Received: 13 July 2022 | Revised: 28 September 2022 | Accepted: 8 November 2022

DOI: 10.1002/mp.16108

RESEARCH ARTICLE

MEDICAL PHYSICS

Fully automated cardiac MRI segmentation using dilated residual network

Faizan Ahmad[1,2] | **Wenguo Hou**[1] | **Jing Xiong**[3] | **Zeyang Xia**[1,4]

[1]Soft Robotics Research Center, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]Medical Robotics and Minimally Invasive Surgical Devices, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences Shenzhen, China

[4]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Shenzhen, China

Correspondence
Zeyang Xia, Soft Robotics Research Center, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
Email: zy.xia@siat.ac.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: U2013205, 62073309, 6210021302; Chinese Academy of Sciences Youth Innovation Promotion Association Excellent Member Program, Grant/Award Number: Y201968; Guangdong Basic and Applied Basic Research Foundation, Grant/Award Number: 2022B1515020042; Shenzhen Science and Technology Program, Grant/Award Number: JCYJ20220818101603008

Abstract

Purpose: Cardiac ventricle segmentation from cine magnetic resonance imaging (CMRI) is a recognized modality for the noninvasive assessment of cardiovascular pathologies. Deep learning based algorithms achieved state-of-the-art result performance from CMRI cardiac ventricle segmentation. However, most approaches received less attention at the bottom layer of UNet, where main features are lost due to pixel degradation. To increase performance, it is important to handle the bottleneck layer of UNet properly. Considering this problem, we enhanced the performance of main features at the bottom layer of network.

Method: We developed a fully automatic pipeline for segmenting the right ventricle (RV), myocardium (MYO), and left ventricle (LV) by incorporating short-axis CMRI sequence images. We propose a dilated residual network (DRN) to capture the features at full resolution in the bottleneck of UNet. Thus, it significantly increases spatial and temporal information and maintains the localization accuracy. A data-augmentation technique is employed to avoid overfitting and class imbalance problems. Finally, output from each expanding path is added pixel-wise to improve the training response.

Results: We used and evaluated our proposed method on automatic cardiac diagnosis challenge (ACDC). The test set consists of 50 patient records. The overall dice similarity coefficient (DSC) we achieved for our model is 0.924 ± 0.03, 0.907 ± 0.01, and 0.949 ± 0.05 for RV, MYO, and LV, respectively. Similarly, we obtained hausdorff distance (HD) scores of 10.09 ± 0.01, 7.25 ± 0.05, and 6.86 ± 0.02 mm for RV, MYO, and LV, respectively. The results show superior performance and outperformed state-of-the-art methods in terms of accuracy and reached expert-level segmentation. Consequently, the overall DSC and HD result improved by 1.0% and 1.5%, respectively.

Conclusion: We designed a dilated residual UNet (DRN) for cardiac ventricle segmentation using short-axis CMRI. Our method has the advantage of restoring and capturing spatial and temporal information by expanding the receptive field without degrading the image main features in the bottleneck of UNet. Our method is highly accurate and quick, taking 0.28 s on average to process 2D MR images. Also, the network was designed to work on predictions of individual MR images to segment the ventricular region, for which our model outperforms many state-of-the-art methods.

KEYWORDS
cardiac ventricle segmentation, convolutional neural network, deep learning

1 | INTRODUCTION

Automated cardiac ventricle segmentation is one of the challenging tasks in medical image processing. A properly performed segmentation can save physicians considerable time by enabling them to segment regions of interest (ROIs) manually. Moreover, this process eliminates the ambiguities associated with human intervention when multiple experts annotate the same region. Therefore, this would be an invaluable step toward automatic disease diagnosis. Specifically, the human heart is susceptible to a range of cardiovascular conditions, which continue to be a significant cause of death worldwide, if not the leading cause.

Cine magnetic resonance imaging (CMRI) is typically acquired for the assessment of cardiovascular health during the relaxing (diastole) and contracting (systole) phase of a cardiac cycle. It is necessary to calculate metrics such as volume, ejection fraction (EF), and strain[1], as well as to investigate local myocardial wall motion irregularities in order to assess cardiac health. Patients with cardiovascular disease are more likely to experience arrhythmias, difficulty holding their breath, or difficulty remaining still during the acquisition of a CMRI. This results in images from MRI scanners potentially containing a variety of image artifacts[2], making it difficult to assess their quality. Clinicians may draw wrong conclusions from imaging data if the data is segmented incorrectly.[3] The process of manual delineation is time-consuming, tedious, and highly unpredictable. Therefore, a semi-automated or fully automatic segmentation pipeline is required to identify the outer wall (epicardium) and inner wall (endocardium) indices.

Deep convolutional neural networks (DCNNs) have achieved state-of-the-art performance in biomedical image segmentation over the past decade. Specifically, the UNet architecture[4] is task-independent and has been applied to various biomedical segmentation tasks with some minor or substantive modifications. A DCNN has been shown to provide better segmentation of the right and left ventricles and the myocardium than the more conventional methods of computer vision.[5] The U-net architecture has led to the development of the most efficient algorithms for ventricular segmentation.

Recently, semi-automatic or fully automatic techniques have gained considerable attention in research for cardiac CMRI segmentation.[6] Generally, these approaches can be categorized into two broad categories based on the prior knowledge used for segmentation: weak or no prior approaches and strong prior approaches. The first group segments the heart structure entirely based on image content or weak assumptions, for example, anatomical details. This work includes image-based methods, such as threshold[7], and deformable models, which uses active contour or level-set methods.[8–10] The second group worked on strong prior knowledge and the geometrical structures of the heart. The method includes deformable models with a shape prior[11], atlas-based techniques[12], and active shape and appearance models.[13] Despite significant progress in automatic segmentation of multi-structure from CMRI, this task remains challenging due to low contrast between adjacent slices, inherent subject variance, and distorted pixel density.

Numerous deep learning approaches for cardiac segmentation have been proposed recently and have shown superior performance compared to traditional techniques. Tran et al.[14] used short-axis MR images to segment LV and RV by deploying a deep fully convolutional network (FCN) architecture for pixel-wise labelling from CMRI. Oktay et al.[15], utilized the image super-resolution (SR) method for segmenting and classifying cardiac pathologies using residual convolution neural networks (CNNs). Bai et al.[16] used FCN to segment the LV and RV using short-axis CMRI, while the left atrium (LA) and right atrium (RA) were segmented using long-axis CMRI. These approaches use pre-processed MRI data slice by slice before applying them to a two-dimensional (2D) CNNs. Ma et al.[17] developed an iterative multi-path fully convolution network (IMFCN) that combines the advantages of two-dimensional (2D) and three-dimensional (3D) techniques for automated cardiac segmentation. Through the use of a multi-path late fusion strategy and arous spatial pyramid pooling (ASPP) module, the network can effectively leverage spatial context information in 2D network.

When comparing 3D CNN methods, 2D methods have the advantage of having more training data, less memory usage, and less training time.[18–20] However, 3D methods consider 3D-context information, and the consistency can be maintained between segmentations of different slices, implying that 3D methods are more accurate and consistent than 2D methods. Nevertheless, cardiac MRI has a low through-plane resolution which limits 3D methods, such as image size reduction and overfitting. Also, 3D methods can restrict the segmentation performance to some extent.[18,21] For example, previous work[22–24] on Automated Cardiac Diagnosis challenge (ACDC) dataset[25] showed that the 3D models did not perform well as compared to 2D models in terms of performance improvement.

In 2D architecture, there have been several methods for enhancing spatial context. For instance, deep networks can be trained with multiple viewpoints to obtain multi-view information. Then this information is combined to produce segmentation results[26–28] using FCN to incorporate the entire stack of 2D slices to enhance the outcome.[29] MRI of the heart provides low-resolution through-plane and high-resolution in-plane. Therefore, correlations within slices are relatively low, except between adjacent slices. In Patravali's study[22], adjacent slice contextual information can directly be achieved using a volumetric segmentation strategy that involves by grouping them as input to the channels

and generating predictions of the targeted slice. This approach treats adjacent and target slices equally while neglecting the previously predicted segmentation information. This is in contrast to the human expert's workflow. When performing manual segmentation slice by slice, clinicians concentrate on the current slice while comparing it with adjacent and previous slices for segmentation results to maintain spatial consistency. Dangi et al.[17] described shape and boundary constraints for FCN-based architecture for segmenting the heart chambers from CMRI to obtain distance maps prediction at the bottleneck of FCN. The distance map regularization improves segmentation performance for both binary and multi-class predictions.

