3D ATTENTION M-NET FOR SHORT-AXIS LEFT VENTRICULAR MYOCARDIUM SEGMENTATION IN MICE MR CARDIAC IMAGES

by
Luojie Huang

A thesis submitted to Johns Hopkins University in conformity
with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland
May, 2021

Abstract

Small rodent cardiac magnetic resonance imaging (MRI) plays an important role in preclinical models of cardiac disease, which is routinely used to probe the effect of individual genes or groups of genes on the etiology of a large number of cardiovascular diseases. Accurate myocardial boundaries delineation is crucial to most morphological and functional analysis in rodent cardiac MRI. However, due to the small volume of the mouse heart and its high heart rate, rodent cardiac MRIs are usually acquired with sub-optimal resolution and low signal-to-noise ratio(SNR). The rodent cardiac MRIs can also suffer from signal loss due to the intra-voxel dephasing. These factors make automatic myocardial segmentation challenging. Manual delineation could be applied to label myocardial boundaries but it is usually laborious and time-consuming. An automatic myocardium segmentation algorithm specifically designed for these data could enhance accuracy and reproducibility of cardiac structure and function analysis.

In this study, we present a deep learning approach based on 3D attention M-net to perform automatic segmentation of the murine left ventricular myocardium. In the architecture, we use dual spatial-channel attention gates between encoder and decoder along with a multi-scale feature fusion path after decoder. Attention gates enable networks to focus on relevant spatial information and channel features to improve segmentation performance. A distance-derived loss term, besides general Dice score loss and binary cross entropy loss, was also introduced to our hybrid loss functions to refine our segmentation contour. The proposed model outperforms previous generic models for segmentation, with similar number of parameters, in major segmenta-

tion metrics including the Dice score (0.9072±0.0187), Jaccard index (0.8307) and Hausdorff distance (3.1754 pixels), which are comparable to the results achieved by state-of-the-art models on human cardiac datasets.

Thesis Committee

Dr. Siamak Ardekani (Primary Advisor)
Assistant Research Professor
Department of Biomedical Engineering
Johns Hopkins University

Dr. Jeremias Sulam
Assistant Professor
Department of Biomedical Engineering
Johns Hopkins University

Dr. Robert George Weiss
Professor of Medicine
Department of Cardiology
Johns Hopkins University School of Medicine

Contents

List of Figures

Chapter 1

Introduction

1.1 Clinical Background

Cardiac magnetic resonance imaging (MRI) is the current gold standard for clinical quantitative cardiac analysis, including the calculation of left ventricular (LV) end-diastolic (ED) and end-systolic (ES) volumes, due to its accurate measurement of both anatomy and function [1, 2], which is crucial to reliable cardiac analysis. Such quantification typically requires the delineation of LV myocardial borders. Beside extensive usage in the clinical scenarios, cardiac MRI is also indispensable in various preclinical research. In the preclinical models of cardiac disease, small rodent cardiac images play an extremely important role, which are routinely used to probe the effect of individual genes or groups of genes on the etiology of a large number of cardiovascular diseases. Accurate myocardial boundaries delineation is crucial to most morphological and functional analysis in rodent cardiac MRI [3, 4].

Traditional segmentation approaches include manual contouring, image processing and machine learning methods, such as active shape models [5] and atlas-based methods [6]. Clinically, most physicians restrict delineation to only the ED and ES phases, which requires approximately 20 minutes to complete manually [7]. Complete delineation across the entire cardiac cycle would be more desirable, as it allows calculation of ventricular filling or ejection rates, but the high cardiac frame-rates make this far too

tedious and time-consuming to be performed manually. Although image processing and machine learning methods are less laborious, they also usually require manual intervention, extensive feature engineering, and prior-knowledge incorporation, which are still not systematically efficient.

1.2 Related Work

1.2.1 Convolutional Neural Networks (CNN)

In the last decade, Deep learning algorithms, especially Convolutional Neural Networks (CNN), have succeeded in various automatic medical image segmentation tasks [8, 9]. A standard CNN is typically made up by an input layer, an output layer and a stack hidden layers to perform a series of nonlinear transformation from inputs to outputs. These hidden layers often consist of convolutional layers, pooling layers (e.g. average-pooling and max-pooling) and/or fully-connected layers. In general, The sequence of convolutional layers effectively extract useful features. A convolutional layer achieves convolution by many convolution kernels and is followed by a normalization layer such as batch normalization, to facilitate model convergence during training and an activation function like ReLU and Sigmoid, to provide a simple non-linear function from which more complex feature maps can be obtained. These feature maps are then downsampled by pooling layers to provide some local spatial invariance to each layer. After that, fully connected layers take features that have been computed from the inputs via convolutional layers and find the most task-relevant features to act efficiently as non-linear classifiers. The output of the CNN model is a fix-sized vector where each element would be a probabilistic score for each category for classification, or a real value for a regression task like left ventricular volume estimation, or a set of values for object detection and localization.

The most significant component of a CNN model is the convolutional layer, which

consists of many convolution kernels for feature extraction. The computational complexity increases exponentially as the kernel size increases. Therefore, the size of each kernel is typically limited, usually 2D 3×3 kernels, so that the model could run more efficiently. While the individual kernels are small, the receptive field can still be increased by increasing the number of convolutional layers, which is also called the depth of a CNN model. For example, a convolutional layer with large 7×7 kernels can be replaced by three layers with small 3×3 kernels. In general, increasing the depth of convolution neural networks to enlarge the receptive field can lead to improved model performance.

The original CNN classification models can also be applied to image segmentation tasks without major adaptations to the network architecture. Ciresan et al. [9] applied 2D CNN to neuronal membranes segmentation in electron microscopic image stacks. The segmentation task is performed by pixel-wise classification by extracting patch around the pixel using a sliding window within each slice. One major drawback of this patch-based approach is the time inefficiency caused by computational redundancy due to multiple overlapping patches in the image and inability to learn global context. For more efficient and end-to-end pixel-wise segmentation, a variant of CNN called fully convolutional neural network is more commonly used.

