IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 16, 2019, accepted February 9, 2019, date of publication February 15, 2019, date of current version March 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2899635

Multi-Depth Fusion Network for Whole-Heart CT Image Segmentation

CHENGQIN YE[1], WEI WANG[2], SHANZHUO ZHANG[1], AND KUANQUAN WANG[1], (Senior Member, IEEE)

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[2]School of Physics and Astronomy, The University of Manchester, Manchester M13 9PL, U.K.

Corresponding author: Kuanquan Wang (wangkq@hit.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61571165.

ABSTRACT Obtaining precise whole-heart segmentation from computed tomography (CT) or other imaging techniques is prerequisite to clinically analyze the cardiac status, which plays an important role in the treatment of cardiovascular diseases. However, the whole-heart segmentation is still a challenging task due to the characteristic of medical images, such as far more background voxels than foreground voxels and the indistinct boundaries of adjacent tissues. In this paper, we first present a new deeply supervised 3D UNET which applies multi-depth fusion to the original network for a better extract context information. Then, we apply focal loss to the field of image segmentation and expand its application to multi-category tasks. Finally, the focal loss is incorporated into the Dice loss function (which can be used to solve category imbalance problem) to form a new loss function, which we call hybrid loss. We evaluate our new pipeline on the MICCAI 2017 whole-heart CT dataset, and it obtains a Dice score of 90.73%, which is better than most of the state-of-the-art methods.

INDEX TERMS CT image segmentation, focal loss, deeply-supervised, multi-depth fusion.

I. INTRODUCTION

Whole-heart CT image segmentation refers to predicting the corresponding category for each voxel in whole-heart CT images, so as to obtain the volume and shape of all cardiac substructures, including pulmonary artery, ascending aorta, right ventricle blood cavity, right atrium blood cavity, left ventricle blood cavity, left atrium blood cavity, and left ventricular myocardium. It is one of the main steps for the diagnosis and analysis of cardiovascular diseases. By getting the whole-heart segmentation results, other functional indicators (ejection fraction, myocardial mass/movement, ventricular volume) can be obtained, which play important roles in the detection of heart failure and congenital heart malfunction [1].

In the past few years, statistical models [2], [3] and atlas-based methods [4]–[6] were widely used on the whole-heart segmentation task. In statistical models, variable parameter models of the heart structure need to be trained from the dataset, and they can be easily over-fitted when the amount of training images is small. Atlas-based methods need to align the target image to one or more template images by using a registration algorithm, then the labels of template images will transmit back to the target image to get segmentation results, which makes these methods highly rely on the accuracy of the registration algorithm. Zhuang *et al.* [7] modified the method by adjusting the weight of each template image through calculating conditional entropy to improve the accuracy. Further in 2016, Zhuang and Shen [8] applied multi-scale and multi-modality atlases to enhance the registration effectiveness. Atlas-based methods have been popular for many years, but they are usually time consuming to achieve accurate segmentation results (from a few minutes to a few hours, such as the method reported in [8] takes 12.58 minutes).

In recent years, the full convolution network (FCN) [9] has been put forward and increasing researchers have been using deep learning in image segmentation tasks. A common network used in medical image segmentation is UNET [10]. Similar to FCN, UNET is made up of an encoding module and a following decoding module, whose decoding module is highly symmetric with the encoding module. Furthermore, UNET has many remote skip connections between the encoding and decoding modules, which are designed to maintain spatial information lost in the encoding process.

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang.