In the past few years, cardiac segmentation from cine-MRI has been achieved using a variety of CNN methods utilizing deep learning. In some proposals, CNNs are exclusively used for segmentation, whereas others recommend CNNs as a part of a multistage network. The CNNs outputs in the latter case are interpreted as raw segmentations that can be further refined through other well-known methods. Avendi et al.[30] utilized a deep learning approach that obtained segmentation tasks from the ground truth labels and applied them to CNNs to determine the LV chamber's features. After the left ventricle's shape was inferred using a stacked auto-encoder, the shape was integrated into deformable models in order to improve the accuracy and robustness of short-axis cardiac MRI.

A study has proposed a method of automated left ventricle segmentation that combines deep learning with level set techniques.[31] The level set method based on shape and appearance is limited in its ability to represent visual variation. Deep learning techniques can be used to model such variations by utilizing small amounts of labeled training data, but these data must be regularized in order to achieve suitable generalization. Another method was used to improve the accuracy of LV segmentation using cost minimization and dynamic programming that combines deep learning and a region-constrained method.[32] Additionally, multi-scale residual DenseNets[33] and multiple bottleneck layers are used as part of the FCN-based method for cardiac ventricle segmentation[34] to enhance overall accuracy.

In this paper, we present a fully automated segmentation approach for cardiac ventricles to address the limitations of the bottom layer of UNet, such as the loss of spatial and temporal information and the difficulty of handling ventricles of different sizes. The main contribution of the study as follows:

1. First, to increase the spatial dimension, we design the dilated residual network (DRN) block and replace it at the bottom layer of the original UNet. Thus, it significantly enhances the spatial and temporal information while maintaining the main features in the image during pixels degradation at the bottom layer of UNet. A residual block is added before every dilated convolution layer to add features from the previous layer. Thus, spatial consistency can be preserved while accuracy can be improved simultaneously.
2. Second, to effectively demonstrate our proposed method, we performed data augmentation to increase the training samples to solve the overfitting and class imbalance problems.
3. Lastly, output from each expanding path is added pixel-wise to improve the training response.

2 | METHODS

The proposed workflow comprises a number of steps from raw cine MR images to the final segmentation result, as shown in Figure 1. In this section, we will explain the two main steps: (i) preliminary data preparation, including preprocessing, cropping, normalizing, and data-augmentation. (ii) The segmentation network architecture consists of a base UNet architecture, dilated residual convolution, hyper-parameters tuning, and inference.

We designed a network that automatically segments the RV, MYO, and LV from raw input images composed of 3D volumes having dimensions $L \times W \times H$. The Raw MR image having volume X is generated by stacking N-2D short-axis slices. A set of 3D label maps of size $L \times W \times H$ are provided. These maps contain Black, RV, MYO, and LV (i.e., black = 0, RV = 1, MYO = 2, LV = 3). Based on the input image X, we extracted short-axis slices beginning at the mitral valves and ending at the apex of the left ventricle.

2.1 | Data pre-processing

The train and test sets of cine-MRI show significant variations in spatial dimensions $W \times H$ and the range of the intensity distribution. We subsequently designed a pre-processing stage in which the MRI data is pre-processed in two steps:

1. The ACDC dataset has a voxel spacing issue. Due to the inability of CNNs to interpret voxel spacing, we preprocess the ACDC dataset by resampling all images to the same voxel spacing of $1.52 \times 1.52 \times 6.35$ mm. Generally, the patch size that can be processed is relatively large. Therefore, the voxel spacing, which directly affects the overall voxel size of the images, also affects the amount of contextual information that the CNNs can extract from the image patches. Furthermore, if the voxel spacing is increased substantially, the image size is reduced to the point where the details are lost. It is imperative to ensure that the trade-off between the amount of contextual information contained in the network patch size and

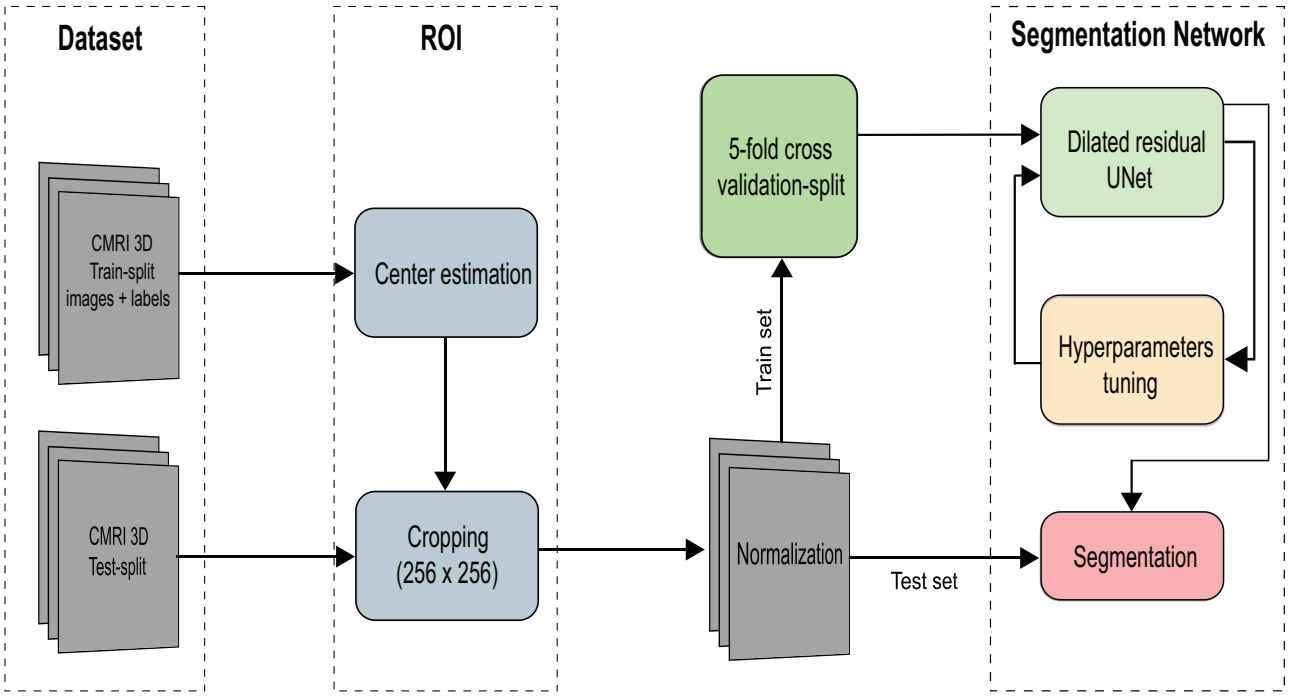

FIGURE 1 Segmentation workflow from raw cine-magnetic resonance imaging data to segmentation.

the amount of detail retained in the image data is optimized for optimum performance. Therefore, we resample all images to a median value of *256×256* pixels for training data.

2. MR images were obtained for the ACDC dataset using multi-slice cine MRI. This procedure involves extracting each patient's 2D-MRI slices and their associated annotations. We performed normalization slice by slice for each time frame.

$$X_{i,j(norm)} = \frac{X_{i,j} - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where $X_{i,j}$ denotes the pixel intensity. The slices $X_{\min}$ and $X_{\max}$ represent the minimum and maximum pixel intensities, respectively.

2.2 | Data augmentation

Data augmentation is used when there are insufficient training images to train the model. It is of even greater significance when working with medical images, particularly for segmentation tasks. Accurately annotating ROI requires expert knowledge and is also time-consuming. Even when the same expert annotates the ROIs twice, there are still ambiguities. Due to the limited training data, the model cannot learn the desired invariance and robustness features, which leads to overfitting. Therefore, we applied various data augmentation techniques to the initial phases of model design and training.[35] A wide range of basic image transformation techniques, including random rotations, random elastic deformations, scaling, flipping, and gamma correction, are frequently applied to increase the number of training samples. When applied to the original training images, they effectively generate several views of the same image, as shown in Figure 2. The detailed parameters for data augmentations are given in Table 1.

