IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 12, 2020, accepted February 29, 2020, date of publication March 31, 2020, date of current version May 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984630

Convolutional-Neural-Network-Based Approach for Segmentation of Apical Four-Chamber View from Fetal Echocardiography

LU XU[1,2,3,4], MINGYUAN LIU[1,2,3,4], JICONG ZHANG[1,2,3,4], AND YIHUA HE[5]

[1]School of Biological Science and Medical Engineering, Beihang University, Beijing 100083, China
[2]Hefei Innovation Research Institute, Beihang University, Hefei 230013, China
[3]Beijing Advanced Innovation Centre for Biomedical Engineering, Beihang University, Beijing 100083, China
[4]Beijing Advanced Innovation Centre for Big Data-Based Precision Medicine, Beihang University, Beijing 100083, China
[5]Department of Ultrasound, Beijing Anzhen Hospital, Capital Medical University, Beijing 100069, China

Corresponding authors: Jicong Zhang (jicongzhang@buaa.edu.cn) and Yihua He (yihuaheecho@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFF0201002, in part by the National Natural Science Foundation of China under Grant 61301005 and Grant 61572055, and in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2019-044.

ABSTRACT An apical four-chamber (A4C) view from early fetal echocardiography is an extremely significant step in early diagnosis and timely treatment of congenital heart diseases. The objective is to perform automated segmentation of cardiac structures, namely, the epicardium, left ventricle, left atrium, descending aorta, right atrium, right ventricle, and thorax, in ultrasound A4C views in one shot in order to assist clinicians in prenatal examination. However, such a segmentation task is often faced with the following challenges: 1) low imaging resolution; 2) incomplete tissue boundary; 3) overall contrast of the image. To address these issues, in this study, we propose a cascaded U-net, named CU-net, with structural similarity index measure (SSIM) loss. First, the CU-net with two branch supervisions helps gain clear tissue boundaries and alleviate the gradient vanishing problem caused by increasing network depth. Second, between-net connections in the CU-net can transmit the prior information from the shallow layer to the deeper layer and obtain more refined segmentation results. Third, the method leverages on SSIM loss to preserve fine-grained structural information and obtain clear boundaries. Extensive experiments on a dataset of 1712 A4C views demonstrate that the proposed method achieves a high dice coefficient of 0.856, Hausdorff distance of 3.33, and pixel accuracy of 0.929, revealing its effectiveness and potential as a clinical tool.

INDEX TERMS Semantic segmentation, convolutional neural network, apical four-chamber (A4C) view, fetal echocardiography.

I. INTRODUCTION

Congenital heart diseases (CHDs) are a series of deformities in the fetal heart structure or function, accounting for functional heart incapacitation, which may result in severe physiology defects [1]–[3]. If CHDs cannot be treated in time, the morbidity and mortality rates of neonates will be high [4]–[6]. Hence, early diagnosis and screening for pregnant women is crucial.

Fetal echocardiography is an elementary low-cost method that does not use radiation and is widely used to detect CHDs by reflecting real-time structures. An apical four-chamber (A4C) view is one of the most important ultrasonic views in fetal echocardiography [7]–[9], because plenty of CHDs could be clearly identified in this view. In prenatal ultrasound examination of CHDs, the diagnostic anatomical structures of A4C views are epicardium (EP), thorax, left ventricle (LV), left atrium (LA), descending aorta (DAO), right atrium (RA), and right ventricle (RV) [2]–[4].

The interpretation of A4C views requires clinicians to have rich theoretical knowledge and clinical experience [10]–[13]. However, doctors may make incorrect decision due to fatigue in long-term diagnosis [14]–[16]. With the development of computer technology, computer-aided diagnosis plays an indispensable role in medical image enhancement, segmentation, and recognition [17]. Accurate segmentation of A4C views can provide pathological information and save clinicians considerable time in observation and measurement.

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

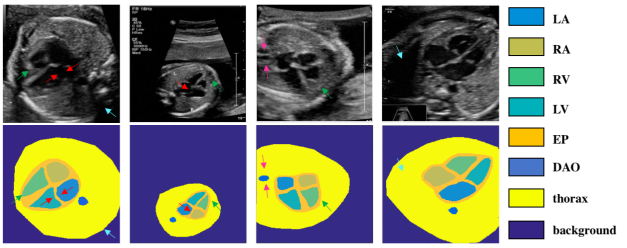

FIGURE 1. Examples of A4C views and their ground truth. The grids in the right denote different structures in labels, distinguished by different colors. The blue arrow indicates the shadow; red arrows indicate the incomplete boundaries of the septum; the green arrow indicates tendinous chordae; and the pink arrows indicate the DAO and pulmonary vein.

For example, from the segmentation of LV and LA, the presence of left ventricular dysplasia can be detected. Accurate A4C segmentation can not only help imaging experts and clinicians avoid medical accidents but also is a key step to prevent future risks of pregnant women and fetuses.

Convolutional neural network (CNN) approaches have achieved state-of-the-art performance in the field of medical image processing. Powered by CNNs, the performance of segmentation has been largely improved, such as in computed tomography (CT) and magnetic resonance imaging (MRI) [18]–[20]. In this work, we try to utilize the CNN for ultrasound image processing to automatically segment RA, RV, LA, LV, DAO, EP, and thorax from A4C views.

The following three main challenges exist in A4C view segmentation: (i) Ultrasound images have low resolution and noise, which result in large artifacts in the processed images, leading to great interference in the segmentation task, such as the blue arrows in Fig. 1. (ii) The boundaries of tissues are incomplete in echocardiography images. For example, because of echo dropout, the mitral valve and tricuspid valve in A4C views may be incomplete, as demonstrated by the red arrows in the first column of Fig. 1. Moreover, the openness of the interventricular septum and atrial septum may lead to incomplete boundaries. The cardiac tendinous chordae in ventricles of heart will blur the boundary, as demonstrated by the green arrows in Fig. 1. These phenomena will render segmentation of the ventricular-atrial boundary difficult. (iii) A4C view segmentation needs to consider the overall contrast of the whole image, rather than local or pixel features. The DAO and pulmonary vein are shown by the pink arrows in the third column of Fig. 1. To obtain accurate segmentation results, the segmentation algorithm must learn the global structural information of the whole image.

