SAUN: Stack attention U-Net for left ventricle segmentation from cardiac cine magnetic resonance imaging

Xiaowu Sun
Division of Image Processing, Department of Radiology, Leiden University Medical Center, PO Box 9600, Leiden 2300 RC, The Netherlands

Pankaj Garg
Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK

Sven Plein
Department of Biomedical Imaging Science, Leeds Institute of Cardiovascular and Metabolic Medicine University of Leeds, Leeds, UK

Rob J. van der Geest[a)]
Division of Image Processing, Department of Radiology, Leiden University Medical Center, PO Box 9600, Leiden 2300 RC, The Netherlands

(Received 7 September 2020; revised 17 January 2021; accepted for publication 19 January 2021; published 4 March 2021)

Purpose: Quantification of left ventricular (LV) volume, ejection fraction and myocardial mass from multi-slice multi-phase cine MRI requires accurate segmentation of the LV in many images. We propose a stack attention-based convolutional neural network (CNN) approach for fully automatic segmentation from short-axis cine MR images.
Methods: To extract the relevant spatiotemporal image features, we introduce two kinds of stack methods, spatial stack model and temporal stack model, combining the target image with its neighboring images as the input of a CNN. A stack attention mechanism is proposed to weigh neighboring image slices in order to extract the relevant features using the target image as a guide. Based on stack attention and standard U-Net, a novel Stack Attention U-Net (SAUN) is proposed and trained to perform the semantic segmentation task. A loss function combining cross-entropy and Dice is used to train SAUN. The performance of the proposed method was evaluated on an internal and a public dataset using technical metrics including Dice, Hausdorff distance (HD), and mean contour distance (MCD), as well as clinical parameters, including left ventricular ejection fraction (LVEF) and myocardial mass (LVM). In addition, the results of SAUN were compared to previously presented CNN methods, including U-Net and SegNet.
Results: The spatial stack attention model resulted in better segmentation results than the temporal stack model. On the internal dataset comprising of 167 post-myocardial infarction patients and 57 healthy volunteers, our method achieved a mean Dice of 0.91, HD of 3.37 mm, and MCD of 1.08 mm. Evaluation on the publicly available ACDC dataset demonstrated good generalization performance, yielding a Dice of 0.92, HD of 9.4 mm, and MCD of 0.74 mm on end-diastolic images, and a Dice of 0.89, HD of 7.1 mm and MCD of 1.03 mm on end-systolic images. The Pearson correlation coefficient of LVEF and LVM between automatically and manually derived results were higher than 0.98 in both datasets.
Conclusion: We developed a CNN with a stack attention mechanism to automatically segment the LV chamber and myocardium from the multi-slice short-axis cine MRI. The experimental results demonstrate that the proposed approach exceeds existing state-of-the-art segmentation methods and verify its potential clinical applicability. [https://doi.org/10.1002/mp.14752]

Key words: cine MRI, segmentation, spatiotemporal information, stack attention, stack model

1. INTRODUCTION

Due to the excellent image resolution and soft-tissue contrast, cardiac cine magnetic resonance imaging (MRI) is considered the reference standard for quantitative assessment of cardiac size and function.[1,2] Typically, imaging is performed in short-axis orientation, and multiple slices and multiple phases are acquired to image the complete left ventricle (LV) over the cardiac cycle. Quantification requires segmentation of many images. Traditional manual segmentation is labour-

intensive and relies on experienced experts. In recent years, the convolution neural network (CNN)-based approaches have achieved immense success in LV segmentation, and many fully automatic segmentation algorithms based on CNN have been proposed. U-Net[3] and fully convolution network (FCN)[4] are the typical CNN models used in medical image analysis due to their capability of multi-scale feature extraction and fusion. Bai et al.[5] used a training set of 4875 subjects (93500 annotated image slices) to build a basic FCN for segmentation of the LV in short-axis MRI and used a fine-tuning approach to enable segmentation in other datasets. This approach required a massive set of images and also labour-intensive manual annotation effort. Isensee et al.[6] integrated the segmentation and classification task into an ensemble U-Net in which geometrical features extracted from the segmentation results were used for pathology classification. Recently, several unsupervised and self-learning strategies have been proposed, most of these methods use multiple branches to explore additional information and then add these branches to the segmentation backbone.[2] Qin et al.[7] proposed a joint model with two branches: one branch introduced an unsupervised Siamese style spatial transformer network to extract motion features, and the other branch was based on the fully convolutional network for segmentation.

A limitation of previous work is that most of the proposed deep learning methods extract image features from a single 2D image only, which implies that potentially relevant spatiotemporal information that can be derived from neighboring slices and phases is not being exploited.[8] In recent literature, the classical optical flow (OF) method[9-11] has been introduced to extract temporal coherence among neighboring phases. For example, Zhao et al.[10] coupled the OF from the specified resolution scale to explore the motion features. Yan et al.[11] computed the OF features between two neighboring phases and integrated those features into a U-Net. However, the OF adopts an iterative method, which is time-consuming. Recently some other deep learning methods have been proposed to detect motion features. Zhang et al.[12] applied an LSTM model to incorporate local motion information by regarding several neighboring frames as input. Desai et al.[13] constructed a multi-channel architecture by stacking several neighboring frames to detect the spatiotemporal features. However, which architecture and input depth are optimal for LV segmentation performance in cine MRI is not fully explored. Hence, we proposed two image stack models to build a multi-channel architecture. One method is called the spatial stack model, combining the target image which is introduced for the segmentation and its neighboring slices from the same cardiac phase. The other method is called a temporal stack, containing the target image and its neighboring phases at the same slice level. Then a stack attention model is proposed to obtain weighted potential cardiac information from the stack. Traditional local image feature extraction, visual saliency detection, and sliding window methods can all be considered as an attention mechanism. However, in a CNN, the attention module is usually an additional brief neural network that can recognize the important parts from the images or assign different weights to different parts of the input. With the development of deep learning, building a neural network with an attention mechanism has been an active topic of research in computer vision.[14-16] Because a neural network can learn the attention mechanism autonomously, the inclusion of an attention mechanism can help the network to understand the image better. Due to its excellent performance, attention mechanism is currently widely used in many fields such as machine translation, speech recognition, image caption, and computer vision.

To improve the accuracy of LV segmentation, our work mainly focuses on the following aspects:

1. We introduce two stack models (spatial stack and temporal stack) as a quasi-volumetric architecture to extend the depth of the input.
2. We propose a stack attention mechanism in which the target image serves as a guide to weigh the features from multiple channels and select the spatiotemporal information.
3. A novel Stack Attention U-Net (SAUN) based on the stack attention and basic U-Net is proposed for automatic LV segmentation.

2. MATERIALS AND METHODS

Different from natural images, MR images only have a single channel (grayscale) and have more complex texture features. Meanwhile, the shape, size and position of the LV only varies slightly between neighboring slices both in the spatial and temporal domain. To address those deformations and contextual information, we will first illustrate how to construct a volumetric architecture using the spatial stack model and temporal stack model, respectively, and then integrate the features from the stack model with a novel stack attention mechanism. Finally, we propose the SAUN model based on stack attention and basic U-Net for segmentation.

2.A. Stack model

Figure 1 illustrates the construction of a stack in a case having 30 cardiac phases and 12 slices. Spatial stack $SSM = \{S_{15,4}, \cdots, S_{15,7}, \cdots S_{15,10}\}$ and temporal stack $TSM = \{S_{12,7}, \cdots, S_{15,7}, \cdots S_{18,7}\}$ can be used to generate an example image stack of dimension $N = 7$ as the input which produces the segmentation result for the central slice $S_{15,7}$.

2.A.1. Spatial stack model (SSM)

We propose a novel method named spatial stack model to combine the target image with its neighboring spatial slices. The stack model for the central slice $S_{p,t}$ can be described as the following, where $S_{i,j}(i = 1,2,\cdots,T; j = 1,2,\cdots,F)$ represents the image from the ith phase jth slice, T and F is the number of phases and slices in the dataset, respectively, and N is the number of the images in the stack.