1.2.2 Fully Convolutional Network (FCN)

Fully Convolutional Network (FCN) [10] and U-Net [11], have been utilized for cardiac segmentation. FCNs are a special type of CNNs without any fully connected layers. FCNs typically have an encoder-decoder structure so that they can take input of arbitrary size and output the prediction of the same size. The input would first be transformed into high-level features by the encoder and then interpreted and recovered by the decoder with detailed spatial information by the decoder. Going through a stack of upsampling and convolution operations within the decoder, The final output

would be a pixel-wise prediction with the size of input. Upsampling can be achieved by either upsampling layers or transposed convolutions such as 3×3 transposed convolution with a stride of 2 to enlarge features by a factor of 2. FCN with the simple encoder-decoder structure may be limited to capture detailed context information for precise segmentation as features would be eliminated by the pooling layers in the encoder. Several variants of FCNs have been proposed to efficiently propagate features from the encoder to the decoder, in order to boost the segmentation accuracy. The most well-known and most popular variant of FCNs for biomedical image segmentation is the U-net. On the basis of the vanilla FCN, the U-net employs skip connections between the encoder and decoder to recover spatial context in the down-sampling path, yielding more precise segmentation. These works, however, mainly focused on 2D slices rather than 3D ones due to the low out-of-plane resolution and motion artifacts in clinical cardiac MR scans. As a consequence, they are not accounting for inter-slice dependencies by performing slice-by-slice workflow. More recent works such as 3D U-net [12], V-net [13], and M-net [14] have focused on network structure refinement to enhance feature learning. Compared to original FCN, these models reuse encoded features from inputs more effectively; therefore, they are widely applied in biomedical image segmentation, especially in low signal-to-noise ratio (SNR) applications.

1.2.3 Loss Functions

Other state-of-the-art approaches to improve network performance include changing supervision methods, modifying loss functions, and introducing attention gates into networks. For example, deep supervision in [15–17] is utilized to regularize networks to capture more meaningful high-level features, especially with a small training dataset. Besides, many loss functions, such as weighted cross-entropy [18], weighted Dice loss [19], focal loss [20] and distance loss function [21], were proposed to overcome imbalanced data issue and to refine the boundaries of segmentation. More recently,

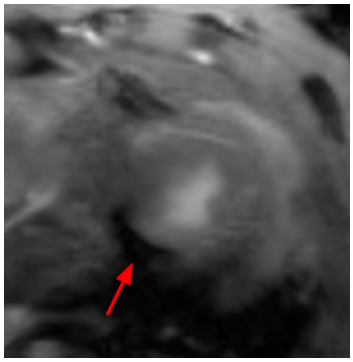


Figure 1-1. Signal loss within myocardium: An example of left ventricular chamber with dark band (Red arrow) due to magnetic field inhomogeneity-induced signal void in myocardium.

attention gates [22] were highlighted in conditioning and regularizing deep learning networks to focus and capture better local features in segmentation application.

1.2.4 Challenges

The aforementioned models aim at human cardiac MR segmentation. Compared to their works, we have focused on developing a deep learning technique that segments left ventricular myocardium in mice cardiac MRI, which is valuable in preclinical studies. In comparison to human cardiac anatomy, a mouse heart is much smaller, resulting in a lower SNR and relatively low spatial resolution, contributing to the partial volume averaging and blurry boundaries. One way to improve the SNR is to use MR scanners with higher magnetic field. However, magnetic field inhomogeneity that is typically observed at this higher field, introduces image artifacts such as signal loss due to intra-voxel dephasing at the air-tissue boundaries in regions where the cardiac wall is in the vicinity of lung parenchyma (Figure 1-1). Similar to human cardiac MR studies, we could also observe slice-to-slice misplacement due to sequential 2D acquisition scheme. To address these challenges, we have proposed a novel pipeline by applying modified M-net with dual attention gates to 3D volume segmentation and introduced a distance-derived loss function for optimal boundary refinement of segmentation.

1.2.5 Proposed Method

The major contribution of this research work is to use a unique in-house data set of dynamic mice cardiac MRI data during a cardiac cycle (cine) that contains both normal and diseased heart (myocardial infarction) to develop a new pipeline for 3D myocardium segmentation based on 3D M-net architecture with attention to a new distance-derived loss function. Our initial analysis indicates that the proposed approach achieves better performance in Hausdorff distance than other current state-of-the-art models with comparable Dice score.

Chapter 2

Methodology

The model architecture detail is depicted in Figure 2-1. The proposed 3D Attention M-net, which is inspired by the original M-net [14], consists of two major parts: encoder and decoder with four levels each. In addition, a dual attention gate is applied between the encoder and decoder at each level to refine the features from the encoder.

2.1 Encoder

Before the encoder, the MRI volumes of size $36 \times 96 \times 96$ are first inputted into a convolution layer to increase the feature channels to 16. Then, it is down-sampled by max-pooling layers as parallel inputs to corresponding encoder levels. This additional input path is inspired by the residual connection, that helps to preserve the original image information in lower levels and prevent some common issues like gradient vanishing as the model gets deeper. As illustrated in Figure 2-1, each encoder level consists of a cascade of 4 convolution blocks following a max-pooling layer for next-level encoders. Each convolution block includes a 3D convolution layer with a Batch Normalization layer and a 'ReLU' activation layer.