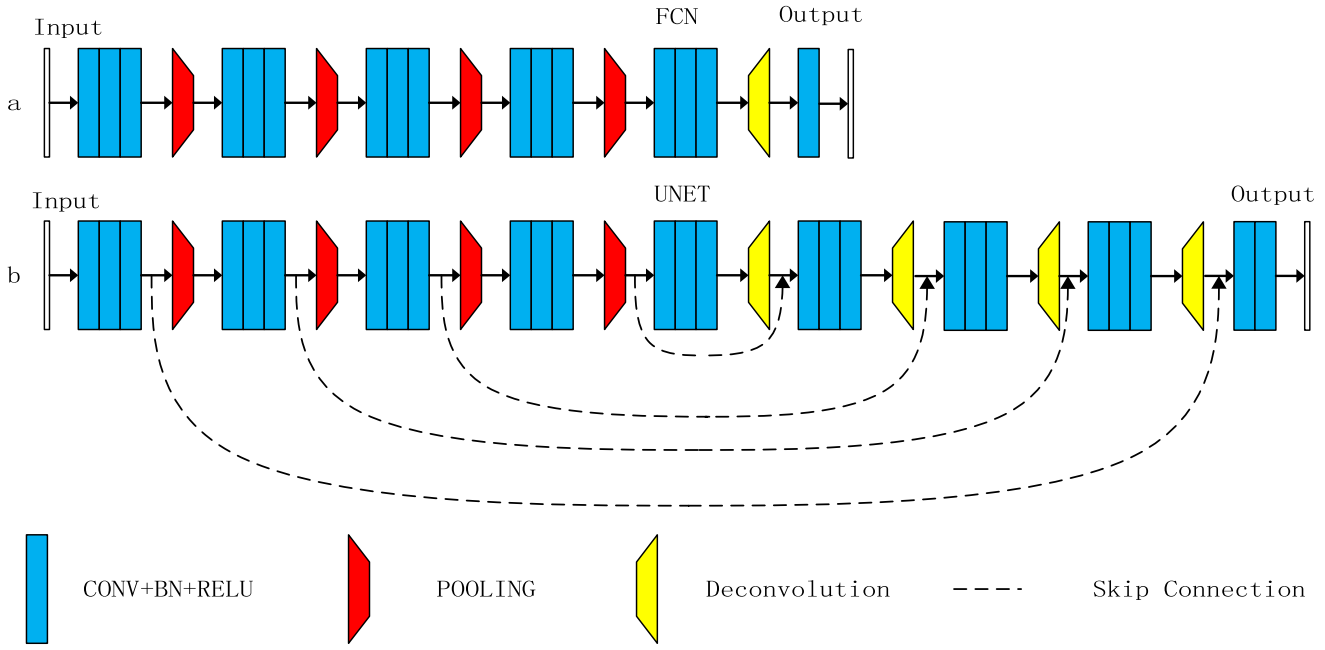

FIGURE 1. Diagrams of the structure of FCN and UNET. (a) The structure diagram of the FCN. The encoding module is composed of convolution modules and pooling modules, which constantly reduce the size of the feature map and increase the number of channels. The decoding module uses deconvolution modules to restore the feature map to the same resolution of the input image, thus obtaining the segmentation result of the input image. (b) The structure diagram of the UNET. The main difference between FCN and UNET is that the encoding module and the decoding module of UNET are highly symmetric. Between the encoding module and the decoding module exists skip connections, which are helpful to recover image information from the encoding result step by step.

The comparison of the structures of the FCN and UNET is shown in Fig. 1.

Whole-heart CT images are often analyzed slice-by-slice on 2D images [11], [12]. Its advantage is the low consumption of storage space and computing time. In addition, pre-trained nets can be easily used for fine-tuning with 2D inputs. However, using 2D images leads to the loss of context information between slices. This problem can be solved by using an LSTM [13] or CRF [14], which can reserve the slice-to-slice context information. Wang and Smedby [15] sliced the whole-heart CT images in three orthogonal directions (axial, coronal and sagittal view) and averaged these three probability graph outputs to make use of the slice-to-slice context information. Since 3D convolution methods keep the context information to the greatest extent, Dolz *et al.* [16] applied 3D CNN to the segmentation of the subcortical structure. Tong *et al.* [17] extracted interested areas through the use of the 3D UNET to focus on useful regions. Yang *et al.* [18] applied 3D fine-tuning to 3D UNET to accelerate convergence. Payer *et al.* [19] adopted a tag transformation network after the UNET network by using the context configuration of different heart substructures [19] and won the first prize of the MICCAI 2017 Multi-Modality Whole Heart Segmentation Challenge. Recently, more advanced networks for medical image segmentation such as DenseVNet [20], VoxResNet [21] and AtriaNet [22] have been proposed. Further description of these networks are given in the second section.

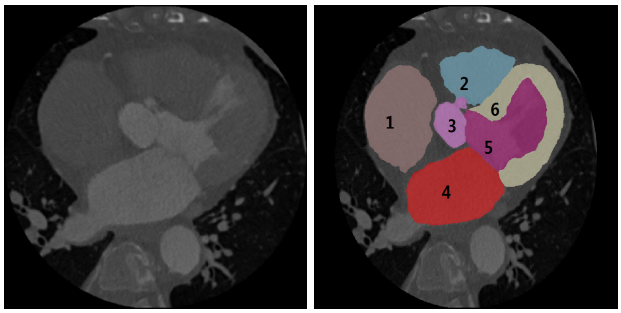

FIGURE 2. A single whole-heart CT image slice (left), and the corresponding segmentation label (right). Area 1 corresponds to right atrium blood cavity, area 2 corresponds to right ventricle blood cavity, area 3 corresponds to ascending aorta, area 4 corresponds to left atrium blood cavity, and area 5 corresponds to left ventricle blood cavity, area 6 corresponds to left ventricular myocardium. The boundaries of different regions are relatively indistinct, so these boundary pixels belong to indistinguishable pixels.

Though many advanced structures of networks have been applied to medical image segmentation, there are still many issues to be resolved. As shown in Fig. 2, the number of background voxels in CT images is far more than that of foreground voxels, and the boundaries between different adjacent tissues are relatively indistinct. In addition, we often cannot find sufficient labelled samples for training and testing networks. People often use Dice loss [23], Jaccard loss [24] or training a localization network to clip out the background region [19] to solve the problem of class imbalance

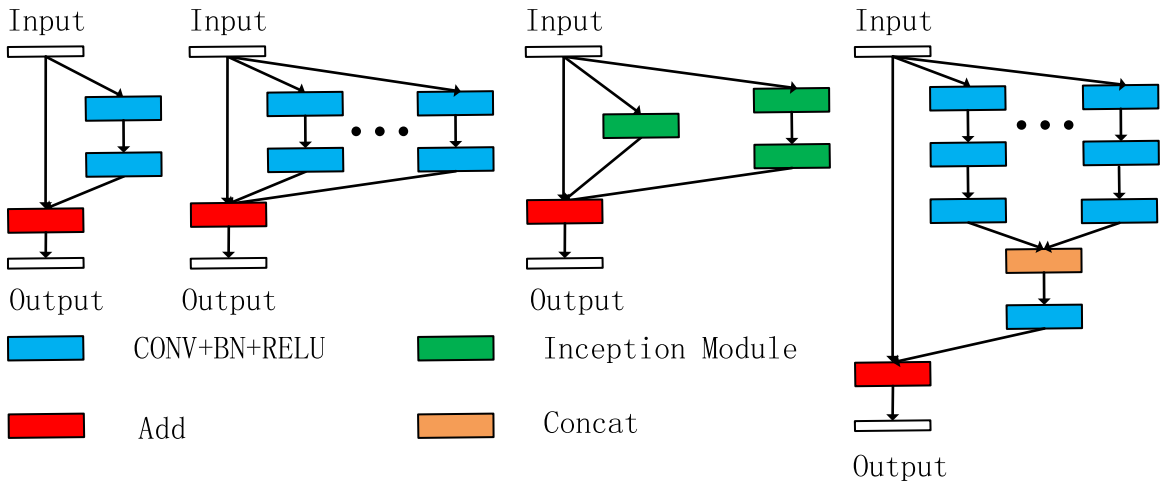

FIGURE 3. The Resnet, Multi-Resnet, PolyNet, ResNext are shown from left to right. Some studies [28] have shown that the residual network performs well is due to a large number of short-path networks. A basic block of Resnet consists of a skip connection and a residual branch. In each basic block, information can optionally be propagated through the skip connection or the residual branch. This resulting in the short path number equal to 2^N (N means to the number of residual blocks). Multi-Resnet proves that increasing the number of parallel residual branches may produce more efficient short paths. ResNext reduces the number of parameters of the multi-residual branches by using 1x1 convolution and group convolution [32]. Finally, Inspired by Inception Network [33], PolyNet shows that using multi-scale structures in multi-residual branches helps improve network performance.