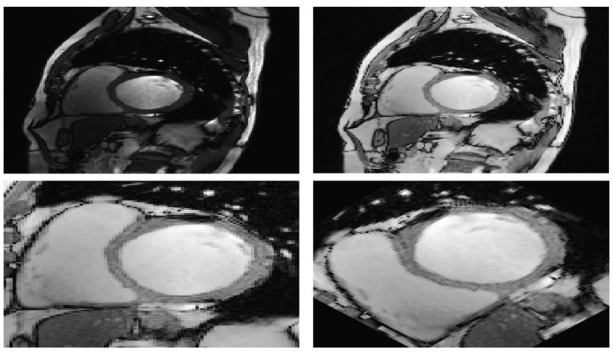

FIGURE 2 Augmented images of training dataset. Original (top left), intensities (top right), cropping (bottom left), and rotation (bottom right).

TABLE 1 Data augmentation parameters during training

Parameters	Values	Probability (%)
Rotation	$[0,\ \pi]$	50
Scaling	[0.98, 1.02]	50
Flipping	[−0.17, 0.17]	50
Deformation	$\alpha = 30,\ \sigma = 4.0$	30

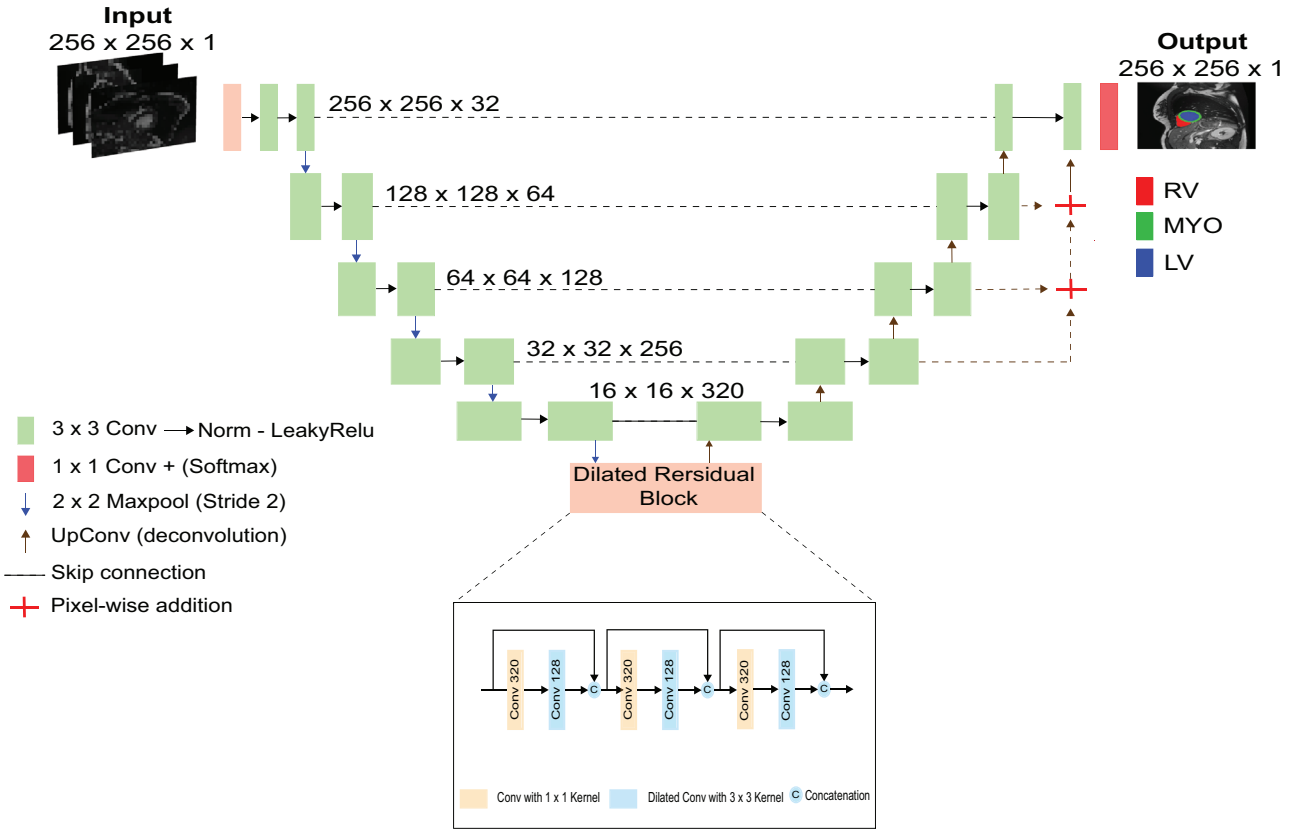

FIGURE 3 Proposed fully automated segmentation pipeline. Dilated residual network is used to recover spatial and temporal information at bottleneck of UNet. The predicted output represented as: (right ventricle (red), myocardium (green), and left ventricle (blue)).

2.3 | Segmentation network

We developed a fully automated segmentation pipeline using a well-known encoder-decoder module, called UNet[4] as base network to segment the RV, MYO, and LV. We proposed a new and lightweight DRN for cardiac segmentation to automate the entire process at full resolution.

Convolution is a powerful operation that extracts features from images automatically by sliding the kernel over the input image. An important characteristic of convolution is that it is translationally unified, meaning that the output remains constant despite a minor shift in the input image. UNet based encoder-decoder method incorporates convolutional layers to extract more high-level semantic features. The convolutional layers output is down-sampled using max-pooling and then restored to the original size using deconvolution. As a result of the pooling operation, the translational unified feature may no longer hold, making the network delicate to minor changes in the input image.[36].

We followed the encoder-decoder approach throughout the segmentation process, from the input image to the final output. The contracting path is constructed using a series of five encoding blocks; each block consists of two convolutional layers with a *3×3* kernel and a *2×2* max-pooling operation with a stride of 2. Initially, 32 convolutional filters are selected. After every max-pooling operation, the filters are doubled, resulting 320 filters in the bottleneck layer of UNet. Similarly, the spatial dimensions of the feature maps are reduced by a factor of 2 through down-sampling operation. Rectified linear unit (ReLU) is replaced with leaky ReLU, and instance normalization[37] is employed rather than batch normalization (BN).[38].

We combine the encoding and decoding path at the bottleneck of UNet through a DRN, which captures the global context and restores the spatial and temporal information without affecting the resolution of the segmentation map. Furthermore, it effectively adjusts the convolution layer's depth without degrading the network. The receptive field is enlarged in the DRN block by employing dilated convolutions with different dilation rates ($d = 1, 3,$ and 5). The previously generated features are then concatenated with the current features through residual connection. A dropout operation is performed to prevent overfitting with a dropout rate of 0.5 after every *3×3* convolution in the DRN block. Thus, the DRN captures contextual image information, high spatial resolution, and multi-textured features. The process for the decoding path is similar to that of the encoding path; however, the order of operations is

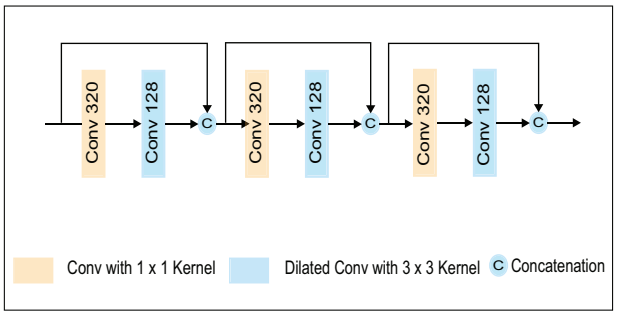

FIGURE 4 Dilated residual block

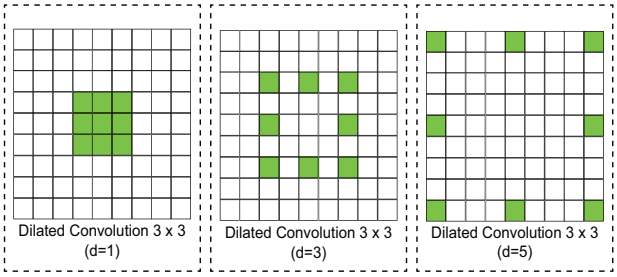

FIGURE 5 Dilated convolution. On the left dilated rate d is 1, middle d is 3, and on the right d is 5. The convolution filter is 3×3 with same number of parameters.

reversed. The deconvolution operation is used during each decoding block, and the *3×3* kernel operation is employed in both convolution layers. The deconvolution gradually increases the spatial dimensions and reduces the feature maps by 2. UNet architecture offers the advantage of reusing encoded feature maps from encoding blocks to their corresponding levels in decoding blocks where the spatial dimensions match. This can be achieved through channel-specific concatenation. A *1×1* kernel projection operation is used at the final level of the decoding path to align the output channel dimension to the segmented classes (RV, MYO, and LV). Finally, we aggregated all expanding path level outputs and performed pixel-wise additions via up-sampling with appropriate projections to enhance the training response. The proposed network is shown in Figure 3.