To overcome the aforementioned challenges, we propose cascaded U-nets (CU-net) with a structural similarity index measure (SSIM) loss function and achieve automated semantic segmentation of the seven structures of fetal heart simultaneously. The proposed CU-net comprises double U-nets within an integrated end-to-end framework. To obtain clear tissue boundaries and mitigate the problem of disappearance of gradients due to increased network depth, we add branch supervision during the training process. To reduce information loss in deeper layers, we design between-net connections to help transmit high-resolution information from shallow layers to the corresponding deeper layers, thus obtaining more refined segmentation results. Furthermore, we present an SSIM loss function to model the spatial information and help the optimization focus on boundaries.

In this study, we proposed a novel image segmentation network, CU-net with the SSIM loss function as a method to achieve automated semantic segmentation. Our experimental results show that this method performs considerably better than other methods in terms of Dice score (DSC), Hausdorff distance (HF), and Pixel accuracy (PA). We improved an end-to-end network, CU-net, by adding branch supervisions and between-net connections for accurate segmentation of the seven structures of fetal heart. Branch supervisions utilize the strategy of coarse-to-fine segmentation. Between-net connections can transmit the prior information from the shallow layer to the deeper layer and obtain more refined segmentation results. Further, we applied the SSIM loss function to ultrasound fetal multi-tissue segmentation and found that it can introduce global information and help the optimization focus on tissue boundaries. This proves the potential and effectiveness of the SSIM loss function in segmentation.

II. RELATED WORK

A. CNNs FOR MEDICAL IMAGE SEGMENTATION

CNNs have been widely utilized in medical image processing [5]. The U-net has been proposed by O. Ronneberger in 2015 to solve the segmentation problem in more complicated scenarios by benefitting from the accumulation of images and higher computational capacity [21].

Since this study, the U-net has been widely used in medical image segmentation [6], [7]. Many improvements of the U-net have also been derived, such as the H-DenseUNet for liver segmentation [22] and the coarse-to-fine U-net for left atrium segmentation [23]. Furthermore, stacked U-net with multiple U-nets has been proposed [8], [9], which helps increase the network depth and the number of trainable parameters, thereby improving the network performance.

Another improvement of the U-net in medical image segmentation is cascaded U-nets, proposed for brain tumor segmentation [38], prostate segmentation [39], and glioma segmentation [40]. For example, [39] proposed two dense U-nets for prostate MRI segmentation. However, excessive network length may result in the appearance of grades in practical training. To address this problem, a skip connection between two U-nets (from the decoder layer of the first u-net to the encoder layer of the second one) was proposed [38], which has also been mentioned in [37]. In this study, we proposed a new between-net connection and compared the two connections in Section V.

B. LOSS FUNCTION FOR CNNs

In most existing segmentation methods, the model is completely monitored by local loss functions at the pixel level,

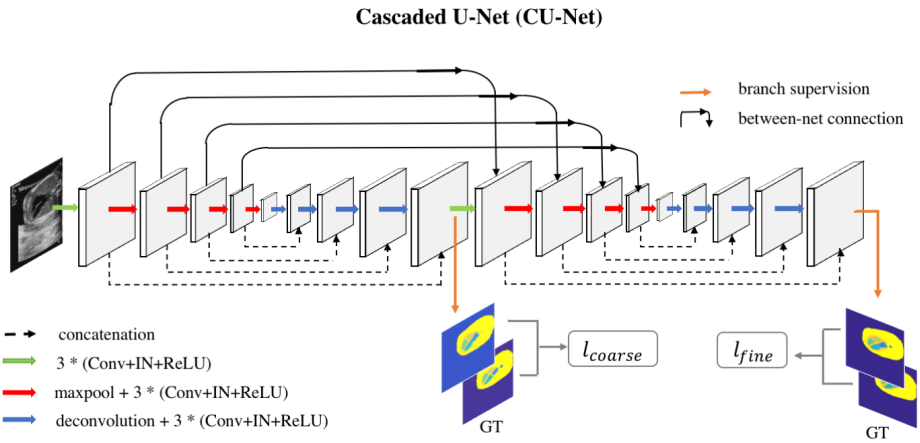

FIGURE 2. Network structures of CU-net, an end-to-end coarse-to-fine segmentation network.

such as cross-entropy loss and dice loss, without utilizing the global dependence and structure information of the output space. Therefore, the global information of prediction results often does not consist of shape priors of the target. The SSIM, originally proposed for image quality assessment [24], measures the similarity between two images. This index defines structural information as independent of brightness and contrast from the perspective of image composition, reflecting the attributes of the object structure in the scene. Hang Zhao *et al.* were the first to use the SSIM in natural image reconstruction [25]. The author believes that the loss of traditional mean squared error (MSE)-based images cannot express the intuitive sense of the human visual system about images. Therefore, because the SSIM combines brightness, contrast, and structural information of an image, it is designed as a loss function. Xuebin Qin *et al.* applied the SSIM loss function to the second-class segmentation of natural images [26].

C. CARDIAC IMAGE SEGMENTATION

With the continuous development of deep learning, CNN models have shown significant advantages in computer vision and image processing problems. Automatic segmentation of cardiac images by CNNs has attracted increasing attention.

Accurate segmentation of cardiac chambers is crucial for diagnosis and prognosis of cardiac diseases. In the recent years, more studies have focused on the segmentation of the left ventricle to calculate clinical indicators of patients, such as left ventricular mass and ventricular volume [27]–[29], while some have also considered right ventricle segmentation [30], [31] to quantify clinical indicators such as ejection fraction. In [32], segmentation of two or four chambers of the heart was performed considering different views. However, all these studies [27]–[32] focus on MRI segmentation. Furthermore, in another study [33], left ventricle segmentation of patients was performed using three-dimensional ultrasound images. As for the segmentation of fetal cardiac structures, Li Yu *et al.* only segmented left ventricle in fetal echocardiographic sequences [34]. In [37], we proposed a DW-net, comprising a dilated convolutional chain (DCC) and a W-net, for A4C segmentation with a dataset of 895 A4C views. This method has the potential to accurately segment complex ultrasound multi-structured images when the data are not large.

III. METHOD

Our proposed method is capable of segmenting seven crucial anatomical structures in A4C views. The architecture of the proposed method consists of a CU-net (see Section 3.1), as illustrated in Fig. 2, and a novel loss function (see Section 3.2).