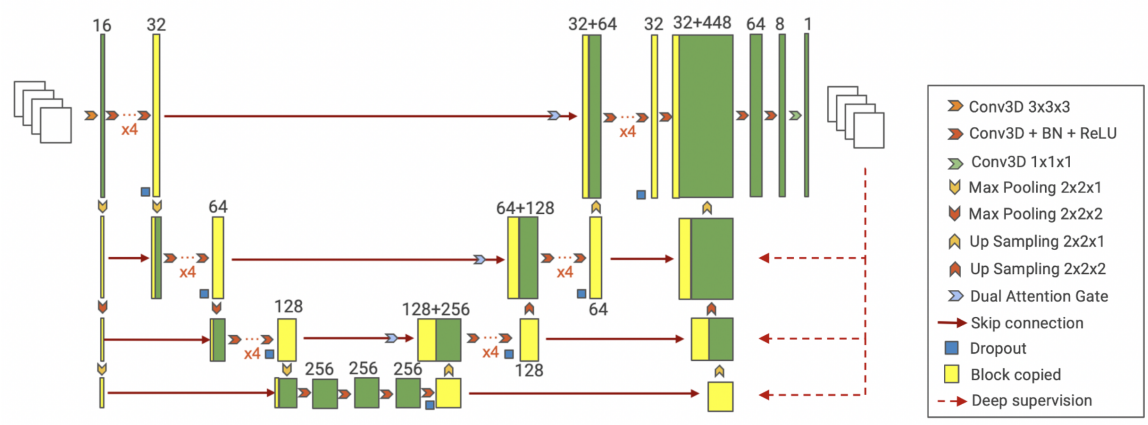


Figure 2-1. Attention M-net Architecture. The model consists of encoder, decoder, and dual attention gates. Architecture of Dual Attention Gate. The dual attention gate, taking both encoder outputs and upsampled decoder outputs as inputs, is made up by two sub-blocks: (b) channel attention block, and (c) spatial attention block.

2.2 Dual Attention Gate

In the original skip connections of M-net, the outputs from encoder are directly concatenated with up-sampled features from the lower decoder level. However, the contextual information extracted from encoder would inevitably include redundant information. To solve this problem, we have introduced the self-attention mechanism to emphasize meaningful information along the spatial and channel dimensions, that are most beneficial features for the final segmentation. Between encoders and decoders at each level, we utilized the dual attention gate proposed by Khanh et al. [22] to regularize our network to focus on extracting meaningful contextual features from encoders and decoders along the spatial and channel dimensions. With the help of the attention gates, the decoders can more efficiently utilize encoded features to generate final segmentation. As illustrated in Figure 2-2, the dual attention gate is made up of a spatial attention gate and a channel attention gate. These two gate are complementary, where spatial attention gate focuses on spatial concepts and channel attention gate concentrates on channel concepts. Therefore, in order to take full advantage of both attention gates, we sequentially multiply the input feature with a spatial attention

map and channel attention map to construct the final refined features, as illustrated in Figure 2-2 (a). The weighted features are then flow into decoder and concatenated with higher-level decoded features for further calculation.

2.2.1 Channel Attention Gate

Although the low-level encoded features are rich in detailed spatial information, they lack semantic information. Due to this large semantic gap, a direct connection between low-level encoder and high-level decoder features adversely affects the prediction outcomes. The channel attention gate is helpful to emphasize the contextual information in the low-level encoded feature, thus narrowing the semantic gap between the encoder and decoder features. The channel attention map combines channel information from both the encoder and decoder, as illustrated in Figure 2-2 (b). And it is constructed based on the channel interdependencies of the convolutional features, which highlighted the most meaningfully discriminative features.

More effective feature fusion between the encoder and decoder is ensured by improving the contextual information of the low-level encoder. First, the spatial features from both encoder and decoder is squeezed by utilizing average pooling and max pooling simultaneously. Second, $1 \times 1 \times 1$ convolution layers are applied on the squeezed feature to capture the dependencies of the channels to generate the squeezed channel attention map. Finally, the squeezed channel attention map are summed up and an additional $1 \times 1 \times 1$ convolution along with the 'Sigmoid' activation function is used to construct the final attention map.

2.2.2 Spatial Attention Gate

Since the encoded features are rich in location information, a more selective focus on salient region is beneficial to target the object and refine the segmentation. A spatial attention gate is used to guide the model to focus more on the detailed structure of

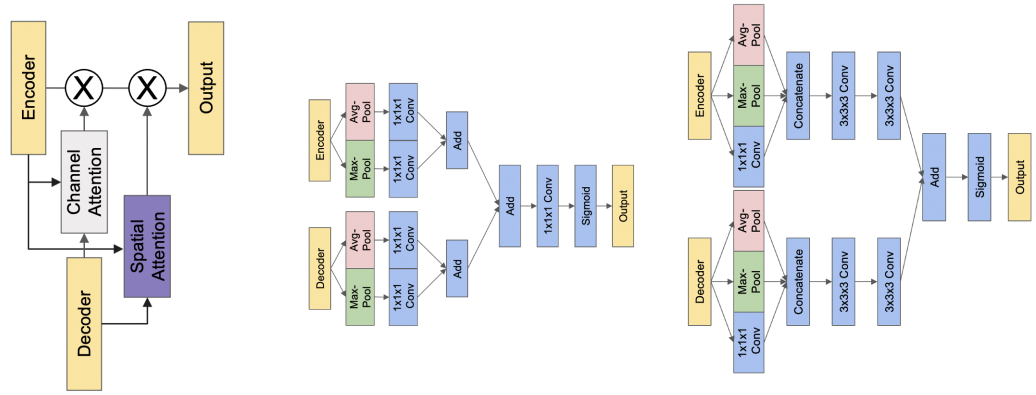
an important region, which is the boundary area in our case. The spatial attention map is constructed based on the interrelation between the spatial information that focuses on the regions in interest.

As illustrated in Figure 2-2 (c), the spatial attention map is also a combination of encoder and decoder features. To compute the spatial attention map at each level, we first apply average pooling, max pooling, and $1 \times 1 \times 1$ convolution through the channel dimension and then concatenate them to construct a capable feature representation. Then, two convolution layers are applied on the concatenated features to produce a self-learned spatial attention map that highlights the region of interest throughout the training process. Finally, the spatial attention map from the encoder feature and decoder feature are summarized to generate the final spatial attention map.

2.3 Decoder

Decoding levels take the concatenation of up-sampled feature from lower level and refined skip connections from corresponding encoder levels as inputs. In each level, the first half part of a decoder mirrors the encoder part, followed by an up-sampling layer to double the size and recover the output to the input volume size.

