IEEE Journal of Translational Engineering in Health and Medicine

Received 5 November 2018; revised 20 January 2019; accepted 14 February 2019. Date of publication 25 February 2019; date of current version 25 March 2019.

Digital Object Identifier 10.1109/JTEHM.2019.2900628

Cardiac-DeepIED: Automatic Pixel-Level Deep Segmentation for Cardiac Bi-Ventricle Using Improved End-to-End Encoder-Decoder Network

XIUQUAN DU[1], SUSU YIN[1], RENJUN TANG[1], YANPING ZHANG[1], AND SHUO LI[2,3]

[1]School of Computer Science and Technology, Anhui University, Hefei 230039, China
[2]Department of Medical Imaging, Western University, London, ON N6A 3K7, Canada
[2]Digital Imaging Group of London, London, ON N6A 3K7, Canada

CORRESPONDING AUTHOR: X. DU (dxqllp@163.com)

This work was supported in part by the National Science Foundation of China under Grant 61673020 and in part by the Anhui Provincial Natural Science Foundation under Grant 1708085QF143.

ABSTRACT Accurate segmentation of cardiac bi-ventricle (CBV) from magnetic resonance (MR) images has a great significance to analyze and evaluate the function of the cardiovascular system. However, the complex structure of CBV image makes fully automatic segmentation as a well-known challenge. In this paper, we propose an improved end-to-end encoder-decoder network for CBV segmentation from the pixel level view (Cardiac-DeepIED). In our framework, we explicitly solve the high variability of complex cardiac structures through an improved encoder-decoder architecture which consists of Fire dilated modules and D-Fire dilated modules. This improved encoder-decoder architecture has the advantages of being capable of obtaining semantic task-aware representation and preserving fine-grained information. In addition, our method can dynamically capture potential spatiotemporal correlations between consecutive cardiac MR images through specially designed convolutional long-term and short-term memory structure; it can simulate spatiotemporal contexts between consecutive frame images. The combination of these modules enables the entire network to get an accurate, robust segmentation result. The proposed method is evaluated on the 145 clinical subjects with leave-one-out cross-validation. The average dice metric (DM) is up to 0.96 (left ventricle), 0.89 (myocardium), and 0.903 (right ventricle). The performance of our method outperforms state-of-the-art methods. These results demonstrate the effectiveness and advantages of our method for CBV regions segmentation at the pixel-level. It also reveals the proposed automated segmentation system can be embedded into the clinical environment to accelerate the quantification of CBV and expanded to volume analysis, regional wall thickness analysis, and three LV dimensions analysis.

INDEX TERMS CBV segmentation, deep learning, encoder-decoder, magnetic resonance images.

I. INTRODUCTION

Cardiovascular diseases cause a great deal of mortality and morbidity globally, which is one of the most common reason of death worldwide [1]. Automatic segmentation of the CBV from MR images is an essential step for the quantitative analysis of the left ventricle (LV), the right ventricle (RV), and it is typically necessary for diagnoses and treatment of cardiovascular diseases [2]. However, the majority of cardiac MR images show that the similar intensity distribution in different regions, thus providing a little edge information. It makes cardiac images segmentation as a very challenging task. Due to the relatively fixed shape of the LV during the complete cardiac cycle, most of the research works only focuses on the segmentation of the LV. In fact, it is more difficult to consider the segmentation of the CBV than the segmentation of the single ventricle. However, when pathologies such as myocardial hypertrophy, myocardial infarction, and ventricular arrhythmias are to be characterized in clinical practice, the CBV segmentation makes more sense. Hence, there is an urgent need for a computational method of LV/RV segmentation simultaneously, namely CBV segmentation.

Currently, many segmentation problems require to employing prior knowledge to improve its accuracy and robustness, as well as in medical image segmentation problems. A large number of segmentation methods based on strong priors have been proposed to segment the LV, RV or CBV, including shape prior [3], active shape and appearance models (AAM/ASM) [4], and atlas-based approaches [5], [6]. Automatic CBV segmentation can benefit from these methods that consist of some assumptions. However, incorrect priori knowledge can lead to unreliable results for the cardiac CBV segmentation methods, and even if the priori information is correct, strong priori information may prevent these CBV segmentation methods from being efficiently used by clinical applications.

To circumvent limitations of these segmentation methods relying on strong priori information, cardiac images segmentation methods with weak or no prior information have grown in popularity. For image-based [7], pixel-based classification methods [8] and deformable models [9], [10], although no methods use strong prior information to segment the CBV, they suffer from a low robustness and accuracy, and require considerable user interaction. These disadvantages will lead to the segmentation accuracy always being at a low level. Therefore, there is still much more room for improvement.

Deep learning has achieved good results in the object semantic segmentation of computer vision in recent years. Inspired by it, many popular end-to-end semantic segmentation techniques of cardiac images such as fully convolutional neural networks (FCN) [11] and U-net [12] are developed. Additionally, it can be found that those approaches [13]–[18] are applied to CBV segmentation field. These methods are all made up of convolutional neural networks (CNN), which can automatically learn complex features and concepts without feature engineering or hard-code a priori knowledge. They are used to classify the pixels in the MR images into the left and right ventricle regions simultaneously. Therefore, these approaches are capable of primly solve the limitation of regression method and traditional segmentation method. However, there has a correlation between the complex shape variables with successive frame images of a cardiac cycle. The temporal dependences between images can extract more accurate and effective features to produce a good segmentation result. Unfortunately, all of the aforementioned methods do not consider the spatiotemporal correlations in the cardiac MR sequences.

In this study, we propose an improved end-to-end Encoder-Decoder network specifically designed to segment the LV, RV and myocardium (MYO) for cardiac MR sequences from pixel-level view, namely Cardiac-DeepIED. This framework is composed of Fire dilated modules, convolution LSTM (Conv-LSTM) and D-Fire dilated modules. It can automatically represent images from the pixel perspective with the spatiotemporal correlations and detect the ventricle region more accurately by leveraging advantages of deep learning to obtain the optimal segmentation result. In summary, our study makes the following contributions:

1) It enables an accurate, robust, and automated segmentation of CBV with less assumption. Hence, with our method, an efficient clinical tool is provided in clinical practice for accelerating the diagnosis and treatment of cardiovascular diseases.
2) It creatively formulates the segmentation task as a multi-classification problem to makes full use of the advantages of deep learning in holistic fashion for optimizing output while maintaining accuracy, robustness, and efficiency.
3) It creatively combines Encoder-Decoder (ED), Conv-LSTM and dilated convolution fusion together in one unified end-to-end framework. The novel integration has super strengths in representation learning and spatiotemporal context learning.