A. DESIGN OF THE CU-NET

As Fig. 2 shows, our segmentation network is a novel end-to-end architecture, mainly composed of two cascaded U-nets, designed to take advantage of coarse-to-fine segmentation. The first-stage U-net performs a coarse segmentation and sends the extracted features to the second-stage U-net for further precise segmentation.

The cascaded structure multiplies the network depth and enhances the ability of the method to extract semantic features. However, deep network may exacerbate the gradient vanishing problem. This may lead to total loss information being lost in long-distance propagation. Therefore, to address this issue, we add an auxiliary supervision of the first U-net. Each U-net has a loss of output: the first U-net has a coarse loss, while the second has a fine loss. In addition, we redesign an inter-network connection. In previous research [37], as mentioned in Section II, we proposed between-net connections (as BNC_DE) from the decoder layers of the first U-net to the encoder layers of the second one. In this study, for the better use of priors, we build between-net connections (as BNC_EE) from the encoder layers of the first U-net to the encoder layers of the second one. We do not

connect from the first decoder because the BNC_EE enables prior information of shallow layers to be preserved and transferred to deeper layers to describe the details of the heart structure, thereby achieving more accurate segmentation.

Each U-net is a typical encoder–decoder neural network structure. Every layer of the encoder comprises three convolution operations, followed by instance normalization, a ReLU activation function, and a max-pooling operation of stride 2. In addition, each encoder layer has 20 convolution filters of size 3×3 and stride 2. The decoder is symmetrical to the encoder, and skip connection is used to connect the feature maps of the encoder with the feature maps of the decoder. The final outputs of both U-nets are produced by a softmax layer.

B. SSIM LOSS FUNCTION

The CU-net is an end-to-end architecture in which two U-nets are trained jointly to ensure the efficiency of data processing. Our training loss is defined as the summation of the outputs of both U-nets:

$$L = \alpha * L_{coarse} + \beta * L_{fine} \tag{1}$$

where α and β are the weighted coefficients; L_{coarse} is the loss between the output of the first U-net and the target;L_{fine} is the loss between the output of the second U-net and the target.

In [25], [26], the SSIM captures luminance, contrast, and structural information of images. Therefore, we integrate it into our training losses to learn the contrast and structural information of the apparent facts of the object.

In previous studies [24], because of image reconstruction, the mean and variance of the whole map often change dramatically over its span. Hence, a sliding window is used to calculate the SSIM of patches under the sliding window with a step size 1, and then the average value is taken as the SSIM of the whole map. Let $\mathrm{S} = \left\{\mathrm{S_{ij}} : \mathrm{j} = 1, \ldots, \mathrm{W}^2\right\}$ and $\mathrm{G} = \left\{\mathrm{G_{ij}} : \mathrm{j} = 1, \ldots, \mathrm{W}^2\right\}$ be corresponding patches cropped from the segmentation result and the ground truth, respectively. Here ii is the number of segmented categories, and W^2 is the size of the patch. The SSIM of S and G is defined as follows:

$$\begin{aligned} SSIM\,(S, G) &= \frac{2\mu_{\mathrm{S}}\mu_{\mathrm{G}} + \mathrm{C}_1}{\mu_{\mathrm{S}}^2 + \mu_{\mathrm{G}}^2 + \mathrm{C}_1} * \frac{2\delta_{\mathrm{SG}} + \mathrm{C}_2}{\delta_{\mathrm{s}}^2 + \delta_{\mathrm{G}}^2 + \mathrm{C}_2} \\ &= l\,(S, G) * cs(S, G) \end{aligned} \tag{2}$$

where μ_{S} and μ_{G} are the means of S and G, δ_{S} and δ_{G} are the standard deviations of S and G, δ_{SG} is their covariance, $\mathrm{C}_1 = 0.01^2$, and $\mathrm{C}_2 = 0.02^2$.

Hence, SSIM loss is defined as follows:

$$L^{ssim} = \sum\nolimits_{\mathrm{c}=1}^{\mathrm{C}} \left[\frac{1}{N}\sum_{n=1}^{N} 1 - SSIM_{cn}(S, G)\right] \tag{3}$$

where c is the number of segmented categories and N is the number of patches.

When the SSIM is used as a measure of image reconstruction, the Gaussian filter is often used to calculate the mean and variance of the image. We use mean filtering to calculate the mean and variance of each patch in SSIM.

Analogous to (3), we may write:

$$\begin{aligned} \frac{\partial L^{ssim}}{\partial S\,(x)} = \frac{1}{N}\sum_{\mathrm{c}=1}^{\mathrm{C}}\sum_{n=1}^{N} -\frac{\partial SSIM_{cn}\,(S, G)}{\partial x} = \frac{1}{N}\sum\nolimits_{\mathrm{c}=1}^{\mathrm{C}}\sum\nolimits_{n=1}^{N} \\ -\left(\frac{\partial l(S, G)}{\partial x} * cs\,(S, G) + l\,(S, G) * \frac{\partial cs(S, G)}{\partial x}\right) \end{aligned} \tag{4}$$

$l(S, G)$ and $cs\,(S, G)$ are the terms of the SSIM (Equation), and their derivatives are, respectively,

$$\frac{\partial l(S, G)}{\partial S(x)} = 2 * \left(\frac{\mu_{\mathrm{G}} - \mu_{\mathrm{S}} * l(S, G)}{\mu_{\mathrm{S}}^2 + \mu_{\mathrm{G}}^2 + \mathrm{C}_1}\right) \tag{5}$$

and

$$\begin{aligned} \frac{\partial cs(S, G)}{\partial S(x)} = \frac{2}{\delta_{\mathrm{s}}^2 + \delta_{\mathrm{G}}^2 + \mathrm{C}_2} \\ * [(G(x) - \mu_{\mathrm{G}}) - cs(S, G) * (S(x) - \mu_{\mathrm{S}})] \end{aligned} \tag{6}$$

In SSIM loss, the mean, standard deviation, and covariance are used as the estimate of brightness, contrast, and structural similarity, respectively.

In our method, the two U-nets use the SSIM loss function as their loss function, which is defined as follows:

$$\mathrm{L_{total}} = \alpha * \mathrm{L_{coarse}^{ssim}} + \beta * L_{\mathrm{fine}}^{\mathrm{ssim}} \tag{7}$$

IV. EXPERIMENT

A. DATASET

The dataset used in this research is provided by the echocardiography department of Beijing Anzhen Hospital, Capital Medical University, Beijing, China. This clinical center specializes in the detection and treatment of fetal congenital heart diseases. Because the hospital collects a large number of data on various complicated cases from all over the country, the dataset is representative and universal. The dataset employed in the research comprises 1712 fetal A4C views. The segmentation ground truth was labeled by experienced doctors from the echocardiography department of the hospital according to clinical criteria.