Moreover, there is an additional up-sampling path for multi-scale fusion to directly introduce lower-level decoded features into the final prediction and generate a smoother segmentation boundary. Moreover, this multi-scale fusion also provides entry for the following deep supervision. The output from a certain level is up-sampled and then directly concatenated with the output from upper level. This additional path allows to introduce deep supervision to lower-level decoder outputs, thus preventing vanishing gradient. Finally, the output from the decoder goes through two more $3 \times 3 \times 3$ convolution layers to reduce the feature size and then are outputted from a $1 \times 1 \times 1$ convolution layer with the 'Sigmoid' activation function.



(a) Dual Attention Gate (b) Channel Attention (c) Spatial Attention

Figure 2-2. Architecture of Dual Attention Gate. The dual attention gate, taking both encoder outputs and upsampled decoder outputs as inputs, is made up by two sub-blocks: (b) channel attention block, and (c) spatial attention block.

2.4 Loss Functions

For our dataset, most errors occur around the segmentation boundaries due to the low SNR and the artifact from signal loss within myocardium. Therefore, in addition to a common loss function such as Generalized Dice Loss (GDL) and Balanced Cross-Entropy (BCE), we propose an additional distance-derived loss (DDL) for contour refinement of the final segmentation.

2.4.1 Generalized Dice Loss (GDL)

Sudre et al. [16] proposed an extension of Dice loss with different weighting for each pixel class, which proved to be effective in unbalanced classes segmentation task.It takes the form:

$$GDL = 1 - 2\frac{\sum_{l=0}^{1} w_l \sum_x \sum_y \left(P_l(x,y) \times G_l(x,y)\right)}{\sum_{l=0}^{1} w_l \sum_x \sum_y \left(P_l(x,y) + G_l(x,y)\right)} \tag{1}$$

where $P_l(x,y)$ and $G_l(x,y)$, respectively, denote the prediction and ground truth of the input slices for class l at position (x,y). In our case, $P_0(x,y) = 1 - P_1(x,y)$ and $G_0(x,y) = 1 - G_1(x,y)$. Additionally, $w_l = 1/(\sum_x \sum_y G_l(x,y))^2$ is used to correct the

imbalanced correlation between region size of different labels and Dice score.

2.4.2 Balanced Cross-Entropy (BCE)

Xie et al. [19] also proposed the BCE loss by adding a modulating factor $\beta_{xy} = 1 - \frac{G_1(x,y)}{H \times W}$, where H and W denote height and width of the 2D slice, in the original binary cross entropy loss to tackle class imbalance:

$$BCE = \frac{-\sum_x \sum_y (\beta_{xy} G_1(x,y) \log P_1(x,y) + (1-\beta_{xy}) G_0(x,y) \log P_0(x,y))}{H \times W} \tag{2}$$

Both of these loss functions were utilized to mitigate the imbalanced voxel classes in our segmentation task due to the relatively small myocardium area in mice compared to the background.

2.4.3 Distance Derived Loss (DDL)

A distance map is defined as the distance between the ground truth and the predicted map. There are two common ways to incorporate distance maps in deep learning models, either create a specific reconstruction head in the architecture along with segmentation to predict the distance map, or induce it into loss function. Caliva et al. [21] created a custom penalty based loss function, using distance maps derived from ground truth to weigh the cross entropy loss. Inspired by their work, we propose a new term of loss function focusing on segmentation contour refinement. We first generate a distance weight map W based on the ground truth mask. For each pixel in the weight map, the weight equals the reciprocal of the closest Euclidean distance to the ground truth mask. Let $G_{pos} = \{g_1, g_2, \ldots, g_n\}$ be the set of valid positions from ground truth mask, where $g_i = (g_{ix}, g_{iy})$. The difference between prediction and ground truth is then multiplied by the weight map $W(x,y) = 1/(\min_i \sqrt{(x-g_{ix})^2 + (y-g_{iy})^2} + 1)$. The voxels closer to the ground truth are assigned larger weights so that this loss term

enables the model to focus more on contours. It is defined as:

$$DDL = \frac{\sum_x \sum_y (W(x,y) \times \mid P_1(x,y) - G_1(x,y) \mid)}{\sum_x \sum_y G_1(x,y)} \tag{1}$$

where $P_1(x,y)$ and $G_1(x,y)$ represent the prediction and ground truth pixels.

2.4.4 Combined Loss

In our model, we have used a combination of three aforementioned losses to address class imbalance (GDL, BCE) and to perform contour refinement (DDL). The combined loss is defined as follows:

$$L_{comb} = \lambda_1 \times GDL + \lambda_2 \times BCE + \lambda_3 \times DDL \tag{2}$$

Moreover, we introduced deep supervision on the decoders at each level in our model. The final loss function is a weighted sum of the losses from the decoders from all 4 levels:

$$L_{deep} = \sum_{n=1}^{4} w_n \times L_{comb_n} \tag{3}$$

Chapter 3

Experiments

3.1 Dataset

As a part of an ongoing project, in-vivo heart MR images of adult male wild type mice and Galectin-3 knockout mice were acquired at different stages of disease following sham surgery or induction of myocardial infarction (pre-op, and days 1, 7, and 56) using MRI spectrometer equipped with a 11.7T magnet and a gradient set capable of developing gradient strengths of 740 mT/m (Bruker Biospin, Germany). The mice were positioned on the MRI 4-channel surface coil and an MRI gating trigger was established via ECG leads and a respirator pillow. Cine MRI was collected (15 frames, echo time = 1.9708 ms, repetition time varied according to the heart rate, slice thickness of 0.8 mm, in plane resolution of 0.1307×0.1307 mm^2, flip angle = 12 degrees, number of excitations = 6) at 8 short axis slices through the LV. The animal protocol was approved by the Institutional Animal Care and Use Committee of the Johns Hopkins University (protocol number: MO19E374).