The rest of this paper is structured as follows: Section II presents an experimental framework to demonstrate and evaluate the efficacy of our Cardiac-DeepIED model on cardiac MR images. Section III introduces the dataset and evaluation criteria. Section IV reports the results and analysis of the evaluation. Section V reports discussions and conclusions.

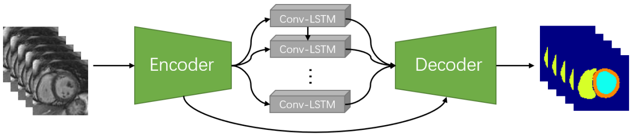

FIGURE 1. The workflow of the Cardiac-DeepIED.

II. METHOD

The workflow of our proposed method is shown in Fig. 1. As we can see, an ED network and a Conv-LSTM module constitute our Cardiac-DeepIED segmentation network. Moreover, skip connections are designed between the encoder and the decoder. In ED network, we replace the standard convolution layers with Fire dilated and D-Fire dilated modules consisting of dilated convolution. The details of the network will be introduced in the section II.B.

The goal of our Cardiac-DeepIED method is to segment the LV, the MYO and the RV of a series of MR image X. This is done by predicting the label map M_n of each MR image X_n, where n represents the n-th MR image. Pixels $p = (i, j)$ of the label map M_n contain a label $M_n^p \in \{$Background, LV, RV, MYO$\}$. Here, we employ the categorical cross-entropy loss function as the objective function of segmentation network. The formula is as follows:

$$L = -\sum_{l \in D} y_l \log(f(x_l; \theta)) \tag{1}$$

where labeled data set D is used for segmentation network, x_l denote the input data, y_l denote the ground truth, and $f(\cdot)$ is the segmentation function parameterized by θ.

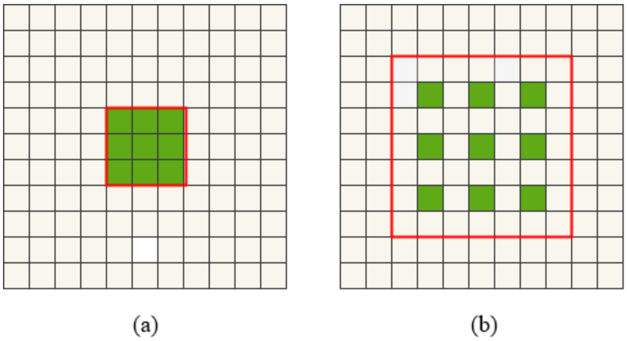

FIGURE 2. Illustration of dilated convolution. Green block indicates the data point of the 3 × 3 filter and the red square indicates the receptive field captured by the 3 × 3 filter. (a) F_0 indicates the 2D image, F_1 is generated from F_0 through a 1-dilated convolution (also called standard convolution); each element of F_1 possess of a 3 × 3 receptive field. (b) F_2 is generated from F_0 through a 2-dilated convolution (dilation rate is 2); each element in F_2 possess of a 7 × 7 receptive field.

A. REVIEW OF DILATED CONVOLUTION

As we know, the Fire dilated and D-Fire dilated module mentioned above is composed of dilated convolution. Before we introduce the details of the proposed method, we will review the principle of dilated convolution. To our knowledge, dilated convolution, also known as convolution with holes or atrous convolution. The dilated convolution is widely used in semantic segmentation [19]–[21], based on the fact that it supports exponential expansion of the receptive field and captures larger context without loss of resolution or coverage.

To better understand the dilated convolution of our network, we first explain the standard 2D convolution layer. Supposed $j(x, y)$ represents a position on the output O, (x, y) represents the spatial coordinate position. w denotes the weight, k denotes the kernel function and b denotes the bias, so the formula of 2D convolution on the input feature map m in stride of 1 is as follows:

$$O_i = f \sum_k m(j + k) * w(k) + b(k) \tag{2}$$

where f represents the kernel-wise nonlinear transformation of Rectified Linear Unit (ReLU). $*$ is a convolve operator. The dilated convolution adds a dilated rate d to each kernel k:

$$O(i) = f \sum_k m(j + d \cdot k) * w(k) + b(k) \tag{3}$$

where the $d \cdot k$ is produced by inserting *d-1* zeros between two consecutive values of each kernel along each spatial dimension [20]. Standard 2D convolution is a special case of dilated convolution with $d = 1$. The receptive field of the dilated convolution is a square of exponentially increasing size as shown in Fig. 2. Our Fire dilated modules and D-Fire dilated modules adopt dilation rates $d = 2$, which allow the Fire dilated modules and D-Fire dilated modules to modify convolution-kernel's receptive fields rather than stacked down-sampling operators.

B. TWO ADVANCED STRUCTURES OF THE SEGMENTATION NETWORK

The ED style network [22]–[25] developed in the recent study is used to segment images and obtain good results by the end-to-end fashion in computer vision. Different from above, our model has integrated the ED and Conv-LSTM in an innovative way into one unified end-to-end network architecture. As shown in Fig. 3, the Cardiac-DeepIED network is composed of an ED structure and a Conv-LSTM structure [26], it aims at extracting more effective information at the pixel level and capturing the temporal dependencies to guide cardiac segmentation. The details of the two advanced structures for the segmentation are as follows.

1) THE ED STRUCTURES

In our model, the ED network takes Fire dilated modules as an encoder and D-Fire dilated modules as a decoder, and three skip connections are added between the encoder and the decoder. Our ED network does not only keep fine-grained information with few parameters and suitable receptive fields but also sufficiently recover spatial dimension and images details information.

The difference between our ED network and existing ED method is that ED network employs Fire dilated modules shown in Fig. 4 (a) as an encoder for efficient images representation and employs D-Fire dilated modules shown in Fig. 4 (b) as a decoder to obtain the full spatial resolution. As illustrated in Fig. 4, the Fire dilated modules and D-Fire dilated modules build a block respectively which is used to encapsulate three dilated convolution layers. The two kinds dilated modules play the role of squeeze network, which capture more image information and enlarge receptive fields more efficiently on limited parameters. Therefore, the ED architecture which consists of Fire dilated modules and D-Fire dilated modules not only definitely solves the high variability of complex cardiac structures from MR images, but also capture more image information and enlarge receptive fields more efficiently without adding extra parameters and computation.