Each label contains seven structures: left atrium (LA), right atrium (RA), left ventricle (LV), right ventricle (RV), epicardium (EP), descending aorta (DAO), and thorax. We divide the training set and testing set in a 3:1 ratio. We randomly selected 1284 fetal A4C views as the training set and the remaining other 428 images as the testing set, and there is no overlap between the two sets. Each image is 256×256 pixels.

B. EVALUATION CRITERIA

We use three measures to evaluate our method: dice coefficient (DSC), Hausdorff distance (HF), and pixel-level accuracy (PA). The DSC of one class is defined as

$$DSC = \frac{2|P_c \cap Q_c|}{|P_c| + |Q_c|} \tag{8}$$

where c is the category of segmentation, P_c is the automated segmentation map of class c, and Q_c is the ground truth of

class c. The range of DSC is [0,1], with the maximum and minimum values being 1 and 0, respectively.

The HF of one class is defined as

$$\mathrm{HF} = \max\{h\left(P_c, Q_c\right), h(Q_c, P_c)\} \tag{9}$$

where P_c is the pixel set of the automated segmentation map of class c, Q_c is the pixel set of the ground truth of class c, and $h\left(P_c, Q_c\right)$ and $h\left(Q_c, P_c\right)$ are given as

$$h\left(P_c, Q_c\right) = max_{(p\in P_c)} min_{(q\in Q_c)} \|\mathrm{p\text{-}q}\| \tag{10}$$

and

$$h\left(Q_c, P_c\right) = max_{(q\in Q_c)} min_{(p\in P_c)} \|\mathrm{q\text{-}p}\| \tag{11}$$

where $\|\cdot\|$ is the Euclidean distance between points q and p. Smaller the value of HF, the more accurate the segmentation results are.

The average PA is defined as

$$PA = \frac{1}{C}\sum_{c=1}^{C}\frac{p_c}{q_c} \tag{12}$$

where c is the number of the class, $\mathrm{p_c}$ denotes the amount of right classified pixels of class c, and q_c denotes all pixels of class c. The range of PA is [0,1], with the maximum and minimum values being 1 and 0, respectively.

In addition, we use a paired t-test to compare the performance between the two methods. As the test sample, the results of seven structure regions obtained from the two experiments are considered. The null hypothesis is that there is no statistical difference between the results of two experiments. A p-value of less than 0.05 indicates a significant difference between the two experiments, while that less than 0.01 indicates a highly significant difference between the two experiments.

C. IMPLEMENTATION DETAILS

The experiment was implemented using Python 3.5 and Tensorflow framework [36], and the hardware used was NVIDIA Tesla K80 GPU.

Our method directly handles clinical A4C views without any data augmentation. We employed the Adam optimization strategy in the training process. The initial learning rate was 0.0004 with a weight decay of 0.1 per 1000 iterations. We trained 100 epochs. As for the loss function, because we used the same dice loss or SSIM loss function for both supervision branches, with the same order of magnitude, we set an equal weight of 1 for both α and β.

V. RESULTS

A. RESULTS OF SEGMENTATION

We performed the experiments using three different network structures, namely the FCN [35], U-net [21], and proposed method, CU-net, to segment seven structures in A4C views, which are EP, LV, LA, DAO, RA, RV, and thorax. We first trained the three methods with dice loss and named them FCN+l^{dice}, U-net+l^{dice}, and CU-net+l^{dice}, respectively. Then, we trained these methods with SSIM loss and named them FCN+l^{ssim}, U-net+l^{ssim}, and CU-net+l^{ssim}, respectively. We designed the patch size of SSIM loss to be 256, and the number of patches for each structure in the loss function was 1. The reason to adopt this value will be discussed at the end of this section.

TABLE 1. Mean DSC, mean HF, and mean PA between the proposed method and the competitive methods. l^{dice} denotes that a model is trained under the dice loss function. l^{ssim} denotes that a model is trained under the SSIM loss function.

	Mean DSC	Mean HF	PA
FCN+l^{dice}	0.756±0.159	3.906±0.949	0.895±0.052
FCN+l^{ssim}	0.777±0.140	3.830±0.918	0.903±0.047
U-net+l^{dice}	0.834±0.117	3.512±0.870	0.918±0.044
U-net+l^{ssim}	0.841±0.113	3.436±0.870	0.923±0.042
CU-net+l^{dice}	0.845±0.109	3.390±0.862	0.923±0.039
CU-net+l^{ssim}	**0.856±0.096**	**3.311±0.805**	**0.929±0.037**

Table 1 presents the PA as well as the mean DSC and mean HF, which are the means of the corresponding values for seven structures. Tables 2 and 3 illustrate the DSC and HF values of seven structures, respectively. The CU-net with SSIM loss obtained an average DSC of 0.856±0.096, average HF of 3.311±0.805, and average PA of 0.929±0.037, indicating high performance in terms of all evaluation metrics.

As shown in Tables 1, 2, and 3 (from the 2nd row, 4th row, and 6th row), we can see that with the same dice loss, the CU-net has a superior segmentation performance. The CU-net significantly outperforms the FCN by 8.9% average DSC, 51.6% average HF, and 2.8% PA. In addition, the CU-net significantly outperforms the U-net by 1.1% average DSC, 11.2% average HF, and 0.5% PA.

From Tables 1–3 (comparing 2nd row with 3rd row, 4th row with 5th row, and 6th row with 7th row), we observe that incorporating SSIM loss into the neural network segmentation models significantly improves results. The CU-net with SSIM loss outperforms the CU-net with dice loss by 1.1% average DSC, 7.9% average HF, and 0.6% PA.

B. VISUALIZATION RESULTS

To further understand the origin of the performance gain, we visualized the segmentation results of some subjects. Fig. 3 shows three example cases for comparing segmentation performances of different networks and application of two different loss functions. In Fig. 3 (1), the A4C view has artifacts at the thorax, which leads to vanishing boundaries of thorax and DAO. In addition, its atrial septum is open, which may influence the boundary between LA and LV. In Fig. 3 (2), cardiac valves are open and RV has artifacts, which is chordae tendineae. Further, in Fig. (3), the boundaries of the four chambers are obscure. The result of the CU-net with SSIM loss shows a robust performance against the above challenges and exhibits better boundaries. Comparison of the 1st column of each group indicates that the CU-net is considerably better than the U-net and FCN. Comparison between the different loss functions indicate that although the performances of the

TABLE 2. DSC for seven structures compared between the proposed method and the competitive methods. l^{dice} denotes that a model is trained under the dice loss function. l^{ssim} denotes that a model is trained under the SSIM loss function.

