LV post infarction, the non-infarcted myocardium wall thickness in infarct hearts is more than the normal hearts. So, max(mean(...)) was chosen to capture the maximum wall thickness seen across slices in long-axis.
   - *ES*[*stdev*(*mean*(*MWT*|*SA*)|*LA*)]: Standard deviation of the mean myocardial wall thickness seen across slices in LA at ES phase.
   - *ES*[*mean*(*stdev*(*MWT*|*SA*)|*LA*)]: Mean of the standard deviation of the myocardial wall thickness seen across slices in LA at ES phase.
   - *ES*[*stdev*(*stdev*(*MWT*|*SA*)|*LA*)]: Standard deviation of the standard deviations of the myocardial wall thickness seen across slices in LA at ES phase.
   - Similar set of four features were estimated for ED phase.

D2. Feature importance study

In order to validate the hypothesis that myocardial features alone were sufficient for distinguishing between MINF and DCM, feature importance study was done using Random Forest classifier trained to classify between MINF and DCM cases using all the features listed in Table D.16. The Fig. D.21((a) & (b)) compares the feature importance ranking by Random Forest (Liaw et al., 2002) when trained to classify all the five groups vs. only DCM and MINF respectively. The expert classifier was trained on myocardial features at end systole phase only.

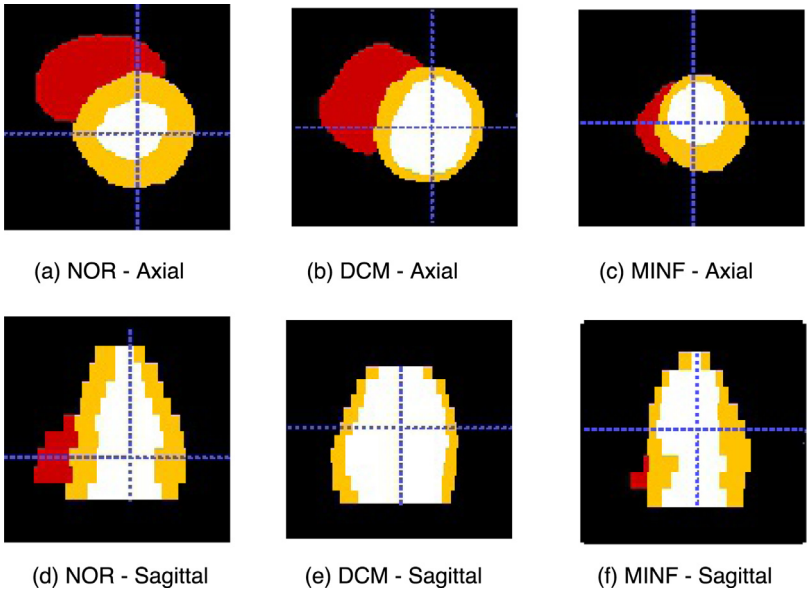


Fig. D.19. The figures shows the short-axis cardiac segmentation of Normal, DCM and MINF cases in axial ((a)-(c)) and sagittal ((d)-(f)) views at end systole phase. The segmentation labels for RV, LV and MYO are red, yellow and white respectively. For the normal case the myocardial wall thickness was uniform through out as seen in axial view and its variation along long axis was also smooth. For DCM the myocardial wall was extremely thin when compared to normal. For MINF, certain sections of the myocardium wall were extremely thin when compared to rest and hence non-uniformity of thickness was seen and also in long-axis the variation was rough. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