Another problem, the resolution of the feature maps and spatial information after three max-pooling layers, need to be solved before our model training, so we add two corresponding up-sampling layer to restore images resolution. However, the lost spatial information is difficult retrieved. To compensate for the loss of resolution caused by the pooling layer, the Cardiac-DeepIED introduces skip connections between its encoder and decoder. Skipping connections helps the decoder path to recover fine-grained information from the encoder path. The higher resolution information is passed by means of standard skip connections between the encoder and the decoder. As a result, it has the advantages of being capable of obtaining semantic task-aware representation and preserving fine-grained information.

In our network, we divided cardiac image input into different semantically interpretable categories. Semantic

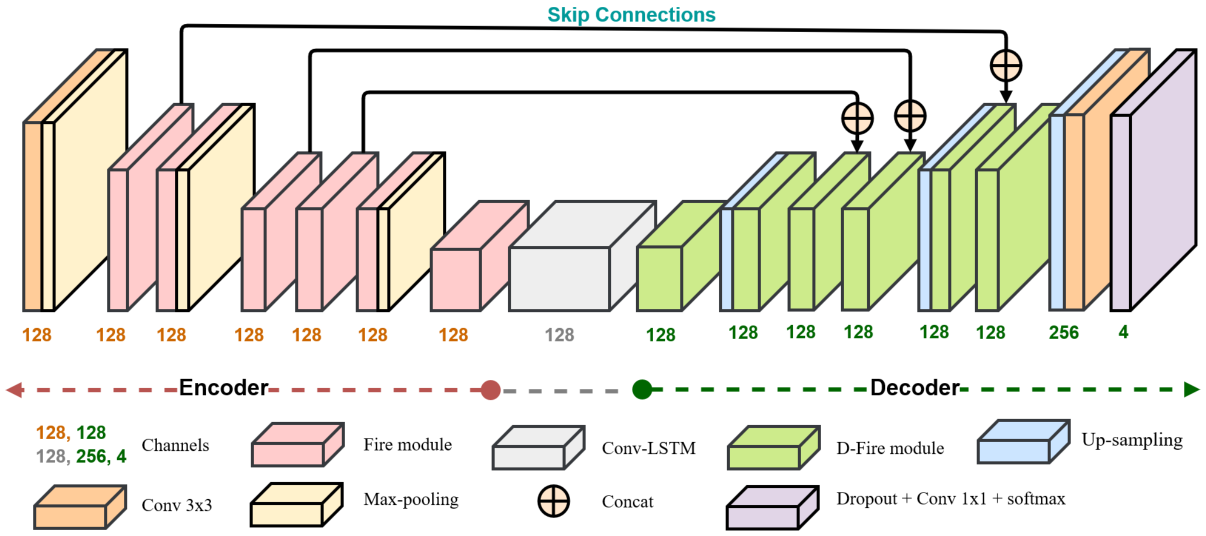

FIGURE 3. The details of the Cardiac-DeepIED network. Blocks of the same color mean they have the same parameters.

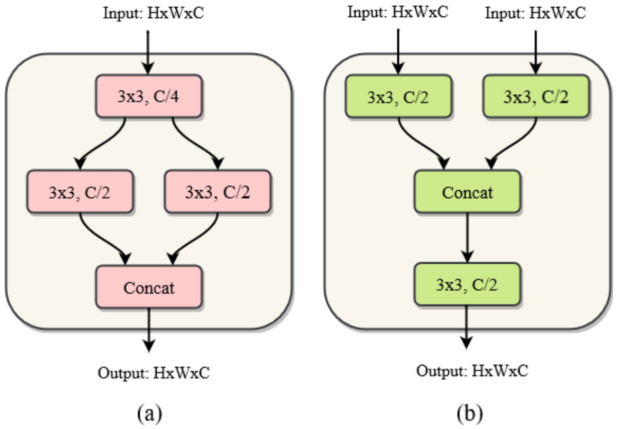

FIGURE 4. (a) Fire dilated module and (b) D-Fire dilated module. Both Fire dilated module and D-Fire dilated module are composed of dilated convolution (dilation rate is 2). In colored rectangles, 3 × 3 represents the kernel size, C/4 and C/2 represent channels of feature maps produced by colored rectangles. Each layer of dilated convolution is followed by a ReLU layer. The input and output of the two modules are the same, and each layer in the modules is the dilated convolution used to increase the receptive field.

interpretability, the classification category, makes sense in real images, and semantic segmentation allows us to understand the cardiac image in more detail. In this architecture, the encoder and Conv-LSTM (Introduced in the next section) can effectively capture local semantic information (pixel-based information) in the image, and the pooling layers in the encoder reduces the size of the feature map, making it a low-dimensional image representation, but with rich semantic information. The decoder receives the image representation, recovers the spatial dimension by up-sampling, and finally, the decoder generates a feature map representing the original image label. These feature maps are then input into a standard convolutional layer with a kernel size of 1 (equivalent to the fully connected layer). The convolution operator outputs a score map, which gives the probability of each class at each pixel, as below.

$$S_k = F(o; k) = \sum_{x=1,y=1}^{J} k_{x,y} o_{x,y} \tag{4}$$

where (x, y) represents the spatial coordinate position, o is the output features of network. k denotes the kernel function and C denotes the number of channels. J is the set of pixel positions. The final score at score map is just summed over all channels of feature maps, as below.

$$\delta(S_c) = \frac{\exp(S_c)}{\sum_{i=1}^{C} \exp(S_i)} \tag{5}$$

where δ is the prediction probability and S is the output of network, and $c \in \{1, 2, \ldots, C\}$. As shown in Equation 4 and Equation 5, the final predicted label is the category with highest probability.