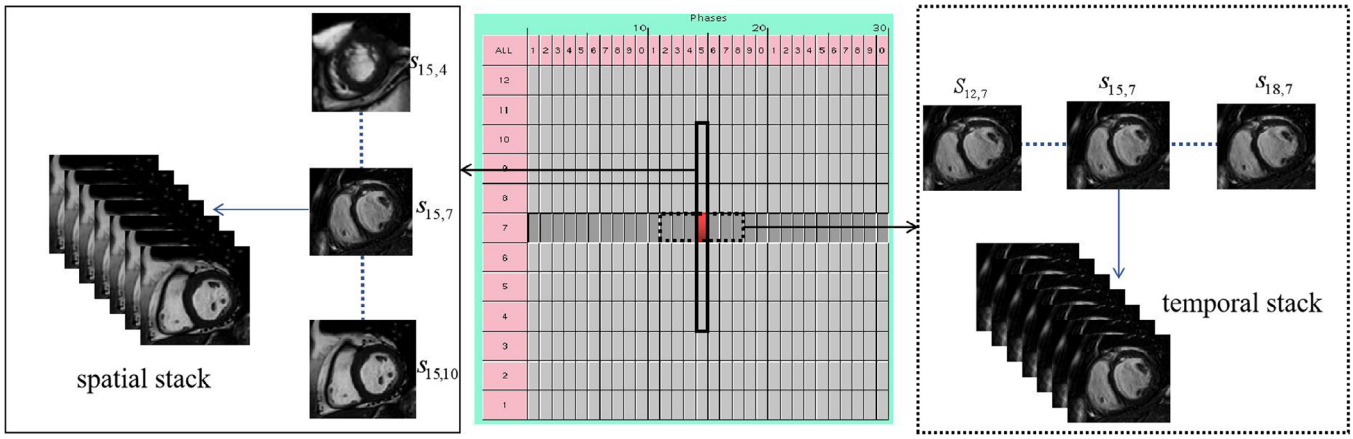

FIG. 1. Example of the construction of a spatial and temporal stack of dimension 7. Slice $S_{15,7}$ is the target slice; spatial stack model uses slices $\{S_{15,4}, S_{15,5}, \cdots, S_{15,9}, S_{15,10}\}$ from the same phase to build the stack model, while temporal stack model introduces slices $\{S_{12,7}, S_{13,7}, \cdots S_{17,7}, S_{18,7}\}$ from the same slice level to construct another kind of stack model. $S_{i,j}$ is the image from the ith phase jth slice. [Color figure can be viewed at wileyonlinelibrary.com]

$$SSM(S_{p,t},N)=\{S_{i,j}|i=p,j=t-(N-1)/2,\cdots,t+(N-1)/2\} \quad and \quad \begin{cases} if \quad j<1, \quad j=1 \\ if \quad j>F, \quad j=F \end{cases}$$

2.A.2. Temporal stack model (TSM)

Similar to the spatial stack, the temporal stack model can be defined as follows.

$$TSM(S_{p,t},N)=\{S_{i,j}|i=p-(N-1)/2,\cdots,p+(N-1)/2,j=t\} \quad and \quad \begin{cases} if \quad i<1, \quad i=i+T \\ if \quad i>T, \quad i=i-T \end{cases}$$

The original MR image is a grayscale image with one channel. The image represented by the stack model can be regarded as a multi-channel image with abundant semantic features. It is important to note that, to our intuition, features derived from images closer (in space or time) to the target image contribute more in segmenting the object in the target slice. Hence, in order to filter out the background noise and extract relevant image information, we further propose the stack attention model.

2.B. Stack attention model

In this part, we introduce the target image as a guide to provide the channel information to fuse the neighboring images into the stack.

In detail, as shown in Fig. 2, we first perform a 3 × 3 convolution with ReLu non-linearity function on the feature maps from the central slice to ensure the number of the feature maps generated from central slice and stack is the same. Then the global spatial information is extracted and squeezed to a vector $P=(p_1,p_2,\cdots,p_C)$ through the global average pooling, which can be described as the following

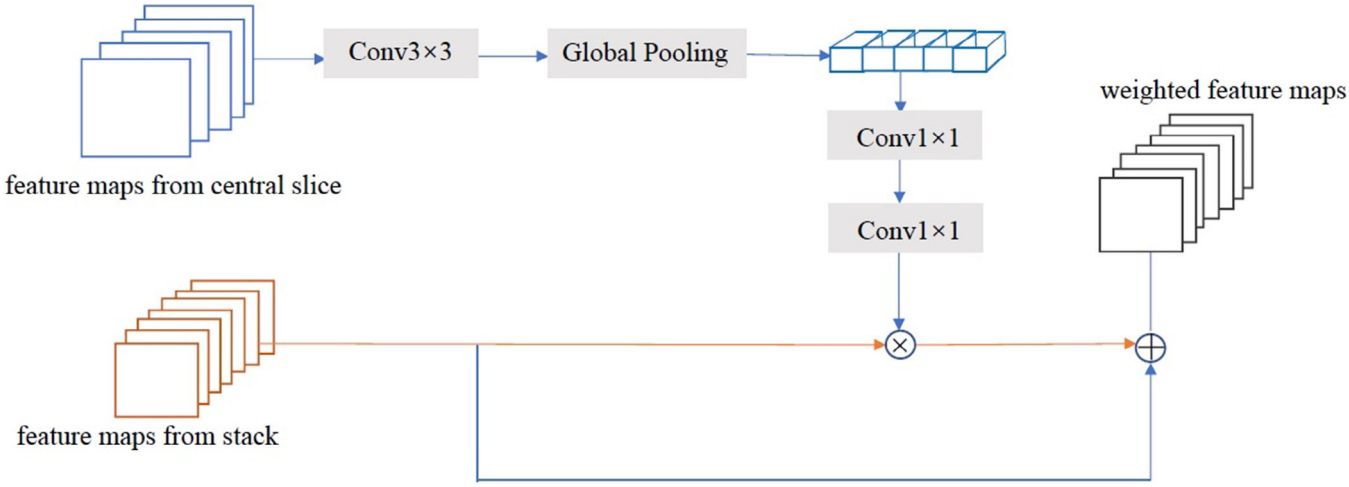

FIG. 2. Fig.Stack attention module structure.

equation where $W \times L$ is the size of the feature map, f_c is the feature map of the cth channel and C is the number of channels which is equal to the number of kernels in the convolution layer.

$$p_c = \frac{1}{W \times L} \sum_{i=1}^{W} \sum_{j=1}^{L} f_c(i,j) \ (c = 1,2,\cdots,C).$$

Two different 1×1 convolutions K_1 and K_2 are applied to further compute the weights of each channel as follows:

$$s' = \sigma(\sigma(P * K_1) * K_2) \cdot s.$$

where $*$ is the convolution operation, σ is ReLu activation function and s is the feature map generated from the stack model. The first convolution K_1 reduces the dimension of vector P from C to $C/2$, and then convolution K_2 resizes the length of vector P into C again. However, the dot production with the weights which range from 0 to 1 repeatedly will degrade the feature values in deep layers, which may lead to negative results. To avoid this problem, finally the weighted stack feature maps are added with the original stack feature maps, which means

$$attS_c(i,j) = (1 + P_c) f_c(i,j) \quad (c = 1,2,\cdots C)$$

where $attS_c$ is the cth channel of the attention stack. When P_c approaches to 0, $attS_c(i,j)$ will approximate to the original features.

2.C. SAUN network architecture

Based on the mentioned stack attention and traditional U-Net, we propose the SAUN for the segmentation task. As shown in Fig. 3, there are two inputs in SAUN, one is the central slice which is the target, and the other one is called the initial stack (either spatial or temporal stack) which is constructed according to $Stack(S,N)$ proposed above. To ensure that the central slice and the stack are at the same feature level, the convolution operation is applied to both of them at the same time.

During training SAUN, we aim to optimize the following loss function, which contains the generalized Dice loss and cross-entropy loss. The loss function can be formulated as.

$$loss = 1 - 2\frac{\sum_{i=1}^{l} w_i \sum_{j=1}^{n} g_{ij} p_{ij}}{\sum_{i=1}^{l} w_i \sum_{j=1}^{n} g_{ij} + p_{ij}} - \sum_{i=1}^{l} \sum_{j=1}^{n} g_{ij} \log(p_{ij})$$

where the second term is the weighted Dice loss for multiple cardiac structure segmentation, and the third term is cross-entropy loss based on pixel-wise classification. Parameters g,p stand for ground truth and prediction results, respectively, l denotes three labels (background, chamber, and myocardium), n is the number of the pixels and w_i is the weight of each label, which were set to $w = [0.1, 0.2, 0.7]$.

2.D. Datasets

2.D.1. Leeds university dataset (LUD)

One of the datasets in this work is from the University of Leeds, UK. This dataset contains 168 post-myocardial infarction patients and 57 healthy volunteers. All subjects were scanned on a Philips Ingenia 1.5T MRI system using a slice thickness of 5.0 mm (or sometimes 8.0 mm) and slice gap of 2 mm. The number of slices ranged from 10 to 20, and 30 phases were reconstructed to cover a complete cardiac cycle. The in-plane image resolution varied from 0.78×0.78 mm^2 to 1.18×1.18 mm^2 and the range of field of view (FOV) from 280×280 mm^2 to 470×470 mm^2. Expert annotations were derived semi-automatically in all cardiac phases and slices by one observer (RG) with 20 yr of experience in cardiac MRI using Mass software (Version V2017-EXP; Leiden University Medical Center, Leiden, the Netherlands), resulting in 6703 annotated images. The subjects' exams were randomly split into three parts with 141, 15, and 69 for training, validation and test, respectively.