in 3D medical images, and use data augmentation or deeply-supervising [17] to alleviate the lack of samples. However, these schemes do not focus on the indistinguishable boundary voxels.

Our work can be summarized as follows:

1. Based on the original 3D deeply-supervised UNET, we proposed a novel network with enhanced segmentation accuracy by continuously combining local and global features through multi-depth fusion.

2. We proposed Hybrid Loss, which incorporated Focal loss into the proposed network to make the model more focus on indistinguishable boundary voxels.

Finally, we achieved an average Dice score of 90.73% on the MICCAI 2017 whole-heart CT dataset, which is better than most of the state-of-the-art methods.

II. RELATED WORK

Building a deeper network has long been considered to play an important role in improving network performance. But a deeper network usually leads to the disappearance of gradients. Resnet [25] alleviates the gradient disappearance problem by using skip connections, and Densenet [26] by using feature reusing. They both achieved deeper networks and showed excellent results. In the field of medical image segmentation, DenseVNet [20] applies DenseNet to 3D FCN, VoxResNet [21] improves segmentation performance by applying residual modules to down-sampling modules. DRINET combines DenseNet and Inception-ResNet [27] to improve performance and becomes one of the most advanced medical image segmentation networks. However, there are also other studies showing that a deeper network is not the only way to improve performance. Recent studies have shown that ResNet is more like a collection of many shallower networks [28]. Based on this idea, many shallower networks with a multi-residual branch like Multi-ResNet [29], ResNext [30], PolyNet [31] have been developed and achieved better results than the original ResNet, which seems to prove that fusing multi-residual branches is more effective than simply increase the network depth. A schematic diagram of a multi-residual branch network is shown in Fig. 3.

On the other hand, combining the local and global information of the images can help improve the segmentation accuracy. AtriaNet [22] first randomly crops the sub-volume of the original volume data as a training sample. Then it crops a larger area centered on the training sample and inputs it into the network to provide context information of the training sample. Combining local and global information can also be achieved by fusing shallow feature maps and deep feature maps, as Densenet does. Yu *et al.* [34] claimed that only using skip connections to fuse feature maps is too rough, and proposed an architecture of hierarchical and iteration aggregation.

Our work was based on the 3D deeply-supervised UNET. We extended the multi-branch residual network and integrated multi-depth fusion to achieve feature aggregation. After that, we applied Focal loss, which could make the network pay more attention to the indistinct boundary voxels. More details are described in the following section.

III. METHOD

The overall structure of the proposed network is shown in Fig. 4. The differences between our network and the original 3D UNET can be understood in the following three main aspects:

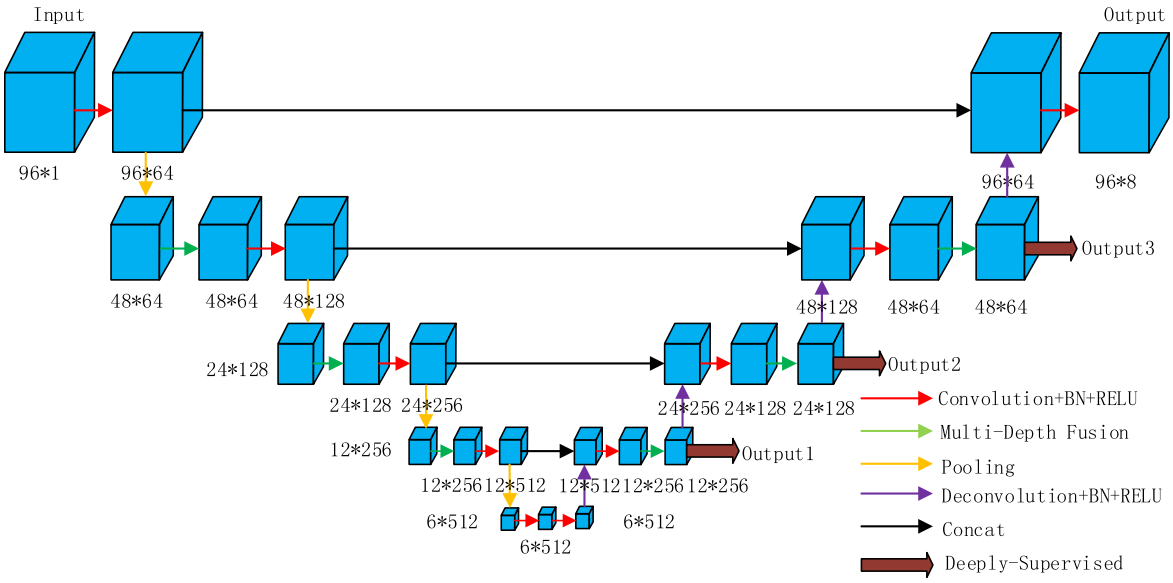

FIGURE 4. The detailed diagram of the network structure. The input layer size is a 96 × 96 × 96 × 1 four-dimensional tensor. We use *A*B* to describe the size of the feature map, where A represents feature map length/width/height and B represents the number of feature map channels. It basically follows the UNET architecture. We introduced a deeply-supervised block and multi-depth fusion block in the network structure to improve network performance.