The myocardium was manually segmented in each short-axis image by two research assistants(inter-rater agreement of 89% based on Dice score) using a free semi-automatic software package in MATLAB called Segment [23]. In principle, manual segmentation of the left ventricle was performed following recommendations by Schulz-Menger et. al [24]. Endocardial and epicardial contours were traced on

short-axis cine images at several time points during the cardiac cycle, with simultaneous viewing of short and long-axis images of the same region, if applicable. Papillary muscles and trabecular tissue were excluded. Most apical region was identified as a section where left ventricular blood cavity pool was visible. Most basilar section was selected at the level of outflow tract. Delineation excluded the aortic valve cusps resulting in myocardial segmentation that resembled crescent shape. The contours were then interpolated across all time points of the cardiac cycle, and manually examined to correct any interpolation errors. Segmentations were then reviewed by an expert with more than 14 years of experience in cardiac MR research. We used 1114 fully annotated volumes of 99 mice that were collected from the repeated acquisitions (each contained multiple time frames per cardiac cycle) of several animals who underwent surgery and followed over the course of disease. All the scanned mice can be divided into 4 subgroup based on the disease stage: (1) 335 volumes are acquired before the infarction surgery; (2) 345 volumes are acquired at day 1 after the surgery; (3) 279 volumes are acquired at day 7 after the surgery; (4) 115 volumes are acquired at day 56 after the surgery. The data were randomly split based on the subgroups into three subsets: 700 for training, 200 for validation, and 214 for testing, to maintain the same disease stage distribution in all three sub-datasets. Selection criterion was based on myocardial volume to ensure similar distribution across all groups. For each volume, since the gap between slices (0.8 mm) is significantly larger than in-plane pixel size (0.1307 mm), one of simplest methods, linear interpolation, was performed to upsample the inter-slice gap to 0.16 mm while maintaining reasonable preprocessing efficiency. Thus, the data size was augmented to $36 \times 96 \times 96 \times 1$.

To improve robustness and generalization to unseen data, we performed data augmentation by applying random shift (up to 15 pixels), uniformly varying rotation (within 15 degrees), and horizontal/vertical flips, with a probability of 0.5.

3.2 Evaluation Metrics

We compare the proposed model with 4 state-of-the-art segmentation models: 2D U-net [11], 3D U-net [12], DeepMedic [25] and 3D FCN [10]. For fair comparison, all models are adjusted to contain similar numbers of trainable variables by tuning the feature numbers of each layer, which is the best strategy to maintain the best prediction performance compared to other adjustments like changing architecture depth. And the models are trained with the loss functions proposed in their original publications. The segmentation performance is quantitatively evaluated by metrics including mean soft Dice Score [26], Jaccard Index [27], Hausdorff Distance [28], Sensitivity and Specificity, Positive predictive value (PPV) and Negative predictive value (NPV). Let P and G represent prediction and ground truth respectively:

- Soft Dice Score are Jaccard Index are both measures of overlap or similarity between prediction and ground truth masks, and sre defined as:

$$Dice\ Score = \frac{2\sum_x \sum_y (P_1(x,y) \times G_1(x,y))}{\sum_x \sum_y (P_1(x,y) + G_1(x,y))} \tag{7}$$

$$Jaccard\ Index = \frac{TP}{TP + FP + FN} \tag{8}$$

- Hausdorff Distance is a symmetric measure of alignment between ground truth and prediction contours by calculating the maximum distance, and is defined as:

$$HD(pc, gc) = \max_i \min_j d(pc_i, gc_j) \tag{9}$$

  where $d(pc_i, gc_j) = \sqrt{(pc_{ix} - gc_{jx})^2 + (pc_{iy} - gc_{jy})^2}$ denotes the Euclidean distance between the pixels on prediction contour pc and ground truth contour gc. It should be noted that the unit of Hausdorff distance in our case is $pixel$.

- Sensitivity (SEN), specificity (SPE), PPV, and NPV are defined as:

$$SEN = \frac{TP}{TP + FN} \quad SPE = \frac{TN}{TN + FP} \tag{10, 11}$$

$$PPV = \frac{TP}{TP + FP} \quad NPV = \frac{TN}{TN + FN} \tag{12, 13}$$

where TP, TN, FP, FN denote true positive, true negative, false positive and false negative.

3.3 Implementation Details

The model is built by tensorflow and Keras in python on an Intel Xeon E5-2620 v4 CPU model with 2 Titan V GPUs. After hyperparameter tuning, the deep supervision loss L_{deep}, as described in the previous section, was set with weights w_i of the first level being 1 and others being 0.5. For the combined loss L_{comb} at each level, we conducted hyperparameter searching to find to best set of weights by setting to searching range within $1.0 - 5.0$. According to our experiments, $\lambda_1, \lambda_2, \lambda_3$ were set to 1.0, 1.0 and 3.0 to achieve the best prediction accuracy on validation set during the training process. The model was optimized by Adam method with a learning rate of 10^{-4} and batch size of 8. For all experiments, we trained the networks from scratch for 150 epochs. The learning rate was automatically adjusted by monitoring callback on plateau, allowing the optimizer to more efficiently reach to a local minimum. The monitored callback is the generalized Dice loss of the final level output with setting parameters *patience* to 8, $factor$ to 0.5 and min_delta to 0.0001.

To test the robustness of our 3D attention M-net model architecture on other dataset, we also trained and tested the proposed model on the ACDC 17 dataset, which is composed of 150 human subjects equally distributed in 5 subgroups: (1) 30 healthy subjects; (2) 30 subjects with previous myocardial infarction; (3) 30 subjects with dilated cardiomyopathy; (4) 30 subjects with hypertrophic cardiomyopathy; (5) 30 subjects with abnormal right ventricle. The dataset is further divided into 100 patients for training and the remaining 50 patients for testing, while maintaining the

same disease type distribution in both sub-datasets [29]. During training, a total of 200 volume scans from 100 patients in the training set were further split into 160 for training and 40 for validation. In addition, we used the same data augmentation strategies, optimizer and loss functions as those in the previous experiments.