2) THE CONV-LSTM STRUCTURE

As a special RNN structure, LSTM was proposed by Hochreiter & Schmidhuber [27] and it had been proved to be stable and powerful in modeling long-term dependencies in previous works [28]–[30]. Subsequently, LSTM was applied to language modeling [31], image caption [32], [33], scene labeling [34] and so on. In the field of the medical image, LSTM was used to evaluate cardiac function [24] via capturing the temporal dependencies between MR image sequences without considering spatial correlations. Recently Shi *et al.* proposed Conv-LSTM that integrated convolution into the LSTM by substituting convolutional filters for the weights for precipitation Nowcasting [26]. Later Conv-LSTM was applied to anomaly detection in the video [35] and autonomous driving [36]. The advantage of the convoluted LSTM is that it does not only propagate the characteristics

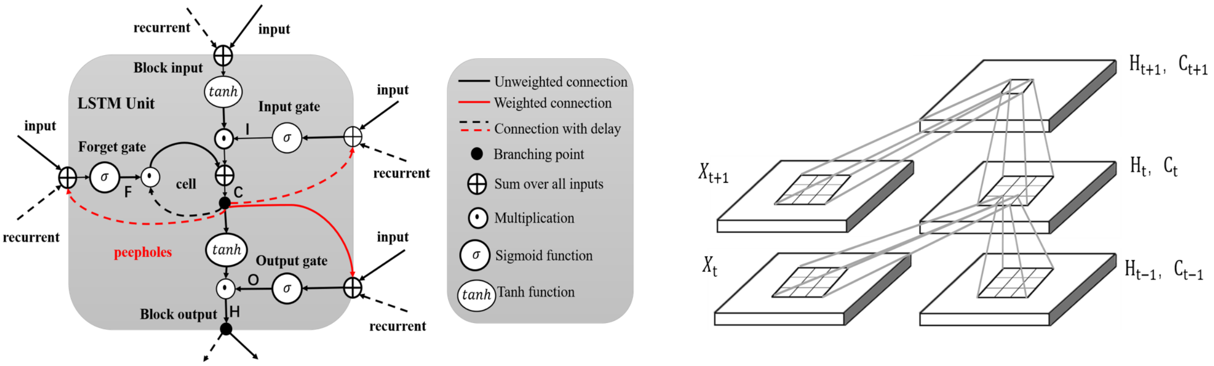

FIGURE 5. LSTM and Conv-LSTM network visualization. In short, the principle of Conv-LSTM is that the vector-to-vector multiplication in the standard LSTM replaced by the previously described spatiotemporal convolution. (a) LSTM unit. (b) Inner structure of Conv-LSTM.

of space but also captures temporal dependences between consecutive frames.

Therefore, Conv-LSTM is more suitable for working with cardiac images than LSTM. The Conv-LSTM structure is explicitly a special LSTM unit and models the potential spatiotemporal correlations between consecutive frame images, which efficiently improve the segmentation accuracy of cardiac dynamic structures with information from consecutive frame images. Therefore, the Conv-LSTM structure is capable of memorizing long-period spatiotemporal correlations from consecutive frame images of cardiac cycles.

The major difference between the proposed method and existing works is that our Cardiac-DeepIED network integrates Conv-LSTM units into ED network for cardiac image analysis. Just because of this, the input and output of all layers are 4-D tensors: the first dimension corresponds to cardiac MR frames, which representing the time step, the two dimensions in the middle represent the length and width of the image, and the fourth indexes different feature channels. When 4D data is fed to Conv-LSTM, it can directly utilize 4D tensors and maintain spatial structure without losing spatial information. In this paper, we use the wrapper Time-Distributed to process 4D data instead of using a 3D network. We adopt the wrapper Time-Distributed to wrap each layer in the network. In fact, we can realize the transition from 2D to 3D through this layer. With the Time-Distributed wrapper, the input of the network is still 4D tensors, while the difference is the size of the first represents the time step. our network parameters and calculations are almost half of 3D networks, which greatly reduces the time cost.

Inner structure of the Conv-LSTM is shown in Fig. 5, given the input, cell outputs and hidden states are represented by $X_{1,\cdots,t}$, $C_{1,\cdots,t}$ and $H_{1,\cdots,t}$ respectively. Input gates, forget gates and cell outputs gates are I_t, F_t and O_t respectively. The formulation of the Conv-LSTM is as follows:

$$I_t = \sigma(W_{XI} * X_t + W_{HI} * H_{t-1} + W_{CI} \circ C_{t-1} + b_I) \quad (6)$$

$$F_t = \sigma(W_{XF} * X_t + W_{HF} * H_{t-1} + W_{CF} \circ C_{t-1} + b_F) \quad (7)$$

$$C_t = F_t \circ C_{t-1} + I_t \circ \tanh(W_{XC} * X_t + W_{HC} * H_{t-1} + b_{CO}) \quad (8)$$

$$O_t = \sigma(W_{XO} * X_t + W_{HO} * H_{t-1} + W_{CO} \circ C_{t-1} + b_O) \quad (9)$$

$$H_t = O_t \circ \tanh(C_t) \quad (10)$$

where $*$ indexes the convolution operator and $\circ$ indexes the Hadamard product. σ is activation function, b index bias vectors. The weighted connections between states are represented by weight matrices W, for example, W_{XF} denotes the matrices of weights from forget to the input, W_{CI}, W_{CF} and W_{CO} are diagonal weight matrices for peephole connections.

The output state controls the amount of information propagated from previous time steps, while the information received by the next time step and layer constitutes the so-called hidden state. The peephole connection not only allows the LSTM unit to access, but also propagates information recorded from the cell state of the previous time step. Just as the convolution filter input to the hidden connection determines the resolution of the feature map created from the input, the convolution filter size hidden to the hidden connection determines the aggregated information received by the Conv-LSTM unit from the previous time step. The state transition between the time steps of the Conv-LSTM unit can then be interpreted as a movement between frames. Larger transition kernels tend to capture faster motions, while smaller transition kernels capture slower motions.

III. NETWORK IMPLEMENTATION

A. DATASETS

We employ 2900 CBV images from 145 subjects to evaluate the segmentation performance of our model. These MR images are 2D short-axis cine. The age of subjects ranges from 16 to 97 years old. Each subject includes 20 frames across a cardiac circle. In each frame, the middle slice is selected according to the standard AHA prescriptions [37] for validation of our proposed Cardiac-DeepIED network. The pixel pitch of the MR images within a small range (0.6836-2.0833 mm/pixel) with the mode of 1.5625 mm/pixel. Two landmarks that junctions of the right ventricular wall with the left ventricular are manually labelled for each cardiac image to provide a reference for cardiac ROI cropping. The cropped images are resized to the dimension of 80 × 80. After the preprocessing, the ground truth is labeled manually by two experienced researchers and two experienced cardiac radiologists check the manually obtained bi-ventricle boundaries from all the cardiac MR images.

B. CONFIGURATIONS

We implement all of the codes using Python on a windows10 desktop with Intel Core i5-7400 CPU. The graphics card is an NVIDIA GeForce GTX 1060. The deep learning libraries are implemented with Keras (Tensorflow). We added dropout layer [38] before the last layer to prevent the network from over-fitting. We choose Adaptive Moment Estimation (Adam) as our optimizer and the softmax function is employed to fine-tuning the model to classify each pixel (ventricle area or not), and leave-one-out cross validation is applied to performance evaluation. Due to the experimental limitation on hardware, we set up the mini-batch size as 1 and the number of training epoch as 40.