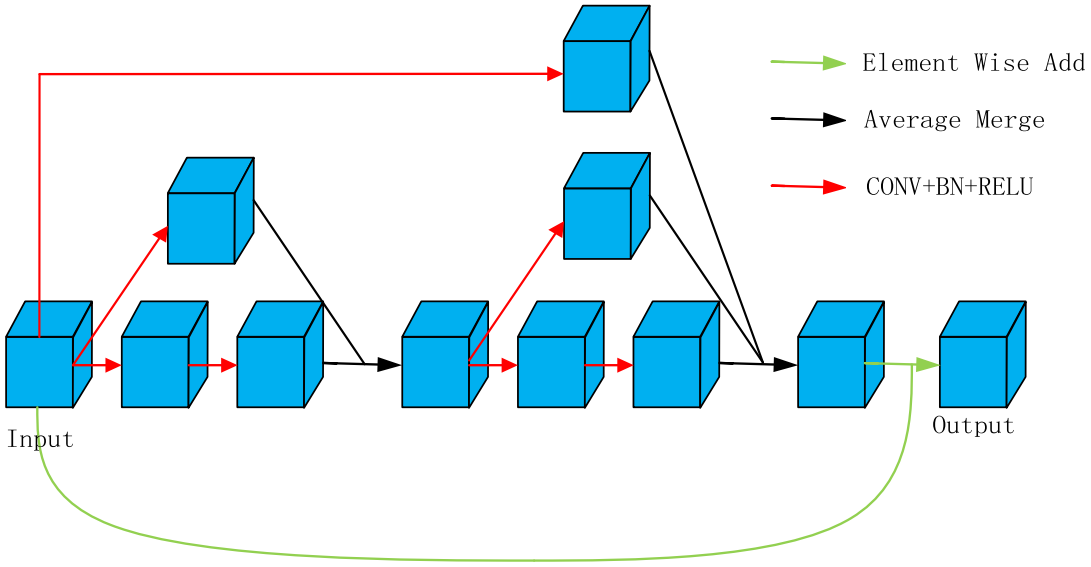

FIGURE 5. The detailed diagram of the multi-depth fusion block. The black line indicates element wise averaging operations to fuse feature maps of different depths. The green line means merging operations by element wise addition, which performs like a residual block.

A. DEEPLY-SUPERVISED MECHANISM

The deeply-supervised mechanism refers to injecting additional auxiliary predictions in the hidden layer. As shown in Fig. 4, we used a total of three deeply-supervised branches in the up-sampling module in the network. In each deeply-supervised branch, the feature map firstly expanded by deconvolution (kernel size 3 × 3 × 3) until the feature map is restored to the same resolution as the input layer. The size of the feature map is doubled after each deconvolution operation. Then we applied loss function to each branch. When training the network, we need to minimize the weighted sum of the loss function in the deep supervisory branch and the main branch.

The deeply-supervised mechanism can play a strong regularization function when the training samples are insufficient, thereby improves the generalization ability of the network and the convergence speed of the network [35].

B. MULTI-DEPTH FUSION

Inspired by Larsson *et al.* [36], we used feature maps fusion on the residual branch. The structure of the multi-depth fusion block is shown in Fig. 5. The size of convolution kernels is

set to 3×3×3 and with stride 1. Since we use the element wise averaging to fuse feature maps, ''same convolution'' was used in our network to keep the same feature map size. Theoretically, the more times we applied the convolution operation, the larger receptive field a feature map would get. For example, if we applied each 3 × 3 × 3 convolution operation with stride 1, then after two convolution operations, each point in the feature map would have a receptive field of 5 × 5 × 5. Since the shallower feature map can capture more detailed information about the image and the deeper feature map can obtain a larger receptive field to capture the context information, by continuously merging the feature maps with various depths, the new feature map would contain both local and global information. Compared to simply merge feature maps of different depths, we used an iterative layered fusion approach. This architecture ensures that detailed information from shallow feature maps is received each time when a deep feature map is generated. Finally, this combination of long and short paths is similar to deeply supervision: shorter paths can quickly obtain predictions and effectively propagate gradients, while deeper paths can achieve finer results.

C. HYBRID LOSS

1) DICE LOSS

As shown in Fig. 2, the number of background voxels in CT images is far more than that of foreground voxels. This leads to a serious category imbalance problem. This problem is generally solved by using weighted cross entropy or Dice loss. Giving the ground truth $G^{w\times h\times d}$ and the predicted probability map $P^{C\times w\times h\times d}$, where w represents the width of the volume data, h represents the height of the volume data, d represents the depth of the volume data, C represents the number of categories of the substructure, and C×w×h×d indicates that the probability map of the output is a four-dimensional tensor whose size is C×w×h×d. Since the whole heart segmentation task belongs to the multi-classification task, it is necessary to encode the ground truth $G^{w\times h\times d}$ into the ''one hot'' form of the C class to form a new ground truth $G^{C\times w\times h\times d}$, then the Dice loss can be expressed as the following form:

$$LmDSC = -\sum_{c=1}^{C} \times \frac{2\times|\sum_{x=1}^{w}\sum_{y=1}^{h}\sum_{z=1}^{d} G^{c,x,y,z}\times P^{c,x,y,z}|}{|\sum_{x=1}^{w}\sum_{y=1}^{h}\sum_{z=1}^{d} G^{c,x,y,z}|+|\sum_{x=1}^{w}\sum_{y=1}^{h}\sum_{z=1}^{d} P^{c,x,y,z}|} \quad (1)$$

where $P^{c,x,y,z}(0 \le P^{c,x,y,z} \le 1)$ refers to the value of the point (c, x, y, z) in the output probability map, and represents the probability that the point (x, y, z) in the input image belongs to the substructure of class c, and $G^{c,x,y,z}$ ($G^{c,x,y,z}$ equals 0 or 1) represents the value of the point (c, x, y, z) in the label map. If $G^{c,x,y,z}$ equals 1, then the point (x, y, z) in the input image belongs to the substructure of class c, otherwise it does not.