2.4 | Dilated residual convolution

Our main idea is to preserve spatial and temporal information at bottleneck of UNet for RV, MYO, and LV cardiac segmentation. Though progressive down-sampling has been very effective at classifying objects; however, the loss of spatial information may have adverse effects on segmentation and can affect other tasks that require spatially detailed image features. Typically, natural images contain many objects whose identities and relative positions are significant for understanding the scene. Additionally, segmentation becomes more difficult when an important object is not spatially prominent, for example, when the required object (class) is small compared to the background. Consequently, the background response may suppress the features of object (class). If the features of the targeted object are lost during down-sampling, it is not easy to recover during training. However, if we maintain high spatial and temporal information throughout the network and provide output features that densely cover the input features, backpropagation can learn important features from smaller and less salient objects. Therefore, we construct the network with a large receiving field, which can extract more spatial information to predict small and dense image features. The discrete dilated convolution[39] as follows:

$$G(p) = \sum_{s+lt=p} F(s) *_l k(t) \quad (2)$$

where, $F, G : \mathbb{Z}^2 \rightarrow \mathbb{R}$ are input and output discrete functions. Let $\Omega_d = [-d, d]^2 \cap \mathbb{Z}^2$ and $k : \Omega_d \rightarrow \mathbb{R}$ are discrete kernel size of $(2d + 1)^2$ and $*_l$ is dilated convolution.

We used three dilated convolution with dilated rate ($d = 1, 3,$ and 5), as shown in Figure 5. Thus, every convolutional kernel $(2d + 1)^2$ having dilation rate d for which the size of receptive field is proportional to d. Increasing the depth of a network will result in significant degradation of performance and training accuracy. To address these issues, we introduce residual blocks, due to the fact that residual mapping is easy to learn than original mapping. Through residual learning, deep networks are trained to improve their accuracy for tasks including

TABLE 2 Training parameters of five-folds cross-validation

Folds	1	2	3	4	5	Average
Training accuracy	0.997	0.998	0.997	0.999	0.999	0.998
Validation accuracy	0.973	0.976	0.961	0.932	0.962	0.961
Training loss	0.035	0.035	0.035	0.035	0.036	0.035
Validation loss	0.086	0.087	0.087	0.086	0.093	0.088
DSC RV	0.964	0.967	0.967	0.968	0.965	0.966
DSC MYO	0.939	0.942	0.938	0.939	0.931	0.938
DSC LV	0.918	0.911	0.914	0.913	0.910	0.913

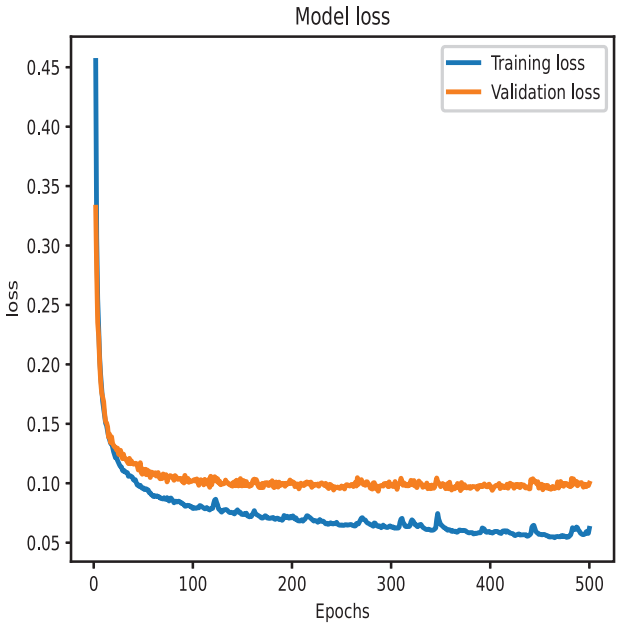

FIGURE 6 Mean training and validation loss of five-folds.

TABLE 3 The Bias$\mu$and limits agreements of cardiac clinical and functional indices

Indices	Bias	Upper limit	Lower limit
LVESV(ml)	2.08	23.61	–19.44
LVEDV(ml)	0.57	20.07	–18.93
RVESV(ml)	–1.81	13.72	–17.34
RVEDV(ml)	–3.46	14.43	–21.35
EFLV(ml)	–0.93	5.05	–6.92
EFRV(ml)	0.18	4.93	–4.56
EDMYO(g)	1.69	16.95	–13.56
ESMYO (g)	0.78	14.05	–12.49

image classification and object detection. Medical image segmentation with a UNet like structure that learns residual maps converges much faster and consistently performs well.

The proposed method only employs five convolution layer blocks to extract the features. In complex medical image analysis tasks, such as multiclass image segmentation, the conventional framework may not be suitable. Although UNet can easily distinguish between the three classes (RV, MYO, and LV), it has a limited ability to distinguish between small features; the apex slice during the cardiac cycle is particularly challenging. Therefore, it is necessary to build deeper network with more layers, particularly for the encoder block. However, when deeper networks converge, a problem will be revealed: as network depth increases, accuracy gets very high and then decrease rapidly. Using skip connection and residual learning, ResNet[40] overcomes the degradation of the network and avoids estimating a large number of parameters generated by the convolutional layer.

We developed a DRN to expand the receptive field better, attain a promising result, and refrain from image degradation at the bottleneck of UNet. Each pixel value in the output of a convolutional network only depends on one specific region of the input. Intuitively, a larger region is capable of capturing more contextual information. To ensure that no essential features are omitted, CNNs must have a larger receptive field. The dilated convolutional layers are based on regular convolution with dilation factors $d = 1, 3,$ and 5. We choose *1×1* kernel for the convolutional layer and *3×3* kernel for dilated convolution as shown in Figure 4.

$$y_{ij}(m,n) = \sum_{i=1}^{M}\sum_{j=1}^{N} x(m + d \times i, n + d \times j)w(i,j) \quad (3)$$

where y_{ij} *(m,n)* represents the dilated convolution of the input x_{ij}*(m,n)* and having a filter of length *M* and width *N*. When $d = 1$, the dilated convolution becomes a standard convolution. In dilated convolution, sparse kernels are used to extend the receptive field without increasing the additional parameters or causing the network to become more complex. Adding more convolution layers can increase the receptive fields but introduce more operations. Nevertheless, dilated convolution is achieved by extending small filter of size $f \times f$ to $(f - 1)(s - 1)$ with dilated stride *s*. For example, the normal convolution gives a receptive field of 3×3, and two dilated convolutions give receptive fields of 5×5 and 7×7, respectively.

3 | TRAINING AND POST PROCESSING

3.1 | Dataset

ACDC[25] contains cine-MRI images of 150 patients acquired by two different magnetic strengths of MR scanners (1.5 T and 3.0 T) at the University Hospital of Dijon. Each time series contains 28–40 individual 3D volumes that partially or entirely cover the cardiac cycle. A series of short-axis slices cover the whole LV from base to apex with a slice thickness of 5–8 mm, inter-slice gap of 5–10 mm, and spatial resolution varies from 1.37 to 1.68 $mm.^2$/pixel. Due to varied breath-hold positions during slice stack acquisition, certain images have significant slice misalignment. Clinical experts manually segmented the ROIs, including the RV cavity, LV myocardium, and LV cavity, during end-diastolic (ED) and end-systolic (ES) phases. The dataset is evenly distributed between four groups of pathologies and one group of healthy patients. These include patients with hypertrophic cardiomyopathy (HCM), abnormal right

TABLE 4 Quantitative ACDC test set results comparison of our model and popular models in term of dice coefficient (DC) and hausdorff distance (HD)

	LV cavit				RV cavity				MYO			
	DC (%)		HD (mm)		DC (%)		HD (mm)		DC (%)		HD (mm)	
Method	ED	ES	ED	ES	ED	ES	ED	ES	ED	ES	ED	ES
Simantiris et al.[47]	0.967	0.928	6.366	7.573	0.936	0.889	13.289	14.367	0.891	0.904	8.264	9.575
Isensee et al.[23]	0.967	0.928	**5.476**	**6.921**	0.951	**0.904**	8.205	**11.655**	**0.904**	**0.923**	**7.014**	7.325
Zotti et al.[42]	0.964	0.912	6.180	8.386	0.934	0.885	11.052	12.650	0.886	0.902	9.586	9.291
Painchaud et al.[43]	0.961	0.911	6.152	8.278	0.936	0.884	13.718	13.323	0.881	0.897	8.651	9.598
Khened et al.[33]	0.964	0.917	8.129	8.968	0.935	0.879	13.994	13.930	0.889	0.898	9.841	12.582
Christian et al.[24]	0.963	0.911	6.526	9.170	0.932	0.883	12.670	14.691	0.892	0.901	8.703	10.637
Baldeon et al.[45]	0.958	0.903	5.592	8.644	0.936	0.884	10.183	12.234	0.873	0.895	8.197	8.318
Yang et al.[46]	0.938	0.900	11.471	10.33	0.872	0.831	19.195	19.020	0.820	0.862	12.499	11.872
DRN (proposed in this paper)	**0.968**	**0.930**	6.310	7.420	**0.953**	0.895	**8.139**	12.050	0.902	0.913	7.190	**7.310**

ventricles (ARV), dilated cardiomyopathy (DCM), or with myocardial infarction (MINF), as well as healthy individuals (NOR). Additionally, height and weight are provided for each patient. The dataset consists of 100 patients for training and 50 patients for testing. Segmentation and classification are performed only on the 100 training cases.