Chapter 4

Results

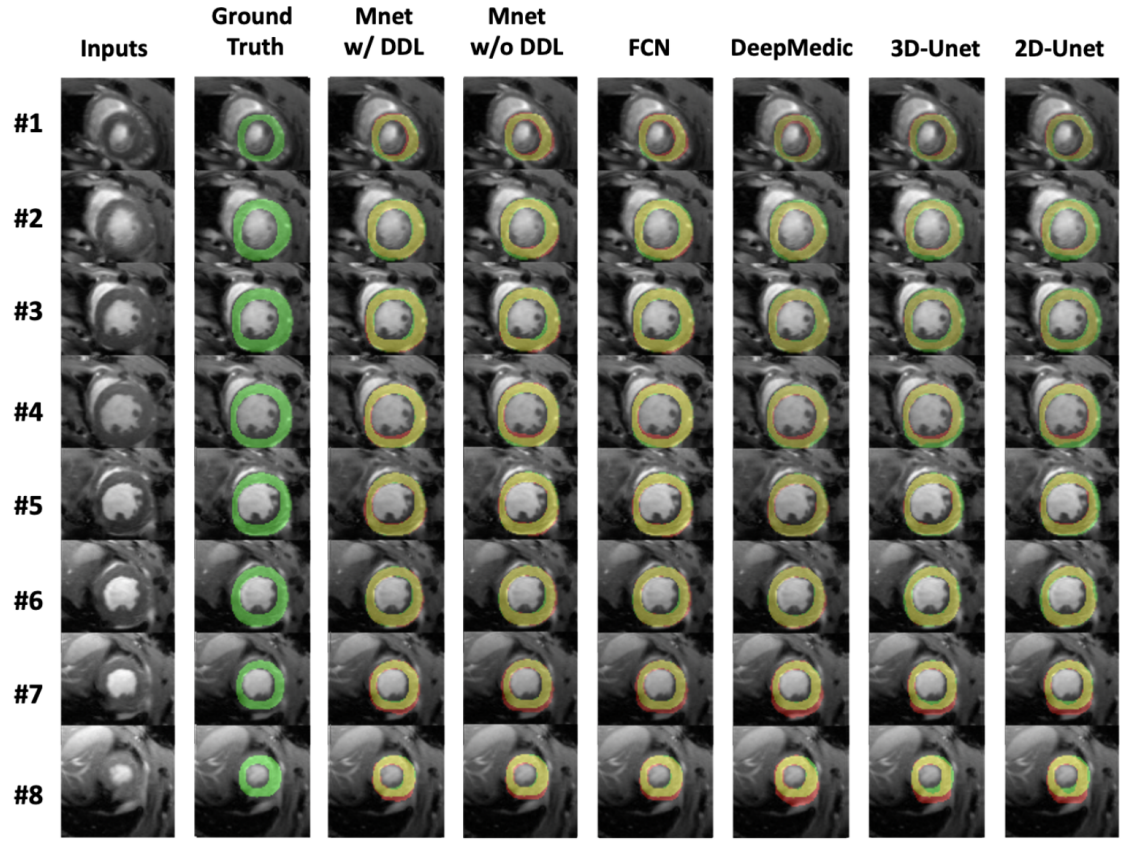


Figure 4-1. Comparison of segmentation results on a selected volume. This figure shows an example of the segmentation ground truth and predictions from all the comparative models discussed in this thesis. Slices 1-8 are the short-axis scans from basal to apical area of the same typical volume selected from the test dataset. Column 1: input MR images; Column 2: manually labeled ground truths; Column 3-8: yellow, green and red masks represent true positive, false negative and false positive, respectively. Mnet: 3D attention M-net; DDL: distance derived loss.

Table 4-I. Segmentation Evaluations of Comparative Models

Model	Parameters	Dice Score	Jaccard Index	HD	Sensitivity	Specificity	PPV	NPV
2D U-net	24,902,996	0.8677±0.0305	0.7675	4.83 (0.6315)	0.876	**0.996**	0.860	0.996
3D U-net	24,502,812	0.8732±0.0284	0.7728	4.62 (0.6037)	0.898	0.995	0.848	**0.997**
DeepMedic	24,505,161	0.8883±0.0262	0.7995	4.59 (0.6003)	0.911	0.981	0.869	0.988
3D FCN	24,128,560	0.8894±0.0206	0.8015	4.32 (0.5641)	0.921	0.987	0.864	0.993
Ours w/o DDL	24,657,034	0.8953±0.0210	0.8111	4.00 (0.5231)	**0.929**	0.987	0.866	0.993
Ours w/ DDL	24,657,034	**0.9072 ±0.0187**	**0.8307**	**3.18 (0.4150)**	0.925	0.989	**0.891**	0.993
Ours (ACDC 17)	24,657,034	0.8550±0.0256	0.7467	7.6158 (9.5198)	0.8950	0.9955	0.8184	0.9976

- 3D Att-Mnet w/o DDL: 3D Attention M-net without distance-derived loss; 3D Att-Mnet w/ DDL: 3D Attention M-net with distance-derived loss.
- HD: Hausdorff Distance. The unit of Hausdorff distance is *pixels*(mm). Smaller Hausdorff distances mean better alignment between contours of ground truth and prediction.
- PPV: positive predictive value; NPV: negative predictive value.

4.1 Comparative Models

Results are summarized in Table 4-I and visualized on Figure 4-1.According to Table 4-I, our proposed models outperform all the state-of-the-art models in dice score, Jaccard index, and Hausdorff distance, with comparable performance in all other metrics. The mean dice score of our 3D attention M-net is 0.9072 ± 0.0187, which is better than other methods: 3D U-net (0.8732 ± 0.0284), DeepMedic (0.8883 ± 0.0262) and 3D FCN (0.8894 ± 0.0206) with similar Sensitivity and Specificity. This indicates that our model has better general myocardium segmentation performance than other models in the mice MR cardiac images. Furthermore, the dice score agreement between our two experienced annotators is 0.89 over 150 samples. Compared to annotators, our proposed 3D attention M-net achieved very comparable results as human being. The proposed model has also demonstrated a potentially equally competent performance on the mice dataset as compared to the state-of-the-art myocardium segmentation models on the human cardiac dataset ACDC17 [29], whose dice scores are around 0.9 on ACDC17 dataset.