C. PERFORMANCE EVALUATION CRITERIA

We evaluate the performance of cardiac segmentation by employing DM, Jaccard coefficient, Accuracy and positive predictive value (PPV). DM measures the overlap between ground truth area and predicted area. The DM value is in the range of 0∼1, and a higher DM denotes a better match between manual contour and predicted contour. The formula for DM is as follows:

$$DM(C_m, C_a) = 2 \times \frac{C_m \cap C_a}{C_m + C_a} \tag{11}$$

Similar to the DM, the Jaccard coefficient also is in the range of 0∼1, and Jaccard coefficient denotes correspondence with the ground truth. The formula for Jaccard coefficient is as follows:

$$J(C_m, C_a) = \frac{C_m \cap C_a}{C_m \cup C_a} = \frac{C_m \cap C_a}{C_m + C_a - (C_m \cap C_a)} \tag{12}$$

where C_m and C_a denote the region of manual and automatic contour, respectively.

In addition, we also use accuracy and PPV evaluate the results of pixel classification. They are defined by:

$$Accuracy_c = \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c} \tag{13}$$

$$PPV_c = \frac{TP_c}{TP_c + FP_c} \tag{14}$$

where TP_c, TN_c, FP_c, and FN_c represent true positive, true negative, false positive, false positive and false negative of class c respectively.

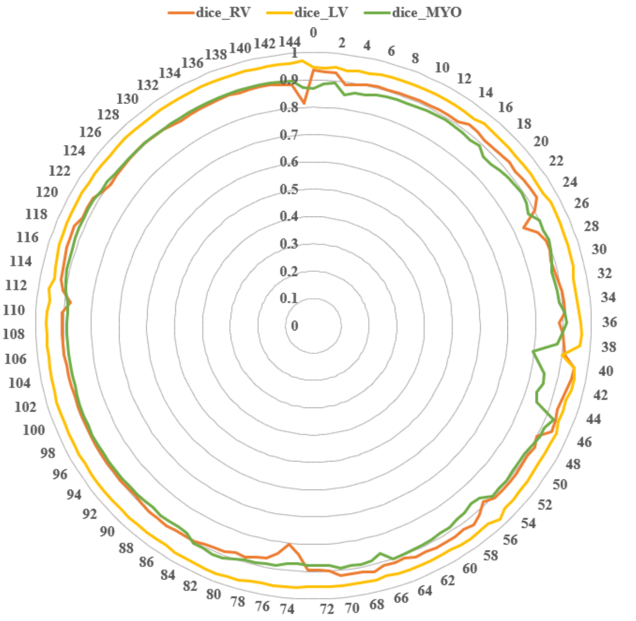

FIGURE 6. Dice value of 145 subjects based on our proposed network.

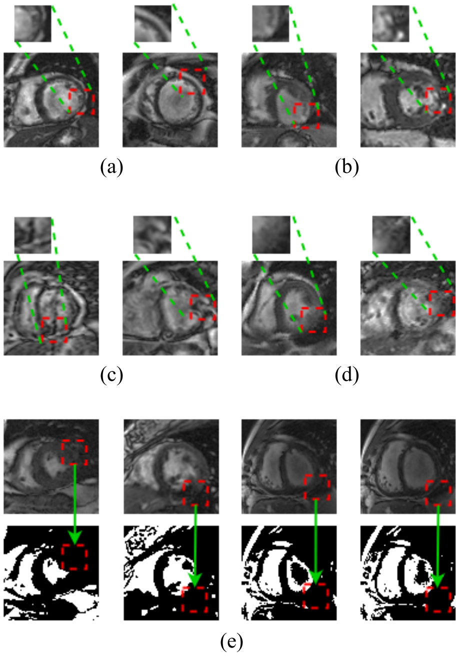
(a)
(b)
(c)
(d)
(e)

FIGURE 7. Some cases in the dataset that may affect segmentation results. In the case of (a, d, e), the red dotted rectangle indicates that the myocardial region provides less boundary information and may lead to poor segmentation performance. In the case of (b, c), the red dotted rectangle indicates that the myocardial boundary region is incomplete, resulting in difficulty in segmentation. (a) Thin. (b) Missing. (c) Interfence. (d) Blurry. (e) Dark and low intensity contrast.

IV. RESULTS

A. THE PERFORMANCE ON 145 SUBJECTS

The DM values of our proposed Cardiac-DeepIED method are measured by comparing our segmentation results with the ground truth. Fig. 6 illustrates the segmentation results that the average DM of each subject (20 images). Although the challenges in segmenting CBV from MR images, our method achieves a high average DM of 0.890 (MYO), 0.960 (LV) and 0.903 (RV). Furthermore, the average DM of each subject exceeds 0.9 for the LV and RV, the average DM of each subject is around 0.890 for the MYO. The above results show

TABLE 1. Comparison between our method and several current segmentation methods (mean value ± standard deviation)).

Methods	DM			Jaccard coefficient		
	LV	MYO	RV	LV	MYO	RV
Conv-Deconv	0.934 ± 0.003	0.846 ± 0.011	0.881 ± 0.009	0.884 ± 0.064	0.733 ± 0.088	0.778 ± 0.109
Seg-Net	0.937 ± 0.009	0.844 ± 0.013	0.878 ± 0.022	0.885 ± 0.061	0.728 ± 0.088	0.779 ± 0.111
FCN	0.950 ± 0.004	0.873 ± 0.005	0.892 ± 0.025	0.907 ± 0.052	0.778 ± 0.078	0.800 ± 0.130
U-net	0.953 ± 0.016	0.887 ± 0.016	0.897 ± 0.026	0.919 ± 0.049	0.800 ± 0.072	0.827 ± 0.095
Cardiac-DeepED *	0.951 ± 0.008	0.870 ± 0.020	0.899 ± 0.018	0.912 ± 0.048	0.775 ± 0.080	0.818 ± 0.099
Our method	**0.960 ± 0.008**	**0.890 ± 0.018**	**0.903 ± 0.026**	**0.923 ± 0.040**	**0.801 ± 0.068**	**0.828 ± 0.097**