3.2 | Evaluation metrices

The objective of segmentation is to detect the target object and draw a contour around it. Automated segmentation contour C_p (predicted) is compared with corresponding label contour C_g (ground truth) to measure the accuracy of proposed method. Furthermore, pixels surrounded by the contours are referred as A_p and A_g.

1. *Dice similarity coefficient (DSC)*: The ratio between the predicted contour and ground truth contour represents the DSC score, which is typically expressed in percentage between 0 and 1. High dice values indicate a good match.

$$Dice = \frac{2|A_p \cap A_g|}{|A_p| + |A_g|} \tag{4}$$

2. *Hausdroff distance (HD)*: HD compares the symmetrical distance between predicted and actual contour and provides a spatial resolution of cine-MRI. The low value of HD distance is evidence of a good segmentation match.

$$HD(C_p, C_g) = \max\left(\max_{i \in C_p}\left(\min_{j \in C_g} d(i,j)\right), \max_{j \in C_g}\left(\min_{i \in C_p} d(i,j)\right)\right) \tag{5}$$

3.3 | Loss function

There is a significant class imbalance in medical images between ROIs and the background. In order to address this problem, different loss functions have been tested, including dice loss and weighted cross-entropy loss. To train our model, we used a dual loss function that incorporated dice loss and cross-entropy loss.

The cross-entropy loss quantifies cumulative error across all pixels by calculating the probability of an error between the predicted output and ground truth label pixel-wise. Let $W = (w_1, w_2, w_3, \ldots w_n)$ are a set of learnable weights, where w_n is the weight matrix of n_{th} layer, and $p(Y_i|X_i, W)$denotes the probability of a pixel X_i. The cross-entropy loss is defined as follows:

$$L_{CE} = -\sum_{c=0}^{C} \log p(Y_i|X_i, W) = \sum_{c=0}^{3} Y(c,x) \log \widehat{Y}(c,x) \tag{6}$$

where C denotes the total number of classes, Y is the targeted label, $\hat{Y}$ is the softmax score of predicted class.

3.4 | Training

We trained our model on Nvidia RTX 3090 (24GB). The architecture was implemented using PyTorch 1.6 framework. Initially, the main aim was to reduce the total loss using Equation (8). All hyper-parameters were reconfigured, includes Adam.[41] optimizer with a learning rate of 3×10^{-4}, dropout of 0.2, and batch size of 8. The 2D MR images were resized to 256×256 dimensions. The proposed model was trained using a five-fold cross-validation approach to obtain and evaluate the performance of the training set. During the cross-validation process, we ensured that the ED and ES images of a patient were equally distributed. We then randomly divided the pre-processed training data into

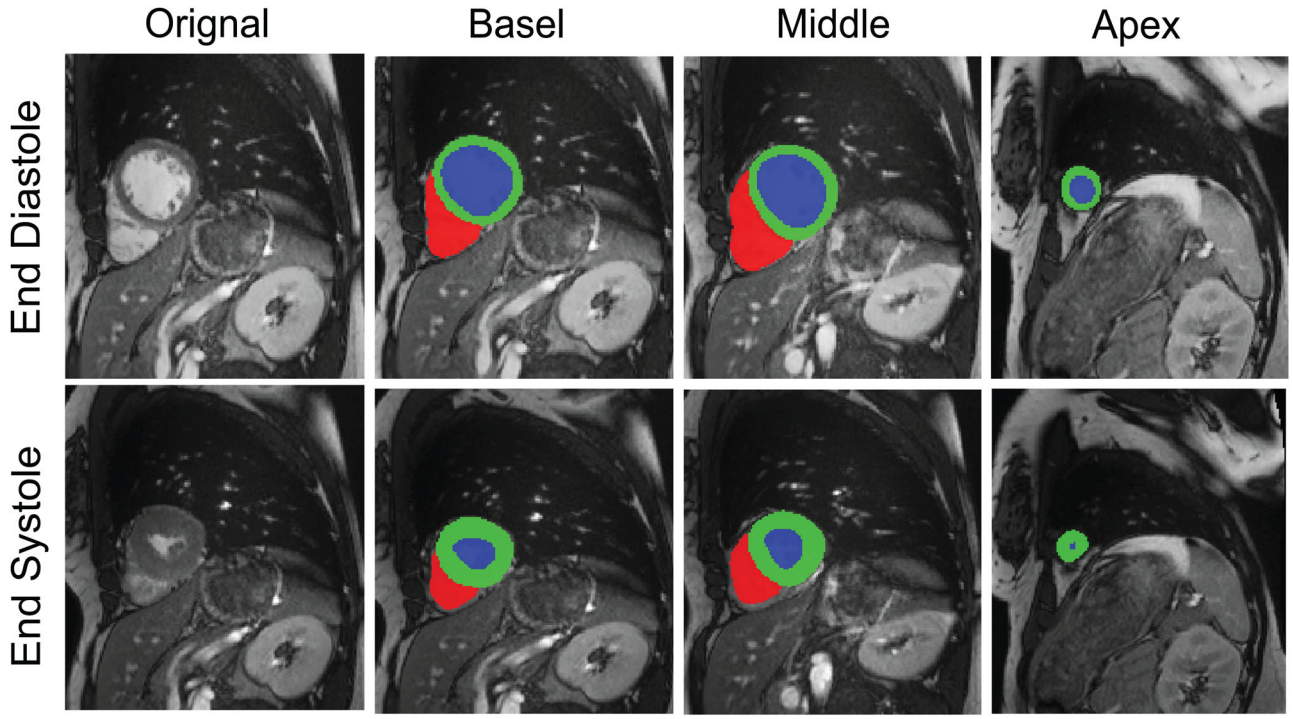

FIGURE 7 Segmentation results on ACDC test dataset. The predictions are shown on End-diastole (ED) and End-systole (ES) cardiac cycle for Basal, Middle, and Apex slices

five-folds, ensuring that each fold contained data from all five groups of pathologies. Each fold of the training data was augmented during the training. The validation data remained unseen during the training stage.

The model was trained for 500 epochs, with 250 iterations per epoch. The batches were created by randomly selecting cases from the training split. In order to resample the patches, we cropped random slices from the training images and evaluated the network after each epoch on validation set. The network is trained using a multiclass variant of dice loss as described in Equation (7). After each epoch, the learning rate *lr* was recalculated, as from Equation (9). Finally, the best model was selected for evaluating the test set to ensure the validation of RV, MYO, and LV classes that achieved a top DSC. Table 2 illustrates the training parameters of our proposed model for five-folds cross-validation. The network provides consistent and stable performance across all folds.

$$L_{DC} = 1 - \frac{2}{|C|} \sum_{c \in C} \frac{\sum_{ik} u_{ik} v_{ik} + \varepsilon}{\sum_{ik} u_{ik} + \sum_{ik} v_{ik} + \varepsilon} \tag{7}$$

Where u and v are the one-hot encoded vector of softmax output and ground truth, respectively. Where $c \in C$is class identifier ($C = 0; 1; 2; 3$) background and three classes, that is, RV, MYO, and LV for ACDC dataset and ε is the small number (constant) for network stability.

$$L_{total} = L_{CE} + L_{dc} \tag{8}$$

$$lr = initial\_learning\_rate\left(1 - \frac{currentepoch}{totalepochs}\right)^{0.9} \tag{9}$$

4 | EXPERIMENTS AND RESULTS

We used the trained weights for the associated folds to calculate the DSC score directly on the validation dataset for RV, MYO, and LV on both cardiac cycle, ED, and ES. Thus, we are able to monitor how each parameter, whether added or changed, affected our DSC score during network training. The mean training and validation loss are shown in Figure 6.