In Table 1, there is a noticeable improvement from 2D U-net to 3D U-net in Hausdorff distances (-0.2142). We also ran a two-sided Wilcoxon rank sum test that tests the null hypothesis that two independent samples are from continuous distributions with equal medians [30]. The p-value we got from the Hausdorff distances of 2D U-net and 3D U-net on the same test dataset are 0.0068, which indicates the rejection of the null hypothesis of equal medians at 0.05 significance level. This implies that the 3D model is able to incorporate additional inter-slice contextual information to achieve better 3D segmentation compared to the 2D model. The contextual information from adjacent slices and whole volume were shared in 3D models to help better segmentation in each slice, especially for the apical slice (the 8^{th} slice in Figure 4-1), which typically contains smaller myocardium and suffers from

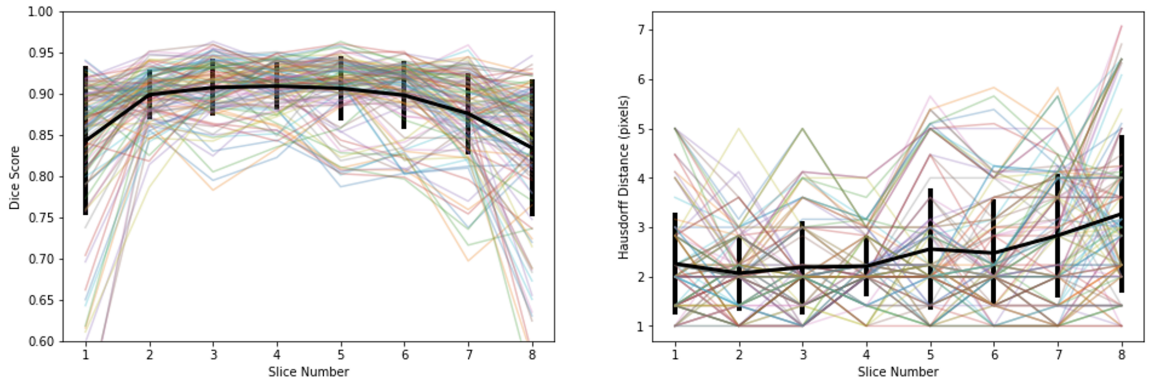
larger signal loss due to its adjacency to the lung parenchyma. In the selected case shown in Figure 4-1, the segmentation of the 2D Unet, compared to all other 3D whole-volume networks, has noticeably worse performance and boundary refinement. DeepMedic also has the same problem at the last slice, because DeepMedic is a local batch-wise 3D network, which does not fully utilize the whole volume contextual information.

4.2 Ablation Study: Distance-derived Loss

Among all 3D models, our proposed model stands out with a noticeably smaller Hausdorff distance. The Att-Mnet without the proposed distance loss has already successfully refined the final segmentation boundary to achieve 4.0026px in Hausdorff distance, which is at least 0.3px less than all other competitive models in Table 4-I. This shows that our proposed M-net with dual attention gates (Att-Mnet) has better segmentation accuracy than others with a similar number of the trainable parameters. Moreover, our Att-Mnet with the proposed distance loss further reduces the Hausdorff distance to 3.1754px. This is a major improvement -0.8272px from our Att-Mnet without the proposed distance loss function. This demonstrates that the proposed distance loss has enabled our model to effectively refine segmentation contours. This will improve the accuracy of global left ventricular volumetric measurements such as end-diastole and ends-stole volumes, ejection fraction, and myocardial mass, as well as the accuracy of computational analysis of ventricular shape and motion.

4.3 Performance Across Slices

In addition to 3D volume segmentation results, we also evaluate the model's performance on each 2D slice. The dice scores and Hausdorff distances across slices are shown in Figure 4-2. Across the whole test dataset, Att-Mnet achieves lower dices



(a) Dice Scores across slices

(b) Hausdorff distances across slices

Figure 4-2. Average dice scores and Hausdorff distances across slices on the test dataset. The black bold lines and bars are the average values and standard deviations of different slices. The fading lines are original results from all the 214 test volumes.

scores with larger standard deviations at the top and bottom slices. According to our observation, there are two main reasons that cause lower dice scores at these slices. On the one hand, the myocardium area is significantly smaller at marginal slices than middle slices, which would result in a worse result even with the same absolute segmentation error according to the calculation of dice score. On the other hand, the segmentation accuracy is also affected by some physiological factors. For example, the artifacts or signal loss from air around the apical section, as well as the complexity of outflow tract such as inclusion of valves at the basilar section, can also be incorporated into low accuracy. And this is also demonstrated by the increase of Hausdorff distance in the Figure 4-2 (b).

4.4 Performance on Human Dataset (ACDC 17)

To evaluate the generalization of the proposed model architecture on human cardiac MR images, we have trained a new model, using the exactly same architecture without further fine-tuning, on one of the most popular human benchmark dataset named ACDC 17. And the results on the test dataset is shown in the last row of Table 4-I. According to the leaderboard of segmentation challenge, the range of dice scores and

Hausdorff distances of the top 15 models are $0.791 - 0.914$ and $7.171mm - 13.434mm$, respectively [29]. Therefore, our model, with a dice score at 0.855 and Hausdorff distance at 9.520mm, has demonstrated the potential to generalize to more kinds of cardiac datasets.

Conclusion

In this thesis, a 3D M-net model with dual attention gates and a distance derived loss term was proposed for myocardium segmentation, tackling the challenges of low image quality and signal loss in mice MR cardiac images. The experiment results indicate that our proposed model not only outperforms other benchmark models, including U-Net and FCN, in dice score and contour agreement (Hausdorff distance), but also has very comparable performance as experienced human annotators. The proposed model enables a fast, objective, and accurate myocardial boundaries delineation of rodent cardiac MRIs, and has strong potential application in morphological and functional analysis for preclinical cardiac models.