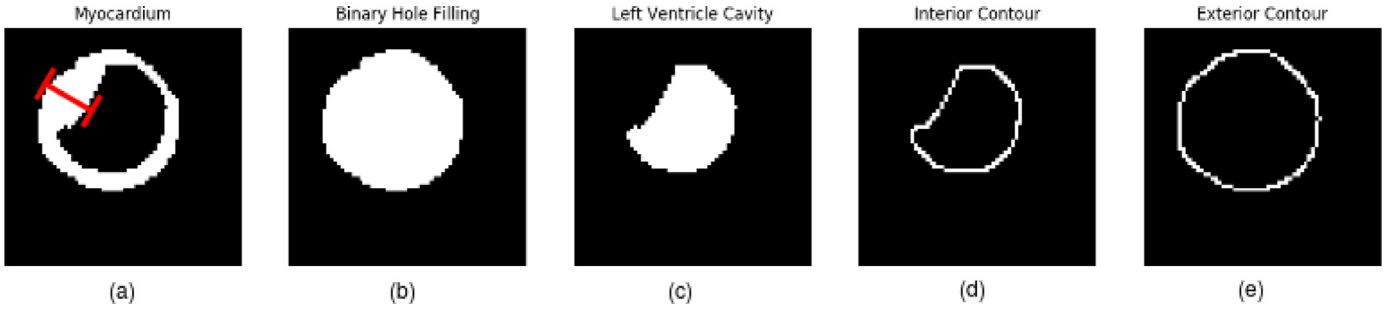
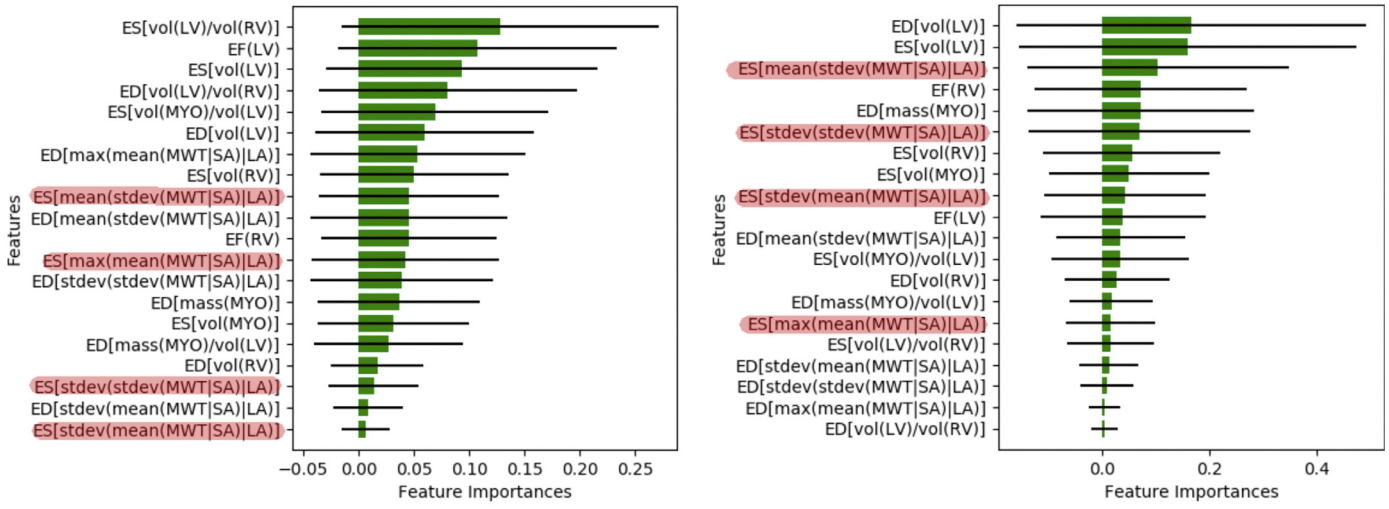


Fig. D.20. The figure illustrates the procedure adopted for estimation of myocardial wall thickness at each short-axis slice. (a) shows the myocardium segmentation and the red cross-bar indicates the wall thickness at that particular location, (b) shows the binary hole-filling operation on (a), (c) shows the left ventricle cavity got by performing the image subtraction operation between (a) and (b), (d)& (e) Canny edge detection with *sigma* = 1 was performed on (b) and (c) to get interior and exterior contours.