Methods	Accurary			PPV		
	LV	MYO	RV	LV	MYO	RV
Conv-Deconv	0.986 ± 0.007	0.965 ± 0.014	0.976 ± 0.010	0.942 ± 0.055	0.840 ± 0.094	0.883 ± 0.111
Seg-Net	0.985 ± 0.007	0.964 ± 0.014	0.976 ± 0.011	0. 930 ± 0.065	0.845 ± 0.090	0.883 ± 0.106
FCN	0.989 ± 0.007	0.972 ± 0.011	0.979 ± 0.014	0.951 ± 0.049	0.878 ± 0.073	0.896 ± 0.112
U-net	0.990 ± 0.006	0.975 ± 0.011	0.982 ± 0.010	0.958 ± 0.048	0.887 ± 0.072	0.909 ± 0.097
Cardiac-DeepED *	0.989 ± 0.006	0.972 ± 0.011	0.981 ± 0.010	0.952 ± 0.051	0.871 ± 0.076	0.907 ± 0.096
Our method	**0.991 ± 0.005**	**0.976 ± 0.009**	**0.982 ± 0.011**	**0.960 ± 0.040**	**0.891 ± 0.066**	**0.913 ± 0.095**

Cardiac-DeepED * represents the our method that replaced the Fire dilated modules and D-Fire dilated modules with standard convolution layers.

that the segmentation results of each patient are at a relatively high level, and the segmentation results between patients do not have large fluctuation amplitudes. This also indicates that our method has powerful generalization ability in segmenting images of patients with large differences.

For the fact that myocardial segmentation results are lower than the left ventricle and right ventricle, we consider the following factors that may limit the performance of the proposed method. As shown in the following Fig. 7 (a), the myocardial region is very thin, and the inner and outer membranes of the myocardium are almost indistinguishable, which makes the segmentation difficult. As shown in Fig. 7 (b), we can see that some region of the myocardium is missing, which is obviously difficult to segment. As shown in Fig. 7 (c), part of the myocardium is interfered by adjacent tissues, causing the myocardial area to be truncated. As shown in Fig. 7 (d), the myocardial region is blurred and it is difficult to define the myocardial boundary. Fig. 7 (e) shows some of the images in patients 38-42. From the images and the corresponding binary image of the images, we can see that the low intensity contrast between myocardium and adjacent tissues, resulting in little border information. These reasons may cause the MYO segmentation to present consistently lower values for MYO segmentation.

B. THE ADVANTAGES OF THE FIRE DILATED MODULES, D-FIRE DILATED MODULES AND SKIP CONNECTIONS

By passing the encoder's feature maps to the decoder, skip connections help the decoder to have more image detail information, thus recovering a better image. At the same time, the skip connection can solve the problem of gradient disappearance in the case of deep network layers and speed up the training process. If there is no skip connection, the decoder cannot recover many important spatial information lost when restoring the spatial resolution, which will also have a certain impact on the segmentation result. It can be seen from the results shown in the first line of Fig. 8 that the Cardiac-DeepIED without the skip connection, the decoder cannot recover important image details and affect the final segmentation result. In fact, the segmentation accuracy of the cardiac bi-ventricle is 0.961 (left ventricle), 0.916 (myocardium) and 0.945 (right ventricle), respectively. This is far worse than the result of the Cardiac-DeepIED. At the same time, this also proves the importance of skip connections.

To verify the importance of the Fire dilated modules and D-Fire dilated modules, we replaced it with standard convolution layers and the segmentation results are displayed in Table 1. Fig. 8 shows some examples of CBV segmentation results, we obviously observe that our proposed method can

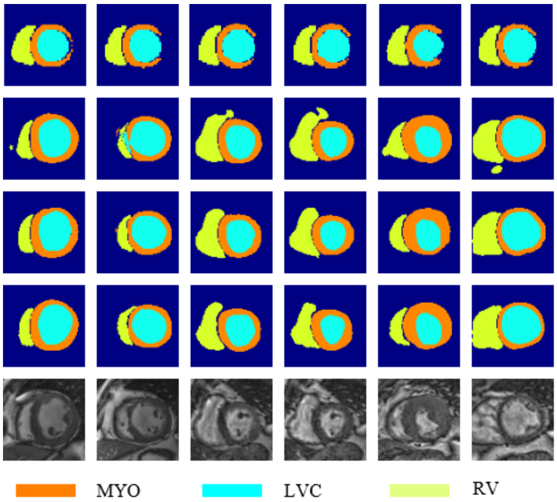

FIGURE 8. Illustration the effectiveness of the fire module and D-Fire dilated module. The first row is the segmentation result of the Cardiac-DeepIED without skip connections. The second row is the segmentation result of replacing fire modules and fire modules with standard convolution layers. The third row is the segmentation result of the proposed method, and the fourth row is ground truth. The fifth row is the cardiac MR image. The last row indicates the corresponding physiological structure of each color region. MYO: Myocardium; LVC: Left ventricular cavity.

effectively make the prediction more consistent. As we have seen, the second row results of the Fig. 8 suffer from background interference, which mainly due to the similar intensity distribution between adjacent tissues and faint cardiac images provide little boundary information. Our Fire dilated modules and D-Fire dilated modules handle the problem effectively. The prediction result is refined by the Fire dilated modules and D-Fire dilated modules shown in the third row of the Fig. 8.

C. QUANTIFICATION ANALYSIS

In order to better demonstrate the effectiveness of our method, we quantify the area of one of the clinical indicators. In Fig. 9 (a), the blue points on the red line indicate that the physiological structure area predicted by our method completely overlaps with the corresponding physiological structure area manually delineation. From the Fig. 9 (a), we can see that the area predicted by the automatic method is very close to the area manually delineation, which demonstrates great correlation between predicted result by automated method and manual delineation. Fig. 9 (b) reports consistency between automated area and the manual delineation area. From the Fig. 9 (b), we can see that most of the points are within the standard line and the mean line is close to 0, which demonstrates the predicted areas by automatic method are highly consist with manual delineation areas.

D. PERFORMANCE COMPARISON BETWEEN OUR METHOD AND OTHER METHODS

To better understand the segmentation results of the Cardiac-DeepIED network, the predicted maps and results of other methods including Conv-Deconv [39], Seg-Net [22] FCN [13] and U-net [12] are illustrated in Fig. 10. It can be seen that our method enables more accurate and robust segmentation of CBV images. This great robustness of our method due to the seamless combination of ED network and Conv-LSTM unit in a unified end-to-end framework. Therefore, the effectiveness and advantages of our method is beneficial to clinical cardiovascular disease diagnosis.