2) FOCAL LOSS

Dice loss can help solve the problem of category imbalance, but it does not solve the problem of indistinct boundary voxels. The intensities of voxels at the boundary of the anatomical structure (hard-to-divide voxels, HPs) may be very close, making it difficult for the network to determine their labels. On the other hand, the labels of the voxels which are far from the anatomical boundary (easy-to-divide voxels, EPs) can be easily decided. The weighted cross entropy loss commonly used in the field of image segmentation does not solve the problem of HPs well. If we use weighted cross entropy loss as the loss function, although each correctly classified voxel only produces a small loss, a large number of EPs will still produce a significant amount of loss, thus affecting the effectiveness of the network.

When the weighted cross entropy loss is applied to a multi-classification task, the loss of a voxel (x, y, z) of the original volume data can be written in the following form:

$$LmCross(x, y, z) = -\alpha c \sum_{c=1}^{C} G^{c,x,y,z} \log P^{c,x,y,z} \quad (2)$$

where α_c is the weighted coefficient which is defined to solve the problem of category imbalance. Let n_c represent the number of voxels belonging to the c^{th} category in the label image. The weighted coefficient for any category c is calculated through (3).

$$\alpha c = 1 - \frac{nc}{\sum_{c=1}^{C} nc} \quad (3)$$

As each voxel in the volume data was only assigned to a single category in the ground truth. After the ''one hot'' encoding process, the tag value corresponding to the real category is 1, and all other tag values are 0. Assuming that the real category of the voxel (x,y,z) in the ground truth is m, then the value of $G^{c,x,y,z}$ is defined as (4):

$$G^{c,x,y,z} = \begin{cases} 0 & c \ne m \\ 1 & c = m \end{cases} \quad (4)$$

Then the total loss defined by (2) can be simplified as (5).

$$LmCross(x, y, z) = -\alpha m \log P^{m,x,y,z} \quad (5)$$

Ignoring the weighted coefficient α_m, the weighted cross entropy loss calculated by (5) as the function of the predicted value $P^{m,x,y,z}$ is shown as the blue curve in Fig. 6. It can be seen that even a voxel is confirmed to be correctly classified ($P^{m,x,y,z} > 0.5$), (5) will still result in a relatively large loss. This is detrimental to the convergence of neural networks to optimal results.

Here, we proposed a multi-class Focal loss [37] to reduce the loss of EPs and named it $L_{mFOCAL}(x, y, z)$, which is

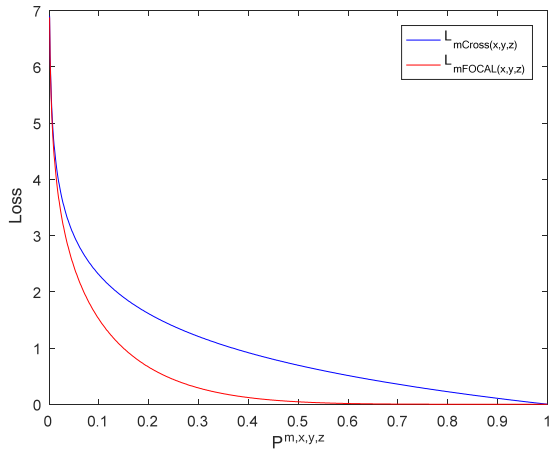

FIGURE 6. Comparison between the weighted cross entropy and Focal loss functions. Assuming that the real category of the voxel (x, y, z) in volume data is m, then in the predicted probability map $P^{m,x,y,z}$ represents the probability with which the voxel belongs to category m. The blue and red curves show the trend of L_{mCross} (x, y, z) and L_{mFOCAL} (x, y, z) as a function of $P^{m,x,y,z}$, respectively.

shown in (6).

$$LmFOCAL(x, y, z) = -\alpha m(1 - P^{m,x,y,z})^r \log P^{m,x,y,z} \quad (6)$$

Compared with the weighted cross entropy loss, Focal loss has one more adjustment term $(1\text{-}P^{m,x,y,z})^r$, where r is the focus parameter (r >=0, we set r to 4 in this work). Using the Focal loss, if an EP results in a higher $P^{m,x,y,z}$, the adjustment term $(1\text{-}P^{m,x,y,z})^r$will quickly decrease to a negligible value. And when a voxel is incorrectly classified, its $P^{m,x,y,z}$ will usually be small, then the adjustment term $(1\text{-}P^{m,x,y,z})^r$ will be close to 1 and the Focal loss will approximate the weighted cross entropy loss. An intuitive comparison is shown in Fig. 6. L_{mFOCAL} (x, y, z) greatly reduces the loss of correctly classified voxels, that is, reduces the influence of EPs and forces CNN to focus on learning features which can reduce the loss of HPs.