Table 4 summarizes the results of our proposed model on ACDC test dataset. For the LV cavity, the highest DSC score was 0.968 and 0.930 on ED and ES phases, respectively. When compared to the previous models, we have seen a significant improvement during both cardiac phases. Similarly, for the RV cavity, our model achieved the highest DSC score of 0.953 on ED, but for the ES phase, we achieved the second-best score of 0.895. Finally, for MYO we achieved the second-highest DSC score of 0.902 and 0.913 on ED and ES, respectively.

The Hausdorff distance (HD) is the second evaluation metric we utilize in our study to assess the effectiveness

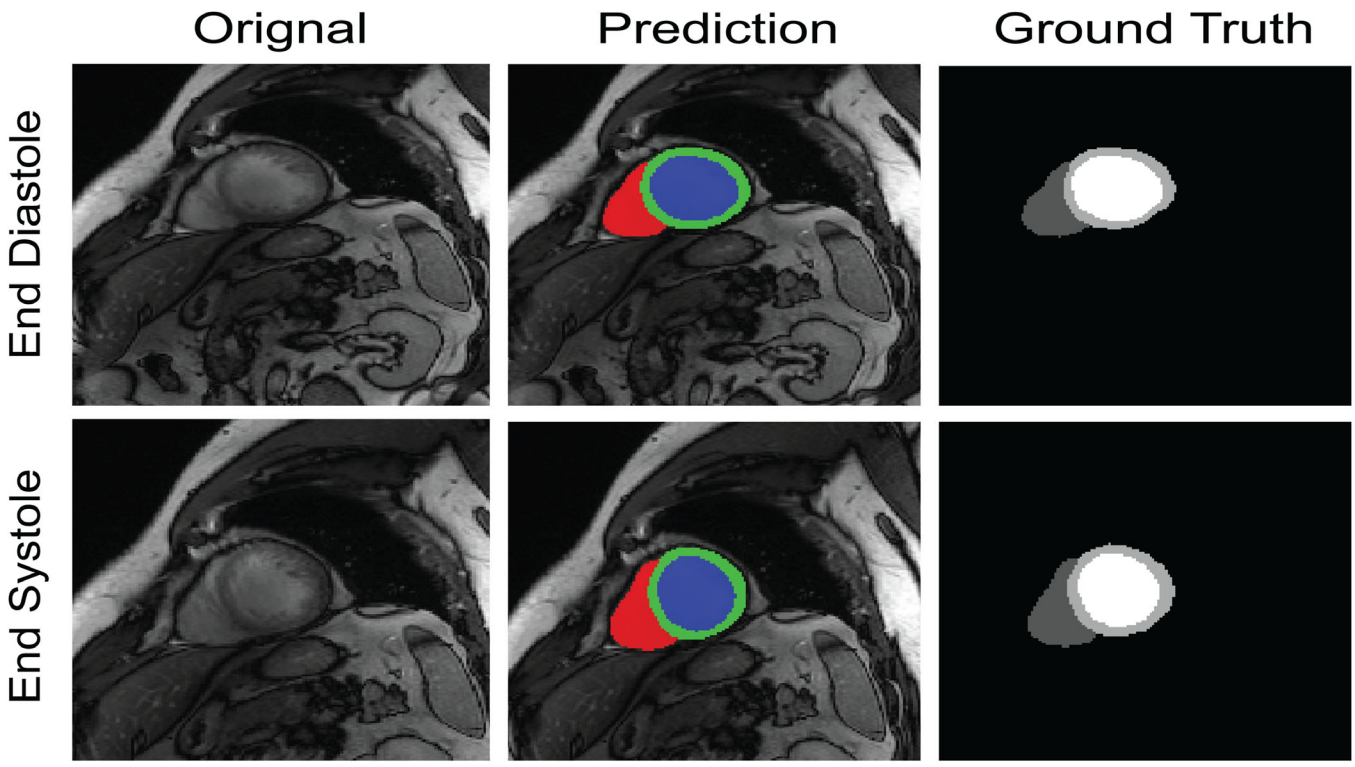

FIGURE 8 Segmentation results on ACDC validation set for End-Diastole (ED) and End-Systole (ES) phase.

of our algorithm. HD identifies the contour distance between the prediction and the ground truth segmentation. Lower values indicate better performance. For the LV cavity on both the ED and ES phases, we achieved HD scores of 6.310 and 7.420 mm, respectively. Likewise, for the RV cavity, we achieved the best HD score of 8.139 mm on ED, and the second-best HD score of 7.420 mm on ES. Similarly, for MYO, ES score of 7.310 mm remains the best, while the ED score of 7.190 mm is the second-best score.

Finally, we analyze and compare our results with those of fully automated and semi-automated methods that appear on the ACDC challenge leaderboard. Isensee et al.[23] used 3D UNet architecture. Khened et al.[33] work on densely connected fully convolutional network (FCN) by not utilizing the skip connections that are used in the UNet architecture. Zotti et al.[42] used a multi-resolution GridNet architecture, which is an extension of the UNet architecture. Painchaud et al.[43] proposed an adversarial variational auto-encoder (aVAE) for anatomically plausible segmentation.

The test set of ACDC includes 50 patients, each with ED and ES frames, a total of 100 images (Figure 7), illustrates the test set results along with our model predictions. However, there are no available ground truth masks for comparison. The predicted results demonstrate a high level of segmentation accuracy for the basal, Middle, and Apex slices during the ED and ES cardiac cycles. Similarly, Figure 8 illustrates the results of our model on the validation set. Visual analyses show that the most challenging part was apex slice due to the small size of the cardiac anatomical structure in that region.

Due to the partial volume effect, the RV and aorta are visible simultaneously on the basal slice. There is no clear distinction between the RV and the pulmonary artery. However, only one or two slices have a median HD greater than 10 mm, while the rest, especially the LV and Myo, have a median HD less than 8 mm.

We also identify and calculated the endocardium and epicardium of the LV and RV. Several clinical and functional indices are derived from these segmentation maps, including left ventricular end-systolic volume (LVESV), left ventricular end-diastolic volume (LVEDV), right ventricular end-systolic volume (RVESV), right ventricular end-diastolic volume (RVEDV), ejection fraction left ventricle (EFLV), ejection fraction right ventricle (EFRV), end-systolic MYO mass (ESMYO), and end-diastolic MYO mass (EDMYO). Moreover, we compared manual assessment with automatic predictions of heart clinical and functional indices. Bland-Altman[44] plots are used to assess the degree of agreement between the manual segmentation values and the predicted segmentation results as shown in Figure 9.

The Bias $\mu$and upper and lower limits are: [2.08 ml, (23.61 ml, −19.44 ml)], [0.56 ml, (20.07 ml, −18.93 ml)], [−1.81 ml, (13.71 ml, −17.33 ml)], [−3.46 ml, (14.42 ml, −21.35 ml)], [−0.93 ml, (5.05 ml, −6.91 ml)], [0.18 ml, (4.92 ml, −4.56 ml)], [1.69 g, (16.94 g,