	LA	LV	RV	RA	EP	thorax	DAO
FCN+l^{dice}	0.722±0.169	0.728±0.171	0.695±0.175	0.707±0.238	0.901±0.049	0.908±0.041	0.628±0.267
FCN+l^{ssim}	0.741±0.143	0.738±0.157	0.709±0.183	0.780±0.167	0.905±0.042	0.916±0.036	0.648±0.254
U-net+l^{dice}	0.822±0.109	0.807±0.139	0.784±0.157	0.843±0.123	0.925±0.037	0.925±0.031	0.729±0.224
U-net+l^{ssim}	0.825±0.116	0.821±0.115	0.806±0.131	0.855±0.129	0.926±0.033	0.933±0.027	0.724±0.237
CU-net+l^{dice}	0.829±0.105	0.819±0.137	0.815±0.133	0.861±0.109	0.925 ±0.043	0.934±0.031	0.730±0.207
CU-net+l^{ssim}	**0.836±0.096**	**0.841±0.097**	**0.834±0.094**	**0.872±0.109**	**0.931 ±0.03**	**0.939±0.027**	**0.736±0.223**

TABLE 3. HF for seven structures compared between the proposed and the previous methods. l^{dice} denotes that a model is trained under the dice loss function. l^{ssim} denotes that a model is trained under the SSIM loss function.

	LA	LV	RV	RA	EP	thorax	DAO
FCN+l^{dice}	3.707±0.916	3.761±0.995	4.179±0.923	3.148±0.952	4.441±0.902	5.820±1.253	2.283±0.701
FCN+l^{ssim}	3.667±0.858	3.685±0.863	3.984±0.932	3.148±0.979	4.395±0.898	5.693±1.210	2.238±0.687
U-net+l^{dice}	3.155±0.789	3.273±0.901	3.577±0.888	2.707±0.761	4.076±0.898	5.738±1.213	2.056±0.645
U-net+l^{ssim}	3.078±0.748	3.232±0.882	3.562±0.9	2.67±0.759	4.063±0.897	5.433±1.259	2.017±0.642
CU-net+l^{dice}	3.075±0.776	3.184±0.873	3.442±0.888	2.634±0.772	3.988±0.907	5.372±1.199	2.036±0.62
CU-net+l^{ssim}	**3.028±0.728**	**3.08±0.802**	**3.354±0.829**	**2.588±0.735**	**3.927±0.787**	**5.213±1.138**	**1.988±0.614**

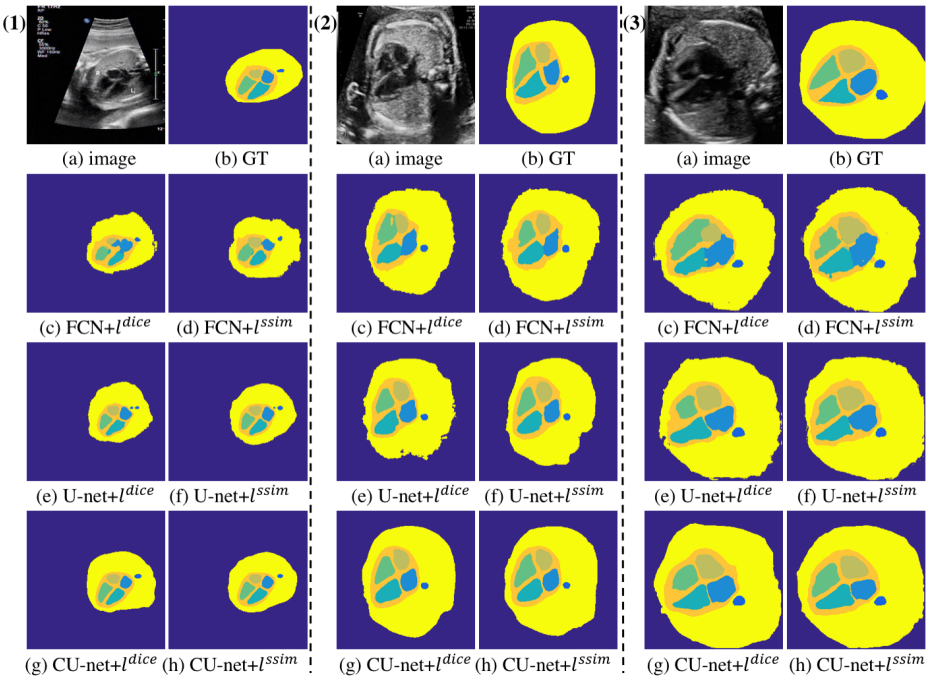

FIGURE 3. Visualization results. Segmentation results on three images are displayed. GT means ground truth.

same network with dice loss and SSIM loss are comparable to each other, SSIM loss results are relatively better.

C. ABLATION EXPERIMENT

In this part, we validate the effectiveness of each key components used in our model. The ablation study involved two parts: ablation experiment on the structure of the CU-net and that on the loss function. All ablation experiments were conducted on the same dataset.

To prove the effectiveness of our CU-net, we report the quantitative comparison results of our model with other related architectures. The results are presented in Table 4. The U-net^2 is a cascaded network of two U-nets. The U-net + BNC_DE [37] connects intermediate layers in the decoder of the first U-net to the encoder of the second U-net based on the U-net^2. The BNC_EE connects the intermediate layers in the encoder of the first U-net to the encoder of the second U-net based on the U-net^2. The Sup appends the

TABLE 4. Ablation study on different architectures.

	mean DSC	mean HF	mean PA
U-net^2	0.831±0.116	3.516±0.891	0.916±0.047
U-net^2+BNC_DE	0.836±0.119	3.451±0.896	0.922±0.041
U-net^2+BNC_EE	0.843±0.108	3.432±0.888	0.922±0.043
U-net^2+BNC_EE+Sup	0.845±0.109	3.390±0.862	0.923±0.039

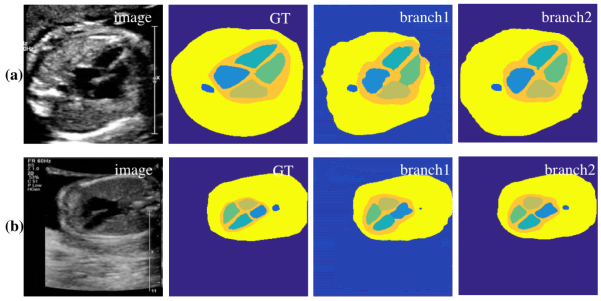

FIGURE 4. Visualization result of two branches for CU-net trained with SSIM loss function. GT means ground truth.