The Focal loss of the whole volume data is the sum of the losses of all individual voxels, as shown in (7).

$$LmFOCAL = \sum_{x=1}^{w} \sum_{y=1}^{h} \sum_{z=1}^{d} LmFOCAL(x, y, z) \quad (7)$$

The Focal loss and the Dice loss were then combined as the final loss function (shown as L in (8)). This final loss function gave consideration to both the problem of category imbalance and segmentation of the HPs.

$$L = 100 \times LmDSC + 1 \times LmFOCAL \quad (8)$$

Finally, the mixed loss function was applied to the network output and three deeply-supervised branches. Assuming that $L_{Out1\sim3}$ represents the loss function of the three deeply-supervised branches, L_{Out} represents the loss function of the network output, the loss function of the whole network is shown in (9).

$$Ltotal = LOut + 0.3 \times LOut1 + 0.6 \times LOut2 + 0.9 \times LOut3 \quad (9)$$

IV. EXPERIMENTS AND RESULTS

A. DATA

We evaluated the proposed network on the MICCAI 2017 whole-heart CT dataset, which includes 60 cardiac CT/CTA that covers the whole heart substructures (20 volume data as training dataset and the other 40 as the testing dataset). All these clinical data were obtained using conventional cardiac CT angiography at Shuguang Hospital, Shanghai, China. Each volume data covers the entire heart structure from the upper abdomen to the aortic arch. The slices were obtained in the axial view, with the in-plane resolution about 0.78 $mm \times 0.78$ mm and the average slice thickness of 1.60 mm. These data were collected in the in vivo clinical environment, so had various image qualities.

Since we cannot obtain labels of the testing dataset, as suggested in [38], we randomly selected 10 volume data as the training dataset and the other 10 volume data as the testing dataset.

B. EVALUATION

To evaluate the segmentation results, two types of measures are usually used: the Jaccard score and the Dice score. The Jaccard score is defined as the ratio of the intersection and union of the predicted and actual results. Given the set of predicted results A and actual results B, the Jaccard score can be calculated from (10).

$$Jaccard(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (10)$$

The Dice score means the Dice similarity coefficient (DSC) and measures the spatial overlap between the predicted and actual results, which can be explained as (11).

$$DSC(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (11)$$

C. EXPERIMENTAL DETAILS

We used Tensorflow to train and test the network. In order to verify the ideas of our network, we use the same partitioning method to divide training dataset and testing dataset as described in the Baseline model [38]. And we used the same way in [38] to do data augmentation: the original volume data was randomly cropped to a sub-volume of $96 \times 96 \times 96$ and rotation was applied to the sub-volume for data augmentation. All training samples were normalized to zero mean and unit variance. We used Adam as the optimizer to update weights of the network, and the batch size was set to 1. The initial learning rate was 0.001. When the accuracy could not converge, the learning rate was halved, and the final iteration was about 60,000 epochs. All experiments were performed on an Intel(R) Core(TM) i7-6800K-based workstation equipped with a 12GB NVidia Geforce Titan X.

D. EXPERIMENTAL RESULTS

To test the effectiveness of multi-depth fusion and Hybrid Loss, we first simply applied multi-depth fusion to the pipeline, then both the multi-depth fusion and the Hybrid

TABLE 1. The comparison between our model and baseline, measured by the Jaccard score.

	PUA	ASA	RVBC	RABC	LVBC	LABC	MLV	Mean
Baseline	71.1%	89.2%	65.9%	70.3%	78.7%	76.1%	69.9%	74.46%
Multi-Depth Fusion	76.3%	92.8%	80.6%	78.7%	87.5%	83.1%	78.8%	82.54%
Multi-Depth Fusion + Hybrid Loss	76.3%	93.6%	81.0%	78.6%	88.7%	84.5%	78.8%	83.07%

TABLE 2. The comparison between our model and the top seven results of MICCAI 2017 Multi-Modality Whole Heart Segmentation challenge, measured by the Dice score.

	PUA	ASA	RVBC	RABC	LVBC	LABC	MLV	Mean
1	84.0%	93.3%	90.9%	88.8%	91.8%	92.9%	88.1%	89.97%
2	83.5%	89.4%	85.7%	87.1%	92.3%	93.0%	85.6%	88.09%
3	78.4%	90.7%	88.3%	83.6%	90.4%	91.6%	85.1%	86.87%
4	80.0%	91.4%	85.6%	83.7%	90.1%	88.4%	84.6%	86.26%
5	67.7%	83.5%	80.6%	85.5%	90.8%	90.8%	87.4%	83.76%
6	69.8%	86.8%	81.0%	81.2%	89.3%	88.9%	83.7%	82.96%
7	73.7%	83.9%	84.9%	79.9%	88.0%	84.5%	81.5%	82.34%
Ours	**86.2%**	**96.7%**	**89.5%**	**87.8%**	**94.4%**	**91.6%**	**88.9%**	**90.73%**

Loss were applied. Along with the Baseline method, Jaccard scores of the three networks were calculated.

The performance of the three networks in the pulmonary artery (PUA), ascending aorta (ASA), right ventricular blood chamber (RVBC), right atrial blood chamber (RABC), left ventricular blood chamber (LVBC), left atrial blood chamber (LABC), myocardium of the left ventricle (MLV) is shown in Table 1. The results demonstrate that our network achieved significantly higher scores compared to the Baseline model.

Finally, we compared the accuracy of our network with those of the top seven participants of the MICCAI 2017 Multi-Modality Whole Heart Segmentation Challenge. Their results were extracted from Payer *et al.* [19], who won the first prize in the challenge. Measured by the Dice score, the comparison of results is shown in Table 2. The results show that we have achieved the most advanced performance.

Some of the segmentation results on the testing dataset are demonstrated in Fig. 7. Snapshots were taken using ITK-SNAP 3.8.