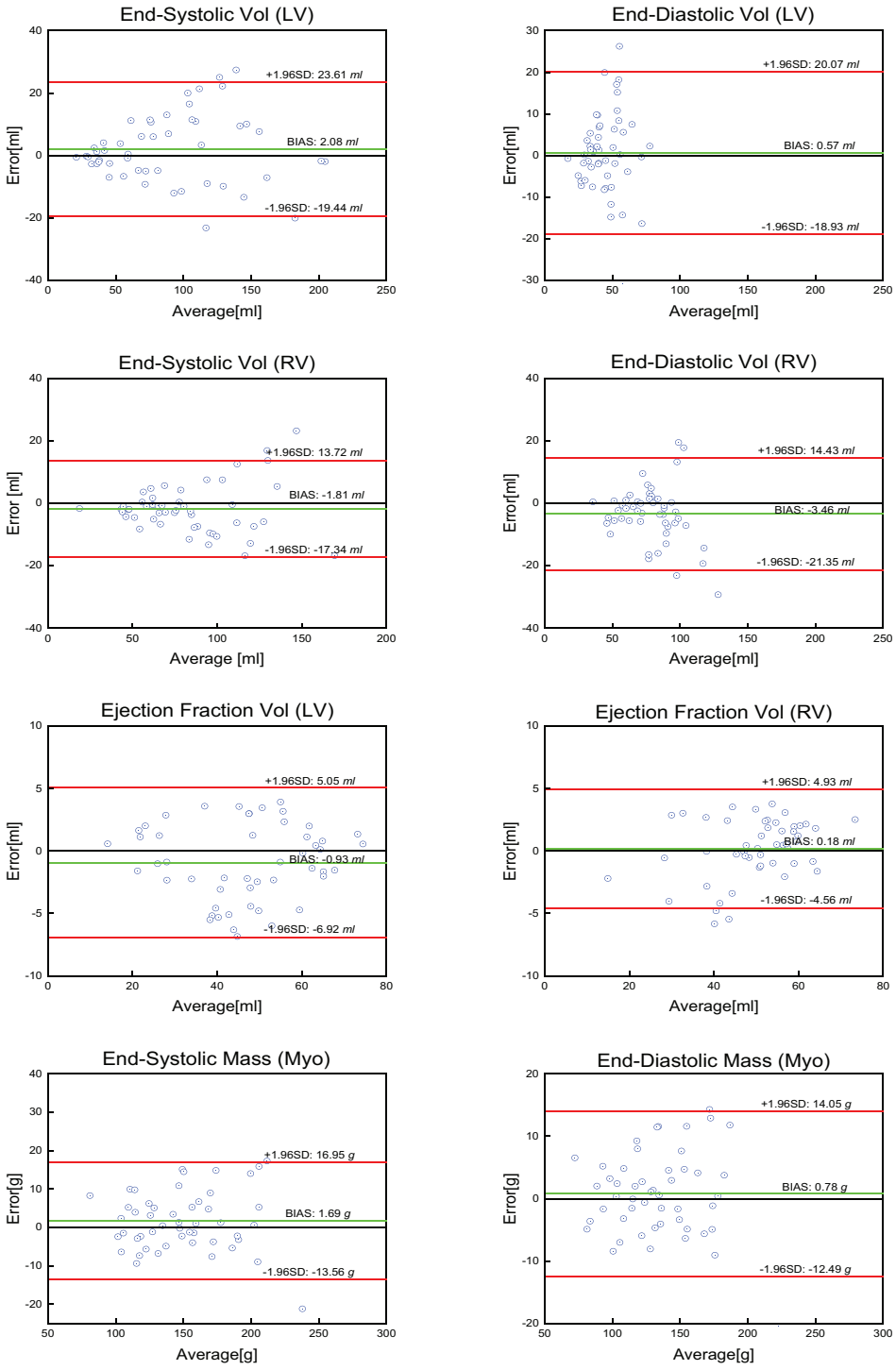

FIGURE 9 Bland-Altman plots for bias and limits agreements (Manual segmentation vs. Automatic prediction) results.

−13.55 g)] and [0.77 g, (14.04 g, −12.48 g)] for LVESV, LVEDC, RVESV, RVEDC, EFLV, EFRV, ESMYO, and EDMYO, respectively. From these results, our automated cardiac ventricle segmentation framework show good limits of agreement with expert level of manual analysis. The clinical indices are mentioned in Table 3.

5 | DISCUSSION

This study aims to design a fully automated segmentation network for cardiac ventricle segmentation. There is a good overlap between our proposed RV, MYO, and LV segmentation methods and the reference frame. The effectiveness of our model was compared to that of

other deep learning approaches with the same objective but different approaches. We find that our model outperforms many other models in terms of DSC and HD scores. Nevertheless, compared to Isensee et al.,[23] we achieved very closed results on DC and HD, respectively. The reason behind this is that the baseline network is UNet for both the studies; however, we utilized the DRN block in our model to capture the minor details and recover the spatial and temporal information, which enhanced the overall performance, and improved the segmentation results. We selected pre-processed input patches based on the lowest possible size that contained RV, MYO, and LV regions in the ACDC dataset. This approach has two main advantages: (i) minimizes the computation time required for training and inference, and (ii) efficiently addresses the problem of class imbalance by excluding unnecessary tissues or background pixels. Most segmentation methods do not explicitly consider spatial and temporal resolution consistency. Specifically, they fail to correctly segment the apex slice due to difficult regions during the cardiac cycle ES and ED, respectively. Our method is significantly more robust than others on these slices. There are 150 different patient data samples, of which 100 were used for training and 50 for testing. Therefore, the algorithm is automatic and can be applied to any new and diverse cardiac CMRI dataset.

Recent research into cardiac motion tracking and image synthesis requires reconstructing a 3D mesh model based on segmentation as a primary component. Without spatial and temporal resolution consistency in the mesh generation, the generated 3D model would be challenging for reconstruction. Moreover, our method is more effective at measuring HD than many other state-of-the-art methods, making it an excellent tool for future studies analyzing cardiac motion. Generally, the smaller the Hausdorff distance between the output and ground-truth contour, the more precisely the path of corresponding structures (e.g., RVC, MYO, LVC) can be captured, and the better motion can be tracked.

6 | CONCLUSION

We designed a dilated residual UNet (DRN) for cardiac ventricular segmentation for short-axis CMRI. The proposed method incorporates two major steps in the original UNet, architecture. Firstly, we replaced the bottom of UNet with dilated convolution and added the residual block before every dilated convolution layer to add the features of the previous layer. As a result, spatial and temporal information is captured by expanding the receptive field without degrading the image main features in the bottleneck of UNet. Secondly, we aggregated all the expanding paths output at a lower resolution and added pixel-wise to the highest output resolution of the decoding blocks. We also used various data augmentation techniques during the training to overcome the class imbalance and overfitting problem. The DRN shows significantly improved results on validation and test set of ACDC dataset. Also, we used a multi-loss function for our network to globalize the heart features during segmentation and improve accuracy. These changes make our approach to predict less false positives and false negatives. We achieved an overall DSC of 0.92 ± 0.02 and a mean Hausdorff distance of 8.06 ± 0.05 mm. Our method is highly accurate and quick, taking 0.28 s on average to process 2D MR images. Also, the network was designed to work on predictions of individual MR images to segment the ventricular region, for which our model outperforms many state-of-the-art methods. This gives us the first step toward the ultimate goal of designing a completely automated network with an error rate lower than manual segmentation.

In the future, we will explore using our automated segmentation pipeline to integrate it with augmented reality based applications to provide a real-time immersive views of cardiac ventricles.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (U2013205, 62073309, 6210021302), and in part by the Chinese Academy of Sciences Youth Innovation Promotion Association Excellent Member Program (Y201968), Guangdong Basic and Applied Basic Research Foundation (2022B1515020042), and Shenzhen Science and Technology Program (JCYJ20220818101603008).

CONFLICT OF INTEREST

The authors have no conflicts to disclose.

REFERENCES

1. Ruijsink B, Puyol-Antón E, Oksuz I, et al. Fully automated, quality-controlled cardiac analysis from CMR. *JACC Cardiovasc Imaging*. 2020;13(3):684-695. https://doi.org/10.1016/j.jcmg.2019.05.030
2. Ferreira PF, Gatehouse PD, Mohiaddin RH, Firmin DN. Cardiovascular magnetic resonance artefacts. *J Cardiovasc Magn Reson*. 2013;15(1):41. https://doi.org/10.1186/1532-429X-15-41
3. Oksuz I, Clough J, Ruijsink B, et al. Detection and correction of cardiac MRI motion artefacts during reconstruction from k-space. In: 2019:695-703. https://doi.org/10.1007/978-3-030-32251-9_76
4. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Cpmputert-Assisted Intervention; 2015:234-241. https://doi.org/10.1007/978-3-319-24574-4_28
5. Chen C, Qin C, Qiu H, et al. Deep learning for cardiac image segmentation: a review. *Front Cardiovasc Med*. 2020;7. https://doi.org/10.3389/fcvm.2020.00025
6. Peng P, Lekadir K, Gooya A, Shao L, Petersen SE, Frangi AF. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. *Magn Reson Mater Physics, Biol Med*. 2016;29(2):155-195. https://doi.org/10.1007/s10334-015-0521-4