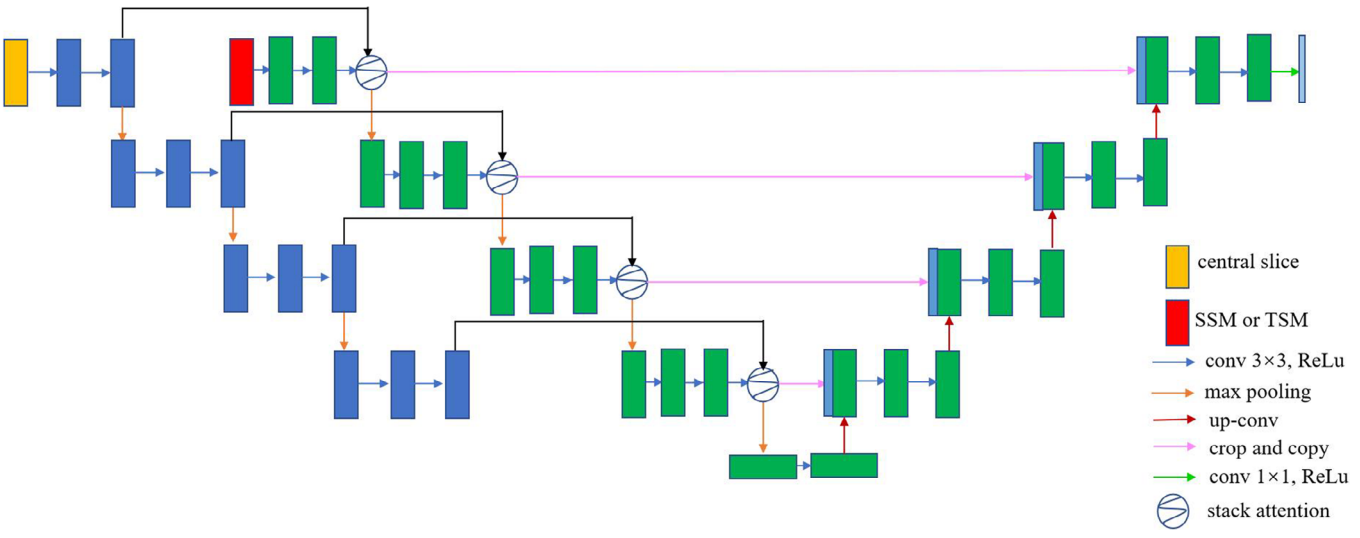

FIG. 3. Segmentation model structure based on Stack Attention and U-Net (SAUN). [Color figure can be viewed at wileyonlinelibrary.com]

2.D.2. MICCAI 2017 Automated cardiac diagnosis challenge (ACDC 2017)

The MICCAI 2017 Automated Cardiac Diagnosis Challenge (ACDC 2017) was organized by the University Hospital of Dijon and the data used in this challenge has become publicly available.[17] The dataset contains short-axis cine-MRI exams of 100 subjects of five patient categories (post-myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, abnormal right ventricle, and healthy subjects). The subjects were scanned on two different scanners (1.5T Siemens Area and 3.0T Siemens Trio Tim) using a typical slice thickness of 5.0 mm (range 5–8 mm), an interslice gap of 5 mm (range 5–10 mm) and pixel spacing ranging from 1.37 to 1.68 mm. For all exams, the manual ground truth annotation was generated by a single clinical expert including contours of the LV cavity and myocardium and the right ventricular cavity in the end-diastolic (ED) and end-systolic (ES) images. In this work, the annotation of the right ventricular cavity was ignored and considered as background in the ground truth. The 100 subjects were randomly divided into five folders, each folder containing five patient categories and each category containing four subjects. We randomly selected three folders to train the network, and the other two folders were chosen for validation and test, respectively.

2.E. Data preprocessing and augmentation

Within the available dataset, the images vary in intensity range, FOV and pixel spacing. The field of view in the LUD data varies from 280 to 470 mm, while the heart as the object of interest typically measures 60 mm, occupying only a small proportion of the whole image. For example, in our LUD dataset, the average proportion occupied by the object relative to the full image is around 2.2%. Hence, several image preprocessing methods were performed to standardize those parameters.

We firstly resample the original images into a common pixel spacing of 1.5 mm, and then the image intensities were normalized according to the following formula where $P_{\min}$ and $P_{\max}$ is the minimum and maximum value of 5% and 95% percentile of image P.

$$p = \frac{P_i - P_{\min}}{P_{\max} - P_{\min}}$$

To solve the label imbalance problem, the YOLO model[18] is applied to localize the region-of-interest (ROI). As illustrated in Fig. 4 each 2D original image is considered as an input, and then YOLO extracts the features from the input to generate the bounding boxes. Lastly, the images are cropped or zero-padded to a uniform matrix size of 128 × 128, centered at each bounding box. Additionally, in order to train a well generalizing network with limited data, data augmentation was employed, including horizontal and vertical clip, image transpose and elastic deformation.

2.F. Evaluation metrics

For quantitative assessment, two aspects, including segmentation and clinical parameter estimation, are proposed to compare the performance among different segmentation methods. All metrics are evaluated on a per-patient basis.

2.F.1. Segmentation accuracy assessment metrics

Dice is introduced to evaluate the overlap between the automatic and manual segmentation mask. In addition, the distance metrics, including Mean Contour Distance (MCD) and Hausdorff Distance (HD) are employed as the segmentation metrics.

MCD and HD are defined as:

$$MCD = \frac{1}{2|C_A|}\sum_{p\in C_A} d(p, C_B) + \frac{1}{2|C_B|}\sum_{q\in C_B} d(q, C_A)$$

$$HD = \max\left(\max_{p\in C_A} d(p, C_B), \max_{q\in C_B} d(q, C_A)\right).$$

where C_A and C_B are the automatic contour and manual contour, respectively, $d(p, C) = \min_{q\in C} d(p, q)$ denotes the minimum distance from point p to contour C.

2.F.2. Clinical metrics

Clinical parameters such as LV volume, LV ejection fraction (LVEF), and myocardial mass (LVM) are another essential aspect of assessing the quality of automatic segmentation. The volume is computed by summation of the number of pixels corresponding to the LV or myocardium binary mask, multiplied by the pixel dimension. Myocardial mass is calculated by the following formula:

$$LVM = Myo - Volume \times 1.05(gram/cm^3).$$

and LVEF is defined as:

$$LVEF = \frac{EDV - ESV}{EDV} \times 100\%$$

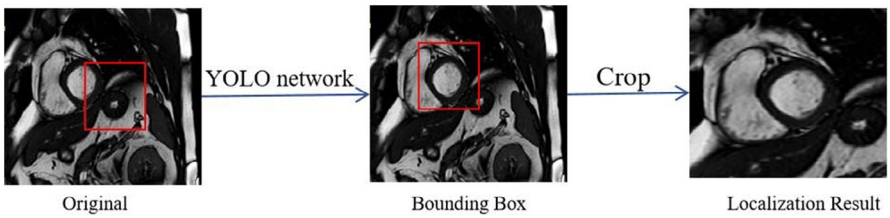

FIG. 4. An example of localization preprocess. In the image, at the left, the red box is at the center of the image initially, but it didn't detect the heart accurately, but after applying the YOLO model, the position of the object can be extracted precisely. Lastly, it is cropped into a fixed size, centered at the red bounding box. [Color figure can be viewed at wileyonlinelibrary.com]

where EDV and ESV are the LV volumes at the end-diastolic and end-systolic phases, respectively.

2.F.3. Statistical analysis

Pearson correlation coefficient (PCC), mean of differences (Bias) and limits of agreement (LOA, 1.96 × standard deviation) are assessed to describe the differences and the agreement between automatically and manually derived segmentation. In addition, Bland-Altman is used to further describe the results.

To investigate the statistical significance of the differences between different segmentation models, the Wilcoxon signed-rank test is used to compare the difference between paired Dice, HD and MCD without assuming the underlying distribution, $P < 0.05$ indicates a significant difference.

3. EXPERIMENTS AND RESULTS

We trained and evaluated our method on both LUD and ACDC datasets. The network is firstly trained on LUD from scratch, and then we performed transfer learning to train the network on ACDC. All the experiments were executed on a machine equipped with an NVIDIA Quadro RTX 6000 GPU with 24 GB internal memory. The networks were implemented using Keras with the following parameters: Adam optimizer, batch size as 50, learning rate as 10^{-5}, 150 epochs, as well as early stopping, to avoid overtraining the network.