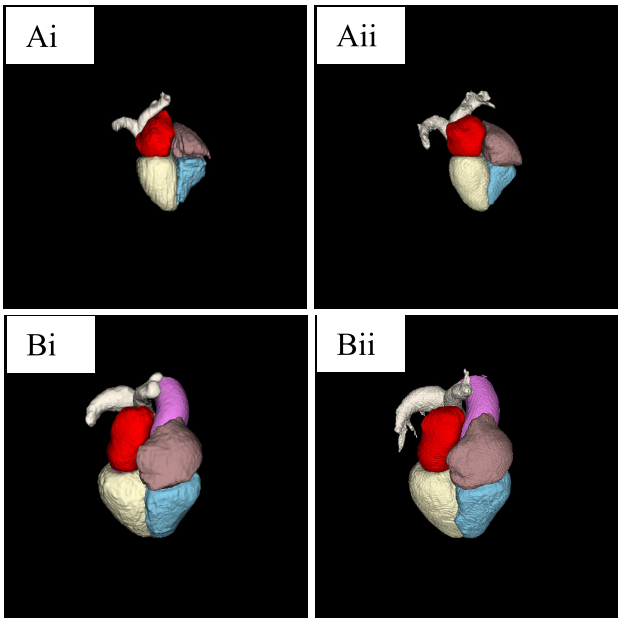

FIGURE 7. Comparison of the ground truth and segmentation results on two volume data. (Ai, Aii) The volume data with the worst segmentation results. (Bi, Bii) The volume data with the best segmentation results. (Ai, Bi) Ground truths of the two volume data. (Aii, Bii) Prediction results of the two volume data.

V. DISCUSSION

Segmenting whole-heart substructures from CT images has always been a challenging task. It is extremely difficult to identify the boundaries of different anatomical substructures because of the blurred voxels in between. In addition, the shape of the heart can be greatly changed during exercise or in diseases, which further increases the difficulty of automatic segmentation of the whole heart. Models trained according to healthy data may not be able to perform well on pathological data [1]. This boundary indistinctness and large variations in the anatomical structure in cardiac CT images force us to capture more advanced features from images such as textures to identify the boundaries. In order to address the indistinct boundaries and anatomical changes, it may be useful to obtain larger context information. In the field of natural images, Deeplab expands the receptive field by using cavity convolution to obtain a larger context [39], however, as Deeplab uses the same dilation rates in each layer, the voxels in high-level feature maps only receive information in a checkerboard fashion, and lose a large portion of information [40]. Combining multi-scale features seems to be more effective than simply expanding the filter's receptive field. Since filters located in deeper layers have greater receptive fields [41], combining feature maps of multiple depths helps to obtain multiple levels of context features. We used the idea of a multi-residual branch network and combined context information by the layered iterative fusion to build a powerful segmentation tool.

Another difficulty in applying deep learning to whole-heart segmentation is that training deep neural networks usually requires a large amount of training data. It is expensive to obtain a large number of whole-heart segmented samples because it takes a lot of time and efforts of human experts to label different structures from medical images. Under these circumstances, the deeply-supervised mechanism shows superior performance when the training data set is small. It has a regularization effect, and can alleviate the over-fitting problem. In addition, we randomly crop the sub-volume from the entire volume as the training sample and apply a rotation transform on it, which also increases the number of training samples. Another idea is to let the data determine the depth of the network. Zhou *et al.* [12] proposed a novel architecture to select appropriate depths in the UNET by pruning the neural network, in order to effectively select the optimal network structure based on the size of an existing training dataset.

For voxels that are far from the boundary, the neural network should be able to accurately determine its class based on contextual information. However, due to a large number of such voxels, the loss value will accumulate to a level big enough to hinder the convergence of the network to the optimal result. Focal loss reduces the loss of EPs and forces the network to focus on dividing indivisible voxels on boundaries without the need for additional boundary correction steps [42]. Compared to the common method which uses the localization network to crop the region of interest for segmentation [17], our method is end-to-end and suppresses category imbalance by using Hybrid Loss.

We also tried to use random paths [36] in multi-depth fusion blocks. Surprisingly, the network using random paths did not show a superior effect (the Dice score decreased by 0.17% in the case of the same training hyper-parameters and number of iterations). A possible reason is that the random path destroys the collaboration of different paths to fuse features. Due to the limit of storage space, we have only tried multiple depth fusion in the up sampling phase and down sampling phase. It may be effective to introduce the multi-depth fusion in the UNET skip connections phase.

Our approach is generic and can be migrated to tasks similar to the whole-heart segmentation, such as the segmentation of brain structures.

VI. CONCLUSION

In this study, we have developed and evaluated a new deeply-supervised UNET for robust whole-heart segmentation from CT images. The network has three key features: the deeply-supervised mechanism, multi-depth fusion blocks, and Hybrid Loss. Our method can generate more advanced features by continuously fusing local and global information, and reduces the loss of EPs, which allows the network to focus on dividing indistinct borders.

We have applied the segmentation network on the MICCAI 2017 whole-heart CT dataset. Results showed that it is superior to most of advanced CNNs proposed recently, and there is a significant improvement in the segmentation of the pulmonary artery, ascending aorta and right ventricular blood cavity. It will surely contribute to the establishment of more robust and accurate whole-heart segmentation methods and assist in the diagnosis and treatment of patients with heart diseases.

ACKNOWLEDGMENT

(Chengqin Ye and Wei Wang are co-first authors.)

REFERENCES

[1] X. Zhuang, "Challenges and methodologies of fully automatic whole heart segmentation: A review," *J. Healthcare Eng.*, vol. 4, no. 3, pp. 371–407, 2013.

[2] K. O. Babalola, T. F. Cootes, C. J. Twining, V. Petrovic, and C. Taylor, "3D brain segmentation using active appearance models and local regressors," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2008, pp. 401–408.

[3] A. Rao, P. Aljabar, and D. Rueckert, "Hierarchical statistical shape analysis and prediction of sub-cortical brain structures," *Med. Image Anal.*, vol. 12, no. 1, pp. 55–68, 2008.

[4] W. Bai, W. Shi, C. Ledig, and D. Rueckert, "Multi-atlas segmentation with augmented features for cardiac MR images," *Med. Image Anal.*, vol. 19, no. 1, pp. 98–109, 2015.

[5] M. P. Heinrich and J. Oster, "MRI whole heart segmentation using discrete nonlinear registration and fast non-local fusion," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, 2017, pp. 233–241.