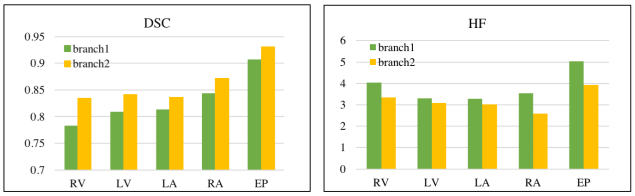

FIGURE 5. DSC and HF values of RV, LV, LA, RA, and EP of two branches for CU-net trained with SSIM loss function.

middle output supervision to the conjunction of two U-nets. Table 4 indicates that the two cascaded U-nets with BNC_EE and supervision architecture achieves the best performance among these configurations. The training time and segmentation time for the four experiments listed in Table 4 are similar, where the training time is approximately 9 h and the processing time for each A4C view is approximately 0.528 s.

For the analysis of the architecture, the results of supervision branch 1 and branch 2 of CU-net trained with SSIM loss are shown in Fig. 4 (3rd and 4th columns, respectively). Fig. 4 (a) presents the images with artifacts generated in RV and thorax. The results of branch 2 show comparatively clearer boundaries than branch 1. Similarly, in Fig. 4 (b), the mitral valve and tricuspid valve of the A4C view are open, which makes segmentation of the boundaries of LA and LV and RA and RV difficult. Through visualization, we can observe that the boundary of branch 2 was better than that of branch 1; then, we compared the DSC of four chambers and epicardium. The DSC and HF values of RV, LV, LA, RA, and EP are presented in Fig. 5. We can observe that all the values of branch 2 are better than those of branch 1.

For loss function analysis, the CU-net with SSIM loss is trained for different values of W, which is the kernel width, and the observations are summarized in Fig. 6. The mean values of DSC and HF, which are useful evaluation metrics to measure segmentation accuracy, are considered. We can observe that the best performance is observed for W = 256. When W increases, the value of DSC increases, whereas that of HF decreases.

D. STATISTICAL ANALYSIS

To further explore whether our method is significantly different from other methods and whether it is effective for improving automated segmentation of A4C views, we conducted a paired sample t-test. The results of the statistical analysis are presented in Table 5.

The statistical results show the effectiveness of our method, and the following three conclusions can be drawn. In the first part (2nd and 3rd rows), compared with the FCN and U-net, the CU-net shows significant differences in terms of the two indexes, HF and DSC.

In the second part (4th to 6th row), the U-net^2+BNC+Sup has a significant effect on improving the performance. Significant differences exist between the U-net^2 and U-net^2+BNC and between U-net^2+BNC+Sup and U-net^2+BNC in terms of DSC and HF. This indicates that the between-net connections and auxiliary supervision are effective.

In the third part (7th to 9th row), significant differences are observed between the different network incorporating SSIM loss and that incorporating dice loss with the p-value being less than 0.05 for both DSC and HF, except for the DSC for U-net incorporating two different loss.

VI. DISCUSSION

In the study, we introduced cascaded U-net with the SSIM loss function for the segmentation of LA, LV, RA, RV, DAO, EP, and thorax from ultrasound A4C views for further extraction of useful clinical indicators. An ultrasound A4C view has three shortcomings: low imaging resolution, obscure tissue boundary, and learning of global structural information. Herein, the proposed model for A4C segmentation focuses on two of these problems, namely, boundary segmentation and utilization of global information.

From Tables 1–3, the following two conclusions can be derived. First, the CU-net performs the best comparing to the FCN and U-net, thus proving its successful design. Second, the SSIM loss function performs better than the dice loss function. The performances of the FCN, U-net, and CU-net are all improved under the constraints of SSIM loss, which demonstrates that SSIM is effective for improving the segmentation performance.

From the experimental results comparing the morphology obtained by the model trained with dice loss and SSIM loss, we observed that the FCN and U-net with SSIM loss are better than the FCN and U-net with dice loss. By constraining the global shape of the target and capturing the global information, a more rounded shape and better segmentation are achieved. Further, visualization of the results indicates that the boundaries of the three subjects are smoother and more accurate, thus confirming that SSIM loss is effective.

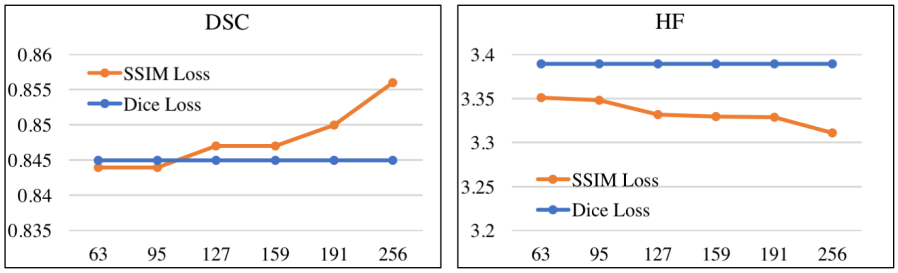

FIGURE 6. Observations of the experiments performed with the CU-net trained with SSIM loss for different W values. The baseline is the CU-net trained with dice loss.

TABLE 5. Results of paired t-test. Bold results indicates the significant differences (p value <0.05). The BNC means the BNC_EE.