7. Liu H, Hu H, Xu X, Song E. Automatic left ventricle segmentation in cardiac MRI using topological stable-state thresholding and region restricted dynamic programming. *Acad Radiol*. 2012;19(6):723-731. https://doi.org/10.1016/j.acra.2012.02.011
8. Gotardo PFU, Boyer KL, Saltz J, Raman SV. A new deformable model for boundary tracking in cardiac MRI and its application to the detection of intra-ventricular dyssynchrony. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*. Vol 1. IEEE; :736-743. https://doi.org/10.1109/CVPR.2006.34
9. Ting Chen, Babb J, Kellman P, Axel L, Kim D. Semiautomated segmentation of myocardial contours for fast strain analysis in cine displacement-encoded MRI. *IEEE Trans Med Imaging*. 2008;27(8):1084-1094. https://doi.org/10.1109/TMI.2008.918327
10. Feng C, Zhang S, Zhao D, Li C. Simultaneous extraction of endocardial and epicardial contours of the left ventricle by distance regularized level sets. *Med Phys*. 2016;43(6Part1):2741-2755. https://doi.org/10.1118/1.4947126
11. Queirós S, Barbosa D, Heyde B, et al. Fast automatic myocardial segmentation in 4D cine CMR datasets. *Med Image Anal*. 2014;18(7):1115-1131. https://doi.org/10.1016/j.media.2014.06.001
12. Bai W, Shi W, Ledig C, Rueckert D. Multi-atlas segmentation with augmented features for cardiac MR images. *Med Image Anal*. 2015;19(1):98-109. https://doi.org/10.1016/j.media.2014.09.005
13. Mitchell SC, Bosch JG, Lelieveldt BPF, van der Geest RJ, Reiber JHC, Sonka M. 3-D active appearance models: segmentation of cardiac MR and ultrasound images. *IEEE Trans Med Imaging*. 2002;21(9):1167-1178. https://doi.org/10.1109/TMI.2002.804425
14. Tran PV. A fully convolutional neural network for cardiac segmentation in short-axis MRI. Published online April 2, 2016. http://arxiv.org/abs/1604.00494
15. Oktay O, Bai W, Lee M, et al. Multi-input cardiac image super-resolution using convolutional neural networks. Medical Image Computing and Computer-Assisted Intervention; 2016:246-254. https://doi.org/10.1007/978-3-319-46726-9_29
16. Bai W, Sinclair M, Tarroni G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson*. 2018;20(1):65. https://doi.org/10.1186/s12968-018-0471-x
17. Ma Z, Wu X, Wang X, et al. An iterative multi-path fully convolutional neural network for automatic cardiac segmentation in cine MR images. *Med Phys*. 2019;46(12):5652-5665. https://doi.org/10.1002/mp.13859
18. Zheng Q, Delingette H, Duchateau N, Ayache N. 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE Trans Med Imaging*. 2018;37(9):2137-2148. https://doi.org/10.1109/TMI.2018.2820742
19. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging*. 2016;35(5):1240-1251. https://doi.org/10.1109/TMI.2016.2538465
20. Zhang W, Li R, Deng H, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuroimage*. 2015;108:214-224. https://doi.org/10.1016/j.neuroimage.2014.12.061
21. Aslani S, Dayan M, Storelli L, et al. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *Neuroimage*. 2019;196:1-15. https://doi.org/10.1016/j.neuroimage.2019.03.068
22. Patravali J, Jain S, Chilamkurthy S. 2D-3D fully convolutional neural networks for cardiac MR segmentation. STACOM; 2017:130-139. https://doi.org/10.1007/978-3-319-75541-0_14
23. Isensee F, Jaeger P, Full PM, Wolf I, Engelhardt S, Maier-Hein KH. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. SATCOM; 2017:120-129. https://doi.org/10.1007/978-3-319-75541-0_13
24. Baumgartner CF, Koch LM, Pollefeys M, Konukoglu E. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. SATCOM; 2017:111-119. https://doi.org/10.1007/978-3-319-75541-0_12
25. Bernard O, Lalande A, Zotti C, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans Med Imaging*. 2018;37(11):2514-2525. https://doi.org/10.1109/TMI.2018.2837502
26. Mortazi A, Karim R, Rhode K, Burt J, Bagci U. CardiacNET: segmentation of left atrium and proximal pulmonary veins from MRI using multi-view CNN. Medical Image Computing and Computer-Assisted Intervention; 2017:377-385. https://doi.org/10.1007/978-3-319-66185-8_43
27. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y. 3D brain tumor segmentation through integrating multiple 2D FCNNs. Medical Image Computing and Computer-Assisted Intervention; 2018:191-203. https://doi.org/10.1007/978-3-319-75238-9_17
28. Xia Y, Xie L, Liu F, Zhu Z, Fishman EK, Yuille AL. Bridging the gap between 2D and 3D organ segmentation with volumetric fusion net. Medical Image Computing and Computer-Assisted Intervention; 2018:445-453. https://doi.org/10.1007/978-3-030-00937-3_51
29. Poudel RPK, Lamata P, Montana G. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In RAMBO; 2016:83-94. https://doi.org/10.1007/978-3-319-52280-7_8
30. Avendi MR, Kheradvar A, Jafarkhani H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med Image Anal*. 2016;30:108-119. https://doi.org/10.1016/j.media.2016.01.005
31. Ngo TA, Lu Z, Carneiro G. 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*. 2017;35:159-171. https://doi.org/10.1016/j.media.2016.05.009
32. Hu H, Pan N, Wang J, Yin T, Ye R. Automatic segmentation of left ventricle from cardiac MRI via deep learning and region constrained dynamic programming. *Neurocomputing*. 2019;347:139-148. https://doi.org/10.1016/j.neucom.2019.02.008
33. Khened M, Kollerathu VA, Krishnamurthi G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Med Image Anal*. 2019;51:21-45. https://doi.org/10.1016/j.media.2018.10.004
34. Abdeltawab H, Khalifa F, Taher F, et al. A deep learning-based approach for automatic segmentation and quantification of the left ventricle from cardiac cine MR images. *Comput Med Imaging Graph*. 2020;81:101717. https://doi.org/10.1016/j.compmedimag.2020.101717
35. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. https://doi.org/10.1038/s41592-020-01008-z
36. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. arXiv preprint, 2017. http://arxiv.org/abs/1710.09829
37. Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: The missing ingredient for fast stylization. arXiv preprint, 2016. http://arxiv.org/abs/1607.08022
38. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint, 2015. http://arxiv.org/abs/1502.03167
39. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv preprint, 2015. http://arxiv.org/abs/1511.07122
40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv preprint, 2015. http://arXiv.org/abs/1512.03385
41. Kingma DP, Ba J. Adam. A method for stochastic optimization. arXiv preprint, 2014. http://arxiv.org/abs/1412.6980

42. Zotti C, Luo Z, Lalande A, Jodoin PM. Convolutional neural network with shape prior applied to cardiac MRI segmentation. *IEEE J Biomed Heal Informatics*. 2019;23(3):1119-1128. https://doi.org/10.1109/JBHI.2018.2865450
43. Painchaud N, Skandarani Y, Judge T, Bernard O, Lalande A, Jodoin PM. Cardiac segmentation with strong anatomical guarantees. *IEEE Trans Med Imaging*. 2020;39(11):3703-3713. https://doi.org/10.1109/TMI.2020.3003240
44. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods for clinical measurement. *Lancet*. 1986;327(8476):307-310. https://doi.org/10.1016/S0140-6736(86)90837-8
45. Baldeon Calisto M, Lai-Yuen, SK. AdaEn-Net: An ensemble of adaptive 2D–3D Fully Convolutional Networks for medical image segmentation. *Neural Networks*; 2020;(126):76-94. https://doi.org/10.1016/j.neunet.2020.03.007
46. Yang X, Bian C, Yu L, Ni D, Heng P-A. Class-Balanced Deep Neural Network for Automatic Ventricular Structure Segmentation. Statistical Atlases and Computational Models of the Heart. *ACDC and MMWHS Challenges*. 2018:152-160. https://doi.org/10.1007/978-3-319-75541-0_16
47. Simantiris G, Tziritas. Cardiac MRI Segmentation With a Dilated CNN Incorporating Domain-Specific Constraints. *IEEE J Sel Top Signal Process*; 2020;14(6):1235-1243. https://doi.org/10.1109/jstsp.2020.3013351

How to cite this article: Ahmad F, Hou W, Xiong J, Xia Z. Fully automated cardiac MRI segmentation using dilated residual network. *Med. Phys*. 2023;50:2162–2175. https://doi.org/10.1002/mp.16108