First, we explored and determined the optimal value of parameter *N* in the spatial and temporal stack. Second, we compared the results of three classical segmentation networks, U-Net, SegNet,[19] and 3D U-Net, with SAUN based on Dice, MCD, HD, LVEF and LVM on both LUD and ACDC datasets. Meanwhile, to further explore the impact of using YOLO for localization and spatial stack for extracting potential features on the segmentation performance, another two networks named YOLO + U-Net (YUN) and SSM + U-Net (SUN) were employed. The cropped images with a uniform matrix size of 128 × 128, centered to the original image, were used as the input of U-Net and SegNet. The input of YUN is presented after localization, and input of SUN and SAUN are preprocessed with localization and SSM. For the input of 3D U-Net, for both datasets, all the 2D slices in the ED or ES phase together are stacked to construct a 3D image. Then, all 3D images were resampled into the same resolution of 2.5 × 2.5 × 5 mm^3 and the signal intensity normalized to (0,1). Lastly, all 3D images were cropped or padded to a size of 112 × 112 × 24 as the input of the 3D U-Net. For the post-processing, the predictions were resampled to their original resolution. All of the networks are assessed using the defined evaluation metrics for different levels of the LV, including apex (25% slices in the apical region and beyond), middle (50% mid slices) and base (25% slices in the basal region and beyond). All of the best performance in the tables are shown in bold case.

TABLE I. Dice of segmentation results generated from different multi-channel architectures with various values of parameter n at lud using saun method. N is the dimension parameter.

	Chamber		Myocardium	
Parameters	SSM	TSM	SSM	TSM
N = 3	**0.95 (0.05)**	**0.93 (0.07)**	**0.86 (0.07)**	**0.84 (0.11)**
N = 5	0.93 (0.11)	0.92 (0.01)	0.84 (0.14)	0.83 (0.13)
N = 7	0.93 (0.12)	0.92 (0.10)	0.84 (0.13)	0.81 (0.14)
N = 9	0.92 (0.13)	0.92 (0.11)	0.82 (0.13)	0.82 (0.12)

3.A. Multi-Channel Architecture

To analyze the impact of the two multi-channel architectures (SSM and TSM) of different dimensions on the segmentation results, we trained SAUN using SSM and TSM with different dimension parameter N as input. The results presented in Table I illustrate the segmentation performance for LV chamber and myocardium.

Results of multi-channel architecture showed four TSM versions (N = 3,5,7,9) achieved stable segmentation performance for LV chamber and myocardium with the best Dice of 0.93 and 0.84, respectively. SSM, however, did work significantly better than TSM with best performance Dice of 0.95 and 0.86 for chamber and myocardium with N set to 3. Hence, SSM with dimension N = 3 is regarded as the optimal input of SAUN.

3.B. Results on LUD

The performance of the SAUN method was evaluated in the LUD testing dataset (69 subjects, 1611 2D images). We

TABLE II. Comparison of the mean and standard deviation (in parenthesis) of Dice metric on LUD for LV chamber and LV myocardium predicted by different networks. (1)U-Net:basic U-Net without localization, (2)YUN: combine YOLO for localization and basic U-Net, (3)SUN: SSM with N = 3 as the input of basic U-Net, (4)SegNet: basic SegNet without localization, (5) SAUN: SSM with N = 3 as the input of proposed SAUN network.

	Apex		Middle		Base		Average	
Networks	chamber	myocardium	chamber	myocardium	chamber	myocardium	chamber	myocardium
U-Net	0.821 (0.210)	0.692 (0.220)	0.939 (0.040)	0.817 (0.086)	0.924 (0.067)	0.800 (0.105)	0.922 (0.120)	0.799 (0.140)
YUN	0.897 (0.100)	0.794 (0.110)	0.945 (0.036)	0.840 (0.069)	0.909 (0.066)	0.793 (0.110)	0.932 (0.066)	0.825 (0.096)
SUN	0.849 (0.180)	0.752 (0.170)	0.949 (0.035)	0.867 (0.058)	0.938 (0.056)	0.839 (0.096)	0.935 (0.100)	0.848 (0.110)
SegNet	0.794 (0.200)	0.654 (0.200)	0.924 (0.039)	0.812 (0.073)	0.919 (0.062)	0.786 (0.107)	0.908 (0.110)	0.788 (0.130)
SAUN	**0.911 (0.080)**	**0.823 (0.080)**	**0.952 (0.034)**	**0.876 (0.042)**	**0.941 (0.046)**	**0.847 (0.069)**	**0.945 (0.053)**	**0.864 (0.066)**

TABLE III. Comparison of the mean and standard deviation (in parenthesis) of HD and MCD metrics on LUD dataset for LV chamber and LV myocardium at apex, middle and base regions predicted by different networks. (1) U-Net:basic U-Net without localization, (2) YUN: combine YOLO for localization and basic U-Net, (3) SUN: SSM with N = 3 as the input of basic U-Net, (4) SegNet: basic SegNet without localization, (5) SAUN: SSM with N = 3 as the input of proposed SAUN network.

	Apex				Middle				Base			
	Chamber		Myocardium		Chamber		Myocardium		Chamber		Myocardium	
Networks	HD	MCD	HD	MCD	HD	MCD	HD	MCD	HD	MCD	HD	MCD
U-Net	5.373 (5.99)	1.584 (1.88)	6.969 (7.68)	1.706 (1.69)	4.785 (3.85)	1.157 (0.52)	5.372 (3.97)	1.311 (0.49)	5.467 (4.28)	1.496 (0.94)	6.782 (5.48)	1.434 (0.72)
YUN	3.677 (3.48)	1.150 (0.51)	4.250 (3.57)	1.228 (0.49)	4.072 (3.13)	1.127 (0.57)	4.382 (2.37)	1.205 (0.45)	5.125 (3.89)	1.528 (0.95)	6.556 (5.28)	1.449 (0.78)
SUN	3.491 (3.11)	1.309 (1.13)	4.254 (3.51)	1.336 (0.91)	2.994 (1.73)	1.028 (0.55)	3.302 (1.46)	1.009 (0.41)	4.254 (3.43)	1.303 (1.01)	5.763 (4.85)	1.187 (0.64)
SegNet	4.978 (3.18)	1.808 (2.98)	6.108 (2.98)	1.785 (0.99)	4.260 (1.46)	1.517 (0.69)	4.815 (1.52)	1.383 (0.36)	5.197 (3.22)	1.639 (0.84)	6.864 (4.59)	1.541 (0.55)
SAUN	**2.913 (2.52)**	**0.942 (0.46)**	**3.817 (3.58)**	**1.067 (0.41)**	**2.796 (1.41)**	**0.978 (0.48)**	**3.185 (1.29)**	**0.950 (0.36)**	**4.032 (2.85)**	**1.234 (0.66)**	**5.367 (4.11)**	**1.148 (0.47)**

TABLE IV. Results of clinical evaluation metrics from all networks against the reference. (1) U-Net:basic U-Net without localization, (2) YUN: combine YOLO for localization and basic U-Net, (3) SUN: SSM with N = 3 as the input of basic U-Net, (4) SegNet: basic SegNet without localization, (5) SAUN: SSM with N = 3 as the input of proposed SAUN network.

	LVEF		LVM	
Networks	PCC (%)	Bias ± LOA (%)	PCC (%)	Bias ± LOA (g)
U-Net	0.969	2.22 ± 5.89	0.957	3.11 ± 28.60
YUN	0.972	0.52 ± 5.76	0.971	1.51 ± 20.29
SUN	0.974	0.22 ± 5.36	0.976	1.47 ± 19.21
SegNet	0.967	3.34 ± 6.01	0.954	4.51 ± 31.57
SAUN	**0.982**	**0.11 ± 4.62**	**0.985**	**0.58 ± 14.24**

compared the segmentation performance for different heart structures between multiple neural networks using the evaluation metrics defined. As the cross-sectional area of the left ventricle at the apical level is very small and the image quality at this level is degraded due to particle voluming, segmentation errors are more likely to occur at this level, although it will have only little effect on the clinical metrics, especially on the LVEF. Hence, we further evaluated the segmentation performance on apex, middle and base level, respectively. Finally, we report the results of the clinical functional parameters.

Tables II and III, respectively, show the Dice and distance metrics (HD and MCD) comparing manual with automatic segmentation. It can be observed that the networks with localization perform better than those without localization, which confirms that localization can filter out the data noise effectively for the label unbalanced data. Moreover, the SAUN method achieved the best segmentation results compared to the other networks on Dice, HD, and MCD. The results for the individual LV levels further indicate that the SAUN model provides much more precise feature maps, leading to the best evaluation metric scores for both LV chamber and myocardium at all LV levels.

The PCC, bias and LOA of the clinical evaluation metrics comparing automated segmentation results with results from manual segmentation are reported in Table IV and Fig. 5. For both LVEF and LVM assessment, the proposed SAUN network achieves the highest PCC, the smallest bias and LOA. Table V summarizes the significance test results between SAUN and the other state-of-the-art methods on LUD, all the *P*-values are smaller than 0.05, which confirms the significantly better results of SAUN compared to the other methods.