(a) 5-Class Feature Importance

(b) 2-Class Feature Importance

Fig. D.21. Feature importance ranking by Random Forest classifier for two different classification tasks. The green bars are the feature importances of the forest, along with their inter-trees variability (standard deviation). The features highlighted in red color indicate hand-crafted myocardial wall thickness features at ES phase. (a) shows the feature importance for 5-class task, it can be seen that highlighted features have been given low importance, (b) shows the feature importance for the 2-class task, clearly it can be seen that the highlighted features have been ranked higher.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al., 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv: 1603.04467.

Albá, X., Lekadir, K., Pereañez, M., Medrano-Gracia, P., Young, A.A., Frangi, A.F., 2018. Automatic initialization and quality control of large-scale cardiac mri segmentations. Med. Image Anal. 43, 129–141.

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

Ayed, I.B., Lu, Y., Li, S., Ross, I., 2008. Left ventricle tracking using overlap priors. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 1025–1033.

Bai, W., Shi, W., Ledig, C., Rueckert, D., 2015. Multi-atlas segmentation with augmented features for cardiac mr images. Med. Image Anal. 19 (1), 98–109.

Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., Zemrak, F., Fung, K., Paiva, J.M., Carapella, V., Kim, Y.J., Suzuki, H., Kainz, B., Matthews, P.M., Petersen, S.E., Piechnik, S.K., Neubauer, S., Glocker, B., Rueckert, D., 2018. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J. Cardio. Mag. Res. 20 (1), 65.

Baumgartner, C.F., Koch, L.M., Pollefeys, M., Konukoglu, E., 2017. An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation. In: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10–14, 2017, Revised Selected Papers, 10663. Springer, p. 111.

Ben-Cohen, A., Klang, E., Diamant, I., Rozendorn, N., Amitai, M.M., Greenspan, H., 2015. Automated method for detection and segmentation of liver metastatic lesions in follow-up ct examinations. J. Med. Imaging 2 (3), 034502.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K.H., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P., 2018. Deep learning techniques for automatic MRI Cardiac multi-structures segmentation and diagnosis: Is the problem solved? PP, 1.

Booz Allen Hamilton Inc, Kaggle, 2015. Second annual data science bowl. Accessed: 06- Dec- 2017. https://www.kaggle.com/c/second-annual-data-science-bowl

Boykov, Y., Jolly, M.-P., 2000. Interactive organ segmentation using graph cuts. In: MICCAI, 1935. Springer, pp. 276–286.

Cetin, I., Sanroma, G., Petersen, S.E., Napel, S., Camara, O., Ballester, M.-A.G., Lekadir, K., 2017. A radiomics approach to computer-aided diagnosis with cardiac cine-mri. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 82–90.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2018. Drinet for medical image segmentation. IEEE Trans. Med. Imag.

Christ, P.F., Elshaer, M.E.A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., DAnastasi, M., et al., 2016. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 415–423.

Clevert, D.-A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv: 1511.07289.

Cocosco, C.A., Niessen, W.J., Netsch, T., Vonken, E.-j., Lund, G., Stork, A., Viergever, M.A., 2008. Automatic image-driven segmentation of the ventricles in cardiac cine mri. J. Magn. Reson. Imaging 28 (2), 366–374.

Cousty, J., Najman, L., Couprie, M., Clément-Guinaudeau, S., Goissen, T., Garot, J., 2010. Segmentation of 4d cardiac mri: automated method based on spatio-temporal watershed cuts. Image Vis. Comput. 28 (8), 1229–1243.

Deng, X., Du, G., 2008. 3d segmentation in the clinic: a grand challenge ii-liver tumor segmentation. MICCAI Workshop.

Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P.-A., 2016. 3d deeply supervised network for automatic liver segmentation from ct volumes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 149–157.

Duda, R.O., Hart, P.E., 1972. Use of the hough transformation to detect lines and curves in pictures. Commun ACM 15 (1), 11–15.

Emad, O., Yassine, I.A., Fahmy, A.S., 2015. Automatic localization of the left ventricle in cardiac mri images using deep learning. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. IEEE, pp. 683–686.

Fahmy, A.S., Al-Agamy, A.O., Khalifa, A., 2011. Myocardial segmentation using contour-constrained optical flow tracking. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 120–128.