Methods			HF	DSC
CU-net	VS	FCN	**<0.01**	**<0.01**
CU-net	VS	U-net	**0.015**	**0.017**
U-net^2	VS	U-net^2+BNC	**<0.01**	**<0.01**
U-net^2+BNC +Sup	VS	U-net^2+BNC	**<0.01**	**<0.01**
U-net^2+BNC +Sup	VS	U-net^2	**<0.01**	**<0.01**
FCN+l^{ssim}	VS	FCN+l^{dice}	**0.011**	**0.028**
U-net+l^{ssim}	VS	U-net+l^{dice}	0.051	**0.030**
CU-net+l^{ssim}	VS	CU-net+l^{dice}	**<0.01**	**<0.01**

Furthermore, the segmentation obtained by the CU-net is more anatomically reasonable than that by the U-net and FCN. As we can see in the first column of each group in Fig. 3, the shape of the chamber segmented by SSIM resembles a circle, with a rough surface. In sum, these results suggest that the proposed CU-net with the SSIM loss function is effective.

To further understand whether the CU-net works in the way it has been designed, we demonstrate the intermediate results of the CU-net. Table 4 confirms that by leveraging the two-branch supervision and the between-net connections from the first encoder to the second encoder, the CU-net can obtain more refined segmentation results. This proves that the BNC_EE is more suitable than BNC_DE [37], [38] for fetal ultrasound multi-tissue segmentation.

In Figs. 4 and 5, we observe that the output of branch 1 is coarser than that of branch 2. The feature output by branch 1 is processed by branch 2, which results in a robust performance as more accurate boundaries are yielded compared to those in branch 1 results.

Our CU-net is more concise compared with other cascaded U-net methods, because of residual blocks in [38], and dense blocks in [39]. As for other fetal echocardiography segmentation methods [37], our CU-net is more less time-consuming. [37] has dilated convolution and highly complex networks and require more segmentation time.

Further, we explored the effect of different kernel widths (W) in SSIM on the performance measured by the dice coefficient and HF. The experimental results confirm that the performance improves with increasing W probably because of the model's ability to grasp global information. Moreover, when W is 256, equal to the size of image, the results are optimum.

The results of the statistical analysis suggest that our method achieved more accurate segmentation compared with two existing methods and the performance improvement due to the use of between-net connections and auxiliary supervision is statistically significant. The results also prove that the CU-net with the SSIM loss significantly outperforms that with the dice loss in fetal ultrasound A4C view segmentation.

Even though the proposed method has been shown to provide good generalization capabilities across the segmentation of images, our work has the following limitation. There is further scope to improve the proposed method, particularly in terms of incorporating clinical prior knowledge into the A4C view segmentation; this will be explored in future.

VII. CONCLUSION

We proposed a novel end-to-end cascaded U-net model, that is, CU-net with SSIM loss, for accurate seven structures segmentation in A4C view. The proposed CU-net is a predict-refine architecture, which consists of two U-nets with branch supervisions and between-net connections. Combined with the SSIM loss function, this method can capture both global information and clear tissue boundaries. Experimental results on A4C view datasets showed that our proposed could achieve 85.6% DSC, 3.311 HF, and 92.9% PA, and performed better than some mainstream methods. Thus, it was demonstrated that our method can assist in early prenatal diagnosis of CHDs. This method can be adapted to semantic segmentation of other organs, and has the potential to be applied to solve segmentation problems in other views of fetal cardiac ultrasound, such as the left ventricular outflow tract view.

REFERENCES

[1] D. A. Lara and K. N. Lopez, "Public health research in congenital heart disease," *Congenital Heart Disease*, vol. 9, no. 6, pp. 549–558, Nov. 2014.
[2] X. Y. Yang, "Incidence of congenital heart disease in Beijing, China," *Chin. Med. J.*, vol. 122, no. 10, pp. 1128–1132, 2009.
[3] V. Miranovic, "The incidence of congenital heart defects in the world regarding the severity of the defect," *Vojnosanitetski pregled*, vol. 73, no. 2, pp. 159–164, 2016.
[4] J. H. Moller and W. A. Neal, *Heart Disease in Infancy*. New York, NY, USA: McGraw-Hill, 1981.