Figure 6 illustrates examples of segmentation results obtained by automated SAUN method and conventional manual method from randomly selected cases from the test data. It shows that the automated results are highly similar to the manual reference at both ED and ES phases.

3.C. Results on ACDC

We also compared our method with other approaches on the public ACDC 2017 dataset, which includes short-axis

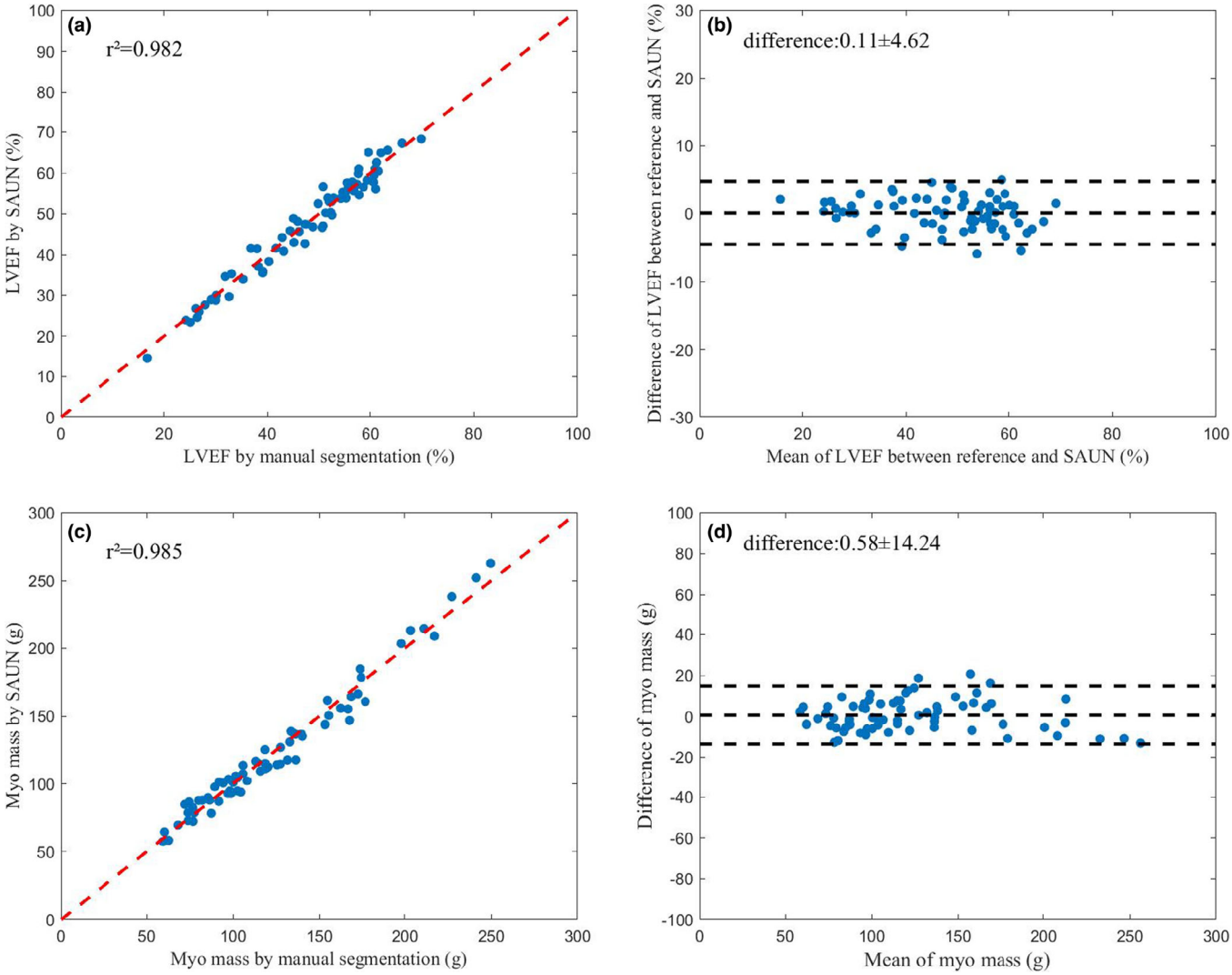

FIG. 5. Correlation and Bland-Altman plots comparing left ventricular (LV) ejection fraction [Figs. 5(a) and 5(b)] and LV mass [Figs. 5(c) and 5(d)] derived from either the SAUN method and manual segmentation on LUD. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE V. Wilcoxon signed-rank test-based significance test results on LUD dataset. (1)W(SAUN,U-Net): Wilcoxon signed-rank test's *P*-value between SAUN and U-Net, (2)W(SAUN,YUN): Wilcoxon signed-rank test's *P*-value between SAUN and YUN(YOLO + U-Net), (3)W(SAUN,SUN): Wilcoxon signed-rank test's *P*-value between SAUN and SUN(SSM stack + U-Net), (4)W(SAUN, SegNet):Wilcoxon signed-rank test's *P*-value between SAUN and SegNet.

	Chamber			Myocardium		
	Dice	HD	MCD	Dice	HD	MCD
W(SAUN,U-Net)	1.36E-05	7.09E-12	0.0129	2.86E-09	1.55E-12	6.22E-04
W(SAUN,YUN)	5.21E-10	5.27E-09	1.37E-08	2.09E-12	1.45E-11	3.00E-10
W(SAUN,SUN)	1.81E-10	6.70E-03	4.87E-07	4.10E-09	2.56E-04	3.24E-07
W(SAUN, SegNet)	1.07E-08	1.36E-12	6.56E-07	1.29E-11	7.99E-13	1.19E-06

1.36E-05 means 1.36×10^{-5}.

Cine MR exams of 100 patients with manual contours. As in this dataset, manual contours are only defined in the ED and ES phases, all results are based on those two phases only.

Table VI summarizes the segmentation results for the ACDC dataset. The best segmentation results on both ED and ES phases are obtained using the SAUN method. In Table VII and Fig. 7, the PCC, bias and LOA are presented and illustrated for the comparison of the clinical parameters. It shows that the prediction results are highly correlated to the reference with a PCC of 0.985 for LVEF and 0.981 for LVM. The Bland-Altman analysis illustrated in Fig. 7 reveals a bias for LVEF and LVM, which is close to zero, while the LOA is