Additionally, we have observed that 3D models had better performance than 2D models by taking inter-slice dependencies into consideration via 3D convolution. The current performance of the proposed model is still partially limited by the sparsity of axial information, since we only have 8 short-axis scans for each volume. In our future work, we plan to further boost segmentation accuracy by utilizing better out-of-plane information. For example, the current out-of-plane information is complemented by the simple linear interpolation due to the efficiency limit, which could be improved to cubic or spline interpolation in the future for a more accurate and meaningful volume reconstruction. Moreover, we also plan to exploit more useful information from potential additional long-axis MR scans to assist with improving myocardium volume segmentation.

Bibliography

1. Marcu, C. B., Beek, A. M. & van Rossum, A. C. Clinical applications of cardiovascular magnetic resonance imaging. *Canadian Medical Association Journal* **175,** 911–917 (Oct. 10, 2006).
2. Pennell, D. Clinical indications for cardiovascular magnetic resonance (CMR): Consensus Panel report? *European Heart Journal* **25,** 1940–1965 (Nov. 2004).
3. Schwitter, J. *et al.* MR-IMPACT II: Magnetic Resonance Imaging for Myocardial Perfusion Assessment in Coronary artery disease Trial: perfusion-cardiac magnetic resonance vs. single-photon emission computed tomography for the detection of coronary artery disease: a comparative multicentre, multivendor trial. *European Heart Journal* **34,** 775–781 (Mar. 7, 2013).
4. Douglas, P. S. *et al.* Outcomes of Anatomical versus Functional Testing for Coronary Artery Disease. *New England Journal of Medicine* **372,** 1291–1300 (Apr. 2, 2015).
5. Petitjean, C. *et al.* Right ventricle segmentation from cardiac MRI: A collation study. *Medical Image Analysis* **19,** 187–202 (Jan. 2015).
6. Tavakoli, V. & Amini, A. A. A survey of shaped-based registration and segmentation techniques for cardiac images. *Computer Vision and Image Understanding* **117,** 966–989 (Sept. 2013).
7. Petitjean, C. & Dacher, J.-N. A review of segmentation methods in short axis cardiac MR images. *Medical Image Analysis* **15,** 169–184 (Apr. 2011).
8. Hesamian, M. H., Jia, W., He, X. & Kennedy, P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging* **32,** 582–596 (Aug. 2019).
9. Ciresan, D., Giusti, A., Gambardella, L. & Schmidhuber. *Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images* in *Advances in Neural Information Processing Systems* (eds Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) **25** (Curran Associates, Inc., 2012).
10. Tran, P. V. A Fully Convolutional Neural Network for Cardiac Segmentation in Short-Axis MRI. *arXiv:1604.00494 [cs].* arXiv: 1604.00494 (Apr. 26, 2017).
11. Ronneberger, O., Fischer, P. & Brox, T. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer International Publishing, Cham, 2015).

12. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (eds Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G. & Wells, W.) 424–432 (Springer International Publishing, Cham, 2016).

13. Milletari, F., Navab, N. & Ahmadi, S.-A. *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation* in *2016 Fourth International Conference on 3D Vision (3DV)* 2016 Fourth International Conference on 3D Vision (3DV) (IEEE, Stanford, CA, USA, Oct. 2016), 565–571.

14. Mehta, R. & Sivaswamy, J. *M-net: A Convolutional Neural Network for deep brain structure segmentation* in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017) (IEEE, Melbourne, Australia, Apr. 2017), 437–440.

15. Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z. & Tu, Z. Deeply-Supervised Nets. *arXiv:1409.5185 [cs, stat].* arXiv: 1409.5185 (Sept. 25, 2014).

16. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (eds Stoyanov, D. *et al.*) 3–11 (Springer International Publishing, Cham, 2018).

17. Huang, H. *et al.* UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. *arXiv:2004.08790 [cs, eess].* arXiv: 2004.08790 (Apr. 19, 2020).

18. Jang, Y., Hong, Y., Ha, S., Kim, S. & Chang, H.-J. in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges* (eds Pop, M. *et al.*) 161–169 (Springer International Publishing, Cham, 2018).

19. Yang, X., Bian, C., Yu, L., Ni, D. & Heng, P.-A. in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges* (eds Pop, M. *et al.*) 152–160 (Springer International Publishing, Cham, 2018).

20. Chen, M., Fang, L. & Liu, H. *FR-NET: Focal Loss Constrained Deep Residual Networks for Segmentation of Cardiac MRI* in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI) (IEEE, Venice, Italy, Apr. 2019), 764–767.

21. Caliva, F., Iriondo, C., Martinez, A. M., Majumdar, S. & Pedoia, V. Distance Map Loss Penalty Term for Semantic Segmentation. *arXiv:1908.03679 [cs, eess].* arXiv: 1908.03679 (Aug. 9, 2019).

22. Khanh, T. L. B. *et al.* Enhancing U-Net with Spatial-Channel Attention Gate for Abnormal Tissue Segmentation in Medical Imaging. *Applied Sciences* **10,** 5729 (Aug. 19, 2020).

23. Heiberg, E. *et al.* Design and validation of Segment - freely available software for cardiovascular image analysis. *BMC Medical Imaging* **10,** 1 (Dec. 2010).

24. Schulz-Menger, J. *et al.* Standardized image interpretation and post-processing in cardiovascular magnetic resonance - 2020 update: Society for Cardiovascular Magnetic Resonance (SCMR): Board of Trustees Task Force on Standardized Post-Processing. *Journal of Cardiovascular Magnetic Resonance* **22,** 19 (Dec. 2020).

25. Kamnitsas, K. *et al.* Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis* **36,** 61–78 (Feb. 2017).

26. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26,** 297–302 (1945).

27. Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytologist* **11,** 37–50. eprint: https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x (1912).

28. Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15,** 850–863 (1993).

29. Bernard, O. *et al.* Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging* **37,** 2514–2525 (2018).

30. Gibbons, J. D. & Chakraborti, S. in *International Encyclopedia of Statistical Science* (ed Lovric, M.) 977–979 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).