[6] G. Galisot, T. Brouard, and J.-Y. Ramel, "Local probabilistic atlases and *a posteriori* correction for the segmentation of heart images," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, 2017, pp. 207–214.

[7] X. Zhuang *et al.*, "Multiatlas whole heart segmentation of CT data using conditional entropy for atlas ranking and selection," *Med. Phys.*, vol. 42, no. 7, pp. 3822–3833, 2015.

[8] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Med. Image Anal.*, vol. 31, pp. 77–87, Jul. 2016.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[10] 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.*, 2015, pp. 234–241.

[11] A. Mortazi, J. Burt, and U. Bagci, "Multi-planar deep segmentation networks for cardiac substructures from MRI and CT," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, 2017, pp. 199–206.

[12] SpringerLink. *UNet++: A Nested U-Net Architecture for Medical Image Segmentation*. Accessed: Jan. 9, 2019. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-00889-5_1

[13] A. A. Novikov, D. Major, M. Wimmer, D. Lenis, and K. Bühler, "Deep sequential segmentation of organs in volumetric medical scans," *IEEE Trans. Med. Imag.*, to be published.

[14] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating FCNNs and CRFs for brain tumor segmentation," *Med. Image Anal.*, vol. 43, pp. 98–111, Jan. 2017.

[15] C. Wang and Ö. Smedby, "Automatic whole heart segmentation using deep learning and shape context," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, 2017, pp. 242–249.

[16] J. Dolz, C. Desrosiers, and I. B. Ayed, "3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study," *NeuroImage*, vol. 170, pp. 456–470, Apr. 2018.

[17] Q. Tong, M. Ning, W. Si, X. Liao, and J. Qin, "3D deeply-supervised U-net based whole heart segmentation," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, 2017, pp. 224–232.

[18] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "3D convolutional networks for fully automatic fine-grained whole heart partition," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, 2017, pp. 181–189.

[19] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Multi-label whole heart segmentation using CNNs and anatomical label configurations," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, 2017, pp. 190–198.

[20] E. Gibson *et al.*, "Automatic multi-organ segmentation on abdominal CT with dense v-networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1822–1834, Aug. 2018.

[21] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage*, vol. 170, pp. 446–455, Apr. 2017.

[22] Z. Xiong, V. V. Fedorov, X. Fu, E. Cheng, R. Macleod, and J. Zhao, "Fully automatic left atrium segmentation from late gadolinium enhanced magnetic resonance imaging using a dual fully convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 515–524, Feb. 2018.

[23] 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. (3DV)*, 2016, pp. 565–571.

[24] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269.

[27] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. (Feb. 2016). "Inception-v4, inception-ResNet and the impact of residual connections on learning." [Online]. Available: https://arxiv.org/abs/1602.07261

[28] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Advances in Neural Information Processing Systems*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2016, pp. 550–558.

[29] M. Abdi and S. Nahavandi. (Sep. 2016). "Multi-residual networks: Improving the speed and accuracy of residual networks." [Online]. Available: https://arxiv.org/abs/1609.05672

[30] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5987–5995.

[31] X. Zhang, Z. Li, C. C. Loy, and D. Lin, "PolyNet: A pursuit of structural diversity in very deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3900–3908.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.

[33] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[34] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[35] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Artif. Intell. Stat.*, 2015, pp. 562–570.

[36] G. Larsson, M. Maire, and G. Shakhnarovich. (May 2016). "FractalNet: Ultra-deep neural networks without residuals." [Online]. Available: https://arxiv.org/abs/1605.07648

[37] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[38] SpringerLink. *Hybrid Loss Guided Convolutional Networks for Whole Heart Parsing*. Accessed: Jan. 9, 2019. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-75541-0_23

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (Dec. 2014). "Semantic image segmentation with deep convolutional nets and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1412.7062

[40] *Understanding Convolution for Semantic Segmentation*. Accessed: Jan. 9, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8354267

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[42] B. Hou, G. Kang, N. Zhang, and C. Hu, "Robust 3D convolutional neural network with boundary correction for accurate brain tissue segmentation," *IEEE Access*, vol. 6, pp. 75471–75481, 2018.

CHENGQIN YE is currently pursuing the M.S. degree with the Research Center of Perception and Computing, School of Computer Science and Technology, Harbin Institute of Technology, China. His research interests include medical image processing, pattern recognition, and deep learning.

WEI WANG received the B.S. and M.S. degrees in computer science from the Harbin Institute of Technology, China. She is currently pursuing the Ph.D. degree in biophysics with The University of Manchester, U.K. Her research interests include computational modeling of cardiac electrophysiology, medical image processing, and the simulation studies of cardiac arrhythmia.

SHANZHUO ZHANG is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. From 2014 to 2017, he visited The University of Manchester, U.K., as a Visiting Research Student. His research interests include medical image processing, in silico modeling of ion channels in cardiomyocytes, using computer models to investigate the interaction of ion channels and drugs, and building a computational systems for cardiac electrophysiological analyses.

KUANQUAN WANG (M'01–SM'07) was the Associate Dean of the School of Computer Science and Technology, Harbin Institute of Technology (HIT), Harbin, and the Dean of the School of Computer Science and Technology, HIT, Weihai, from 2011 to 2014. He is currently a Full Professor and a Ph.D. Supervisor with the School of Computer Science and Technology and the Director of the Research Center of Perception and Computing, HIT. He has published over 300 papers and six books and holds more than ten patents. His main research areas include image processing and pattern recognition, biometrics, biocomputing, modeling and simulation, virtual reality, and visualization. He is a Senior Member of the China Computer Federation, the ACM, and the Chinese Society of Biomedical Engineering. He received the Second Prize in National Teaching Achievement.