[5] J. P. McGahan, "Sonography of the fetal heart: Findings on the four-chamber view.," *Amer. J. Roentgenology*, vol. 156, no. 3, pp. 547–553, Mar. 1991.
[6] Y. Singh and L. McGeoch, "Fetal anomaly screening for detection of congenital heart defects," *J. Neonatal Biol.*, vol. 5, no. 2, pp. 100–115, 2016.
[7] C. O. Fernandez, C. Ramaciotti, L. B. Martin, and D. M. Twickler, "The four-chamber view and its sensitivity in detecting congenital heart defects," *Cardiology*, vol. 90, no. 3, pp. 202–206, 1998.
[8] R. D. Greenwood, A. Rosenthal, L. Parisi, D. C. Fyler, and A. S. Nadas, "Extracardiac abnormalities in infants with congenital heart disease," *Pediatrics*, vol. 55, no. 4, pp. 485–492, 1975.
[9] L. Allan, "Prenatal diagnosis of structural cardiac defects," *Amer. J. Med. Genet. C, Seminars Med. Genet.*, vol. 145C, no. 1, pp. 73–76, Feb. 2007.
[10] E. Molins, F. Macià, F. Ferrer, M.-T. Maristany, and X. Castells, "Association between Radiologists' experience and accuracy in interpreting screening mammograms," *BMC Health Services Res.*, vol. 8, no. 1, p. 91, Dec. 2008.
[11] G. C. Kagadis, A. Walz-Flannigan, E. A. Krupinski, P. G. Nagy, K. Katsanos, A. Diamantopoulos, and S. G. Langer, "Medical imaging displays and their use in image interpretation," *RadioGraphics*, vol. 33, no. 1, pp. 275–290, Jan. 2013.
[12] E. P. M. van Vliet, J. J. Hermans, W. De Wever, M. J. C. Eijkemans, E. W. Steyerberg, C. Faasse, E. P. M. van Helmond, A. M. de Leeuw, A. C. Sikkenk, A. R. de Vries, E. H. de Vries, E. J. Kuipers, and P. D. Siersema, "Radiologist experience and CT examination quality determine metastasis detection in patients with esophageal or gastric cardia cancer," *Eur. Radiol.*, vol. 18, no. 11, pp. 2475–2484, Nov. 2008.
[13] A. P. Brady, "Error and discrepancy in radiology: Inevitable or avoidable?" *Insights into Imag.*, vol. 8, no. 1, pp. 171–182, Feb. 2017.
[14] S. Li and E. Brantley, "Malpractice liability risk and use of diagnostic imaging services: A systematic review of the literature," *J. Amer. College Radiol.*, vol. 12, no. 12, pp. 1403–1412, Dec. 2015.
[15] N. Stec, D. Arje, A. R. Moody, E. A. Krupinski, and P. N. Tyrrell, "A systematic review of fatigue in radiology: Is it a problem?" *Amer. J. Roentgenology*, vol. 210, no. 4, pp. 799–806, Apr. 2018.
[16] S. Waite, S. Kolla, J. Jeudy, A. Legasto, S. L. Macknik, S. Martinez-Conde, E. A. Krupinski, and D. L. Reede, "Tired in the reading room: The influence of fatigue in radiology," *J. Amer. College Radiol.*, vol. 14, no. 2, pp. 191–197, Feb. 2017.
[17] N. Goel, A. Yadav, and B. M. Singh, "Medical image processing: A review," presented at the 2nd Int. Innov. Appl. Comput. Intell. Power, Energy Controls Their Impact Humanity (CIPECH), Nov. 2016, doi: 10.1109/cipech.2016.7918737.
[18] L. O. Hall, A. M. Bensaid, L. P. Clarke, R. P. Velthuizen, M. S. Silbiger, and J. C. Bezdek, "A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain," *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 672–682, Sep. 1992.
[19] J. Jiang, P. Trundle, and J. Ren, "Medical image analysis with artificial neural networks," *Computerized Med. Imag. Graph.*, vol. 34, no. 8, pp. 617–631, Dec. 2010.
[20] Z. Shi, L. He, K. Suzuki, T. Nakamura, and H. Itoh, "Survey on neural networks used for medical image processing," *Int. J. Comput. Sci.*, vol. 3, no. 1, p. 86, 2009.
[21] 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.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[22] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.
[23] S. Jia, "Automatically segmenting the left atrium from cardiac images using successive 3D U-Nets and a contour loss," in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*. Cham, Switzerland: Springer, 2019, pp. 221–229.
[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[25] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
[26] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.
[27] M. R. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI," *Med. Image Anal.*, vol. 30, pp. 108–119, May 2016.
[28] M. Nasr-Esfahani, M. Mohrekesh, M. Akbari, S. M. R. Soroushmehr, E. Nasr-Esfahani, N. Karimi, S. Samavi, and K. Najarian, "Left ventricle segmentation in cardiac MR images using fully convolutional network," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1275–1278, doi: 10.1109/embc.2018.8512536.
[29] W. Yan, Y. Wang, Z. Li, R. J. van der Geest, and Q. Tao, "Left ventricle segmentation via optical-flow-net from short-axis cine MRI: Preserving the temporal coherence of cardiac motion," in *Proc. Med. Image Comput. Comput. Assist. Intervent (MICCAI)*. Cham, Switzerland: Springer, 2018, pp. 613–621.
[30] J. Chen, H. Zhang, W. Zhang, X. Du, Y. Zhang, and S. Li, "Correlated regression feature learning for automated right ventricle segmentation," *IEEE J. Transl. Eng. Health Med.*, vol. 6, pp. 1–10, 2018.
[31] M. R. Avendi, A. Kheradvar, and H. Jafarkhani, "Automatic segmentation of the right ventricle from cardiac MRI using a learning-based approach," *Magn. Reson. Med.*, vol. 78, no. 6, pp. 2439–2448, Dec. 2017.
[32] 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.
[33] S. Dong, G. Luo, K. Wang, S. Cao, Q. Li, and H. Zhang, "A combined fully convolutional networks and deformable model for automatic left ventricle segmentation based on 3D echocardiography," *BioMed Res. Int.*, vol. 2018, pp. 1–16, Sep. 2018.
[34] L. Yu, Y. Guo, Y. Wang, J. Yu, and P. Chen, "Segmentation of fetal left ventricle in echocardiographic sequences based on dynamic convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1886–1895, Aug. 2017.
[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[36] Tensorflow. (2019). *An Open Source Machine Learning Framework for Everyone*. [Online]. Available: https://www.github.com/tensorflow/tensorflow
[37] L. Xu, M. Liu, Z. Shen, H. Wang, X. Liu, X. Wang, S. Wang, T. Li, S. Yu, M. Hou, J. Guo, J. Zhang, and Y. He, "DW-Net: A cascaded convolutional neural network for apical four-chamber view segmentation in fetal echocardiography," *Computerized Med. Imag. Graph.*, vol. 80, Mar. 2020, Art. no. 101690.
[38] H. Liu, X. Shen, F. Shang, F. Ge, and F. Wang, "CU-Net: Cascaded U-Net with loss weighted sampling for brain tumor segmentation," in *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy*. Cham, Switzerland: Springer, 2019, pp. 102–111.
[39] S. Li, Y. Chen, and S. Yang, "Cascade Dense-Unet for Prostate Segmentation in MR Images," in *Proc. Int. Conf. Intell. Comput.*, 2019, pp. 481–490.
[40] D. Lachinov, E. Vasiliev, and V. Turlapov, "Glioma Segmentation with Cascaded Unet," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham, Switzerland: Springer, 2019, pp. 189–198.

LU XU was born in Haicheng, Liaoning, China, in 1996. She received the B.Eng. degree from Northeastern University, Liaoning. She is currently pursuing the Ph.D. degree with Beihang University, Beijing, China, under the guidance of Prof. J. Zhang. Her research interest includes the AI application in medical image processing.

MINGYUAN LIU was born in Shenyang, Liaoning, China, in 1995. He received the B.Eng. degree from Beihang University, Beijing, China, where he is currently pursuing the M.Eng. degree under the guidance of Prof. J. Zhang. His research interest includes the AI application in medical image processing.

JICONG ZHANG received the B.S. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, in 2003 and 2006, respectively, and the Ph.D. degree from the University of Florida, Gainesville, FL, USA. He is currently a Professor with the School of Biological Science and Medical Engineering, Beihang University, Beijing, China. His research interests include mathematical modeling, machine learning and AI applications in medical image processing, and biomedical signal processing.

YIHUA HE received the B.S. degree from Harbin Medical University, in 1993, the M.S. degree from Dalian Medical University, in 2004, and the Ph.D. degree from Xi'an Jiaotong University, in 2012. She is currently a Chief Physician with the Echocardiography Department, Beijing Anzhen Hospital, Capital Medical University, Beijing, China. Her research interests include fetal echocardiography, adult echocardiography (coronary heart disease, valve disease, cardiomyopathy, congenital heart disease, and so on), and transesophageal three-dimensional echocardiography.