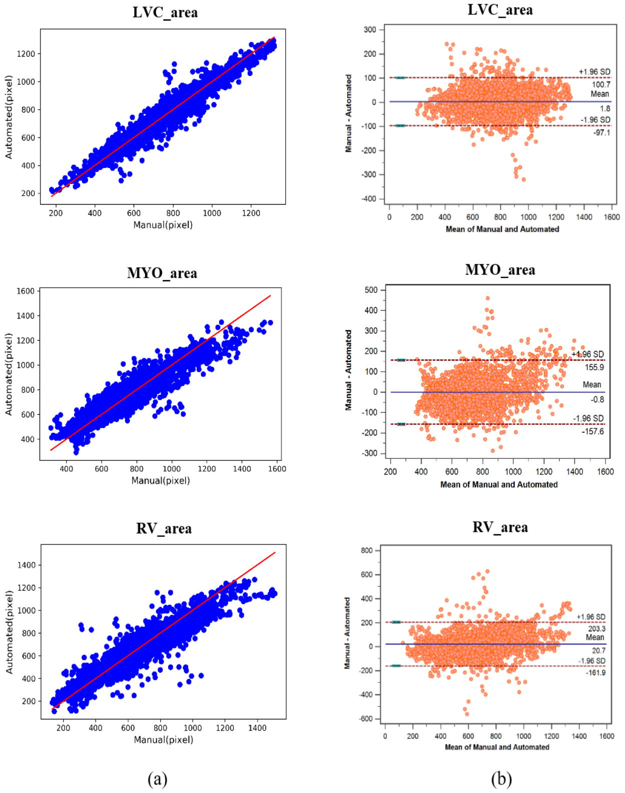

FIGURE 9. The comparison between predicted areas by automated method and manual delineation areas. (a) Linear regression analysis. (b) Bland-Altman analysis. The solid line in the middle is the mean line of the difference between the aforementioned two measurement methods, and the upper and lower dashed lines represent ±1.96 standard deviation (SD) of the difference. LVC: Left ventricular cavity; MYO: Myocardium; RV: Right ventricle.

To evaluate the effectiveness of our method, we compare it with other methods including Conv-Deconv [39], Seg-Net [22], FCN [13] and U-net [12]. From Table 1, we can see that the segmentation performance of all of these compared methods are lower than ours, which demonstrates that our improvements are resulted from adopting the proposed Cardiac-DeepIED method to represent cardiac images at pixel level. In summary, we can conclude that our method is generally outperforms the previous method with higher segmentation capability for segmenting the CBV. We can clearly see that the proposed model is a more effective method to segment CBV compared to existing methods. Therefore, it makes us more convinced that our method is capable of being a useful tool for the diagnosis and treatment of cardiovascular diseases in clinical practice.

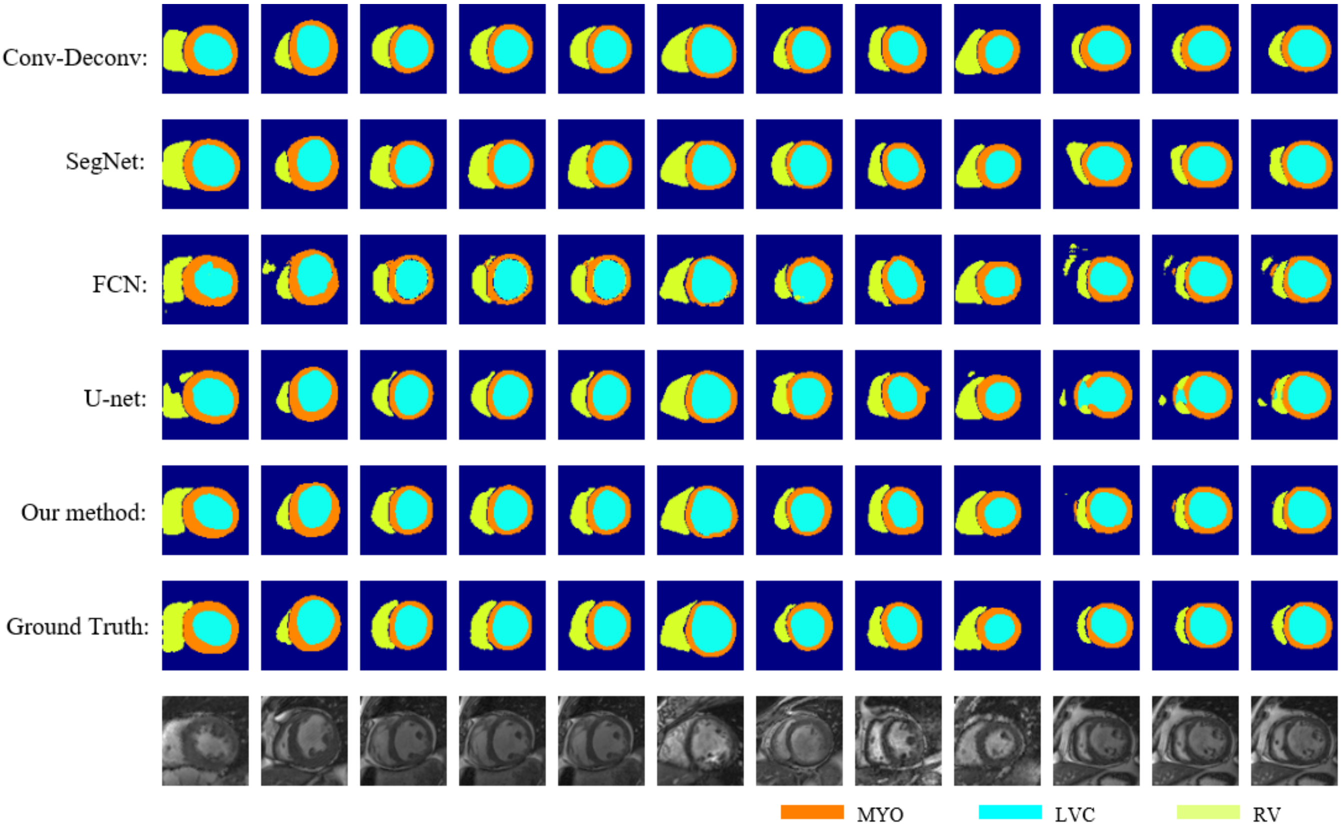

FIGURE 10. Visualization of segmentation results.