Fradkin, M., Ciofolo, C., Mory, B., Hautvast, G., Breeuwer, M., 2008. Comprehensive segmentation of cine cardiac mr images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 178–185.

Frangi, A.F., Niessen, W.J., Viergever, M.A., 2001. Three-dimensional modeling for functional analysis of cardiac images, a review. IEEE Trans. Med. Imag. 20 (1), 2–5.

Fukushima, K., 1979. Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. IEICE Technical Report, A 62 (10), 658–665.

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

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 1, p. 3.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456.

Isensee, F., Jaeger, P., Full, P.M., Wolf, I., Engelhardt, S., Maier-Hein, K.H., 2017. Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features, 120–129.

Jang, Y., Hong, Y., Ha, S., Kim, S., Chang, H.-J., 2017. Automatic segmentation of lv and rv in cardiac mri. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 161–169.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, pp. 1175–1183.

Jolly, M.-P., 2006. Automatic segmentation of the left ventricle in cardiac mr and ct images. Int J Comput Vis 70 (2), 151–163.

Jolly, M.-P., Guetter, C., Lu, X., Xue, H., Guehring, J., 2011. Automatic segmentation of the myocardium in cine mr images using deformable registration. In: STACOM. Springer, pp. 98–108.

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.

Karamitsos, T.D., Francis, J.M., Myerson, S., Selvanayagam, J.B., Neubauer, S., 2009. The role of cardiovascular magnetic resonance imaging in heart failure. J. Am. Coll. Cardiol. 54 (15), 1407–1424.

Katouzian, A., Prakash, A., Konofagou, E., 2006. A new automated technique for left--and right-ventricular segmentation in magnetic resonance imaging. In: Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE. IEEE, pp. 3074–3077.

Khened, M., Alex, V., Krishnamurthi, G., 2017. Densely connected fully convolutional network for short-axis cardiac cine mr image segmentation and heart diagnosis using random forest. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 140–151.

Kingma, D., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv: 1412.6980.

Kong, B., Zhan, Y., Shin, M., Denny, T., Zhang, S., 2016. Recognizing end-diastole and end-systole frames via deep temporal regression network. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 264–272.

Korshunova, I., Burms, J., Degrave, J., Dambre, J., 2016. Diagnosing heart diseases with deep neural networks. Accessed: 01- Nov- 2017. http://irakorshunova.github.io/2016/03/15/heart.html.

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.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Li, B., Liu, Y., Occleshaw, C.J., Cowan, B.R., Young, A.A., 2010. In-line automated tracking for ventricular function with magnetic resonance imaging. JACC 3 (8), 860–866.

Liaw, A., Wiener, M., et al., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.

Lieman-Sifry, J., Le, M., Lau, F., Sall, S., Golden, D., 2017. Fastventricle: Cardiac segmentation with enet. In: International Conference on Functional Imaging and Modeling of the Heart. Springer, pp. 127–138.

Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv: 1312.4400.

Lin, X., Cowan, B., Young, A., 2006. Automated detection of the left ventricle from 4d mr images: validation using large clinical datasets. Adv. Image Video Technol. 218–227.

Lin, X., Cowan, B., Young, A., 2006. Model-based graph cut method for segmentation of the left ventricle. In: Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the. IEEE, pp. 3059–3062.

Lorenzo-Valdés, M., Sanchez-Ortiz, G.I., Elkington, A.G., Mohiaddin, R.H., Rueckert, D., 2004. Segmentation of 4d cardiac mr images using a probabilistic atlas and the em algorithm. Med Image Anal 8 (3), 255–265.

Lötjönen, J., Kivistö, S., Koikkalainen, J., Smutek, D., Lauerma, K., 2004. Statistical shape model of atria, ventricles and epicardium from short-and long-axis mr images. Med Image Anal 8 (3), 371–386.