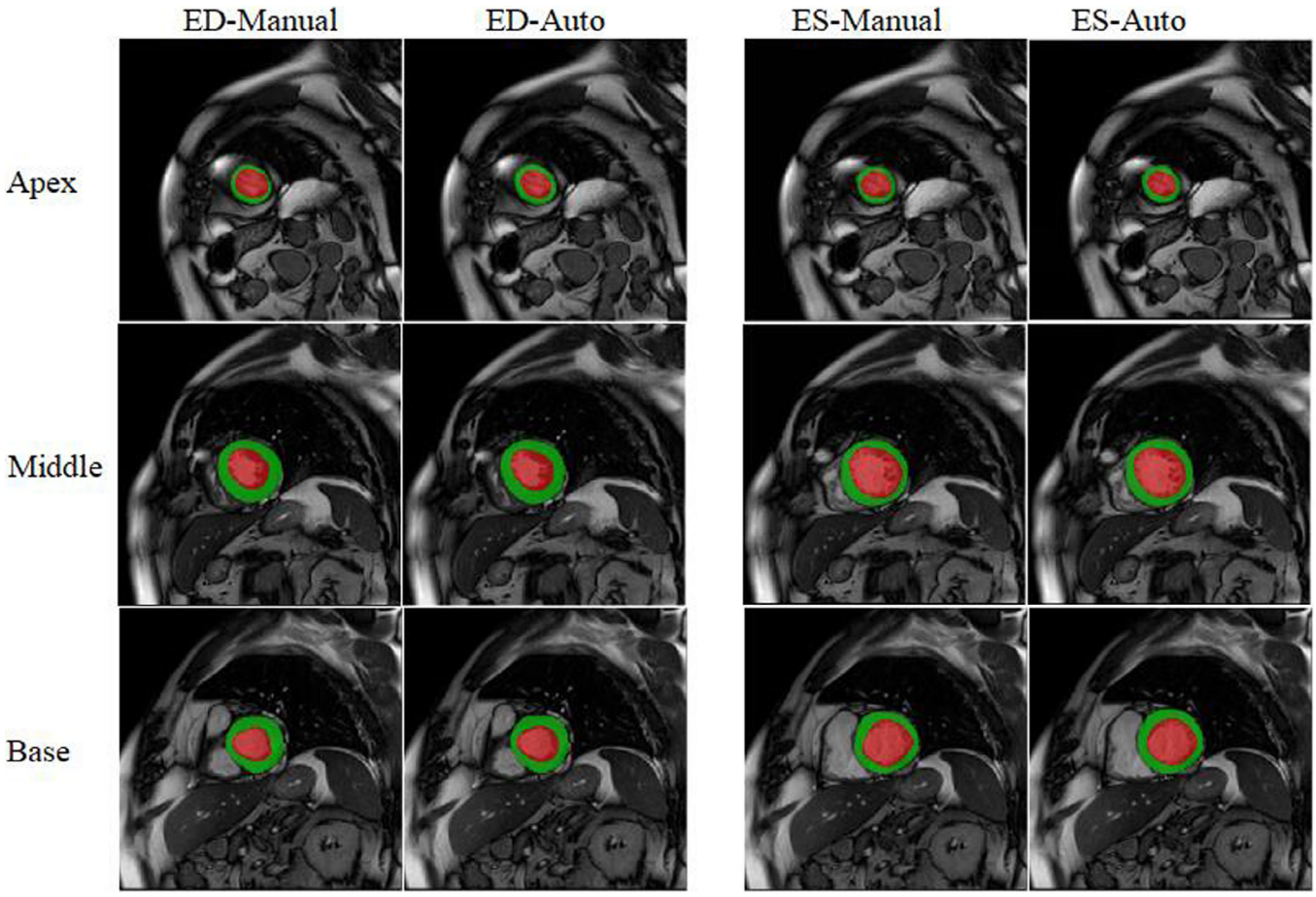

FIG. 6. Examples of the segmentation results from the SAUN method. The left two columns show end-diastolic images, and the right two columns show images of end-systolic phase. For each phase, images at the apex, middle and base levels are shown. [Color figure can be viewed at wileyonlinelibrary.com]

less than 5% for LVEF and less than 6 g for LVM. Table VIII reports almost all the *P*-values between SAUN and U-Net, SegNet and 3D U-Net on ACDC dataset are smaller than 0.05, which confirms there is a significant improvement of SAUN compared to the other state-of-the-art methods. Figure 8 shows the example segmentation results of two randomly selected cases from the test set.

4. DISCUSSION

To explore more spatiotemporal information for automatic cine MRI segmentation, we proposed two stack models to construct a multi-channel architecture, then introduced a segmentation network based on a stack attention mechanism to weight the feature maps from different channels. The method was evaluated on an internal and a public dataset demonstrating competitive results compared other typical CNN networks.

4.A. Multi-channel architecture comparison

Our results demonstrate that, when the spatial stack was used to combine the target slice and its neighboring slices from the same phase together as the input of the network, the performance improved in the test data. The segmentation results were found to be sensitive to the dimension of the spatial stack model. For both spatial and temporal stack the optimal value for the dimension parameter N was found to be 3. However, the use of temporal stack had a negligible impact on the cardiac segmentation results. It also can be observed that all of the evaluation metrics from the spatial stack and SAUN are much better than those predicted from basic U-Net and SegNet whose input is a single 2D image, which illustrates the spatial stack model can provide more useful information than a single MRI slice and temporal stack. The images in the temporal stack are similar to each other and provided comparable features for the network. Whereas, the images from the spatial stack vary obviously with the heart region, and when combining the target slice and its neighboring spatial slices together as the input of the network, the spatial stack contains more information about position, size and shape of the heart. However, including more slices in the stack does not necessarily result in better segmentation results. This was clearly demonstrated by the multi-channel architecture comparison experiment, which showed that when the parameter N was set to a value higher than 3, which implies introducing more spatiotemporal information, the performance degraded. In addition, the limited difference between the stack network and 2D network is only at the first convolution layer. The stack network regards a ($W \times L \times N$)

TABLE VI. Comparison of the mean and standard deviation (in parenthesis) of segmentation results on ACDC dataset for LV chamber and LV myocardium segmentation by different networks. (1) U-Net:basic U-Net without localization, (2) YUN: combine YOLO for localization and basic U-Net, (3) SUN: SSM with N = 3 as the input of basic U-Net, (4) SegNet: basic SegNet without localization, (5)3D U-Net: basic 3D U-Net without localization (6) SAUN: SSM with N = 3 as the input of proposed SAUN network.

	ED						ES					
	Chamber			Myocardium			Chamber			Myocardium		
Networks	Dice	HD	MCD	Dice	HD	MCD	Dice	HD	MCD	Dice	HD	MCD
U-Net	0.940 (0.051)	10.780 (7.29)	0.596 (0.39)	0.836 (0.089)	11.538 (6.72)	0.791 (0.47)	0.841 (0.074)	11.854 (7.94)	1.529 (0.96)	0.812 (0.083)	13.338 (11.12)	1.358 (1.15)
YUN	0.942 (0.067)	10.163 (7.06)	0.571 (0.45)	0.847 (0.057)	11.144 (6.78)	0.714 (0.31)	0.861 (0.076)	11.255 (6.47)	1.057 (0.76)	0.838 (0.061)	12.478 (6.18)	1.206 (0.78)
SUN	0.941 (0.058)	11.347 (4.97)	0.614 (1.33)	0.824 (0.130)	11.975 (7.88)	0.746 (0.84)	0.841 (0.069)	13.259 (5.91)	1.618 (0.98)	0.808 (0.096)	13.750 (7.31)	1.724 (0.69)
SegNet	0.932 (0.056)	10.240 (6.11)	0.577 (0.58)	0.833 (0.081)	11.241 (7.74)	0.757 (0.64)	0.839 (0.084)	11.311 (6.32)	1.088 (0.86)	0.798 (0.092)	12.673 (8.68)	1.188 (0.81)
3D U-Net	0.935 (0.048)	10.634 (7.42)	0.645 (1.67)	0.811 (0.087)	12.217 (6.37)	0.849 (1.32)	0.847 (0.089)	11.281 (7.71)	1.106 (1.94)	0.792 (0.067)	13.681 (9.69)	1.922 (1.06)
SAUN	**0.956 (0.031)**	**9.759 (3.45)**	**0.541 (0.28)**	**0.877 (0.064)**	**10.192 (4.37)**	**0.672 (0.189)**	**0.887 (0.061)**	**10.132 (5.35)**	**1.024 (0.52)**	**0.873 (0.058)**	**11.711 (7.36)**	**0.939 (0.62)**

TABLE VII. Results of clinical evaluation metrics from all networks against the reference. (1) U-Net:basic U-Net without localization, (2) YUN: combine YOLO for localization and basic U-Net, (3) SUN: SSM with N = 3 as the input of basic U-Net, (4) SegNet: basic SegNet without localization, (5)3D U-Net: basic 3D U-Net without localization (6) SAUN: SSM with N = 3 as the input of proposed SAUN network.

	LVEF		Myo Mass	
Networks	PCC	Bias ± LOA	PCC	Bias ± LOA
U-Net	0.956	7.40 (12.06)	0.965	6.06 (9.51)
YUN	0.972	**2.04** (15.64)	0.974	2.60 (7.08)
SUN	0.962	0.29 (11.09)	0.953	−2.69 (10.56)
SegNet	0.947	8.17 (11.59)	0.958	9.26 (10.97)
3D U-Net	0.948	9.51 (17.61)	0.963	5.37 (8.96)
SAUN	**0.985**	2.39 (**4.61**)	**0.981**	**-0.42** (**5.74**)

volume as an input, and the 2D network accepts a single slice ($W \times L \times 1$) as an input, while the other parts of the network are the same, which leads to stack networks having more parameters only at the first convolution layer.

Unlike the stack model transferring stack input into multiple 2D feature maps, Çiçek[20] and Perslev[21] proposed a 3D network for the segmentation. A 3D network will introduce more parameters to extracts the depth-wise features through the entire network than 2D and stack networks. During the training process, in order to fit the scan volume in memory, Arjun[13] set the batch size to 1, resulting in less stable feature regularization. Other researchers set the number of filters at the initial convolution layers into a low value to reduce the number of parameters, but a lower number of filters will likely contribute to inferior feature representation and in turn cause less accurate segmentation. Another disadvantage of repeated pooling and convolution operation is the loss in spatial information in cases with fewer slices. Whereas, the spatial stack network can maintain the spatial information and keep the approximately same number of parameters as a 2D network resulting in improved segmentation performance.

4.B. Effect of stack attention

SUN achieved better performance than YUN in LUD; however, in ACDC SUN did not get as good results as YUN. Because the slice gap (5mm or 10mm) in ACDC is larger than in LUD (2mm), the recognizable variance between the spatially neighboring images, such as the shape, size or outline of LV is larger in ACDC, which will confuse the network. Meanwhile, when we combine neighboring slices having imbalanced labels to build the spatial stack, the proportion of the background will increase, compared with a single 2D image, which results in the spatial stack generating more data noise. This issue is overcome by employing the proposed attention mechanism which weighs and fuses the feature maps of different channels from the spatial stack and balances the noise.