V. CONCLUSION

In previous works, some effective methods of the CBV from cardiac images have proposed. However, these methods have one or more limitations, such strong prior, low robust and accuracy and so on. Because of the limitations, they are no more applicable in real clinical applications. Therefore, we proposed an automatic segmentation network based on deep learning to segment CBV at the pixel-level. Our method combines the advantages of the powerful Fire dilated modules, D-Fire dilated modules with ED for the semantic fine-grained representation, leverages the specialty of the Conv-LSTM structure for modeling spatiotemporal correlations between consecutive frame images. After leave-one-out cross-validation on the 145 subjects, our method achieves an accurate segmentation. Experimental results show that the proposed method performed significantly well in segmenting the CBV. This automatic method of the CBV segmentation has paved a great way for other medical image segmentation, such as spine image segmentation, brain image segmentation and so on.

In our study, we only pay attention to cardiac bi-ventricle segmentation in single magnetic resonance modality. Multi-modalities cardiac magnetic resonance of the cardiac segmentation will be considered for future work. Besides, our method can't completely get rid of the radiologist, it can only help them to reduce the heavy workload to a certain extent. In further research, we will consider automatically generating a diagnostic report on the multi-modal cardiac magnetic resonance image, which can speed up the initiation of many specific treatments and help save time related to cardiology.

REFERENCES

[1] W.-Y. Low, Y.-K. Lee, and A. L. Samy, "Non-communicable diseases in the asia-pacific region: Prevalence, risk factors and community-based prevention," *Int. J. Occupat. Med. Environ. Health*, vol. 28, no. 1, pp. 1–7, Dec. 2015.

[2] X. Zhen, Y. Yin, M. Bhaduri, I. B. Nachum, D. Laidley, and S. Li, *Multi-task Shape Regression for Medical Image Segmentation*. Athens, Greece: Springer, 2016.

[3] J. Senegas, C. A. Cocosco, and T. Netsch, "Model-based segmentation of cardiac MRI cine sequences: A Bayesian formulation," *Proc. SPIE, Med. Imag.*, vol. 5370, pp. 432–443, May 2004.

[4] H. Zhang, A. Wahle, R. K. Johnson, T. D. Scholz, and M. Sonka, "4-D cardiac MR image analysis: Left and right ventricular morphology and function," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 350–364, Feb. 2010.

[5] M. Lorenzo-Valdés, G. I. Sanchez-Ortiz, A. G. Elkington, R. H. Mohiaddin, and D. Rueckert, "Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm," *Med. Image Anal.*, vol. 8, no. 3, pp. 255–265, Sep. 2004.

[6] X. Zhuang *et al.*, "Robust registration between cardiac MRI images and atlas for segmentation propagation," *Proc. SPIE, Med. Imag.*, vol. 6914, pp. 91408–91408, Mar. 2008.

[7] A. Katouzian, A. Prakash, and E. Konofagou, "A new automated technique for left- and right-ventricular segmentation in magnetic resonance imaging," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, Feb. 2006, pp. 3074–3077.

[8] D. T. Gering, "Automatic segmentation of cardiac MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Montréal, Canada, Nov. 2003, pp. 524–532.

[9] G. Hautvast, S. Lobregt, M. Breeuwer, and F. Gerritsen, "Automatic contour propagation in cine cardiac magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1472–1482, Nov. 2006.

[10] M. Sermesant *et al.*, "Cardiac function estimation from MRI using a heart model and data assimilation: Advances and difficulties," *Med. Image Anal.*, vol. 10, pp. 642–656, Jun. 2006.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 79, Jun. 2015, pp. 3421–3440.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 9351. Cham, Switzerland: Springer, Nov. 2015, pp. 234–241.

[13] P. V. Tran. (2016). "A fully convolutional neural network for cardiac segmentation in short-axis MRI." [Online]. Available: https://arxiv.org/abs/1604.00494

[14] H. B. Winther *et al.* (2017). "ν-net: Deep learning for generalized biventricular cardiac mass and function parameters." [Online]. Available: https://arxiv.org/abs/1706.04397

[15] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-mrivia time-series segmentation and domain specific features," *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, vol. 10663. Springer, Mar. 2018, pp. 120–129.

[16] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P. M. Jodoin, "GridNet with automatic shape prior registration for automatic MRI cardiac segmentation," *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, vol. 10663. Canada: Springer, Mar. 2018, pp. 73–81.

[17] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "Ω-Net (Omega-Net): Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks," *Med. Image Anal.*, vol. 48, pp. 95–106, Aug. 2018.

[18] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, Oct. 2016, pp. 565–571.

[19] F. Yu and V. Koltun. (2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122

[20] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). "Rethinking atrous convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1706.05587

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla. (2015). "SegNet: A deep convolutional encoder-decoder architecture for image segmentation." [Online]. Available: https://arxiv.org/abs/1511.00561

[23] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent CRF for real-time roadobject segmentation from 3D LiDAR point cloud," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2018, pp. 1887–1893.

[24] W. Xue, A. Islam, M. Bhaduri, and S. Li, "Direct multitype cardiac indices estimation via joint representation and regression learning," *IEEE Trans. Med. Imag.*, vol. 36, no. 10, pp. 2057–2067, Oct. 2017.

[25] J. Lieman-Sifry, M. Le, F. Lau, S. Sall, and D. Golden, "FastVentricle: Cardiac segmentation with ENet," in *Functional Imaging and Modelling of the Heart*, vol. 10263. Cham, Switzerland: Springer, Apr. 2017, pp. 127–138.

[26] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 802–810.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] R. Pascanu, T. Mikolov, and Y. Bengio. (Nov. 2012). "On the difficulty of training recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1211.5063

[29] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, Sep. 2014, pp. 3104–3112.

[30] A. Graves. (2013). "Generating sequences with recurrent neural networks." [Online]. Available:https://arxiv.org/abs/1308.0850

[31] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. INTERSPEECH*, 2012, pp. 601–608.

[32] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 3337–3345.

[33] R. Vedantam, S. Bengio, K. Murphy, D. Parikh, and G. Chechik, "Context-aware captions from context-agnostic supervision," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 251–260.

[34] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with LSTM recurrent neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3547–3555.

[35] J. R. Medel and A. Savakis. (2016). "Anomaly detection in video using predictive convolutional long short-term memory networks." [Online]. Available: https://arxiv.org/abs/1612.00390

[36] L. Chi and Y. Mu. (2017). "Deep steering: Learning end-to-end driving model from spatial and temporal visual cues." [Online]. Available: https://arxiv.org/abs/1708.03798

[37] M. D. Cerqueira *et al.*, "Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart. A statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association," *Circulation*, vol. 105, no. 4, pp. 539–542, 2002.

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[39] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.