Lu, Y., Radau, P., Connelly, K., Dick, A., Wright, G.A., 2009. Segmentation of left ventricle in cardiac cine mri: an automatic image-driven method. In: International Conference on Functional Imaging and Modeling of the Heart. Springer, pp. 339–347.
Lynch, M., Ghita, O., Whelan, P.F., 2006. Automatic segmentation of the left ventricle cavity and myocardium in mri data. Comput. Biol. Med. 36 (4), 389–407.
Lynch, M., Ghita, O., Whelan, P.F., 2008. Segmentation of the left ventricle of the heart in 3-d+ t mri data using an optimized nonrigid temporal model. IEEE Trans. Med. Imag. 27 (2), 195–203.
Margeta, J., Geremia, E., Criminisi, A., Ayache, N., 2011. Layered spatio-temporal forests for left ventricle segmentation from 4d cardiac mri data. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 109–119.
Miller, C.A., Jordan, P., Borg, A., Argyle, R., Clark, D., Pearce, K., Schmitt, M., 2013. Quantification of left ventricular indices from ssfp cine imaging: impact of real-world variability in analysis methodology and utility of geometric modeling. J. Magn. Reson. Imag. 37 (5), 1213–1222.
Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on. IEEE, pp. 565–571.
Nambakhsh, C.M., Yuan, J., Punithakumar, K., Goela, A., Rajchl, M., Peters, T.M., Ayed, I.B., 2013. Left ventricle segmentation in mri via convex relaxed distribution matching. Med. Image Anal. 17 (8), 1010–1024.
Ngo, T.A., Lu, Z., Carneiro, G., 2017. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Med. Image Anal. 35, 159–171.
Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., ORegan, D.P., et al., 2018. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imag. 37 (2), 384–395.
Ordas, S., Oubel, E., Leta, R., Carreras, F., Frangi, A.F., 2007. A statistical shape model of the heart and its application to model-based segmentation. Prog. Biomed. Opt. Imaging Proc. SPIE, 6511, K65111.
Paragios, N., 2003. A level set approach for shape-driven segmentation and tracking of the left ventricle. IEEE Trans. Med. Imag. 22 (6), 773–776.
Patravali, J., Jain, S., Chilamkurthy, S., 2018. 2d-3d fully convolutional neural networks for cardiac MR segmentation. In: Pop, M., Sermesant, M., Jodoin, P.-M., Lalande, A., Zhuang, X., Yang, G., Young, A., Bernard, O. (Eds.), Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. Springer International Publishing, Cham, pp. 130–139.
Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., Glocker, B., 2018. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers, 10670. Springer, p. 450.
Pednekar, A., Kurkure, U., Muthupillai, R., Flamm, S., Kakadiaris, I.A., 2006. Automated left ventricular segmentation in cardiac mri. IEEE Trans. Biomed. Eng. 53 (7), 1425–1428.
Peng, P., Lekadir, K., Gooya, A., Shao, L., Petersen, S.E., Frangi, A.F., 2016. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. Magn. Reson. Mater. Phys., Biol. Med. 29 (2), 155–195.
Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in mri images. IEEE Trans. Med. Imag. 35 (5), 1240–1251.
Petitjean, C., Dacher, J.-N., 2011. A review of segmentation methods in short axis cardiac mr images. Med. Image Anal. 15 (2), 169–184.
Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.-N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.-C., et al., 2015. Right ventricle segmentation from cardiac mri: a collation study. Med. Image Anal. 19 (1), 187–202.
Poudel, R.P., Lamata, P., Montana, G., 2016. Recurrent Fully Convolutional Neural Networks for Multi-slice Mri Cardiac Segmentation. In: Reconstruction, Segmentation, and Analysis of Medical Images. Springer, pp. 83–94.
Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A., Wright, G., 2009. Evaluation framework for algorithms segmenting short axis cardiac mri. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge 49.
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.
Rupprecht, C., Huaroc, E., Baust, M., Navab, N., 2016. Deep active contours. arXiv: 1607.05074.
Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (4), 640–651.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556.
Sonka, M., Bosch, J.G., Lelieveldt, B.P., Mitchell, S.C., Reiber, J.H., 2003. Computer-aided diagnosis via model-based shape analysis: cardiac mr and echo. In: International Congress Series, 1256. Elsevier, pp. 1013–1018.
Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J.Mach.Learn.Res. 15 (1), 1929–1958.
Suinesiaputra, A., Ablin, P., Alba, X., Alessandrini, M., Allen, J., Bai, W., Cimen, S., Claes, P., Cowan, B.R., Dhooge, J., et al., 2018. Statistical shape modeling of the left ventricle: myocardial infarct classification challenge. IEEE J. Biomed. Health Inform. 22 (2), 503–515.
Suinesiaputra, A., Cowan, B.R., Al-Agamy, A.O., Elattar, M.A., Ayache, N., Fahmy, A.S., Khalifa, A.M., Medrano-Gracia, P., Jolly, M.-P., Kadish, A.H., et al., 2014. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac mr images. Med. Image Anal. 18 (1), 50–62.
Suinesiaputra, A., Frangi, A.F., Kaandorp, T.A., Lamb, H.J., Bax, J.J., Reiber, J.H., Lelieveldt, B.P., 2009. Automated detection of regional wall motion abnormalities based on a statistical model applied to multislice short-axis cardiac mr images. IEEE Trans. Med. Imag. 28 (4), 595–607.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.
Tan, L.K., Liew, Y.M., Lim, E., McLaughlin, R.A., 2017. Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine mr sequences. Med. Image Anal. 39, 78–86.
Tavakoli, V., Amini, A.A., 2013. A survey of shaped-based registration and segmentation techniques for cardiac images. Comput. Vision Image Underst. 117 (9), 966–989.
Tran, P.V., 2016. A fully convolutional neural network for cardiac segmentation in short-axis mri. arXiv: 1604.00494.
Üzümcü, M., van der Geest, R.J., Swingen, C., Reiber, J.H., Lelieveldt, B.P., 2006. Time continuous tracking and segmentation of cardiovascular magnetic resonance images using multidimensional dynamic programming. Invest. Radiol. 41 (1), 52–62.
Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23 (7), 903–921.
Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I., 2018. Automatic segmentation and disease classification using cardiac cine MR images. In: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. Springer International Publishing, Cham, pp. 101–110.
Xue, W., Brahm, G., Pandey, S., Leung, S., Li, S., 2018. Full left ventricle quantification via deep multitask relationships learning. Med. Image Anal. 43, 54–65.
Yang, H., Sun, J., Li, H., Wang, L., Xu, Z., 2016. Deep fusion net for multi-atlas segmentation: Application to cardiac mr images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 521–528.
Zhang, H., Wahle, A., Johnson, R.K., Scholz, T.D., Sonka, M., 2010. 4-D cardiac mr image analysis: left and right ventricular morphology and function. IEEE Trans. Med. Imag. 29 (2), 350–364.
Zhang, L., Gooya, A., Dong, B., Hua, R., Petersen, S.E., Medrano-Gracia, P., Frangi, A.F., 2016. Automated quality assessment of cardiac mr images using convolutional neural networks. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 138–145.
Zhao, F., Zhang, H., Wahle, A., Thomas, M.T., Stolpen, A.H., Scholz, T.D., Sonka, M., 2009. Congenital aortic disease: 4d magnetic resonance segmentation and quantitative analysis. Med. Image Anal. 13 (3), 483–493.
Zheng, Q., Delingette, H., Duchateau, N., Ayache, N., 2018. 3-D consistent & robust segmentation of cardiac images by deep learning with spatial propagation. IEEE Transactions on Medical Imaging 37 (9), 2137–2148.
Zhu, Y., Papademetris, X., Sinusas, A.J., Duncan, J.S., 2010. Segmentation of the left ventricle from cardiac mr images using a subject-specific dynamical model. IEEE Trans. Med. Imag. 29 (3), 669–687.
Zotti, C., Luo, Z., Humbert, O., Lalande, A., Jodoin, P.-M., 2018. GridNet with Automatic Shape Prior Registration for Automatic MRI Cardiac Segmentation. In: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. Springer International Publishing, Cham, pp. 73–81.