During fusion of the features, the features from the target slice should be regarded as the primary components,

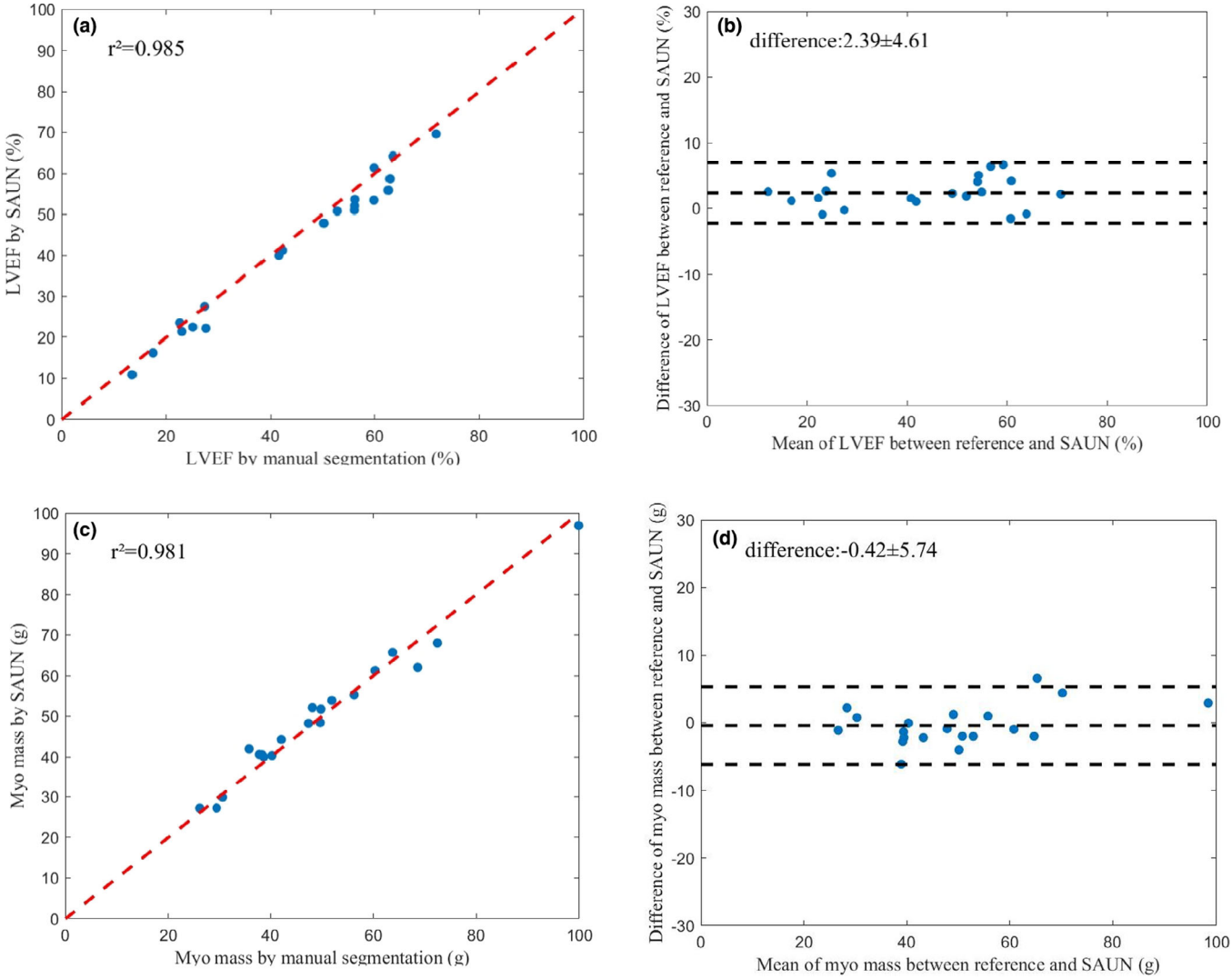

FIG. 7. Correlation and Bland-Altman plots comparing left ventricular (LV) ejection fraction [Figs. 7(a) and 7(b)] and LV mass [Figs. 7(c) and 7(d)] derived from either the SAUN method and manual segmentation ACDC dataset. [Color figure can be viewed at wileyonlinelibrary.com]

and the others from the neighboring slices should be considered as the additional information. In the stack attention mechanism, the target slice serves as a guideline to keep the primary features, and the global pooling is used to compute the weights of different channels to select the feature maps generated from the target slice. Therefore, the stack attention can not only reserve the primary feature information but also balance the importance of different channels to pick up the more important maps. Figure 9 illustrates the process of SAUN method extracting the feature maps from a random sample taken from the LUD dataset. The first row illustrates the features for the LV chamber, and the second row is the features for the myocardium. The first column is one test case, the last column is the ground truth segmentation, and the middle four columns represent feature maps from the low, middle, high level and final layer.

It can be observed from the performance of Dice on LUD that the segmentation predicted by SAUN for the apical level is much better than the other approaches. When comparing the results from SUN and SAUN, it can be found that the LOA from the SAUN is further improved. The clinical evaluation results on ACDC illustrate that the PCC, bias and the limit of agreement computed by SUN is inferior compared to the other networks. The evaluations predicted by SAUN achieve best with the attention mechanism. The Bland-Altman plots show almost all of the subjects from LUD and ACDC distribute between the upper bound and lower bound, which confirms that in the clinical measures the automated method is almost unbiased to the manual results. The experiments demonstrate that the proposed stack attention mechanism performs well in filtering out data noise during integrating neighboring spatial information, weighting, and confusing the feature maps of various levels as well.

Our proposed method has several limitations. It ignores the right ventricle (RV) and only provided segmentation for the left ventricle and myocardium. If more annotation information about the RV is provided for the network, the segmentation results could become more accurate. In the current implementation, we separately trained the localization and

TABLE VIII. Wilcoxon signed-rank test-based significance test results on acdc dataset. (1)w(saun,u-net): wilcoxon signed-rank test's *P*-value between saun and u-net, (2)w(saun,yun): wilcoxon signed-rank test's *P*-value between saun and yun(yolo + u-net), (3)w(saun,sun): wilcoxon signed-rank test's *P*-value between saun and sun(ssm stack + u-net), (4)w(saun,segnet):wilcoxon signed-rank test's *P*-value between saun and segnet, (5)w(saun,3d u-net):wilcoxon signed-rank test's *P*-value between saun and 3d u-net.

	ED						ES					
	Chamber			Myocardium			Chamber			Myocardium		
	Dice	HD	MCD	Dice	HD	MCD	Dice	HD	MCD	Dice	HD	MCD
W(SAUN,U-Net)	3.65E-03	0.475	5.58E-03	3.62E-05	0.189	1.34E-05	2.10E-04	1.02E-03	7.08E-04	3.81E-06	3.22E-04	7.08E-04
W(SAUN,YUN)	2.61E-04	0.0241	3.28E-03	1.34E-04	7.30E-03	3.22E-04	6.48E-03	1.34E-05	8.31E-03	3.12E-04	6.29E-05	3.65E-03
W(SAUN,SUN)	1.49E-06	0.0583	2.10E-04	5.72E-06	0.0441	1.33E-06	0.0215	0.0355	1.43E-03	9.54E-06	0.1893	1.91E-06
W(SAUN, SegNet)	3.97E-05	4.22E-03	1.91E-06	2.91E-06	3.65E-03	1.19E-05	9.54E-06	1.68E-04	1.69E-03	7.01E-05	1.05E-04	3.95E-04
W(SAUN,3D U-Net)	1.89E-04	0.0124	6.81E-04	8.86E-05	2.20E-03	1.40E-04	0.0407	0.232	5.93E-04	1.03E-04	0.0479	1.63E-04

3.65E-03 means 3.65×10^{-3}

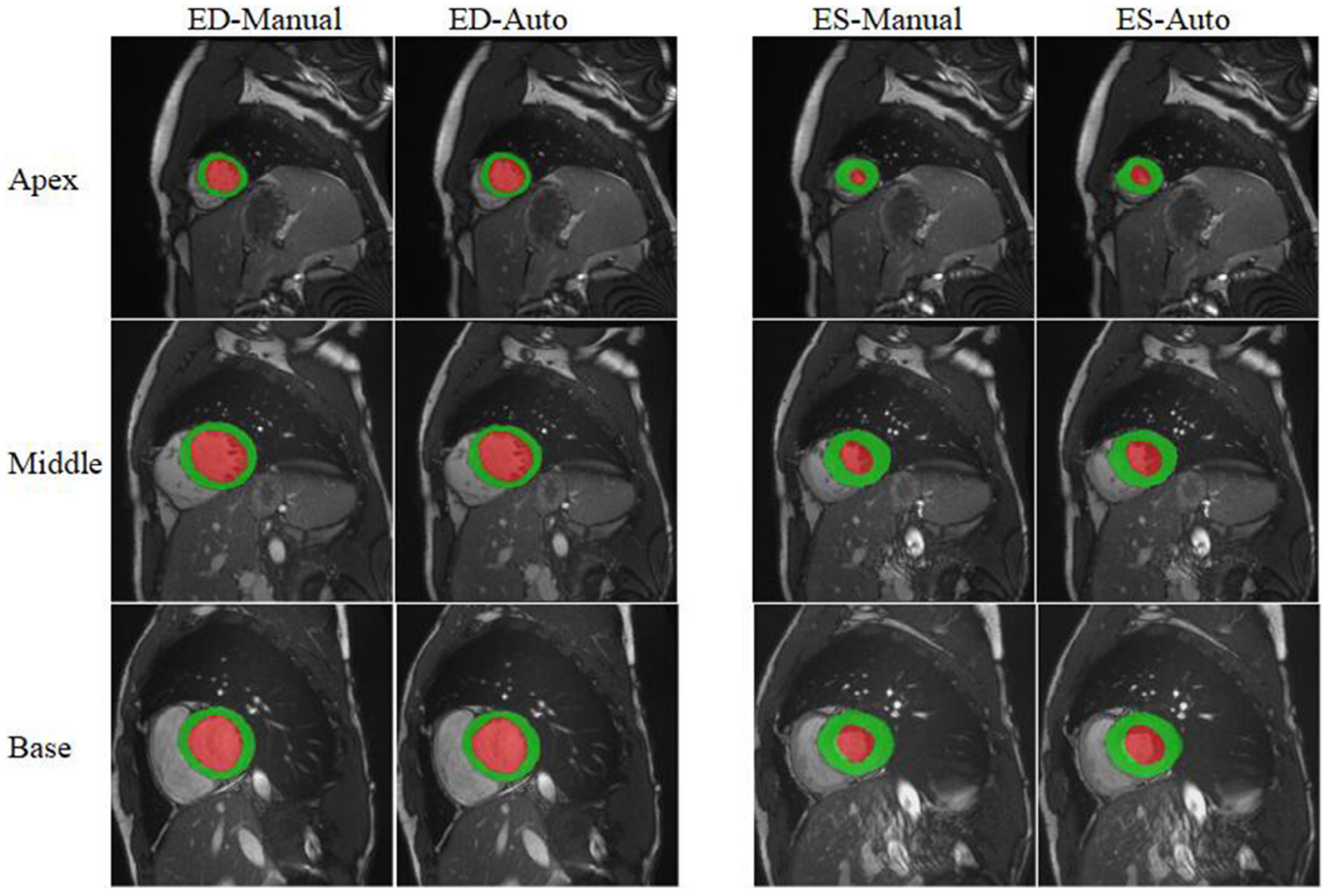

FIG. 8. Examples of segmentation from the SAUN method from two randomly selected cases from the ACDC dataset. The left two columns show end-diastolic images, and the right two columns images of end-systolic phase. For each phase, images at the apex, middle and base levels are shown. [Color figure can be viewed at wileyonlinelibrary.com]

segmentation networks. As for both tasks, the MR image features need to be explored; integration of both tasks into a single network would result in improved efficiency of the segmentation algorithm.

5. CONCLUSION

In this work, we proposed a Stack Attention U-Net-based method for automatic LV segmentation in short-axis cine

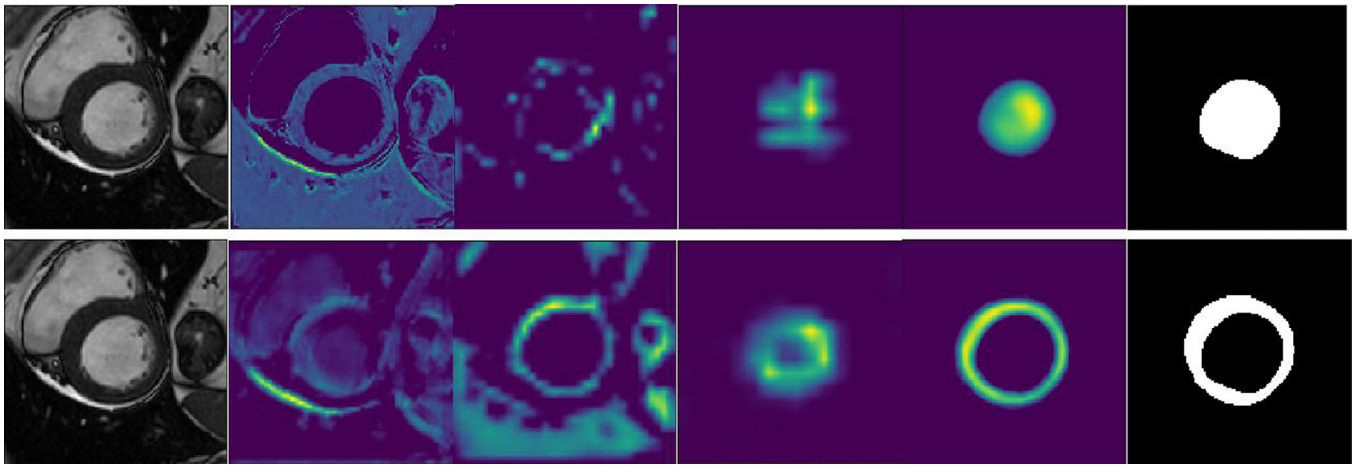

FIG. 9. Feature map visualization of SAUN. There are 42 convolutional layers in SAUN, we did the visualization for each convolutional layer. The first and last columns are the original image and the ground truth, the other four columns represent the feature maps from low, middle, high levels (from 3rd, 18th, 32nd convolutional layer) and the output of the final layer. [Color figure can be viewed at wileyonlinelibrary.com]

MRI and confirmed its benefits in integrating more information from neighboring spatial images by employing an attention mechanism to weight each channel of the feature maps. The experimental results demonstrate that the proposed approach exceeds existing state-of-the-art segmentation methods and verify its potential clinical applicability.

ACKNOWLEDGMENTS

Prof. Sven Plein from the University of Leeds is acknowledged for granting access to the image data used in this work. We also would like to acknowledge the organizer of ACDC 2017 challenge to collect and public the dataset. X. Sun is supported by the China Scholarship Council No. 201808110201.

CONFLICT OF INTEREST

The authors have no conflict of interest to disclose.

Data Availability Statement

The LUD datasets generated and analyzed during the current study are not publicly available, due to the nature of this research, participants of this study did not agree for their data to be shared publicly. The ACDC data that support this study are openly available at https://acdc.creatis.insa-lyon.fr/description/databases.html [https://doi.org/10.1109/TMI.2018.2837502].

[a]Author to whom orrespondence should be addressed Electronic mail: R.J.van_der_Geest@lumc.nl.

REFERENCES

1. Kaus MR, Von BJ, Weese J, Niessen W, Pekar V. Automated segmentation of the left ventricle in cardiac MRI. *Med Image Anal*. 2004;8:245–254.
2. 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.
3. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention; 2015.
4. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE CVPR*. 2015;3431–3440.
5. Bai W, Sinclair M, Tarroni G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson*. 2018;20:65–77.
6. Isensee F, Jaeger PF, Full PM, et al. Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features, in STACOM, 2018:120–129.
7. Qin C, Bai W, Schlemper J, et al. Joint learning of motion estimation and segmentation for cardiac MR image sequences. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2018:472–480
8. Tao Q, Yan W, Wang Y, et al. Deep learning–based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. *Radiology*. 2019;290:81–88.
9. Cheng J, Tsai YH, Wang S, Segflow YMH. Joint learning for video object segmentation and optical flow. IEEE international conference on computer vision; 2017:686–695.
10. Zhao N, O'Connor D, Gu W, Ruan D, Basarab A, Sheng K. Coupling reconstruction and motion estimation for dynamic MRI through optical flow constraint. SPIE: Image Processing; 2018.
11. Yan W, Wang Y, Li Z, Van Der Geest RJ, Tao Q. Left ventricle segmentation via optical-flow-net from short-axis cine MRI: preserving the temporal coherence of cardiac motion. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2018:613–621.
12. Zhang N, Yang G, Gao Z, et al. Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine MRI. *Radiology*. 2019;291:606–617.
13. Desai AD, Gold GE, Hargreaves BA, Chaudhari AS. Technical considerations for semantic segmentation in MRI using convolutional neural networks. arXiv preprint arXiv:1902.01977. 2019.
14. Abraham N, Khan NM. A novel focal tversky loss function with improved attention u-net for lesion segmentation. International Symposium on Biomedical Imaging; 2019:683–687.
15. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. 2018.
16. Huang Q, Yang D, Wu P, Qu H, Yi J, Metaxas D. MRI reconstruction via cascaded channel-wise attention network. International Symposium on Biomedical Imaging; 2019:1622–1626

17. 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;37:2514–2525.
18. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. IEEE conference on computer vision and pattern recognition; 2016:779–788.
19. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2017;39:2481–2495.
20. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention; 2016:424–432.
21. Perslev M, Dam EB, Pai A, Igel C. One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2019:30–38.